The Research Road We Make: Statistics for the Uninitiated

Saviour Formosa PhD

Sandra Scicluna PhD

Jacqueline Azzopardi PhD

Janice Formosa Pace MSc

Trevor Calafato MSc

2011

Published by the National Statistics Office, Malta

Published by the
National Statistics Office
Lascaris
Valletta VLT 2000
Malta
Tel.: (+356) 25997000
Fax:(+356) 25997205 / 25997103
e-mail: nso@gov.mt
website: http://www.nso.gov.mt

CIP Data

The Research Road We Make: Statistics for the Uninitiated – Valletta: National Statistics Office, 2011
xxii, 279p.

ISBN: 978-99957-29-14-1

For further information, please contact:

Unit D2: External Cooperation and Communication
Directorate D: Resources and Support Services
National Statistics Office
Lascaris
Valletta VLT 2000
Malta
Tel: (+356) 25997219

NSO publications are available from:

Unit D2: External Cooperation and Communication
Directorate D: Resources and Support Services
National Statistics Office
Lascaris
Valletta VLT 2000
Malta
Tel.: (+356) 25997219
Fax: (+356) 25997205

Contents

	Page
Preface	vii
Foreword	viii
Acknowledgements	ix
Abbreviations	xi
Glossary	xiii
Imagery	xix
Introduction	xxi
Chapter 1 What is Statistics? An Intro for the Uninitiated!	1
Why Statistics?	3
The Tower of Babel Syndrome or Valhalla?	4
The 'fear of stats'	5
Myths and Realities	5
Chapter 2 Research Methodology	7
The Research Design	9
Social Scientific Research Methods	10
Research Problems	11
Sampling	11
Causality, Association and Correlations	12
Methods of Research	13
Chapter 3 DIKA	25
Why is Research Necessary?	29
Forms of Research	31
Types of Research	32
How Research is Done	33
Techno-Centric or Socio-Technic Approach?	35
Use and abuse of statistics	36
Avoiding Research	38
Data	39
Information	41
Knowledge	42
Chapter 4 Structuring Your Research	45
A Datacycle Approach	47
Design	47
Choosing the correct mining/trawling tools	49
Matrixing	50
Data gathering	58
Analysis	59
Reporting	64
Chapter 5 From Concept to Tangibility	71
Basic Concepts	73
Data issues – the structures	74
Data Measurement: The Scales	80

Page

Chapter 6 Data Acquisition and Data Quality 85

Data Categories 87
What is a Metadata? 88
Data Sourcing 90
Data Capture 92
Quality 92
Error 92
Primary, Secondary and Tertiary Sourcing 92
Capture Modes 94

Chapter 7 Visualisation 97

Graphing 103
Charting tools 107
Mapping 108
GIS as a tool for scientific research 109
GIS tools 119

Chapter 8 Mind Mapping 121

What is a model? 123
Who are the players? 130
Conceptual Modeling 130
Content Analysis 132
CRISOLA Model 134

Chapter 9 Tools 143

Which Tools are Available? 145
Spreadsheets 147
Macros 149
Dedicated Statistical Software 149
Quantitative 149
Qualitative tools 160
Geo-statistical tools 164
Online Tools 166

Chapter 10 IT/IS and Databases 171

Moving from basic tools to databases 173
From a Mind to an EAR 176
Querying language: SQL 179

Chapter 11 Statistical Testing 183

Generic Statistics Publications sorted by Publication Date 185
Thematic Publications sorted by Theme 187
Basic Statistics 189
Statistical Tests 203
Spatial Statistics 204

Chapter 12 Case Studies 207

A taste for working with Census data 209
Using the Census for research 212
In an Archive 221

	Page
Chapter 13 Data Sources	227
Analogue – a library/archive approach	229
Specialised libraries	230
Digital – online	231
Chapter 14 Ethical Issues	235
What is Ethics?	237
Criteria for Ethical Research	237
Referencing	238
Plagiarism	238
How to compile a reference section	239
Bibliography	243
Appendix – Questions and Answers	249
Ending	279

Preface

Today the use of statistics has evolved far beyond its genesis. The use of statistics by individuals and organisations has become ubiquitous, leading to a wider informed knowledge-base that starts from the process to gather and understand the data up to the end phase where informed decisions need to be made. This is the case across the different thematic domains ranging from the natural and social sciences, medicine, business, and other areas. Statistics can be used and unfortunately, abused; the latter leading to the generation of serious errors in both description and analysis, in turn leading to misleading interpretations. Few are immune to such error generation and the need is felt to ensure that we are aware of the responsibilities researchers have in generating statistics.

Therefore, when the National Statistics Office was approached, early in 2010, by a group of researchers to assist in publishing a book on statistics intended for persons who are non-conversant with this discipline, I readily accepted, as they had found a niche I believe was untapped locally. This idea was appropriately launched during a World Statistics Day seminar, marked for the first time ever on 20 October 2010. The feedback was positive and encouraging.

This book introduces statistics to higher education students who are not necessarily mathematically oriented, and whose perception of statistics may be tinged with fear. Its scope is not to review equations and formulae, but to outline the importance of the strict rules that govern research methodology and the sound interpretation of results. In other words, it identifies uses and abuses of statistics, and is a useful text even for those who may eventually specialise in statistics. In many respects, therefore, it is different from other mainstream statistics textbooks.

While the views expressed by the book's authors do not necessarily reflect NSO's official stance, such initiatives are to be commended as they help promote statistics, which are so vital in today's day and age. The ultimate aim of this publication is to educate, and to promote statistical literacy. This forms an intrinsic part of NSO's long-term vision.

Michael Pace Ross
Director General
National Statistics Office

August 2011

Foreword

Statistics is a vital part of social science training. In the United States, statistical understanding is required for undergraduate academic degrees across the social sciences. It is considered an indispensable component of research methods and a prerequisite for postgraduate study, which typically includes further study in advanced and specialised statistical topics. In the United Kingdom, statistical analysis is fast becoming a priority for postgraduate training although undergraduate degrees in sociology and related fields tend to place less emphasis on quantitative methods. British undergraduates gain experience in use of qualitative research techniques through completion of the undergraduate dissertation. In my experience, as a visiting lecturer at the Institute of Criminology, the University of Malta tends to blend the best of both traditions. Maltese students have the opportunity to learn statistical techniques at the undergraduate level, and further, to incorporate these in methodological designs for undergraduate dissertations.

The Research Road We Make: Statistics for the Uninitiated represents a perfect example of the Maltese tradition of pursuing the best of both worlds. The authors begin with an explanation of the research process, which situates statistical analysis within overall project design. Next, the discussion examines the types of data suitable for statistical techniques and provides a guide to data collection. The authors then move on to issues of interpretation of findings and presentation of results; this discussion includes an explanation of the very latest methods of visual projection. The final sections explain computer-assisted techniques and tools for data management. Overall, the text provides an engaging, substantive explanation of statistical analysis with reference to interesting, real-world issues important for social research.

In addition, the authors provide a bridge between concepts and techniques, many of which have been developed in Europe and North America, with applications to the Maltese context. To carry out effective and meaningful social science research, it is important to consider Malta's history, culture, and geography, and the discussion here achieves this with a range of examples from the Maltese Islands. *The Research Road We Make* represents an indispensable resource for success in statistical research.

Paul Knepper, PhD
Department of Sociological Studies
University of Sheffield

Acknowledgements

A word of thanks goes to the following persons who authored the book.

Saviour Formosa and **Sandra Scicluna** for authoring the book based on their work in the quantitative/spatial and qualitative fields respectively.

Jacqueline Azzopardi who had the unenviable task of reviewing the chapters, making sense of the runaway sections, drafting all the changes as well as authoring the Questions and Answers chapter. Not an easy job that one! Especially for an expert in the qualitative field faced with formulae and spatial reviews!

Janice Formosa Pace for authoring the Ethics chapter and reviewing the referencing throughout the book.

Trevor Calafato for authoring the qualitative section of the Tools chapter.

Ramon Azzopardi for the chapter title graphics; a talented budding artist!

Michael Pace Ross (Director General NSO) and **Catherine Vella** for book review and editing, **Claire Meli** for book setting and the following NSO personnel for their continuous support: **Silvana Mizzi, Rose Marie Portelli, Stephania Farrugia Dimech, Jesmond Galea, Margaret Bugeja, Joanna Bonnici** and **Shawn Borg**.

The book is dedicated to the Staff of the Institute of Criminology at the University of Malta and the Staff at the National Statistics Office. The front image, entitled *kampane 3 weeks* was created by Saviour Formosa as part of the analysis on the process that renders 3D models of a mathematical structure created through fractal and chaos scientific methodology. Kampane refers to a dedication to the late father of a relative and the three weeks to the actual time it took a Pentium 4 to render the image (a literal 504 hours). This imaging process reflects the way research must be tackled; with patience, creativity and a tinge of hyperactivity.

Biographies

Dr. Saviour Formosa

B.A.(Hons) Soc, Dip.AppSocSt., M.Sc. GIS (Hudd), Ph.D. (Hudd), FRGS

Saviour Formosa holds a Ph.D. in spatio-temporal environmental criminology. He is a senior lecturer within the Institute of Criminology, University of Malta and lectures in various faculties at the University of Malta. His main expertise lies in the implementation of cross-thematic approaches and uses to the data cycle and management with emphasis in the thematic, social and spatial data structures, GIS, visualisation, modelling, web-mapping, analysis and dataflow management and reporting. His bridging the gaps between hi-end technology and the social sciences has help to create new methods for analysing the impacts of spatial developments. He is a Member of the Applied Criminology Centre at the University of Huddersfield. Dr. Formosa has developed the Information Resources processes at the Malta Environment and Planning Authority and represents Malta in various international fora as are GEO, ESPON, and the EEA. He was also the Project Leader for ERDF project "Developing National Environmental Monitoring Infrastructure and Capacity".

Dr. Formosa has developed the www.crimemalta.com website which covers ongoing news and crime-related statistics in Malta. He also produced a series of interactive applications for the local and international markets inclusive of the interactive Census maps for the NSO.

Dr. Sandra Scicluna

B.A. (Gen.), B.A. (Hons), Dip.PS, M.Sc. (Leic.) Ph.D. (Leic.)

Dr Sandra Scicluna holds a Ph.D. on research about prison rehabilitation in Malta. She is a senior lecturer with the Institute of Criminology, University of Malta. And lectures in the areas of transnational

crime, punishment, substance abuse, the world of corrections, dealing with foreign offenders and organised crime. She has past experience working as a Probation Officer and was elected as member of the CEP Board. She is also a member of the Police Academy Board and has acted as an assistant to prisoners on the Prison Appeals Tribunal. In addition, she acts as a consultant to the Ministry of Justice and Home Affairs and is Deputy National Science Correspondent for CEPOL.

Dr. Scicluna has produced and contributed to various publications on topics which include: substance abuse, domestic violence and the development of probation and prisons in Malta.

Dr. Jacqueline Azzopardi
B.Ed. (Hons), Dip.PS, M.Sc. (Leic.) Ph.D. (Leic.)

Dr Azzopardi holds a first degree in Education, a postgraduate Diploma in Probation Services, a Masters in Criminal Justice (Leicester University – UK) and a Ph.D. (Leicester University – UK) in policing. Dr Azzopardi is the Director of the Institute of Criminology and a senior lecturer. Dr Azzopardi is a member of the Police Academy Board and has also acted as an assistant to prisoners on the Prison's Board of Appeals for about five years. She is the current Maltese National Research and Science Correspondent for Cepol. Dr Azzopardi has published and contributed to publications/articles on: culture, policing, policewomen, police culture, violence, women and politics as well as youths and delinquency.

Dr Azzopardi is also involved in the Dingli community - where she serves as a local councillor and the President of the Dingli Primary School council.

Ms. Janice Formosa Pace
B.Psych., Dip.PS., MSc (Leic)

Ms. Formosa Pace obtained her first Degree in Psychology and also a Post Graduate Diploma in Probation Services at the University of Malta. Ms Formosa Pace completed her Masters Degree in Forensic and Legal Psychology at the University of Leicester in 2003. Ms. Formosa Pace is a full time teacher of Personal and Social Education at St Clare College Boys' Secondary School Gzira. She is also a visiting lecturer within the Institute of Criminology, University of Malta. Her main areas of interests are the psychological approaches to understanding crime, the role of psychology in criminal investigation and courtroom interactions.

Ms. Formosa Pace is currently a PhD Candidate at the University of Leicester specialising in the incidence of crime transference across the generations.

Mr. Trevor Calafato
B.A.(Gen.), Dip.PS, M.Sc. (Leic.)

Mr Calafato joined the Institute on the 1st April 2009 as a full-time assistant-lecturer. Having achieved a BA in Psychology and Communications from the University of Malta, he continued his studies in Post Graduate Diploma in Probation Services. Between 2003 and 2009 he served as a Probation Officer within the Department of Correctional Services. Mr. Calafato also followed a number of courses at the Occupational Health and Safety Authority (OHSA) and the Institute of Health and Safety (IHS).

In 2004 Mr. Calafato read for an MSc degree in Security and Risk Management at the University of Leicester. The MSc dissertation focused on the preventions and possible reactions of emergency and security services to respond effectively to a major terrorist incident in Malta, whilst investigating the Maltese populace trust in the local authorities in terrorist contingencies. Later his interest in terrorism led to reading an E-Learning Certificate in Terrorism Studies, at the University of St. Andrews, Scotland.

Mr. Calafato is currently a PhD Candidate at the University of Sheffield researching terrorism.

Abbreviations

2D	Second Dimension
3D	Third Dimension
APA	American Psychological Association
CAQDAS	Computer-Assisted Qualitative Data Analysis Software
CCP	Corradino Civil Prison
CDR	Common Data Repository
CRISOLA	Crime Social Landuse Model
Dbf	Database file format
DBMS	Database Management Systems
DEMs	Digital Elevation Models
DG	Directorate General
DIKA	Data – Information – Knowledge - Action
Doc	Microsoft Word File Format
DRIPS	Data Rich Information Poor Syndrome
EAR Diagram	Entity Attribute Relationship Diagram
EAS	Census Enumeration Areas
EEA	European Environment Agency
ESRC	Economic and Social Research Council
ESS	European Statistical System
EU	European Union
FAO	Food and Agriculture Organisation
GE	Google Earth
GIGO	Garbage In Garbage Out
GIS	Geographical Information Systems
GPS	Global Positioning System
GUI	Graphic User Interface
GWR	Geographically Wieghted Regression
HTML	Hyper Text Markup Language
HU	Hermeneutic Unit
ICT	Information Communications Technology
IR	Information Resources
IS	Information Systems
IT	Information Technology
KWIC	Key Words in Context
LAU	Local Area Units
Max	Maximum
MEPA	Malta Environment and Planning Authority
Mif/Mid	MapInfo Export File Format
Min	Minimum
NAM	National Archives of Malta
NNA	Nearest Neighbour Analysis
NNH	Nearest Neighbour Hierarchical Analysis
NOIR	Nominal (N), Ordinal (O), Interval (I) and Ratio Scales (R)
NSO	National Statistics Office
OCR	Optical Character Recognition

OECD	Organisation for Economic Cooperation and Development
OODBMS	Object-Oriented Database Management System
OPAC	Online Public Access Catalogue
ORDBMS	Object-Relational Database Management System
PASW	Predictive Analytics SoftWare
PDA	Personal Digital Assistant
QDAS	Qualitative Data Analysis Software
RDBMS	Relational Database Management System
SAGE	Spatial Analysis in a Geographical Environment
SAS	Statistical Analysis System
Sde	Standard Deviation Ellipse
SGeMS	Stanford Geostatistical Modelling Software
SPSS	Statistical Package for the Social Sciences
SQL	Standard Query Language
SSDA	Statistical Spatial Data Analysis
STAC	Space and Temporal Analysis of Crime Software
StatDB	National Statistics Office Statistical Database
SWOT	Strengths, Weaknesses, Opportunities and Threats
Tab	MapInfo File Format
UN	United Nations
VRML	Virtual Reality Modeling Language
W6H	Who, what, when, how, why, where, why not?
WHO	World Health Organisation
WHOSIS	WHO Statistical Information System
WWW	World Wide Web
WYSIWYG	What You See Is What You Get
Xls	Excel File Format

Glossary

3D Map	A map that extrapolates the data of maps into 3d format.
Accuracy	The extent to which an estimated value approaches the true value.
Action	The result of policy making as emanating from knowledge creation. "Action" is the implementation of policy.
Adduction	With reference to archival research, finding the best explanation of a set of data' (Josephson and Josephson, 1994:157).
Analogue	The study through physical material means (such as hardcopies or books).
Anonymity	The identity of individuals and organisations/institutions needs to be kept anonymous (British Psychological Society, 1997).
ANOVA	Analysis of variance, also known as the anova, determines the existence of differences in datasets that contain two or more sample means.
Applied form of research	An inquiry of a scientific nature designed for and conducted with an operational and practical application as its goal.
Archival data	Original records that are gathered by the researcher or another researcher. This data is in its original state and has not been interpreted by others.
Area Charts	Area charts are ideally used for data that requires depiction of individual variables in relation to a total.
Attribute	Column in a spreadsheet/database containing a number of cells population by data.
Authenticity	Verification that the documents are original.
Bar Charts	Bar charts are composed of bars separated by spaces. Ideal for displaying the distribution of variables measured at the nominal level.
Basic/pure form of research	Research that is theoretical in nature and does not aim at the immediate provision of tangible results. The aim of such research is the acquisition of new information and the development of the scholarly disciplines.
Bias	Systematic deviation from a norm or from the truth.
Bimodal	A type of mode (the other being unimodal). It has two peaks with the highest point in a distribution indicating the most frequent score.
Causality	A change in 'a' brings about a change in 'b'.
Chi Squared	A critical test that investigates, looking for the frequencies of category (nominal) presence in a sample and analyses whether they represent the predicted frequencies in the total population.
Choropleth map	A map that depicts data based on ranges.
Close-ended questions	Questions which have limited responses. The most simple require a yes/no answer. Others might require the respondent to make a choice out of several answers, while others still use a 4 or 5 point scale model such as: always, sometimes, rarely, never and I do not know.
Completeness	Extent to which data is supplied for all component parts and time periods.
Composition	Process that allows researchers to develop concepts, visualise variables, create the lineages between the variables and also to identify the statistical measurements required for each linkage and how they link within themes and across themes.
Conceptual model	A first-run model based on an idea (concept) which would later be developed into a full model.
Conceptualisation	The nailing down the elusive variable – moulding an idea into something that can be measured.
Confidence interval	An estimate used to indicate the reliability of an estimate resulting in a range within a percentage of how far from reality the results are: example +/- 3%.
Confidentiality	Personal information acquired through research should not be divulged. Only with the permission of the subjects can confidentiality be broken.
Context	Normally referring to the social activity, the physical reality, the environmental design within which that activity occurs.
Content analysis	Refers to the analysis of a written material ex: a political speech where words are analysed (within historical/political context) in an attempt to envisage the meaning of the writer.

Correlation map	A map that depicts correlations between two variables in polygon format.
Credibility	Includes an assessment of potential and actual sources of error and distortion.
CRISOLA	CRIme SOcial LAnduse. The main area of study is the interaction between the crime characteristics, the social characteristics and the physical characteristics (land use).
Data	Information that is coded and structured for subsequent processing, generally by a computer system (British Computer Society, 1989).
Data analysis	The diverse ways that one can employ to make sense of the data in conjunction with the findings extracted from the literature review. Data analysis deals with the interpretation of the data within the context under study.
Data mining	The process taken to gather the data either manually or through initiated processes such as a questionnaire.
Data trawling	The automatic process whereby data is gathered by machines or sensors which will review the availability of that data, extract it and store it in a specific location for the researcher's perusal.
Database	A collection of information that is organised so that it can easily be accessed, managed, and updated. In one view, databases can be classified according to types of content: bibliographic, full-text, numeric, and images (SearchSQLServer).
DBMS (Database management systems)	General purpose computer programmes aimed at making a database work.
Deception	The unacceptable practice of providing participants with misleading information or with partial information about the study they have consented to participate in as subjects (British Psychological Society, 1997).
Deductive method	The theoretical interpretations and logically interpretive prepositions to start a study.
Demography	The study of populations.
Dependent variable	Also known as criterion variables in that they are tested for changes that occur when a value in the independent variable has changed.
Descriptive research	Describes a situational construct in time and space. This type of research allows researchers to understand what something is.
Descriptive statistics	Describes a dataset quantitatively through summarisation rather than through the usage of probability analysis.
Digital	Information that is stored in an electronic form as against an analogue form (hardcopy).
Distance statistics	Analysis of values based on proximity.
Dot density map	A map that depicts data based on randomly-located dots representing numbers of cases.
DRIPS (Data-rich-information-poor-syndrome)	Turning data into information within a context.
Entitation	The way we recognise, understand and define the entities about which we wish to collect data.
Entity	An item that can be described and measured in statistical analysis.
Errors	The difference between the captured data and the real data that exists.
Ethics	Also known as moral philosophy. They provide us with recommended guidelines of what is right and what is wrong (International Encyclopedia Of Philosophy, 2010).
Ethnography	Or participant observation. A research based on the researcher's observation and analysis of group behaviour.
Explanatory research	Explaining why something occurred and the causes behind it. This method allows the researcher to explain behaviour and counteract the problem under study.
Focus groups	Facilitated and organised group or a specific topic.
Frequency distribution	A list of options is shown together with the number of respondents.
F-Test	Test comparing the ratio of the two variances which, if equal, should result in a value of 1.

Geographic information	Information which can be related to specific locations on the earth (UK Department of the Environment, 1987).
GIS (Geographic Information Systems)	A group of procedures that provide data input, storage and retrieval, mapping and spatial, and attribute data to support the decision-making of the organisation. (Grimshaw, 1994).
Graduated map	A map that depicts data as a series of graduated points (which could also comprise pie-charts).
Histogram	Very similar to bar charts but depict a distinct difference. Adjacent bars used to display the distribution of quantitative variables. These variables vary along a continuum with no gaps.
Hotspot analysis routines	Depict data based on the concentration of values in a spatial location.
Hypothesis	A testable phrase aiming to explain a phenomenon using scientific means to test it.
Independent variable	Or Predictor variables, cause changes in other variables when value is changed and is manipulated.
Indicators	A list of statistical values set up to be sure that a particular data set is valid, and can be repeatedly gathered, analysed and processed.
Inductive method	Research using the observation of reality without any theory.
Inferential statistics	Data allowing analysis through probability tests that allows one to come to conclusions on a population. These are also called inductive statistics.
Information	The meaning given to data by the way in which it is interpreted (British Computer Society, 1989).
Informed consent	Researchers need to carry out their studies after obtaining consent from participants.
Insitu data gathering	Gathering processes which occur in the field as against from remote (far-away) locations.
Internet	An international network of computers and technologies that provide services such as the WWW (World Wide Web), emails and ftps, among others.
Interpolation statistics	When primary data is used to predict the probable value of areas within a boundary relative to the location of the primary data.
Interval scale	A Scale referring to amounts (numbers) but not as actual amounts but as a representation of that amount.
Interviews	Questions asked orally (in person or by phone). Qualitative tool.
K-Means	Data based on statistical clustering of related data points.
K-Means clustering map	A map that depicts data based on statistical clustering of related data points.
Knowledge	"Knowledge" represents and fits within the social reality under study. Knowledge serves as the tool that extracts the meanings given to the data trawlers and changes that to policy.
Likert scale	A method of ascribing quantitative data to what are essentially qualitative data such as the analysis of agreement or disagreement to something.
Line Charts	Charts composed of lines along an axis. This type of chart allows multiple variables to be depicted in the same chart.
Lineage	A document that holds a step by step record of the process employed to reach the end result.
Logical consistency	Suitability of commands, operations and analysis.
Macros	Digital tools that allow researchers to input their data in specified cells and run the resultant measure accordingly thus drastically reducing the need for repeated work.
Matrix	A collection of cells that serve as an aid to structure data according to set columns and records in what can best be described as a spreadsheet.
Maximum	The largest number in the dataset.
Mean	The score located at the mathematical centre of a distribution and represents the arithmetic mean which is also called the average.
Measures of central tendency	The values that are either at the middle point of a set of data or are typical of that type of data.
Median	The score located at the 50th percentile. The median allows researchers to identify that middle value which serves as a divider between the two halves of a dataset.

Metadata	The process of how one ensures that data has a context within which it was created and that it serves as a veritable identity card/passport for that particular dataset. It is data about data. It provides a description of what a dataset is composed of.
Mind map	A tool to clarify one's mind and helps visually draft the process from concept to tangible measuring. It is a tool that enables the researcher to build an idea of: (a) what exists; (b) what should exist; (c) how best to come up with a method to identify the links/research questions.
Minimum	Refers to the smallest number in the dataset.
Mnemonics	A technique used to develop and assist memory using codes.
Mode	Refers to the score that occurs most frequently. There are two types of mode: the unimodal and the bimodal.
Model	A representation of a structure which has within it the requirements which allow for the investigation of that same structure. Either a theory, a law, a hypothesis or a structured idea. It can be a role, a relation or an equation. It can be a synthesis of data. Most important, from the geographical viewpoint, it can also include reasoning about the real world by means of translations in space (to give spatial models) or in time (to give historical models) (Chorley and Haggett, 1967; 21-22).
Multipurpose research	Research somewhere in between the basic form and the applied form. It is both conceptual and factual (incorporated in one study).
Nearest neighbour analysis	The aggregation of the data on the proximity of an activity to the nearest location of another activity.
NNA (Nearest neighbour hierarchical analysis) map	A map that depicts data showing hotspots in the form of ellipsoids.
Nominal scale	A scale not referring to amounts (numbers) but to tags that serve as an identification.
NUTS (Nomenclature of territorial units)	A European common classification of territorial units for statistics.
Objectivity	The carrying out of research that is value free i.e. the social scientist has to be aware of his/her own values and perspectives and not allow them to impinge on the research findings.
OODBMS	A database that builds its model around objects, each of which has its own set of attributes and behaviours that have complex information built into them.
Open-ended questions	Questions which require an elaborate answer.
ORDBMS	Similar to the relational database but incorporates an object-oriented database model.
Ordinal scale	A scale that refers to amounts (numbers) but not as an actual amount but as a representation of an amount.
Parameter	The whole population under study as against the study through a sample.
Percentage distribution	A list showing percentages indicating options with 100% as total.
Percentages	Refer to the same method as proportions but expressed as a figure out of 100.
Pie charts	Circles (pies), as in the case of bar charts display their data in the form of slices. All the slices make up a cake or a pie!
Pilot study	An initial test run used by researchers to test the way respondents react to the questionnaire/survey.
Plagiarism	"The unauthorised use or close imitation of the language and thoughts of another author and the representation of them as one's own original work" (http://dictionary.reference.com/browse/plagiarism).
Point map	A map that depicts data based on points representing the actual location of an activity.
Polygon-based cluster analysis	A map that depicts cluster data ranged across polygons (areas).
Population pyramid	A tool that depicts population structures based on sex and age in a mirror-like bar chart.
Precision	Level of recorded detail.

Predictive research	Helps to establish future actions by modelling the potential futures based on the testing of scenarios.
Primary key	Unique identifier in a database.
Primary sources	Sources that point at data gathered first-hand.
Proportions	Fractions of the total.
Qualitative research	Research based on small samples. The tools used are usually observations, interviews, documentary analysis, in-depth interviews with a small number (especially with difficult to access populations).
Quantification	Process that converts an entity into a measurable structure.
Quantitative research	Researcher based on a large sample. Yields statistical data which is usually analysed through a statistical tool. Quantitative research assigns numbers.
Range	Defined as the difference between the two extremes in the data range: the minimum and maximum.
Ratio Scale	A scale referring to actual and true amounts (numbers).
RDBMS	A relational model that links the different elements in a study through the relationships between the attributes in each element.
Referencing	Acknowledging the use of the work of others. This is essential to avoid plagiarism.
Regression analysis	The line of best fit. Regression is used to establish the existence of a linear relationship.
Repeatability	The extent to which independent users can produce the same data or output.
Representativeness	The extent to which a document is typical of another document from the same context.
Research question	The target question a researcher employs in a study to analyse a theoretical construct.
Sample	The representativeness of the cohort being studied vis-à-vis the population from which they are drawn.
Scale and resolution	The smallest size that can be displayed (for spatial datasets).
Secondary sources	Sources that are based on the findings of others, such as those made available in academic journals.
Semantic accuracy	Quality with which objects are described.
Socio-technic Approach	An approach that has proven successful in its strive to bring technology to the uninitiated, particularly those in the social-domain.
Spatial distribution	The spread of values around a spatial mean.
Spatial information	Means by which information can be related to a specific position or location (Shand and Moore, 1989).
Spreadsheet	The electronic version of the graph paper. Composed of multiple cells in what are described as rows (records) and columns (attributes).
Standard deviation	A widely used measure to calculate the deviation (dispersion) of the data around the mean.
Statistics	Science that deals with the collection, analysis and interpretation of quantitative data.
Stereotype	An oversimplified conception of what an entity, person or group should be, based on a subjective interpretation.
Subjectivity	An interpretation process based on the results of emotional feelings and opinions as against objective and empirical study.
Surrogate	The substitution of one variable by another corresponding or similar variable.
Techno-centric approach	An approach that is mainly concerned with improving the technologies, sometimes to the detriment of the social and human factor involved in the implementation of the same results from this approach.
Temporal consistency	Repeated elements of the data handling process.
Tertiary sources	Sources that are not directly linked to an author or editor.
Triangulation	The process of research actuation through the employment of more than two methods which would allow double checking of the results.
T-tests	Employed for testing standard deviations when the population is normally distributed. It is a random interval or ratio sample, where the standard deviation is computed from the sample data.

Unimodal	One of two types of mode (the other being bimodal). Unimodal type has one peak with the highest point in a distribution, indicating the most frequent score.
Variance	Defined as the sum of the squared deviations from the mean, divided by n-1. It is computed as the average squared deviation of each number from its mean.
Versioning	This occurs when multiple persons use the same document concurrently and the software ensures that all edits are inserted.
Virtual reality	The digital world creation of an artificial reality based in a computer system such as the internet.
Visualisation	Visualisation with an “s” refers to the actual mental image that one can see within one’s mind.
Visualization	Visualization with a “z” refers to the process that occurs when one converts data into a graphic representation.
W6H	Parameters for social research: who, what, when, how, why, where, why not?
Z-Score	The distance that the sample value has from the mean (always in terms of standard deviations).

Imagery

Saviour Formosa

Front Cover

Ending

Ramon Azzopardi

Introduction

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Chapter 10

Chapter 11

Chapter 12

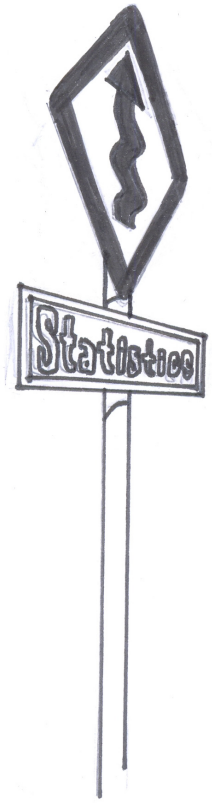
Chapter 13

Chapter 14

Bibliography

Appendix

Introduction



Clarke's Second Law: The only way of discovering the limits of the possible is to venture a little way past them into the impossible.

Arthur C(harles) Clarke

Profiles of the Future: An Enquiry into the Limits of the Possible (1962, rev. 1973), 21.

Why a book on statistics for the Uninitiated?

Twenty years since starting analysing data, 15 years since initiating research projects in a leading agency and 10 years since lecturing on research methodology, it still strikes the authors on how students, particularly those reading for social sciences, change their colours gradually to lighter shades on the recognition that they need to carry out statistical analysis!

Having realised the need for an introductory book on research methodology, the hope of understanding statistical issues without shedding tears is still a dream for many. If one aspires to be successful at tertiary level, one certainly needs to know how to conduct research. However, fear of numbers is more widespread than most may realize. The mere 'auditory tickling' of a numerical equation is enough to send persons into fits. Thus the short-circuiting of research into other than quantitative methodology!

The quantitative–qualitative diatribe is there to stay. Perhaps finding the adapt research method might not be a Herculean challenge, once one follows courses in research methods. However, admittedly, the choice is not an easy one and it is not rare for students to find themselves in a dilemma …bewildered and panic-stricken, before they even start their dissertations.

This book has been written with these students in mind. Yet, it attempts to bring statistics and research methods into the everyday realm not only of the student, but of the budding researcher and of the seasoned researcher/project manager. It is not an attempt to teach statistics and equations – there are hundreds of books specifically dealing with these. Instead, the aim here is to ensure that the reader understands what constitutes a statistic and what the actual research process is. It sets out to explain research methodology – the process that one needs to follow until, at the end, a numerical figure is reached and interpreted.

There are tuples, attributes, spatio-temporal issues, mind maps and a myriad of other items tackled in this book. It is hoped that this book's list of contents did not send any neo-researcher running in panic. Once readers start their research journey through this book, they should realise that every attempt was made by the authors to ensure readability and the presentation of research methods in a light-hearted, informal style. In fact, the authors strive to enable the reader to come to grips with the realities of statistics without fearing the worst. All that is required is an open mind, a willingness to learn, a strong doze of determination and a mug of coffee.

Aims and Objectives

- **Aim**:

  This publication aims to introduce statistics to higher education students who are not mathematically oriented and who have a perceived 'fear' of statistics. The book covers the basics as taken from the point of view of humanities/social and spatial sciences and delves into an 'operational process-aiding research'.

- **Objectives:**

  - To introduce statistics to the uninitiated
  - To discuss research methodology employing the datacycle and DIKA processes
  - To understand visualization, mind mapping and conceptual modelling
  - to understand database theory, information technology and systems
  - to review the available tools and experience their employment
  - to serve as a indicator for data sources

Chapter 1
What is Statistics?
An Intro for the Uninitiated!

I have yet to see any problem, however complicated, which, when you looked at it in the right way, did not become still more complicated.

Poul (William) Anderson

Quoted in William Thorpe, 'Reduction v. Organicism,' *New Scientist*, 25 Sep 1969, **43**, No 66, 638. In Carl C. Gaither, *Statistically Speaking: A Dictionary of Quotations* (1996), 187

Why Statistics?

Up to a few years ago, access to data was something nearing the impossible. Official statistics was available in state and national libraries while research outputs were only available in hard-to-find expensive books, mostly out of reach from most young researchers' pockets.

Three relatively recent upheavals changed it all!

I) The advent of Personal Computers in the 1980s
II) The launching of the internet and the worldwide web in the 1990s
III) The availability of raw real-time data in the 2000s.

From an era of hoarding and dearth of data we now live in a reality where the problem lies with unravelling the most-relevant dataset required for one's study.

Computers are here to stay. They will become smaller and integrated within all kinds of tools and appliances. Possibly, they will become almost invisible and unobtrusive. But in the past (say, for example 25 years ago) life was very different! There was even a time when people feared computers and predicted that these machines of artificial intelligence would replace human employees. Of course, this would have brought about a disastrously, soaring unemployment rate. Thankfully, these technophobic prophets were wrong. Their apocalyptic prophecy proved to be wrong. In fact, not only did not computers replace human workers but the advent of computers even created more jobs as, for example, in sectors such as gaming, social networking, research and academic work.

People today are comfortable with a visualisation module that allows them to interact with their digital tools. We now live in an era of digital networking, virtual lives and on-demand access to information. Away from a text-based operating system, we started getting WYSIWYG[1] awe-inspiring screens. Fast-forward a few years and one is frowned upon if terminologies are questioned: a mouse is not a cheese-guzzler, a monitor is not a warden, a keyboard is not a musical instrument! This reality was brought shockingly to the attention of the authors very recently when a 65-year old described a computer to them as: 'a television with a piano and a box'. Some of us never realise that it is not a box anymore, especially when one can run complex software on a 10cm Smartphone which carries a spatial locator (known as GPS for surveying), a camera for target recording, a microphone for interviewing, all the spatial and statistical tools possible and of course a phone…. pity mobile telephones have yet to start brewing coffees and baking cookies! Thus processing power and miniaturisation process initiated the radical change necessary to ensure that research be made more accessible.

The next step was even more interesting and had an impact on the social, economic and physical lives we live in: the internet! The internet changed it all! Actually the World-Wide-Web and a genius by the name of Tim Berners Lee did! Basing himself on technologies bringing together interconnected computers as far back as the 1960s within military circles, Mr. Lee created something that was not forbidding to the non-tech as well as easy to understand: an interactive and linked interface. Big words but a simple concept lies behind all this; a network of hotlink, which activate once one clicks on them to open new information nodes.

No longer restricted by mere text, users can now create their own online surveys, run online searches, carry out statistical analysis, employ graphing tools and eventually draft reports using such interfaces. The 1990s have yet to tell us their story on how civilisation was changed in a few years. From the fear of immersion as promoted in the 1990s film "The Lawnmower Man' to today's real-time Social Networks such as Facebook[2] and Second Life[3], the 'new' online world is beckoning researchers to perpetually take the next step.

---

[1] WYSIWYG – What You See Is What You Get
[2] http://www.facebook.com
[3] http://secondlife.com/

The third pivot was brought about by such legislative tools as the Data Protection Act[4], the Aarhus Convention[5] and the Freedom of Information Act[6], which have enabled both the solidification of ethical and regulatory issues as well as provided access to information way beyond the dreams of researchers in the 1990s.

This said, the main problem today is one of access to information. The issue is not necessarily the result of sourcing of data, since that has been solved by both the transmission mode (internet) and access issues (legislation), but more of understanding what the data being used stands for: what does its metadata hold, where does the data reside, what is its currency, can it be compared to other data? There is much one needs to know before delving into the usage of said data or information. The problem encountered today is more one of too much data as against a situation where little data was available. Today, it is more a question of how to interpret such data than where to access it from.

The Tower of Babel Syndrome or Valhalla?

Should researchers today face the reality of having to deal with vast amounts of data? This is a reality that cannot be ignored in information processing. Society is today facing a situation unprecedented in history, where data and information is readily available, it is easy to decipher and is accessible to all…

Or is it?

Are we going down the Babel way?

Technologies used by researchers today face the same problems as the ancient architects who got together and created the brightest inventions and plans for the tallest tower. There is a problem that data may not be readable, not comparable, not of a reliable format, not current and/or does not follow standard regulation. If the data is not decipherable, not comparable, not of a consistent format, not up-to-date and does not adhere to standard research rules, then research would crash as its output would not be credible.

The tower of Babel was not concluded since at the end there were too many languages and they could not communicate between themselves, which situation eventually killed the project. Data faces a similar dilemma… too much data, easily-abused choices of statistical measures, and over-reliance on online data and technologies can all lead to non-credible outputs.

Tower of Babel by Lucas van Valckenborch in 1594[7]
Source: http://commons.wikimedia.org/wiki/Main_Page

---

[4] http://idpc.gov.mt/ (The Data Protection Act of 2001 was enacted in Malta on the 14th December 2001)
[5] http://ec.europa.eu/environment/aarhus/
[6] http://europa.eu/legislation_summaries/environment/general_provisions/l28091_en.htm
[7] http://en.wikipedia.org/wiki/File:Tower_of_Babel_cropped_square.jpg

The 'fear of stats'

The study called statistics calls for red-eyed rimmed nights for many a student who has few notions that numbers can be fun! Having seen troops of students calling on "San Pawl tal-iStatistika ghini" (Saint Paul of Statistics – if ever there was one – help me) to "Madonna tal-hlas ehlisni" (Our Lady of Parturition grant me release) just before an open book exam always mystified us. The *fear of stats* was there 20 years ago in our student days and still seems to be one of the major phobias afflicting current students reading for the social sciences. The same idea that students can run statistical analysis has yet to filter through in some disciplines, particularly due to the entrenched schools of thought leaning towards the qualitative, which unfortunately has not helped students to realise that the difference between the two is steadily growing leaner, particularly since the import of new technologies made available tools that help students understand qualitative aspects through quantitative ones, giving a better understanding of the movement towards triangulation studies, where both methods are employed, aimed at enhancing the final study result. One could argue that the fear of stats is unfounded since the main target should not be centred on the numeric and statistical measures (forming the dreaded equations!) but on the process of getting there.

The scope of this publication is not to review those same equations (though a few have been inserted for ease of reference) – there are numerous books to this effect. Some research methods books are even area-specific. For example, one finds 'statistics for sociologists or psychologists' with others taking the further step of concentrating on the topic and the particular software for that same theme. Echoing the notorious ancient Chinese curse: researchers today live in highly interesting times! This publication strives to take an abstract view of the whole research process as well as to aid students in their research journey.

Myths and Realities

There are:

LIES
DAMNED LIES and
STATISTICS[8]

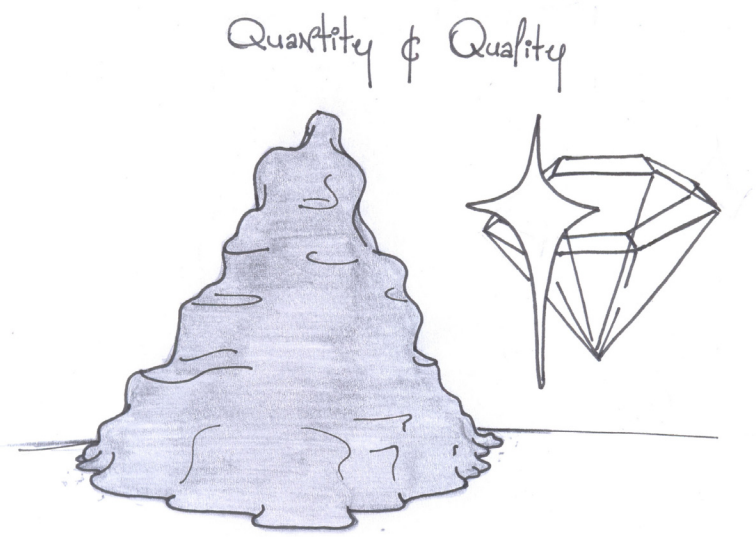
…. so the adage goes and one need only take a look at the phenomenon of media and bloggers' interpretation of what numbers represent. Statistics has fallen in the same dilemma as that perhaps faced by a person holding a glass of wine half full: "What makes a glass half full or half empty?" Statistics has been driven into a situation where numbers have been given a life of their own and have ended up representing themselves rather than the thematic target. Society has become inundated with huge volumes of raw numbers, data and information snippets. Overload is surely due! And in turn, this situation has led to the use and abuse of statistics.

Mythology has become the fulcrum for statistical review with numbers isolated from the reality within which they reside, percentages are drafted without any reference to the absolute numbers, surveys are misquoted, in cases emphasising the importance of a minority over the majority (of a 10% having more weight than a 90%), of a sample stating that 50 questionnaires represent the whole nation or that it is easy to run a survey, often starting from the end rather than the beginning (actually drafting a questionnaire before conceptualising the whole study)…

The realities are highly different: research methodology follows strict rules and that is why it is termed a science. Banking on such steps across a wide range of disciplines, research methodology has become the foundation of solid physical and social dimensions. From Durkheim's Rules of Sociological Method to Codd's database rules, from simple summing equations to complex modeling structures, from base data to action processes; such has been the journey travelled.

[8] attributed by Mark Twain to the British Prime Minister Benjamin Disraeli (1804–1881)

Chapter 2
Research Methodology



I told him that for a modern scientist, practicing experimental research, the least that could be said, is that we do not know. But I felt that such a negative answer was only part of the truth. I told him that in this universe in which we live, unbounded in space, infinite in stored energy and, who knows, unlimited in time, the adequate and positive answer, according to my belief, is that this universe may, also, possess infinite potentialities.

Albert Claude

Nobel Lecture, *The Coming Age of the Cell*, 12 Dec 1974

A number of decisions need to be made, prior to conducting research. The first step would be to decide on the topic or area of study. While doing this, keep in mind two things:

1. That the area must interest you – you will spend a number of hours for months or years reading about the topic; and
2. That you have access to the subjects which you wish to study.

For example, if you would like to research the programmes that are available in a particular prison to help prisoners re-integrate into society and you have no access to that prison, then this research would not be possible. The topic interests you, but you have no access to the subjects.

Often, inexperienced researchers have a problem in narrowing down their area of study. This might be the consequence of two main factors: fear of not finding enough information and failure to acknowledge the complexity of the problem. Let us say you want to research young people's behaviour during the weekends. This research question is very broad. You will need to narrow it down by deciding the following: which type of behaviour would you look at (e.g. criminal, sexual, drinking and driving, taking drugs etc.); which locality would you be analysing (it could also be necessary to narrow down the locality); and during which weekends (e.g. summer/winter/feast season etc.). All these variables will influence your results.

Next on the list is the research question. The research question can be formulated using two approaches: the deductive method or top-bottom approach, and the inductive method or the bottom-up approach. The deductive method uses theoretic interpretation and logically interpretive propositions to start the research. Conversely, when the researcher uses the inductive method, the researcher uses the observation of reality without any theory. What the researcher does is construct a number of indexes on which the theory will be built. These methods are used in everyday life as well. The deductive method starts with *why* certain behaviours occur moving to *whether* they will occur, while the inductive method moves from the *whether* to the *why* (Maxfield and Babbie, 2006:21). For example, in deductive reasoning you will reason that, since the Mediterranean climate exposes Mediterranean countries to dry summers, it will not rain in July in Malta. Alternatively in inductive reasoning, one would reason that as it has never rained in July in Malta, next July Malta will get no rain.

The Research Design:

The classical research design starts by **studying the existing theories and works** on the topic. You should consult the most recent documents. You should have a mixture of books and journals. Remember that journals can be richer in material and more actual as a journal article takes less time to go to print.

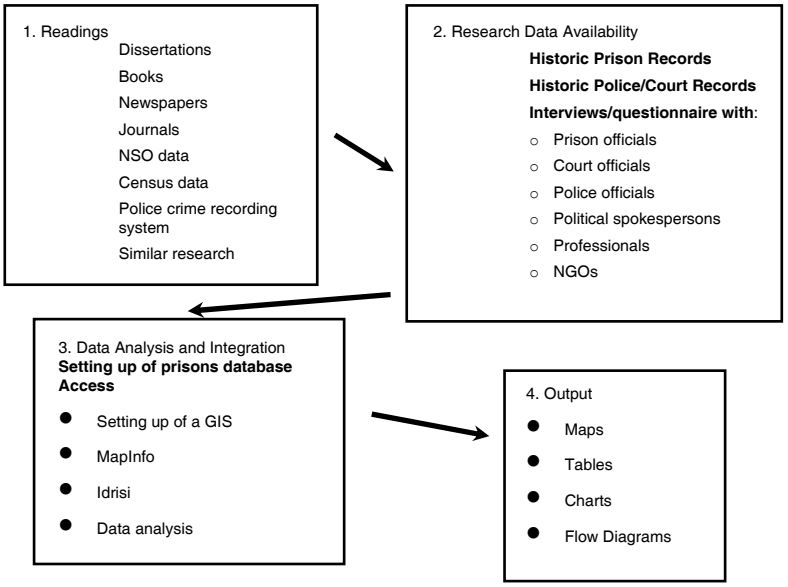
Once you have read as much as possible about your topic, you will now be ready to **define your study**, delineating your aims, goals, time-constraints, costs, human resources, outputs, etc. Depending on how big your research is, you might need to have interim aims and goals in order to check your process. Once your study is defined you should start on your **review of the literature**. This is important because it grounds your work in theory and helps you **formulate your hypothesis**. A hypothesis could read "Men are more likely to be prosecuted for violent crimes than women". This hypothesis will be tested in your research. It could be accepted or rejected. Rejecting a hypothesis does not mean that your research is invalid. If all hypotheses were accepted there would be no reason for research.

Once the literature review has informed your hypothesis it is now the time to turn to your **research design**. You need to assess which type of research tool would be best to help you in accepting or refuting your hypothesis. If we take the above hypothesis, our best method of research would be to look at court data as this would reveal the type of crime and the gender of the offender. You should ask yourself: "If I opt for a different research design, would I get all the necessary information?" Let us say that you opt to use interviews with police officers, members of the judiciary and lawyers. What you would discover is their perception on men/women and violent crimes but you would not be able to conclude if the statement is true or not.

This leads you to the next two steps which are **data collection and data analysis**. Data collected must be analysed. In case of a large quantity of data, this is usually done through statistical analysis. In smaller quantities this could be done manually. All or part of the relevant data gathered needs also to be

interpreted. The final stages of any research include the **drafting of the report** and **the presentation of results.** Figure 2.1 gives us a graphical representation of the research process.

Figure 2.1: The Research Process



Social Scientific Research Methods

Social scientific research or empirical research can be defined as a method applied to understand social reality using logic and observation (Hagan, 1997). In this definition there are three important words – method, logic and observation. These are interrelated. Your choice of methodology needs to be a logic choice, while any observation you do needs to have logic behind the techniques you are utilising.

Research is based on objectivity. In any study, the researcher must be aware of any inherent biases that exist. It is not a good idea to choose a research area in which you are emotionally involved. For example a researcher whose life-partner is undergoing drug rehabilitation, should not try to analyse drug rehabilitation programmes, because she/he would be too close to the situation to be able to assess the state of affairs objectively, in an unbiased way.

Empirical research attempts to either prove or disprove a conclusion about human beings. Social science research establishes trends and fashions in society. Any conclusion is not set in stone as it is more a probability than a fact. For example, after conducting research we might have concluded that a child who grows up in a criminogenic neighbourhood is likely to become a criminal. We are not certain of this conclusion, and the child might grow up to be a law-abiding citizen. As human behaviour is based on free and individual choices, social science conclusions can never be 100% accurate, contrary to the natural sciences where conclusions are based on hard facts.

In empirical studies, conclusions are based on interpretation. This is a subjective analysis of the data acquired. This subjective interpretation is the major difference between the social and natural science and represents a stumbling block for social scientific research. Interpretation makes us see why people take certain actions, whether the information is accurate, which policy can be formulated and we can then predict the most useful method to be used in achieving the aims of the proposed policies.

The data is presented in the form of graphs, numbers, lists and maps. Sometimes it is necessary to omit certain data. This is a subjective decision which could result in biasness as the researcher must provide the readers with all the available data so that they can reach their own conclusions. The idea is to allow readers the freedom of making their own interpretations. However this is only possible in an ideal situation. Subjectivity must come into play therefore and some form of evaluation is necessary. The researcher must define which data is correct and necessary. To enable an in-depth study of the data

chosen and to get a wider view of things, the researcher must focus on certain data and ignore other. Thus it is important to collect and choose the right data.

Research Problems

Any researcher must take the following rules into account when conducting a research:

1. The Rule of Reliability – This is secured when repeated measurement of the same thing gives the same results.

2. The Rule of Validity – This refers to the actual method. The question to be asked is: "Does this method actually measure the concept under study?" This rule asks you to consider whether the discrepancy between your operational definition (i.e. the problem you have set out to research) and the theoretical definition (you will find this in the literature review) which you started with, are comparable. If they are different, you need to adjust either your theoretical framework or your research tools. One must never forget that the data is not always valid (Reeve, 1997).

3. The Rule of Credibility – This refers to the fact that questions must: be directly related to research and that they make sense (are credible). Overall, if this is secured, the research tool looks good. This rule is extremely important in archival data. Some researchers contend that one of the most famous forgeries portrayed to be true are Hitler's Diaries. In 1983, the *Stern* magazine published parts of the so-called Hitler's diaries, which were later revealed to be fakes.

4. The Rule of Causality – This rule points to the fact that one thing may affect another that is not present. For example if we are studying age and crime we have to keep in mind that, as one grows older, one may tend to commit less crimes and that this has nothing to do with any policy implementation that might be going on.

5. The Rule of Representation – In research a sample is usually chosen. Census researches are very rare because they are time consuming and costly. This leaves us with having to choose a sample for our study. It is important that our sample represents the whole population. For example if we take a random sample of about 1,000 people from the Maltese Islands, then the results obtained would be true for a large number of the Maltese population. Certain studies are not representative. This is especially true if we choose to conduct studies under controlled conditions (experimental studies). These represent only the people undergoing the experiment. No generalisations can/should be made in these studies.

Sampling

When we are conducting a study we very rarely have the time and resources to ask each individual of a population for his/her input. Therefore we resort to sampling. Explained in very simple terms, before we dive into a pool, we don't check the temperature of every water molecule. Instead, we just immerse our big toe in the water. In most times, we quickly retrieve our toe and claim that the water ... all the water ... is cold. In other words, we sampled the water and took it for granted that the volume of water we immersed our toe in represents all the water in the pool. That is our sample – our representative sample. The most scientific sampling is one in which each person has the same chance of being chosen. This is called **Simple Random Sampling**. In random sampling your data can come from different sources, e.g. the electoral register, the telephone directory, computer databases and so on. The more comprehensive the database, the better the results. Samples are representative of the population from which they are drawn.

When we chose a sample there will always be a sampling error. This happens because in the research not everyone was included. The greater the sample the smaller the sampling error, however do not make the mistake of thinking that by doubling the sample the sampling error will be halved. You will need to quadruple the sample for this to occur. A sampling error of 5% is acceptable. That is, your results are accurate ± 5%. Other causes of error, such as wrong or untruthful answers, or missing data, do not help to eliminate entirely the sampling error.

It is not always desirable to have a total random sample. This happens when you need to target a group of the population. For example if you want to see what type of leisure activities youth are engaged in, you should used the so-called **Purposive or Systematic Sampling**. In this case you would limit your target group to youth only. From that group you would randomly choose your sample. If you had used

the random sampling method, most of your sample would not have been able to give you an accurate answer.

In other cases you would need to use the **Cluster Random or Stratified Sample.** This is used when you have to choose a number of people from the same place to minimise costs. This is less accurate but effective. This is useful when you have large countries like the USA.

Sampling is usually used to get a representative answer of the population. However, sometimes one has to adopt a **Disproportionate Stratified Sample.** By now you are saying it does not make sense to go for a sample that is obviously non-representative when you want your results to represent the whole population. Consider a situation where you have small pockets in the population that risk not being picked up in a random sampling exercise because of their small numbers. Let us take the example of police officers in England and Wales in 2008. There were 103,565 white police officers compared with 1,087 minority group police officers. The latter group is further divided into the diverse minorities. If the researcher wants to give a voice to the minority group it will be necessary to force a disproportionate number from the minority group officers into your sample.

Snowball Sampling is used when access to the sample proves difficult. If you decide to study the life of organised crime bosses you might have some difficulty approaching your research subjects. You certainly cannot access a list of the bosses with their home address from where you can choose your sample randomly. The method that is usually chosen in these cases is one of snowball sampling where you ask your first contact to put you in contact with other bosses and then ask each new subject to do the same until you reach the desired number of subjects or you run out of contacts. This method is not random but in certain cases, it is the only possibility to study difficult-to-access subjects. We suggest that you stir clear of studying such bosses, but this method is used to study certain populations such as prostitutes, ex-inmates, drug addicts on the street and so on.

Causality, Association and Correlations

An important factor in research is causality. Causality happens when a change in “A” brings about a change in “B”. For example, there is a change in sentencing policy - people who commit a second crime are sent to prison. This will result in more people in prisons. Therefore a change in policy (A) has caused a change in rate of imprisonment (B). Here, causality is said to have occurred. However, sometimes it is difficult to establish what happened first…the so called “chicken and egg situation”. Let us take the above example. It appears that prison rates increased due to the change in policy, however other variables could come into play. The questions to ask is what happened before the policy change. Was there an increase in crime? Was there a surge in unemployment? Was there an influx of people into the country or province? All these could explain why the imprisonment rate increased. All these are called variables.

Variables are vital to determine causality. Variables must be transformed into measurable and observable concepts. Variables are divided into two types: independent and dependent. Independent variables are those variables that are affecting the dependent variable. Therefore a change in the independent variable will produce a change in the dependent variable.

For example – High alcohol intake leads to high violence rates.

Independent Variable – Alcohol

Dependent Variable – Violence

In this example alcohol is the independent variable because the higher the alcohol intake the more violence occurs (dependent variable), i.e. a change in alcohol consumption effects violence.

In variables there are certain things that should be looked for before concluding that one variable is affecting the other. The **Temporal Order Rule** is the rule that applies when the cause precedes the effect. Another problematic issue is to **Rule out Alternative Causes.** For example before drafting a housing policy, one must be sure that the main factor of a rise in the demand for more houses is an increase in resident population and not foreign buyers. In normal social circumstances there are rarely cases where only one independent variable (cause) influences a dependent variable (effect). These are called **Multiple Causes**. In these cases one must see which variables have the greatest direct effect and look at the total greater effect (both direct and indirect) on the dependent variable. For example an increase in the demand of medical care can be linked to an aging population, a more sedentary

population, an obese population, an educated population and therefore an increase in the demand of preventive medicine, a virus, risky behaviour and so on.

Another example of causality is the following: an increase of two households in a hamlet brings about an increase of two cars. This is a direct **perfect positive relation**. However, we can have an association of the variable. This happens when a change in "A" brings about a change in "B" but the change may not be in the same direction, and this would be a **perfect negative relation**. For example an increase of two households in a hamlet may bring about an increase in cars; alternatively, the number of cars may remain the same or even decrease, because, since there is now a lack of space, people have decided not to buy cars. In some cases there could be no relation between the variables.

Perfect Positive Relationship	**A↑ B↑**
Perfect Negative Relationship	**A↑ B↓**
No Relationship at all	**A↑ B↑↓**

Correlation is said to occur when a change in variable "A" brings about a change in variable "B", however this relation is not necessarily even. For example an increase in "A" of 2 units brings about a change in "B" of 4 units in the same direction - an increase of two households in a hamlet brings about an increase of 4 vehicles (2 personal cars and two tractors) in the hamlet. We have seen an increase, but to a greater extent.

When one is conducting research, one often needs to consult official data (unobtrusive data). Here, one may find some problems to access the data. Furthermore, the researcher has to work with what is available. Sometimes this may not be the data the researcher actually needs but remember that this data was not collected for the purpose of the researcher's needs. At other times the data which is officially available may not be comparable with the data the researcher was collecting. For example if one is researching crime victimisation one might have asked about larceny, however the statistics held by the Malta police force will not have larceny listed as it does not constitute a crime under Maltese law.

Methods of Research

Social science research could be divided into two categories: the qualitative approach and the quantitative approach. Jupp (1989:28) sums up the difference between the two. He explains how quantitative research "assign[s] numbers" while qualitative research "report[s] observations".

Qualitative research takes small samples and uses in-depth methods of study to generate the data. An example of qualitative research would be in-depth interviews where the researcher interviews a small number of people to analyse a social reality. This method is usually preferred when dealing with a difficult-to-access population (for example terrorists, prostitutes and murderers). In the quantitative approach, one bases the research on a large sample. This would yield statistical data which is usually analysed through a statistical package tool such as the Statistical Package for the Social Sciences (SPSS). The choice of the method is based on your type of study.

There exist a number of different research tools in the social sciences. These can be used either alone or in a combination, depending on the researcher's needs. For example, if you decide to study the *Sette Giugno* riots in Malta, you would be combining archival research (which is research that is historical in nature) with case studies (an in-depth study of a particular case). The techniques adopted by qualitative research vary from observations, to interviews to documentary analysis.

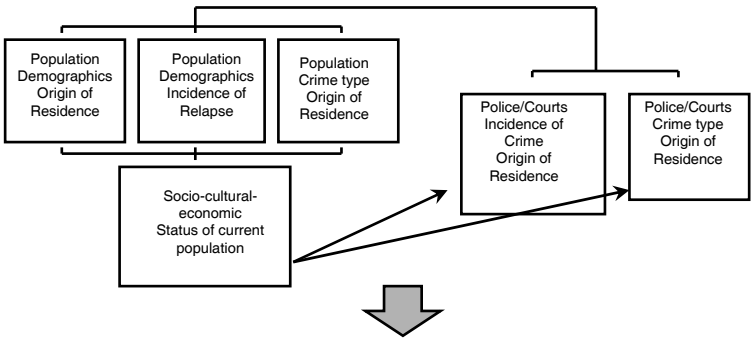
Archival Research

Denzin (1970, cited in Macdonald and Tipton, 1994:199) maintains that triangulation is of paramount importance for archival research to be valid. The first type is data triangulation. Data needs to be triangulated for time, place and author. All three variables need to be coherent. The second type of triangulation is investigator triangulation. This means that data from different sources should be used to confirm initial findings. Denzin calls the third type of triangulation 'theory triangulation' where theories are used to interpret a social phenomenon. The final triangulation he calls 'methodology triangulation', which refers to the method used.

Figure 2.2 gives you an idea of the complexity of historical research. It is important that when you are analysing historical documents you look at more than one source of information.

Figure 2.2: 20th Century Historical Demographic Analysis

Corradino Correctional Facility



Main Issues of Crime during 20th Century
(to research in relation to periods of specific crime duration and measures of control)
e.g. thefts, murders, drugs

Another variable in documentary research is found in adductive inference (Josephson and Josephson 1994:6), where one can construct a theory. We use adductive inference in everyday life when our observations and past experiences are used to generate meaning. Adduction inference is also used to support our historical knowledge. For example if you are conducting historical research and you spot that a prisoner's age as 2 years old, it is an obvious mistake, and you therefore point it out as a copying error. Adductive reasoning can also be used to create a theory. Although you take the historical documents at face value you have to also subject them to critical analysis. For example if you are analysing an official document created for public consumption, it may portray a picture which the government of the time wanted to show to the public. Therefore when an obvious skewed picture is being portrayed, it should be pointed out. Adduction is the whole process of generating information, criticising it and the possible acceptance or rejection of the hypothesis (Josephson and Josephson 1994:9). The best definition for adduction is 'finding the best explanation of a set of data' (Josephson and Josephson 1994:157).

The aims and purposes of the original writer are of paramount importance and require elaboration (Scott, 1990:13). Official documents are created for a particular purpose and not for future research. Hakim (1983:489) divides official documents into three types: routine, regular and special. Routine documents are central in administration. They form the backbone of the bureaucracy and they are usually extensive, consistent and factual. An example of prison documents that fall into this category are the admissions registrars. Regular documents are those produced for everyday purposes. Their use is purely internal. They are usually less important than the former, but they aid routine work. Special documents are those articles produced for a specific reason, such as annual reports. These offer accountability and transparency by being public.

Eley (1980:60) points out that a critical point in archival research is that, often it generates facts without interpretation. This happens because the huge amount of material one discovers in archival research makes it somewhat difficult to integrate it with theory. A common mistake that inexperienced researchers do is that they get lost telling the story and forget to interpret and analyse the data. History is made through the accumulation of pieces of valid and reliable data that leads to the construction of a story of historical narrative (Parker, 1980:422), but this does not explain change over time. Historical research is both chronological and topical. For example, Marx's political and economic determinism is well suited to Northern Europe of the 1840s, but further theorising is necessary as the focus moves to another place and another time (Parker, 1980:424). The most appropriate approach is to avoid narrative history (telling

the story), avoid the so-called Annales group (who tell the story but notice contradictions) and avoid the Marxist approach (focusing mostly on contradictions). The judicious by constant use of theory allows 'the interrelating of the story of human beings in everyday happenings and events with the movement of ongoing variables and structures' (Parker, 1980:428).

The status and standing of the archive material has four sequential dimensions (Scott, 1990:6-8). The first relates to authenticity and includes verification that the documents are original rather than fraudulent. The second relates to credibility, including an assessment of potential and actual sources of error and distortion. The third relates to representativeness, or whether a given document is typical of another from the same context. The final task relates to the attribution of meaning. After assessing the documents another problem is whether some documents have been distorted or if there had been some storage problems (Scott, 1990:8).

A further problem in archival research is access and restrictions in seeing the data. Not all data would be available for research. For example Maltese law makes official and personal data only available for research after a certain number of years have passed from their collection. For example prison data has a 30-year moratorium on official ledgers and an 80-year moratorium on personal data.

Case Studies

A case study is an in-depth study of a particular site, individual or occurrence in order to find some common interpretation or principle (Johnson and Christensen, 2008). In these instances an event or individual is studied over a period of time. There are no set rules but the researcher takes notes. Usually notes are taken keeping a format in mind. Choosing the case to be studied requires some thought. It could either be a particular individual – i.e. an individual who has certain characteristics or has done certain things; characteristics and experiences which you will not easily find in others. It could also be a particular event that has had an effect on society. The term 'case study' could also refer to a cluster of events, people or sites. For example you could decide to use the case study approach to study prostitutes. In this case you might be able to contact five prostitutes who will be willing to participate in the research and analyse their lifestyles.

Case studies are usually classified as qualitative research designs. Although this is true because of the small number of cases studies, many research strategies use sophisticated statistical analysis (Yin, 1994 cited in Maxfield and Babbie, 2006: 145), such as Nvivio, to link the variables together and create data sets.

The single case study is problematic when it comes to generalisation.

Surveys

Survey research can be divided into two – interviews and questionnaires. While in interview research the researcher asks questions orally (either in person or by telephone), in questionnaires the researcher presents the respondents with a set of written questions and expects them to give an answer (Harris, 1998). Interviews are usually classified as a qualitative methodology while questionnaires are seen as a quantitative tool.

Interviews

Using the interview as a research methodology is useful when one wants to get the story behind a person's experience. This methodology allows the researcher to dig out hidden or in-depth information about the subject. The questions used are usually open-ended as these elicit the most information.

Before you start your interviews you need to plan what you will be doing. After identifying your sample (or interviewees) you need to plan the interview. Interviews could either be totally planned (i.e. structured) or semi-structured. If you go for structured interviews you will have a set of questions which you will use, in the same order, with every interviewee. In a semi-structured interview you will construct a general outline of the interview. You will use this guide to ensure that you will ask the same questions to all interviewees. Another method is to have an informal conversation with your subjects. This is the least preferred method as there is a risk of forgetting certain questions and not asking all the information which you asked the others.

Once you have the interview schedule ready, you must contact your subjects, explaining to them the purpose of the interview, the confidentiality parameters and the possibility of stopping the interview at any time. The interviewee also needs to be told how long the interview will take, so that they can plan to be available during the required time. Remember that interviews are time consuming and that you might need to reschedule interviews because your subjects cancel the appointment, even at the last moment. Therefore be prepared and do not give up.

It is important that you chose your setting for the interview. Sometimes you will be constrained by the interviewee to go to their office. However if you can chose the setting, chose a quiet setting with as little distractions as possible and select a place you think will make your interviewee comfortable. You should once again explain what the aim of the interview is, confidentiality issues and that they are free to stop at any time. It is advisable to have a consent form which you will ask the interviewee to sign.

It is unlikely that you will remember all the information from the interviews; therefore you should at least take notes while interviewing. Interviews are sometimes recorded. This means that you will have to transcribe the interviews which have been recorded. When asking questions, ask one question at a time and give your respondents time to answer.

Certain questions are sensitive in nature. Be careful not to start with these questions as you risk putting your interviewee off. If you are going to ask anything controversial, two things are important. The first is that you ask about some facts before going on to controversial or sensitive material, the second is that you ask the question in a neutral manner. For example before asking about personal crime victimisation you start by asking a generic question about crime. This serves as a warm up question and puts your interviewee at ease. You will continue with a gradation of severity in crimes. Therefore you will ask whether they were victims of car vandalism, theft from cars, and theft of cars before asking about house burglaries, domestic violence and rape. The last question can be a generic question where you ask your interviewees if they want to add anything else and/or forward any suggestions.

When the interviews are ready, they have to be analysed. You need to look for common ideas, phrases, etc. and code them. This would enable you to pick out common trends and ideas. Many researches today use computer-assisted qualitative data analysis.

OPEN-ENDED questions – Questions which require an elaborate answer.

CLOSE-ENDED questions – Questions which have limited responses. The most simple require a yes/no answer. Others might require the respondent to make a choice out of several answers, while others still use a 4 or 5 point scale model such as: always, sometimes, rarely, never and I do not know.

Questionnaires

A questionnaire is made up of a number of written questions which the respondent is expected to answer. Questionnaires should:

- Be as concise as possible
- Include only indispensable questions
- Be thought out in a way as to allow the researcher to acquire the best possible information.

In an ideal world I would tell you to have open-ended questions which respondents will answer fully. However this is not an ideal world and using open-ended questions will drive you crazy when you start analysing and your respondents will give up when they are half way through your questionnaire.

The advantages of using questionnaires are that:

- They can be filled in the respondent's free time
- They can reach more people
- They can be filled in privacy
- The respondents remain anonymous

The disadvantages of questionnaires are:

- Response rates are low – about 30% return rate
- The respondents cannot ask for clarifications

Questionnaires should not take more than 20 minutes to fill. Questions should be clear, readable, easy to answer, stimulating and brief. Once drafted, questionnaires should be piloted (tested) in order to make sure that the intended meaning and the way people understand the questions are the same. Some questions would need revising while others would be discarded.

While constructing the questions (and this is also valid for interviews) avoid leading questions i.e. question which make your respondents agree with your statement e.g. "Would you support the government in the campaign against child abuse?" Such a question will invariably get you a "yes" answer. A better question would be "How efficient do you think that the government's child abuse campaign is: Very efficient, efficient, somewhat efficient, not efficient, and very inefficient".

You should also avoid including questions which have two statements, therefore require two answers as respondents might agree with a part of the statement and disagree with another part. For example: "Some women volunteer sexual favours to obtain career advancement. Thus policemen have learned to expect this sexual attention". Another common error is to use questions that are misleading, e.g. if you want to measure the likelihood of victimisation you do not ask the question "How crime-prone are the following: elders, middle-aged, youth and children?" as this measures the propensity to commit crime.
Avoid inserting questions that sound ridiculous such as "Executions teach a lesson even to others" and always check that your options match the questions. For example

A. Some people think that gays/lesbians should not be allowed to join the police force. To what extent do you agree?
   1. Strongly agree,
   2. Agree,
   3. Disagree,
   4. Strongly disagree
   5. I do not know

B. Some people think that gays/lesbians should not be allowed to join the police force. To what extent do you agree?
   1. Always
   2. Sometimes
   3. Rarely,
   4. Never,
   5. I do not know.

The options under B are not right.

When a researcher decides to use questionnaires, the aim is to get as large a number of responses as possible. Questionnaires are coded and the data is entered into a computer, usually using the Statistical Package for the Social Sciences (SPSS) software. SPSS enables you to analyse data using a Matrix (see Chapter 4).

The answers given in a survey have a meaning. Yes and No are the most straightforward. The answer "Depends" could be value loaded; for example, it could mean: "According to the time of passing of policy, law etc". The "No Opinion" category often means that the persons answering do not want to give their opinion, therefore one should divide this between the yes and no answers. Another problem is the missing values. The researcher is frequently faced with the question: Why did the respondent not answer the question?

Survey results need to be presented. The simplest method is to use frequency distribution – where a list of options is shown together with the number of respondents. In these instances N would refer to the total. Percentage distribution is a list showing percentages indicating options with 100% as total, while frequency and percentage distribution unites both frequency and percentage. Alone, they could give an unclear picture but when united, they give a complete account. In the following fictitious example, after one considers the table, one realises that if you were English you were much more likely to be granted a

pardon than if you were Maltese. Even from the numbers, one realises this. However, the percentages show this bias much more as shown in Table 2.1.

Table 2.1 Frequency distribution of the requests for pardon in 1941

	Refused		Accepted		Reformatory		Total	
	N	%	N	%	N	%	N	%
Maltese	75	87.21	4	4.65	7	8.14	**86**	**100**
English	11	55	9	45	-	-	**20**	**100**
Total (N)							**107**	**100**

In a survey, data is collected. There are four types of data falling within groupings called measurement scales: Nominal data, Ordinal data and Interval/Ratio data. **Nominal data** is data that fits distinct categories, for example Male/Female. It is the least sophisticated of the four, and only measures of central tendencies such as frequencies can be used.

Measures of central tendencies are the mean (average), the median (the middle point of the data) or the mode (the most common occurring value).

Ordinal data are ordered in categories for example: strongly, moderately, slightly, etc. An example would be the Likert Scale. It also includes groupings of data into cohorts (example 5 to 10 years)

The Likert Scale is a measuring scale used in questionnaires where respondents are asked to show up to what extent they agree with a statement. Respondents are asked to choose from a 4 or 5 point scale such as: I totally agree, I agree, I somewhat agree, I disagree, and I totally disagree.

Interval (and Ratio) data refer to the base data which allows for a way of grouping the data. An example is grouping individual ages of people to have a number a categories instead a whole list of ages. Therefore age groups could be divided into 0-5 years; 6-10 years, 11-15 years etc.

These data types are discussed under the section Measurement Scales in Chapter 5.

Surveys

This method has practically no interaction at all with the human target under study since it constitutes locational or remote data gathering processing. Thus, there are physical on-the-ground modes of data gathering and remote or desktop-based database surveying.

One can gather data in the field or from a desk-based study following initial snapshot recording.

The methods here are varied:

i) consider a surveyor who needs to review how many apartments are located in a specific street. One can go physically to the location and count the number of apartments, one can count the number of bells/intercoms/letter boxes in an apartment block or tick off against the list published online through a mapserver or an electronic register such as the electoral or postcodes online databases. The end result is the same: the number of apartments can be acquired, though one must compensate for those still under construction and those still vacant which would render the totals different in the online than the physical surveys.

ii) One can identify a location and wait for the target group to approach. An example would be the logging (counting) of the number of cars travelling through a major road, which data would serve as a surrogate for road-based pollution. The data gathering mode may constitute the ticking of marks on a paper or the click of physical counters or inputting in a digital hand-held mapping device.

iii) One can travel around a place and mark the vacant lots in a town. Such an exercise allows for mobility.

One can state that whatever the level of interaction, though restricted in a survey, does elicit the obligatory candidate who just has to come over and query the researcher on what his/her purpose is, why s/he is gathering the data and why that particular patch of ground…

Focus Groups

"Focus groups are group discussions organised to explore a specific set of issues" (Ketzinger, 1994:1) such as work problems, crime and health issues. Any topic could be addressed using this method. This type of research focuses on topics that discussion would help generate more information. This tool could be useful if, for example, a prison governor would like to address problems that prison officers face at work. Groups of 8 to 15 officers would discuss various prison officers' work problems and come up with suggestions. This method is not without problems. There is the problem of generalisation of the findings, if you are not addressing a small enough population where a census-like research can be carried out. A further problem could be that people will be reluctant to talk about certain issues in a group.

Group work is very important for "grounded theory development", where theory is generated through the use of the subjects' experiences. This is different from the traditional theoretical development where theory is first developed and later tested (Martin and Turner 1986). The use of focus groups has the following advantages:

- It helps the researcher identify the participants' priorities and language;
- It promotes discussion between participants even when the subject is embarrassing - usually reactions are more spontaneous;
- It helps identify group norms and the working of the group; and
- It helps people listen and reflect on each other's ideas, encouraging new ideas resulting in more accurate information.

Focus groups have some disadvantages. The major problem is the non-response rates, as those who choose not to participate, especially in sensitive topics, could turn your grounded theory upside down had they participated. Another problem is that the moderator of the discussion can introduce biases in the discussion procedures especially if he/she gives his/her opinion or does not ask all the questions. Furthermore well-trained and skilled moderators are not easy to find and their services are usually expensive.

In focus groups, analysis of the data is carried out in a similar method as in interviews and in participant observation.

Ethnography/Participant Observation

Participant observation is a qualitative approach where the researcher spends time analysing and observing a group. This tool was developed by anthropologists. By using the word "ethnography", the researcher indicates the wish to describe a cultural group. Participant observation can be conducted either overtly (i.e. the participants in the group know about your role as a researcher) or covertly (i.e. you hide your research purpose from the group and you portray yourself as one of them). The latter yields truer and richer data but it has huge ethical issues.

Participant observation is used in developing grounded theory. Grounded Theory is developed by using only field data. Contrary to the classical method of constructing theory, this type of theoretic development ignores existing theory and constructs theory from the findings. As data is continuously being discovered, this feeds the theory thereby changing the latter continuously (Jupp, 1996:59). Participant observation has numerous advantages, such as enabling the researcher to understand a reality which is foreign to his/her culture. McNeill (1994:83) claims that the major advantage is that subjects may be observed in their "natural setting" and that it enables researchers to conduct a "study of social process" instead of being restricted to a mere "snapshot or series of snapshots".

Mc Neill maintains that the main disadvantage of ethnographic research is that it is difficult for the researcher to remain detached from the situation especially in covert research. Furthermore participant observation is also difficult, time-consuming, expensive and unreliable. The major problems are that this type of research cannot be empirically tested and it is very difficult for the researcher to remain detached and unbiased. Another problem with participant observation is that the presence of the researcher might alter the groups' behaviour and by definition this research cannot be generalised.

There are various contenders to this throne but all have a common element: the observer is somehow involved in the process. There is no actual hiding behind the proverbial bush, but some interaction is carried out. This type of interaction can be termed as participative at various diverse levels. The next steps identify such levels of activity:

- **Uncontrolled/Naturalistic Observation**

  This method elicits visions of tropical paradise scenarios with bare-backed tribal natives being observed by the solitary be-moustached anthropologists in the sweltering, humid, mosquito-infested field. Anthropologists use this method to its highest sophistication levels. Jeremy Boissevan (1964, 1980) and Adrianus Koster (1984) employed it in the Maltese context, the former dedicating his life to the Maltese nation where his works spanned decades.

  Such methods might seem archaic in modern times, but the world is a large place and such studies can be carried out in such modern setting:

  Check out "People Watching" by Desmond Morris (2002) for a comprehensive read on such an activity.

  The method has been taken to new levels through the advent of the internet and the world-wide web with the resultant virtual worlds' observation.

  Stop for a second and think of activities that one can partake to on or through the internet that subscribes to this process:

  a. CCTV with live feeds on the net. (such can be used for both academic and also for practical studies as could be used by intelligence agencies to observe people movements). On the other hand, intranet systems (those confined to the virtual boundaries encompassing an agency network and not disseminated to the whole world) are normally used to observe against shoplifting (such as evidenced in most markets, malls and parks);
  b. Blogging and studying the comments by certain persons who periodically feel the urge to put in their penny's worth;
  c. Social networks, such as Facebook, where even if you have been invited to join a "friends" list and then observe the activities that that person writes about or does, such as going to events, writing about one's mood, interacting with other friends and a myriad of other day-to-day activities;
  d. Real-time 3D world interactions such as "Second Life" where one can move around different worlds and observe the interactions therein.

  Other networks can also serve as a tool for such activities. Bring up the reference to the ECHELON Project within the European Parliament 2001 Report (http://www.statewatch.org/news/2001/sep/echelon.pdf) which uses remote technologies to operate a 'global interception system' that observes communications around the globe.

  Though as yet still far from a 'big brother scenario", where all activities are monitored by the state in such dramas as Nineteen Eighty Four (George Orwell wrote that classic in 1949) the tools used through the employment of modern technologies are vast and open to debate on how sophisticated this data gathering method has become.

Complete Participant

This method enables the highest-intervention mode where the investigator becomes an intrinsic component of the target under observation. This is also consonant with full-immersion activities where the person becomes part of the same topic under investigation.

Think infiltrators
Think undercover
Think inside man/woman

The action is less glamorous than depicted in the media or in movies. It elicits a well thought out project, a long-term strategic approach and a high-risk approach to research activity. The approach requires the researcher to give up certain liberties and may even cause him/her to become part of the same problem/issue under investigation.

The principal point concerns the fact that the researcher must conceal his/her identity from those he/she is observing. This *modus operandi* also requires the full participation in the activities of group.

There are very serious issues that need to be tackled in this observation mode. These relate to ethical issues that impinge on the same study. This process may extract very sensitive and controversial points, which are dealt with in Chapter 14. This mode may appear as very unethical and deceiving as there is no informed consent.

As an example consider the case of a student who wished to use this method to observe an extreme religious group; a sect known for its non-conformity to social rights and obligations and which operates through extreme indoctrination. The student requests a guarantee from the university that after a specific period the supervisors would extract the researcher from that group. This cannot be granted due to the fact that it is not legal to extract someone from a group when that same person (an adult in this case) refuses to do so; practically the researcher has become one with the group and has rights under law to maintain that relationship even if she had signed an extraction letter years before. The process of extraction could become even more dangerous as that person may be exposed and placed in a potentially life-threatening position.

Not an easy case!

There is also the issue here that the period of observance requires vast amounts of time dedicated to unlearning or desensitisation to the activities that had become the norm during the period under immersion (full-participation in the group activity).

Participant as observer

This method calls for the researcher's full immersion in the activities of the target but goes a step beyond and identifies his/her role as a researcher.

This mode helps the researcher in terms of safety and in terms of ethical issues, however it has its intrinsic disadvantages. Among these:

i) the participants are aware of the observation and act accordingly within such knowledge thus contaminating the process. They act in order to impinge upon the observer those actions that they want the observer to record, and not others that show otherwise or even normal processes that they abide by;
ii) the observer can become subjective due to the same group dynamics that are resultant from team-building efforts, risking the research's validity and becoming prejudiced.

For example, if the research is about drug abuse during parties, the group under observation may choose to stay away from drugs while the researcher is around. Of course, the group could indulge in drugs again once the researcher is away.

Observer as Participant

This method enables the researcher to observe a group but the immersion is not total. The researcher identifies his or her identity to the targets and observes the group on an ongoing basis.

The interactions possible within this method are those deemed as constituting episodes that are both formal and sporadic.

An example could include participation in a sports group with interventions targeted around new sports rules where the researcher drops hints periodically covering the inclusion of new regulations for sports events, which the group then reacts to.

Complete Observer

This method, though apparently similar to the uncontrolled methods discussed earlier, does the same work but the target group is still aware of the observer's presence.

The researcher thus is totally detached from the study group and s/he observes their activities. However, s/he does not become part of that same group.

For example: observing youths' behaviour during the screening of a violent movie.

Questions (refer to Appendix for the answers)

1. What are the two main points that you should keep in mind before deciding what your research topic should be?
2. The research question is formulated using two approaches: the deductive method and the inductive method. Briefly describe these two methods.
3. List the nine main steps of research design.
4. What is the empirical research (social scientific research) method?
5. Good research is based on objectivity. Very briefly explain this.
6. Briefly describe one major difference that exists between the social sciences and the natural sciences.
7. Why is it very important for the researcher to collect and choose the right data?
8. List the five main rules that researchers must take into account when conducting research.
9. What do you understand by "sampling".
10. List the five main types of sampling.
11. What do you understand by "sampling error"?
12. What do you understand by "causality"?
13. When does a perfect positive relationship between variables occur?
14. When does a perfect negative relationship between variables occur?
15. What conditions enable the researcher to claim that there is a correlation between variables?
16. List the four main problems associated with using formal official data.

17. Very briefly describe the two main categories of research: qualitative and quantitative.

18. Triangulation is of paramount importance for archival research to be valid. List the four main types of triangulation.

19. Very briefly explain what you understand by “adduction” (with reference to archival research).

20. List the three main types of official documents (with reference to archival research).

21. Eley (1980) warns about a critical point in archival research. What is it and why does it happen?

22. Scott (1990) claims that the status and standing of the archive material has four sequential dimensions. List them.

23. List the two main problems associated with archival research.

24. Briefly explain what case studies are and state their main problem.

25. Survey research can be divided into two main categories: interviews and questionnaires. Briefly describe these two categories.

26. List three main advantages of interviewing research participants.

27. List the four main advantages of using questionnaires and the two main disadvantages of using questionnaires.

28. Why should questionnaires be piloted (tested)?

29. List the three main types of data and very briefly describe each one, even if by simply providing an example.

30. What are “focus groups”?

31. List the four main problems associated with conducting focus groups and list the four main advantages reaped by conducting focus groups.

32. Briefly describe ethnography/participant observation.

33. List the main advantages of conducting ethnography/ participant observation.

34. List the main disadvantages of conducting ethnography/ participant observation.

Chapter 3
DIKA

The modern research laboratory can be a large and complicated social organism.

J. Michael Bishop

How to Win the Nobel Prize: An Unexpected Life in Science (2004), xii.

What is DIKA and why should one bother with trying to make sense of a research process? Isn't it easier to just start gathering data and analysing the result? Why should one bother with trying to follow rules and regulation? Why do so many researchers use mnemonics? Such argumentation has seen many a study hit the brink and end up in the dumps.

Let's start off with a look at research and the process taken to understand the way one goes about using the fruits of research labour. This chapter takes the research process and identifies those thought and applied steps required to understand how and why a study is carried out. Consider DIKA for a few moments:

DIKA is a mnemonic with high-powered strength in research methodology. Consider its implications in terms of a report, a survey, a dissertation, a thesis or any other studies. The mnemonic is easy to remember and the explanations then fit in terms of one's particular endeavour as explained in Table 3.1.

Table 3.1 DIKA

	Meaning	**Description**	**Form taken**	**Research Parallel**
D	Data	The process required to understand what data is required, how to acquire it and how to transpose it to understandable forms.	Normally in numeric or coded forms. Main issue here is NUMBERS.	Consider this as the process employed to review literature and translate the establishment of a methodology chapter followed by the data gathering, inputting, cleaning and testing.
I	Information	The meaning given to that data. Requires a context within which that data resides, the context normally referring to the social activity, the physical reality, the environmental design within which that activity occurs. The context is king in information management: 100,000 persons in Malta is a veritable number but those same numbers are practically insignificant in China or India, thus effecting studies on population density, economy, migration, amongst others.	Textual forms but can be expressed in a wide range of formats inclusive of graphics, maps, imagery, 3d structures, and other forms. Main issue here is CONTEXT.	Consider this as the analysis and description chapters.

	Meaning	Description	Form taken	Research Parallel
K	Knowledge	The uniqueness of research that generates new forms of awareness, literature and scientific results, which can be used for implementation exercises and feedback into the same society from which and within which the results were sourced. The output should also enable the recommendations for future policy-making, decision-making and academic research.	Textual forms and presentation methodologies required for dissemination processes, whether analogue or digital. Main issue here is Translating the context into POLICY.	Consider this as the discussion, conclusion and recommendations chapter.
A	Action	The operational processes taken after the knowledge has been used to implement change in the society/entity under study. The action can be physical, political and scientific. It can be implemented at the different levels of administration whether at super-national, national, regional and local levels.	Translated/transposed to legislation at the different levels, which brings into focus the implementation of such entities as required to operationalise the separation of powers: The Legislative (decision-makers - Parliament), which passes laws based on policy-makers' recommendations; The Executive (Ministries and Departments/Entities), which implements the day-to-day running of the legislative's decisions; and The Judiciary (Legal Entities), which ensure that the operations follow the legislative protocols. Main issue here is IMPLEMENTATION.	Consider this as the post-document implementation of recommendations and review of results. Requires ongoing or period monitoring and maintenance.

Each of the elements of the DIKA process shall be dealt with later on in the chapter, however prior to delving into an in-depth study description of each element one needs to bring the focus back to the concept of research itself: what it is, why it is necessary, and how does one go about implementing it?

It is not easy to come to terms with why research is necessary. One can argue that some important discoveries are made accidentally. Cases in point are Newton's accidental resting-time under the apple tree 'caused' the **Universal Law of Gravitation** to come to fruition and the veritable comical "Eureka" moment when the Greek scholar Archimedes stepped into his bath and when the water overflowed, the bright spark discovered the **Principle of Displacement** and eventually ran naked in the streets of Syracuse.

However, both the above and many others strove to analyse their studies in measured and repeatable ways. They tried hard to ensure that their method was sound and only after many years of trying does uniqueness of research efforts shine through.

Why is Research Necessary?

While brainstorming exercises may bring to the fore an image of white-clad, bespectacled and wild-haired old men as nominal researchers, reality posits a different image that may be different than the stereotype.

Research is carried out for a variety of reasons stemming from the innate human need to discover new knowledge, the need to understand the mechanisms of the knowledge and to implement the findings of that knowledge. One may opt for all three modes but no researcher is bound by the entirety, since there may be specialists in each of the fields who concentrate on one of the factors in isolation. However the findings of each helps strategists to implement the action required in a constructive way that feeds back to the theorists and researchers.

The most important assets and issues that one needs to be aware of in approaching the long-road to research are:

- **Curiosity – wanting to know about a problem**

Curiosity is a staple for any research activity. Whether it initiates through the exploration of new activities to a pure human need to know, it is important for every researcher to be very aware that a healthy fascination with the world around us and beyond is maintained throughout life. This has to be coupled with knowledge of the process required to change that fascination into fact. Intuition and feelings have an initial phase but must give way to factual analysis. The discoveries encountered during and at the end of the study enhance further curiosity.

- **Problem identification – what is the issue under study and why is it a problem?**

The identification of an issue leads to the exploration of methods that ensure the investigation is complete and exhaustive. One must not turn issues into a layman's understanding of the work problem, but as an instance/occurrence that seeks to understand and provides potential solutions. One must not state that every issue is a problem, as this may imply that something is 'wrong' or dysfunctional. In effect, problem identification in research implies the need to understand functionality, linkages and pointers to solutions to the issue at hand.

In addition, one must not seek to carry out research that is aimed at simply proving one's hunches or that is directed at establishing one's theoretical dominance. Such is not science and should be eliminated from the process.

- **Development and testing of theories – pure and applied research**

Once the curiosity and the problem identification have been satisfied and given a structure for implementation, the third fulcrum looks at the underlying and established theories that have been developed and those that are emergent. The development and testing of theories is crucial for

research processing since the basis of such brings in the use of statistics through the data-cycle method. Pure and applied research follows their own trails but leads to the eventual understanding of the realities (physical and virtual) being investigated.

This brings into fruition the need to be knowledgeable of the W6H planets revolving around the research sun (CMAP[1], 2002). Each of these establishes where researchers should point to and how they should employ their study process.

- **W6H**
  - **WHO** (Target Group)
  - **WHAT** (Research Question)
  - **WHEN** (Indicator)
  - **HOW** (Modus Operandi)
  - **WHY** (Linkages)
  - **WHERE** (requires extra spatial queries but allows outputs to visualization tools for locational analysis)
  - **WHY NOT?** (Why not carry out a study even if it is brand new and totally ground-breaking or controversial? Cross-thematic links are investigated)

Having established some basic concepts, one can understand that there exists a need for rules to be established so as to ensure that a study can be replicated, that it can be measured and it can identify the linkages between the items under study, herein called variables.

BUT where does one start a study?
AND
What is research really?

To quote Dantzker and Hunter, 2000, p. 10, research is…

"The conscientious study of an issue, problem, or subject

… empirical research that yields scholarly results

… the scientific investigation into or any phenomenon linked to any or all aspects of a theme".

Therefore, in a few words research strives to understand a problem (a topic of study) through empirical research (using a scientifically sound method). The researcher must follow a few basic steps in order to initiate research and follow it through.

These steps are investigated throughout this book; however we have to start from somewhere:

i) Does the researcher have the drive to actuate such a study?
ii) Is s/he aware of the time it will take up?
iii) Is the data/information available?
   a. If yes, are there problems of access?
   b. If, no, what steps are required to gather that information?
   c. Can surrogates be used?
iv) Is the target group willing to cooperate?
   a. How will they be contacted?
   b. What type of methodology will be used?
      i. Group activity
      ii. Individual interviews
      iii. Questionnaires
      iv. Online surveys…

[1] http://www.nlectc.org/cmap/

v) Will software be used?
   a. Do I need a simple spreadsheet tool, complex statistical programme or a hi-end spatial information system?
   b. Is it available? And expensive?
   c. Are there alternatives such as online modules?
   d. Does it have a steep learning curve?
   e. Will graphics be used?

vi) Does one need to go back for clarifications post survey?

vii) Is the drafting language too technical?

viii) Do the results reflect the original aims of the study?

ix) Does the referencing conform to the established protocols?

These and many other questions need to be considered prior to initiating any study. The researcher must come up with a sequential process that makes sense of them, in a way, that one can follow through. What comes first: the questionnaire or the data sourcing: the mapping or the tabulation? Many questions but with a few initial steps that requires a researcher to follow such preoccupations will soon disappear and the processes will fit into place. Chapter 4 outlines the way forward to tackle such processes and also how one can follow basic steps in a research study.

Forms of Research

The next section outlines the forms of research that can be undertaken, mainly pivoted on three forms:

BASIC
APPLIED
MULTIPURPOSE

While each has its own distinct characteristics, the three forms are all important for research and one is not any more important than another as shown in the third form which actually integrates the other two.

Basic Form

Basic or 'pure' research is a form that drives towards an understanding of concepts, being more theoretical in form and is not targeted at the immediate provision of tangible results. This form of study is aimed at the acquisition of new information and with the development of the scholarly disciplines. It takes on a descriptive approach to research. The results are more conceptual in nature and may not reach the stage where the concepts start taking on a factual form and end up analysed through statistical measures.

Applied Form

The Applied research form has a more immediate or real-world timeframe as it is an inquiry of a scientific nature designed for and conducted with an operational and practical application as its goal.

Examples of Applied research would include the following:

- Policing – stress, patrol effectiveness, use of force, response times;
- Juvenile Courts – types of sentencing, jury versus judge verdicts;
- Migration – number of persons migrating, type of migration, age cohorts, natural balance;
- Environment – dust particles smaller than 10μm, wind direction, atmospheric pressure;
- … and so on.

The above examples depict a form of evaluative research, focused on answering questions that evoke answers dealing with the real world. Some questions which may be asked include the following:

- Is the programme effective?
- If not, how is it deficient?
- How can it be improved?
- Should it be continued as is, changed or discontinued?

Multipurpose Research

In a real world, however, research occurs in between both forms. While Basic and Applied forms may be seen as the two extreme forms of research, the multipurpose form takes on board both forms and ensures that both the conceptual and the factual are integrated into one study.

Multipurpose research thus aims to launch scientific inquiries into issues or problems that could be both descriptive and evaluative, both theoretical and empirical. The multipurpose form initiates its function as exploratory and delves into the exploration of operational and applicable results.

As an example one can bring up the following scenarios:

- Merton's Theory of Deviance and social cohesion in small states;
- Disorganisation Theory and modern service-industry giants; and
- Youth Services provision and job satisfaction.

The DIKA fits ideally within such a form but still banks on the findings of the Basic form in its conceptualisation function.

Types of Research

Having covered the forms of research, the next step is to understand the roads such a research could take, ranging from seeking between what something is, why something is, what something will/could be to what can be done in the meantime. Each type of research holds its own since the choice of type is dependent on the availability of information and pre-established knowledge. The intricate filaments established between the W6H elements mentioned earlier and the types identified here are needed for the researcher to understand which research method should be employed prior to establishing a research process.

- **Descriptive** – describes a situational construct in a time and space. This type allows researchers to understand what something is.

This type of study is carried out when information is scant either because there are few cases available or data is restricted. This study can also take the form of a qualitative or post-modern construct since the research pool is constrained to review only those cases/individuals under study which may not necessarily represent all society but only their own specific situation.

It is very difficult if not impossible to carry out statistical measures in such cases. However, referencing to similar studies that had established parallel knowledge-bases is to be encouraged as they serve as a basis until the numbers under study increase to enable statistical analysis.

As an example one can take the case where a study on incarcerated females who had been involved in the gambling industry results in very few cases that can be investigated.

- **Explanatory** – tries to tell us why something occurs, the causes behind it.

This method allows the researcher to explain behaviour and counteract the problem under study. It is imperative that researchers do not mistake the wood for the trees since no study exists without its linkages to the context it appears in, something emphasised throughout this book.

Explaining an action requires an understanding of the cause and effect, a realisation of the realities within which that activity occurs and delves into the sequential steps that led to the activity itself, as influenced by and influencing the linkages around it.

Imagine a situation where a number of persons have been incarcerated in a specific part of a town. What was the cause for such a high number of offenders compared to other areas? What are the linkages to the social construct? What happened to their children? These and other questions help elicit many queries on the relationships that would help explain why such an activity occurred.

- **Predictive** – helps to establish future actions by modelling the potential futures based on the testing of scenarios.

This method helps to identify what could happen in the future. To do this it is essential to establish what the present is and have a concrete foundation on the past. No predictive analysis can occur without a realistic understanding of the past, the present and the intervening state between the present and the future.

As an example a study on natural growth can only model growth/decline if past trends on fertility and births/death have been recorded. Another example could be found in a research on the utility of boot camps. How boot camps could be set up and their influence on future sentencing. This could be based on trends in juvenile crime, the outcomes of similar processes in similar countries/communities and the knowledge gained during other research activities.

- **Intervening Knowledge** – allows one to intercede before a problem or issue gets too difficult to address.

The most high-powered research with operational actions being taken up during the actual study process is one where the intermediate or initial results are deemed too sensitive to leave dormant until the study process is concluded. This type of intervening knowledge has immediate effect on the DIKA process since the action phase takes precedence due to the issues related particularly to safety and security across the social and physical domains.

As an example one can bring up the issue of viral research which results in the identification of a new type of pandemic-related entity which requires immediate intervention by the health entities.

Another example could be the identification of hotspots which show that crimes occur in areas that have darkened alleys and which with simple intervention from a local council can be mitigated quite easily through lighting and community policing initiatives.

In all the above the most important issue to keep in mind is that of **CONTEXT**. Any study has to be carried out in a specific time and space or across time and space which entails a full understanding of the social, physical, cultural, economic and structural constructs within which that study is occurring.

How Research Is Done

In the previous chapter we have experienced the research process. In this chapter we will expand on the steps mentioned previously. Irrespective of the form or type of research undertaken, whether applied or basic, descriptive, explanatory, predictive or intervening, we must follow certain first steps before embarking on research.

Step 1:

- **Identifying the Research Problem**

What is going to be studied? Will it be a qualitative or a quantitative study where different methods will be required? Has the research problem been identified beforehand or is it a totally new field?

All this requires a logical approach which is essential in all studies. An apprentice guide would follow the few steps identified below.

The 3x3 Structure

We call this the **3x3** structure: quite different from a 4x4 vehicle but it definitely gets you there in research terrain!

This preliminary listing will be further explained in Chapter 4.

i) **A**. - Create an AIM which should be expressed as a statement that shows the topic of study and the direction you wish your research to take.
ii) **B.** - The first of the 3x – create 3 OBJECTIVES based on the AIM

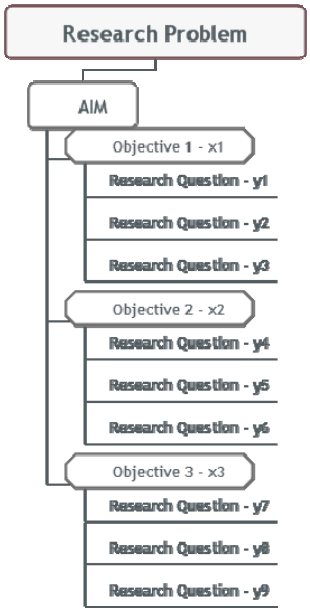
i. The first objective (x1) should describe that you wish to understand a specific topic through the literature review;
ii. The second objective (x2) should state what you want to achieve;
iii. The third objective (x3) should state how you aim to achieve those results;

iii) **C.** – For each of the objectives identify **3 research questions.** This section closes the 3x3 loop as it allows for the creation of up to 9 research questions. These could take the form as an example...

i. For Bx1 – the first Research Question (y1) could state that the researcher will investigate whether a relationship exists between the theoretical exponents and the realities that occurred over the last century;
ii. For Bx2 – the first Research Question (y4) could state that the study will aim to identify the linkages between one theme and another, for example whether there is a relationship between social disorganisation and delinquency;
iii. For Bx3 – the first Research Question (y7) could state that the researcher will investigate whether a relationship exists between the spatial data gathered and the demographics of migration.

Figure 3.1 depicts the above structure in a simple image.

Figure 3.1: The Aims, Objectives and Research Questions



One has to create a sequence of Aims-Objectives and Research Questions that ensure a flow between each section, as well as leading the reader to an understanding of the link between the literature read and the research. One must not fall into the trap that the literature is there simply for report-writing's sake but one must realise that it serves as the glue which provides cohesion to the research, yet allows it to flow. The literature review anchors the theory into space and time with the context under study.

Step 2:

- **Research Design**

Once the logical mechanism on how one is to proceed has been set out, the next step serves the role of identifying which 'ideal' method should be employed in order to actuate such as study. In fact, this process will lead to the creation of a "blueprint" for the study. However, one must also consider the numerous hurdles researchers are expected to surmount, namely data access issues, access to persons, and non-completing of data gathering process, among others. The ideal state should however be identified through the creation of a 'mind map' dealt with in Chapter 8. This helps the researcher to build an idea of what exists, what should exist and how best to come up with a method to identify the linkages emerging from the Research Questions.

This process is not easy to come by, especially since it requires knowledge of the different methods (Chapter 2 explained the different methods of research) one could employ, such as surveys, field work, experimental (observational), life histories or case studies, archival studies, record studies, content analysis (documentation), and others that may include remote and real-time data.

Step 3:

- **Data Collection**

The third step looks at the methods used to collect the required data. This is the most crucial of all the research work since it will take up more than 80% of all the time available for the research. Whether physical or online, whether analogue or digital, the process is time-consuming and seemingly never-ending. These are not exclusive to any method but concern all research types including: surveys, interviews, observations, previously existing data, new data sourced from machines and sensors as well as archival research.

Remember data collection can take up more than **80%** of the time allocated for your research!

Few first-time researchers realise this, and from experience most students and researchers initiate this process in the final quarter of the study, resulting in rushed analysis, discussion and reporting. For a 10-month study the data process will entail eight months for data gathering, data inputting, data cleaning, and overall data management.

Step 4:

- **Data Analysis**

The fourth step takes into account the post-data collection process. It looks into the diverse ways that one can employ to make sense of that data in conjunction with the findings extracted from the literature review.

This process entails the issue of interpretation of the data within the context under study, of the identification of the lacunae in data availability and how this will affect the analysis, the running of statistical tests to seek relationships between the variables and the eventual translation of those statistics into readable and understandable text.

This step also focuses on the employment of statistical and other research tools which aid the researcher to reach more informed and reliable results.

Step 5:

- **Reporting of Results**

The final step brings together the diverse findings from the analysis in line with the findings from the literature review and the context under study. The results cannot be isolated from both these fulcra as they ensure that the 'local' findings are formed within established methodologies as identified in other research studies.

The results can take various forms, mainly reports, articles in journals, books, computer presentations and other multimedia outputs. The introduction of distributed and worldwide sites, such as the internet and more specifically the world wide web, has enabled a more real-time and ongoing output in the form of interactive content. This has enabled users to query online data gathered from research processes and to carry out cross thematic analysis from anywhere in the world. This phenomenon combined with virtual libraries has transformed research and its structures. Chapters 12 and 13 give an overview of case studies and data sources, which include online data sources.

Techno-Centric or Socio-Technic Approach?

Having established the need to use diverse processes to deliver our research concept to fruition, the next immediate question which comes to mind is that of the decoupling between social sciences and the tools that are used.

Researchers may opt to use or ignore the tools currently available for research analysis. The debate in information circles concentrates on whether to take research into purely techno-centric approaches, possibly at the expense of the social dimension under study, or whether it should enhance the emergence of the socio-technic approach to encompass the requirements of the social dimension (Reeve, 1997).

The debate rages on how best to approach research tool creation on whether the process should concentrate on the development of technology and its delivery in isolation to the socio-physical dimensions as two track rails that only meet at the perceptual horizon. This ICT-based approach ensconces users as the secondary 'potential' market and may not be functional to the purposes of this book's targets.

On the other hand, the debate focuses on the need for concentration on the social implications of that technology which would ensure that any tool can be used by society at large rather than merely specialists in specific topics. This person-based approach with ICT as the 'conveying tool' has taken on rapid momentum and is evident in such data dissemination tools that have enabled immediate access to information otherwise too difficult to decipher. Take for example the issue of online maps (such as Google Maps[2] and Multimap[3] - refer to Chapter 7 for a comprehensive list), which have enabled users to interpret and generate data previous inaccessible in highly numeric and voluminous forms. This is veritable evidence that research has become more accessible to more people, and that the need now exists for formal training in the use of such tools to ensure that research is carried out only on that data that has a reliable metadata attached to it which follows international standards on reliability, sourcing, measurement, and other attributes.

This process has enabled most scholarly activity to realise that as people get more used to discrete and unobtrusive technology the move from techno-centric towards socio-technic will occur and in effect the research base will expand. This said there is again use and abuse of such tools, but researchers need to move beyond a fear of technology in the social sciences and need to investigate how technology can serve their purposes for quicker and more reliable results.

Use and abuse of statistics:

As mentioned earlier, there are "Lies, Damned Lies and Statistics", thus any data can take different forms that lead to vastly different outcomes to the study under construct. However, such malpractice can only occur through a deliberate effort to use the results available for one's own needs as against the needs of society at large.

Let us take a look at the use and abuse of statistics through a series of examples.

Use

i) The study of demographics – an analysis of population change aimed at mitigating an ageing time bomb. This use of statistics enables policy and decision makers to prepare for potential changes in a society's structure and plan ahead for any required legislative changes or initiatives. Take for example the case of Western states that have experienced a drastic decline in the fertility rate which in turn has caused a shrinking population growth that will result in fewer persons in the working-age cohorts.

Such a situation has resulted in initiatives to encourage people to have more children, commonly through financial incentives, the opening of immigration controls or quotas, the increase of the retirement age and various other measures.

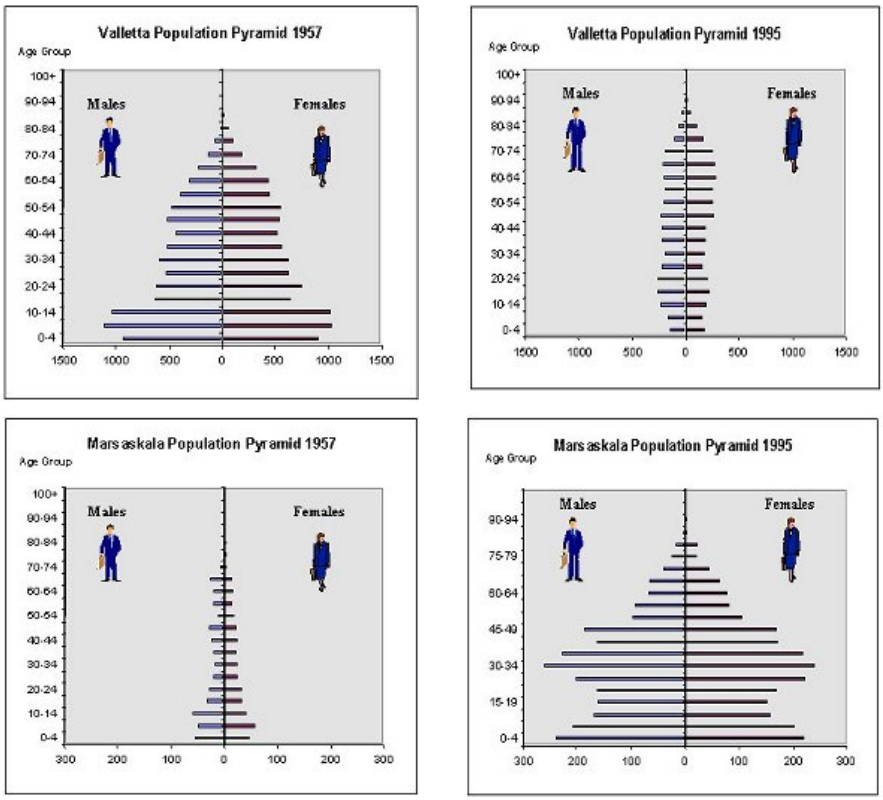
ii) Another use related to the study of past and current trends in population and housing which allows for the creation of reliable projections for future planning purposes could be identifying a figure on future trends on housing that would allow policy-makers to identify the land-take required, the number of units required, the slack that is needed to ensure population movements as well as plan for future services as the population in areas increases or declines.

---

[2] http://maps.google.com/
[3] www.multimap.com/

Take for example the population increase of Marsascala between 1957 and 1995 as against a decrease in Valletta's. (Figure 3.2). Use of statistics allows for projections even up to 2050 for national figures which in turn can help both policy makers and decision makers to take a strategic approach in their actions, rather than short-term or fire-fighting decisions.

Figure 3.2: Population Changes: 1957 and 1995 (Valletta and Marsascala)



Source: Formosa S., (2000)

Abuse

Statistics and research emanating from studies having ulterior motives are rarely endorsed, but state a lie repeatedly enough using high-end state-of-the-art tools with massive data backing and that same untruth takes a life of its own.

i) Take for example the pre-election polls which, if not carried out stringently enough, may show one party far ahead than another prior to an election. Trust such surveys to be held within strongholds which offer little suffrage for voter-opinion manipulation and the results reflect those same defective processes.

Another abuse is in the employment of sample sizes that are not relevant to the representation of a whole population. A sample of 100 will not deliver reliable results, neither will 300, as the confidence intervals are very low; that is the higher the sample number the higher the representation rate. A 51% outcome with a +/- 3% confidence interval still leaves voters with a 48% to a 54% range; highly different from a 50%+1 minimum requirement.

The latter examples of research that goes wrong is testament to the adage that:

'A bad research is worse than no research at all'

Avoiding Research

Prior to initiating a study, researchers should ask themselves the following questions. If the answer to even one of these questions is "yes", then the research topic should be avoided!

- Does the research problem involve question(s) of value rather than fact?
- Is the solution to the research question predetermined, effectively annulling the findings?
- Is it impossible to conduct the research effectively and efficiently?
- Are the research issues vague and ill-defined?

The research process described can in turn be consumed through the implementation of the DIKA elements.

Each of the elements of DIKA was described earlier in the Chapter with descriptions showing that each phase has its own construct. Let us revisit a summary of the table described earlier in the Chapter Table 3.1 and Table 3.2).

Table 3.2: DIKA summarised

	Meaning	Form taken
D	Data	Normally in numeric or coded forms. Main issue here is **NUMBERS.**
I	Information	Textual forms but can be expressed in a wide range of formats inclusive of graphics, maps, imagery, 3d structures, and other forms. Main issue here is **CONTEXT.**
K	Knowledge	Textual forms and presentation methodologies required for dissemination processes, whether analogue or digital. Main issue here is Translating the Context into **POLICY.**
A	Action	Translated/transposed to actions on the ground. Main issue here is **IMPLEMENTATION.**

The process is a straightforward one, leading the researcher through a series of elements that flow into each other.

Data ⇨ **Information** ⇨ **Knowledge** ⇨ **Action**
Numbers **Context** **Policy** **Implementation**

The flow allows researchers to move through a process where each step is essential. Each element builds on the previous element. Implementation cannot occur without a solid policy base structured on sound methodological information analysis. Information cannot exist without reference to a real context within which the policies can be implemented but no information can be gathered without building on a solid data foundation from data gathered through various methods.

The next steps take each of the elements in isolation and review their constructs and functionality.

Data

What is data? Is it in the singular or in the plural? Data is commonly used for both, however datum should refer to the singular. Irrespective of the terminology, one can go back in time to extract a definition which stands the test of time. The processing of data has changed drastically since the 1989 British Computer Society quote but the fundamental description remains constant.

Data are information coded and structured for subsequent processing, generally by a computer system (British Computer Society, 1989)

Thus data is termed as coded information. This seems like a chicken and egg situation since information derives from data and here we are stating that data is coded information! A nice way to initiate descriptions of a complex item….one would say! Let us go about it in another way…. Focus on the word code and one can start from there.

Identify with a set of numbers

1
10
30
1000

What do these numbers signify?

Euros? Chickens? Thousands of persons?

In fact they signify nothing except numbers on a white sheet of paper.

Now let us code the list and see if they make better sense:

Kilometres
1
10
30
1000

Immediately the list takes a life of its own. The numbers actually refer to something tangible, at least in this case kilometres, which in itself has a construct which can be referred to as something in terms of 1,000m, 10,000cm, 100,000mm and so on.

The mere designation of a title has given meaning to the list. From absolute raw data it has grown into something that one can work with. However, before venturing further it may be wise to apply brakes on the process under discussion and review possible data problems.

Data Problems

One needs to keep their eyes open for the following data-related problems:

- Data has been and is sold sometimes at very expensive rates, especially where creators need to recoup costs. Cost effectiveness is creating cost-recovery measures through the use of copyright and access restrictions which is hindering access to researchers that have limited budgets within which to operate;

- Access can be restricted at various levels from administrative to archival status protection (deteriorating records) to human interpretation of laws. This has to change and various legislations have been enacted such as the Data Protection Act, the Aarhus Convention and the Freedom of Information Act. The main issue here rests with implementation and enforcement of

the relative legislation to ensure that data is made available within the remits of that relative legislation;

- Data can be extensively hoarded due to thematic territoriality where data is restricted on the observance that it may be interpreted wrongly by non-experts. The philosophy should be inverted where this concept is turned upside down through the argument that the more data is disseminated, the higher the accessibility for students to specialise in that field. One must note at this stage that in some very rare cases, due to the need to protect the data of an entity which would be the only one in existence in a country and such data dissemination can lead to competitive collapse, such sensitive data cannot be published but may be aggregated at more abstract levels to ensure confidentiality;

- In the case of specific datasets, some countries have yet to develop such fundamental datasets, as are the referenced address point datasets that exist, which would allow for the analysis of housing, industry and other land-related spatial analysis. A disseminated common database that allows various-level access to that data could help mitigate this important issue. How can a researcher base his/her studies on an analysis of poverty in a high-dwelling density area when the addresses are not mapped?

- Various countries have zip or post codes that are either not reliable enough or not mapped, rendering the same problem as the previous point:

- Researchers requesting data from third party agencies should be aware that that data was gathered for the purposes of that agency and not for the researcher (unless specifically requested as a new query) and as such the data may not be wholly or even partly relevant to the study in question:

- Take great care to ensure that the dataset origin is known, the date it was gathered, as well as other check-list items that need to be reviewed as described later in the metadata section in Chapter 6. It is imperative that such datasets are accurate, up-to-date (have the latest currency), are complete and are tagged with appropriate metadata:

- It is important to decide earlier on at which level the datasets are required and this refers to the so-called NUTS levels (see Chapter 12). Most data is available at national levels, but researchers may need data at regional, locality, sub-locality, street-level or even point data (at the exact location it occurs). Note that most data, apart from the national levels is available at local council level:

- Versioning is rarely employed so ensure that data has a stamp that it is the latest version. Many data operators save versions and copies of versions in various locations, with potentials for loss of one version or other. Real versioning occurs when multiple persons use the same document concurrently and the software ensures that all edits are inserted: and

- Ensure that a lineage exists. A lineage is a document that holds a step by step record of the process employed to reach the end result. This includes file names, location of saved files, the statistical measures used, the queries programmed, as well as the difficulties and potential errors identified.

Having reviewed the problems faced in the data structuring process, one can conclude that the most important aspects that make data vital for one's study include relevance, timeliness and accuracy.

Bringing back to mind the issue where a title to the attribute (column) had given life to the original nameless list, one can state that titling, together with the non-existence of the problems mentioned above, allows the researcher to move one step further.

Even though the meaning of the list changed through the titling effect, by itself the list says nothing which can eventually lead to implementation, unless… but that is an issue for the section on Information.

Information

What is information? It is a term much abused across the different disciplines. It is mistaken for data and is also given a terminology more suited to textual discussion.

As we move from a passive to an active mode of the DIKA process, information bridges the gap between the coded data and the links that the same data has to the reality it appears in.

By definition, again taking the 1989 quote, **information is the meaning given to data by the way in which it is interpreted (British Computer Society, 1989)**. In other words data becomes information when it is given a meaning that ensconces it into a construct whether thematic, locational or otherwise. What is meaning? This seemingly philosophical word refers to the second life the data takes when placed in a CONTEXT. The interpretation of that data within the context allows researchers to create scenarios that lead to the drafting of policies that should be implemented within that context under study.

One thing to note, however, is that the Data Cycle is suffering from **DRIPS.**

Data-Rich Information-Poor Syndrome

A large number of research activities stop at the data level, with very little priming towards the information element. With myriads of data trawlers and sensors currently gathering raw data, the data could end up as an **end in itself** rather than a **means to an end**. The 'fixation' to gather data in its entirety takes a life of its own and moves from fascination to obsession with researchers ending up concentrating on descriptive rather than analytical studies.

Going back to the raw data and coded data example, the best way to illustrate the context is to imagine that a researcher is carrying out a health-related study. To carry out such a study one cannot simply investigate one variable in isolation, but needs to clash that variable against one or more other variables. The example below reviews such a comparison.

The study investigates the ability of humans to breathe at different altitudes (vertical distances from a point on earth, nominally at sea level). Thus the title Kilometres takes on a new form precisely because it was placed in a context that stated that the departure point was at 0 Km at sea level. The other aspect of the context is that the distance is calculated as that point perpendicular to the earth plane.

1
10
30
1,000

Kilometres
1
10
30
1,000

Kilometres Above Sea Level
1
10
30
1,000

Therefore this study shows that investigating ability to breathe was given a context by the departure location, the distance travelled and the direction of target. Bringing in the theme of breathing immediately depicted the potential for interesting results. Enhance it with readings for oxygen molecules by volume and one can render a specific cut-off point where breathing becomes impossible.

Kilometres Above Sea Level	Ability to Breathe
1	High
10	Difficulty
30	Not Possible
1,000	Not Possible

Prior to moving to the final two elements of DIKA, one can use the above example to understand the existence of two more levels of information that go beyond single theme or variable analysis.

The first level covers **geographic information**: information which can be related to specific locations on the earth (UK Department of the Environment, 1987). Relate this to the example above: it does not

simply investigate the existence of oxygen at the different levels but takes on the issue of breathing above the different points on earth. People can breathe on a hill in Sicily (1 km), they would have difficulty breathing at 10km on top of Mount Everest and impossible at the other heights. Therefore two data points had an earthly (geographic) tag while the others have a space-related tag.

The second level covers the concept of **spatial information**: the means by which information can be related to a specific position or location (Shand and Moore, 1989). Further investigation into the breathing issue and the social construct enables the researcher to note that at 1 km most societies thrive due to the readily available oxygen levels, while this is not the case at the other levels, with progressively lower levels of oxygen. Therefore, the fact that one investigates the social construct and the relationships between those same constructs and oxygen brings in new study parameters.

Let us quickly depict another example. A chair exists as an aggregate of wood and metal (raw data). It is located in a point in space that nothing else can occupy at the same time (information). It is located in an exact x and y coordinate (unique longitude and latitude) which explains the need for geographic location. The chair, however, takes on a special role (context) when that XY coordinate happens to be the location of the President's chair during audience sessions; a very important activity. The chair can be a common chair, the floor could be any floor, but the mere fact that that chair is located in the geographical location within a wider socially-interactive arena gives that spatial location significance. Only because the location happened to be in the spatially-relevant location did it become an important item of furniture. Thus the spatial construct enhanced the meaning given in that context.

Knowledge

Knowledge and the creation of a unique output foundation based on research is the Valhalla for policy making. The jump between information and knowledge is not an easy one to make. This is due to the fact that one must ensure that information both represents and 'fits' within the social reality under study. The information can enhance a society and also propose new ways to deviate from the perceived norm since the findings indicate trend shifting where policies are no longer tenable.

The uniqueness generates new forms of awareness, literature and scientific results, but also bases itself on the established knowledge prevalent to date. Take for example a study based on the welfare gap debate which had reviewed prior demographic trends, population structures and projections. The shrinking welfare gap was seen as the fulcrum for the eventual transformation of policies leading to the extension of the retirement age.

The past and present sustain knowledge building, but new research helps to investigate if the *status quo* has been maintained or if the trends show otherwise. If it is the latter, then that same research serves as the launching pad for policy-making or for policy amendments.

As stated earlier, the outputs from a study should help identify both recommendations for future policy-making, decision-making and academic research. Also, in a professional world, one does not necessarily give just one outcome, but has to keep in mind that society is much more complex. This complexity in itself allows for different scenarios to be built, which help decision makers to be informed of potential issues prior to transposing that policy to legislation and operational tools.

In summary, knowledge serves as the tool that extracts the meanings given to the data trawlers (data gatherers) and changes that to POLICY.

Action

The final level is both distinct from and at the same time integrated with the data-information-knowledge elements. It basis its outputs on the implementation of the policy recommendations set out by the knowledge element. Action requires a very solid operational enterprise that ensures the implementation of the need for installing and maintaining structures that can take up those requirements necessary to follow it through.

This element must not be seen as solely related to legislation passing or the setting up of bodies to manage the outcomes of the legislation, but it also has to prepare for capacity-building in terms of employment opportunities. This in turn can only happen if training is provided and courses are launched to ensure that same capacity-building. This concluded, the enforcement issue has to be ensured as this

is where most legislation falls through, with a lack of enforcers or review personnel being given the necessary tools to operate in.

Once implemented and running, the researcher has to maintain the running of either continuous or periodic monitoring surveys that review the on-the-ground changes in trends or social interactionism. Situations might exist where the monitoring exercise may show that the action failed or, at the other end, brought about very rapid change. This might require legislation change once more.

In summary Action does not stop here, but requires a continuous feedback loop.

Questions (refer to Appendix for the answers)

1. What 3 major developments changed the process of conducting research? Were research problems solved, or were they merely replaced by new problems? Mention some of these modern research-related problems.

2. What are triangulation studies?

3. What is DIKA and what do the letters stand for?

4. What is the W6H in relation to conducting research? How are the W6H elements helpful?

5. What is research and why do we need it?

6. What are the main questions that need to be asked to a potential researcher?

7. Mention the 3 main forms of research and briefly explain each one.

8. List the 4 main types of research.

9. The choice of type of research depends on 2 factors. Name them.

10. When conducting research, the most important issue to keep in mind is context. What do you understand by “context”?

11. Before starting off on a research project: one needs to be clear on what is going to be studied; one needs to decide whether to adopt a quantitative or a qualitative approach and to ascertain whether the research problem has been identified. This requires a logical approach and the authors recommend the 3x3 structure. What is this 3x3 structure?

12. What quality should a sequence of Aims-Objectives and Research Questions have and why is it important?

13. Why is the literature review important?

14. What are the main research hurdles researchers are prone to encounter?

15. How would creating a “mind map” assist the researcher?

16. What percentage of the total research time does data collection take up?

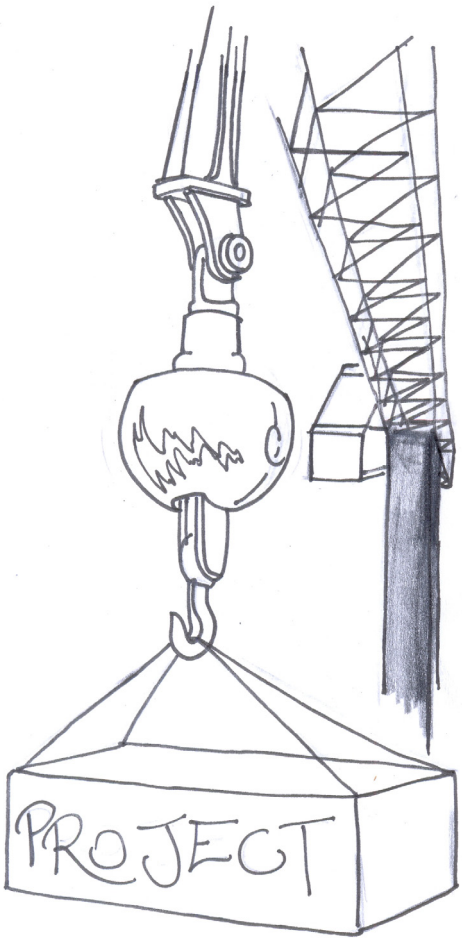
17. What do you understand by “data analysis”?

18. What does reporting of results entail?

19. How did technological advances in research tools affect research?

20. Explain the difference between the use and abuse of statistics.
21. Before embarking on research, the researcher should ask herself/himself 4 questions. List these questions.
22. List the 10 major data problems.
23. What are the 3 most important aspects that make data vital for one's study?
24. Define "information", in relation to research.
25. Define "meaning", in relation to research.
26. What do researchers mean when they say that the data cycle is suffering from DRIPS?
27. Describe geographic information as opposed to spatial information.
28. Define "knowledge", in relation to research.
29. Define "action", in relation to research.
30. What do researchers mean when they claim that action requires a continuous feedback loop?

Chapter 4
Structuring Your Research



The resolution of revolutions is selection by conflict within the scientific community of the fittest way to practice future science. The net result of a sequence of such revolutionary selections, separated by periods of normal research, is the wonderfully adapted set of instruments we call modern scientific knowledge.

Thomas S. Kuhn

The Structure of Scientific Revolutions (1962), 171.

The previous chapters had the task to describe the theoretical issues and the concepts behind the research process. This chapter takes a more hands-on approach in describing how best to carry out the research design, what software to use and how to choose which questions are really needed for the study. This is followed by the issues identified during the data gathering process, the data analysis part, the reporting process and how to ensure that ethics are recognised and adhered to.

The final sections revisit the drafting of the aims and objectives and the eventual research questions and review how they fit within the datacycle. Finally, the hypothesis issue is tackled to ensure that all this fits together.

A Datacycle Approach

The Datacycle self-explains its *raison d'être* through its procedural function. Earlier on in the book, it was identified that research requires data; however data gathering is not without its own set of requirements. The process initially entails a clear design of the method that will be undertaken, the choice of the tools to be used and the drafting of a tool called a matrix (discussed later on in this chapter) to help identify which questions are really required. Only when these have been established can the actual data gathering process start, followed by the analysis process with its inherent querying and recording methodologies. Where issues deemed problematic to the process are identified, this loop is flexible enough to allow the research to go back and re-initiate either the whole process or individual elements.

Design

The design process takes up the bulk of the cycle's work. One has to have a clear idea of what is required from the study. This can be ascertained through the drafting of a clear aim, a set of objectives and a set of research questions. The latter will help identify which variables are required.

Assuming that the literature review process has been concluded and that one has chosen which research method will be employed, the next step is to come up with a series of phases that will be undertaken.

Draft a **checklist** which covers the following:

Methodology Issues

- Is the researcher in a position to initiate studies using a particular methodology?
    a. Will a qualitative approach be taken?
    b. Will a quantitative approach be taken?
    c. Will a mixture of the two be taken?
- Has the literature review brought up very specific data requirements?
    a. If yes, are they available for gathering?
    b. If not all could be accounted for, have surrogates been considered (for example income data is absent therefore a surrogate could be used that indirectly reflects that variable, such as the number of cars in a household)?
    c. Does the data have a timestamp before it can be accessed? For example, national data has a 30-year moratorium as per Chapter 477 of the Laws of Malta.
- Will the researcher need to carry out archival research, interviews, and surveys or will s/he use readily accessible distributed data (for example from the net databases)?
    a. If archival, is the material physical (analogue – hardcopy)? Is it too deteriorated to read? Is it accessible?
    b. Is it in the country of residence or in another country whence costs have to be factored in?
    c. If based on surveys, are the methods understood? Does one need to send online emails or physical mail shots with return envelopes? In addition, has a prize been included for respondents and have the necessary permits been acquired?

Operational Issues

- Have costs been factored in?
    a. Some agencies charge a unit cost for the data.
    b. Others charge a nominal cost irrespective of the query type.
    c. Others charge an hourly rate for the time executed during the running of the researcher's query but do not charge for the data.
    d. Others do not charge at all due to legislative constraints or due to the fact that the queries were already prepared.
- Have the contacts been made?
    a. Have you allowed enough time to enable the contacts to fit your sessions in their timetable?
    b. Will communication technology be employed? Will you use internet communication tools such as basic emails, online questionnaires, video conferencing or otherwise? Or will phones be employed?
        i. Have you ensured that the system works?
        ii. Have you ensured that you have enough backup power to record the sessions?
        iii. Have you prepared for a contingency just in case of power failure (whether direct electricity or backup power)?
    c. Will the sessions be recorded?
        i. If yes, will permission be requested from the interviewee?
        ii. How will you record the sessions and how have the ethical issues been accounted for?
        iii. Have you accounted for enough time to transpose the interviews into text? This transcription method should ensure that the written text is exactly loyal to that in the recorded media.
    d. How will you store the files? Always keep multiple backups of the digital files and at least one copy of the analogue (hardcopy) material.
        i. Keep digital copies in a CD/DVD format;
        ii. If possible keep one in a secure online location; and
        iii. Ensure that the formats are readable in more than one document format in case of software malfunction.

Technical Issues

- If the files are highly sensitive, how will they be stored?
        i. Where will the files be kept – is a secure place available, a site that cannot be compromised?
        ii. How will the names of the persons interviewed be protected? Have you created a system that enables the conversions of those names to codes, and is the code document stored in a separate place?
- Has error checking been given the relative weighting?
    a. Are errors accounted for especially during the input cycle?
    b. How will you verify that the data is still sound especially following such exercises as sorting, which tend to be the main sources of data misalignment?
    c. What types of error checking will be carried out once the data has been inputted? (This could include a number of summation exercises).
- Is the researcher skilled in the use of analytical software?
- What process or language will be used to record the queries based on the variables? (This is highly crucial in that a query carried out at a certain time and the same query carried out at a later stage should deliver the same result). Simple logic, no? However, this is the main source of error generation and the inability for results to be replicated. This may be due to a variety of factors:
    a. The data may have been edited in the meantime (importance is due to the version being used);
    b. The data may be sourced from a live database that could be updated on an ongoing basis, inclusive of the archival data. Take dwelling permits as an example: a file may be superseded and the original data referring to the number of units would subsequently be removed. This would cause a change in the variable pertaining to number of units permitted by year.

- Has a lineage system been prepared?
    a. Will you be able to backtrack should an error be identified late in the study and the researcher needs to go back and rectify that error?
    b. Have the processes been recorded in detail so that all the successive steps can be taken again?

IS A PLAN B AVAILABLE?

This is one major concern for any researcher! What happens should the topic under study prove to be impossible to research?

There are various issues why this can occur:

- Data is not available;
- Contacts do not cooperate;
- The topic is too sensitive and feedback is limited;
- The topic was superseded by new legislation;
- The topic was overtaken by events;
- The literature review led to a deviation from the aim of the study…

Thus the need for a PLAN B, but what does this signify in reality?

Researchers have to be aware that due to the above, they may have to change their research topic; therefore the best option is to draft an alternative topic early in the proposal drafting stage. This is cumbersome work but it helps ease the trauma of a difficult situation, especially at a moment when time pressure is high.

Note that there is no need to have a plan B that calls for a full topic alternative, but it could be one that changes part of the topic or even the methodology.

An example of a plan B would be a quantitative study that proposed to study migration to and from Malta which data recording was disrupted due to legislative changes and the termination of recording systems. A plan B that would adhere to the topic could propose to conduct a qualitative study on those returned migrants who have been registered as such prior to the termination of recording.

Choosing the correct mining/trawling tools

This section, though covered in Chapters 6 and 9, is best introduced at this point since it will lead to the issue of matrixing. How will one know how to choose the correct data gathering tools? How will one structure the interview/survey/questionnaire forms to enable easy data input? Are specific tools needed?

The answers vary based on the type of study proposed, however this section will focus on the concept behind the need. Let us take a look at the steps one needs to take in order to structure the mining and trawling processes.

Mining: the process taken to gather the data either manually or through initiated processes such as a questionnaire.
Trawling: the automatic process whereby data is gathered by machines or sensors which will review the availability of that data, extract it and store it in a specific location for the researcher's perusal.

Step 1: How will the data be gathered?

- Manually – *in-situ*
- Automatically - remotely

Step 2: What forms will be used?

- Pre-prepared forms
- Open-ended – no formal forms

Step 3: Which tools will be used?

- Analogue – paper/clipboard style – may need maps for *in-situ* analysis
- Digital – using PDAs or a laptop with a scanner or recorder

Step 4: Will the forms have all the variables inserted?

- Yes and includes all the sub-categories
- Partial – allows for the inclusion of new variables and new types of archival input

Example of a checklist matrix:

Target: To gather data on the number of fish caught close to the shore through the use of in-situ measurements and the taking of photographs using open-ended methodology.

	Forms	In-situ	Analogue	Digital	Full Variables	Part Variables
Manual	X	X				
Automatic			X	X		X

Matrixing

Once the process on how to develop the study has been concluded the next phase to be employed should concentrate on creating a matrix that helps the researcher to review if the questions prepared for the data analysis process can be reviewed against the findings from the literature review.

To recap:

i) The literature review identified a number of relationships that the researcher would like to analyse in a local context;

ii) Each relationship is transposed into a variable (an element that can be analysed, whether based on keywords, codes and text);

iii) The variables list is converted to a value-free question if using the interview/questionnaire option or to a spatial variable if using the in-situ survey mode;

iv) The questions are drafted in sequence to ensure that they flow smoothly and not go back to a previous section. Make sure the number is manageable to reflect the real needs of the study. Extra questions are just baggage and tend to burden a study unnecessarily:

   i. Initial section – background information (on person/area/location under study);
   ii. Thematic section - (topic-relevant variables);
   iii. Feedback section – allows the interviewee to state issues not tackled in the survey. This is not always carried out or even needed since only if the statements are relevant to the topic should they be included in the analysis;
   iv. If using a pro-forma template (particularly for archival research the variables need to reflect the material being inputted such as background information, the individual variables listed in the particular ledger/book/inventory.

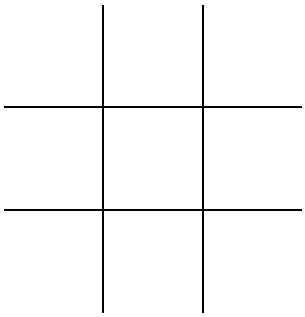
Once this is done, it is time to create a matrix to allow the researcher to form an idea of the eventual relationships that can be investigated within the study.

So….. What is a Matrix?

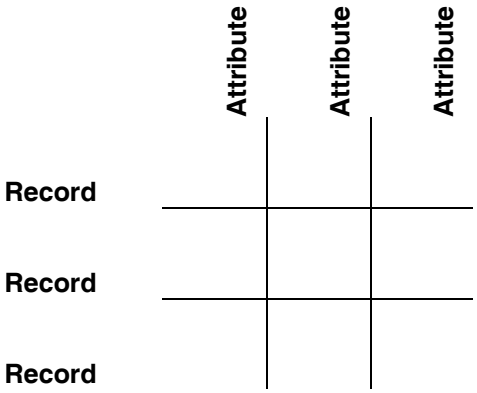
Forget science fiction movie infiltration and the complex worlds created within such virtual worlds…

A matrix is a collection of cells that serve as an aid to structure data according to set columns and records in what can best be described as a spreadsheet.

Imagine a Tic-Tac-Toe game (also referred to as OXO)



This matrix is empty – let us identify the cells
Columns – attributes
Rows – records



A Tic-Tac-Toe game that has been concluded may look like this:

	Attribute 1	Attribute 2	Attribute 3
Record 1	**X**	O	O
Record 2		X	O
Record 3	X	X	O

The matrix tells us that the game was won when 3 Os were placed in sequence under attribute 3 which has a relationship with Records 1, 2 and 3.

Creating the matrix in step-by-step sequence

Taking the Tic-Tac-Toe concept further and visualise a questionnaire that has 10 questions.

Sample questions are listed below:

1. Sex
   - i) Male
   - ii) Female

2. Age: _____

3. Status
   - i) Single
   - ii) Married
   - iii) Legally Separated
   - iv) Divorced
   - v) Widowed

4. Residence Locality

   _____

5. Employment

   _____

6. Do you think that the current way of communication (from management to employees) is effective?
   - i) Yes
   - ii) No

7. What makes a communication system effective?
   i) a cohesive team
   ii) target-oriented *modus operandi* (goal oriented)
   iii) mutual understanding between the parties of each other's skills
   iv) knowledge of information systems that actuate and enhance communication styles

8. Which communication method is the most effective?
   i) Telephone
   ii) E-mails
   iii) Meetings
   iv) Letters
   v) Notices

9. From your experience in management, do employees perform better when there is effective communication?
   i) Yes
   ii) No

10. From your past experience when communication maybe was not as effective, did workers perform effectively?
   i) Yes
   ii) No

Step 1: Create the Matrix shell as shown below

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

Note that the question numbers are reflected on both the top Attribute cells and on the left Record cells. It is essential that the matrix is created in this way, the reason for which is given in the next Steps.

Step 2: Highlight those numbers that correspond to the same question numbers

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

The diagonal cells show the ones that fall within the same question number within the attributes and records. These variables can be analysed through descriptive statistics as they cannot be clashed against themselves.

Step 3: Fill in all the cells below in grey

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

This step ensures that any two questions are not analysed against each other twice in the same study. Comparing Question 1 with Question 2 is the same as comparing Question 2 with Question 1. A simple enough process, but one can be surprised how many times such double entries are made!

Step 4: Create a list of those questions which will be compared/cross-analysed.

Example from the questionnaire listed above:

Q1 vs. Q2

Step 5: Mark the corresponding cells with an X

	1	2	3	4	5	6	7	8	9	10
1		X								
2										
3										
4										
5										
6										
7										
8										
9										
10										

Add also those for the next 3 comparisons:

Q1 vs. Q5
Q2 vs. Q8
Q6 vs. Q9

	1	2	3	4	5	6	7	8	9	10
1					X					
2								X		
3										
4										
5										
6									X	
7										
8										
9										
10										

In a real research scenario the matrix would include a significant number of **X**s. Note that one must not saturate the cells with Xs as that defeats the purpose of the whole exercise. The scope is to ensure that those cells which are relevant to the Research Questions identified in the previous exercises are really and truly included and also to ensure that no extra (potentially unnecessary but definitely time-consuming) comparisons are made. Every research encounter the authors have had with a large number of students and professionals brings up the surprising number of cross-analysis that was included in the analytical chapter/sections but never relevant to the study and not used at all. In some cases, even the relative questions in the questionnaire/interview sheet/input list would have been discarded prior to the launching of the sessions. One might argue that the pilot study would highlight this issue, however pilot studies rarely tackle potential comparisons between the variables, but only concentrate on whether the questions are understood by the interviewee/respondent.

Step 6: Identify the Measurement Scale which will be used (Refer to Chapter 5 for a description of the Scales). In this section the answers have been pre-prepared based on the questionnaire drafted above.

As an ice-breaker, there are four types of measurement scales as used by the matrix, which define the mathematical levels of precision with which the values of a variable are expressed. These are:

- **Nominal Scale**
- **Ordinal Scale**
- **Interval Scale**
- **Ratio Scale**

In the Matrix Model, the NOIR is used: Nominal (N), Ordinal (O), Interval (I) and Ratio Scales (R).

Now, identify under which scales the relative questionnaire categories fall:

Question No.	Measurement Scale
1	N
2	R
3	N
4	N
5	N
6	O
7	N
8	N
9	O
10	O

In order to ease analysis matters, it is best to change the interval to a range which in turn from Nominal becomes an Ordinal. This issue is best explained in terms of age categories. Imagine analysing question 1 and 2 based on the single age replies as listed in the questionnaire. If the ages ranged between 10 and 100, then there is a possibility that for every age-year one could have an input such as 2 males and 3 females. Since it is too cumbersome to analyse, the best method is to translate the single years to cohorts. The best way to do this is to use the basic demographic 5-year age cohorts, which would be classified as follows:

Age Cohorts (years)
0 – 4
5 – 9
10 – 14
15 – 19
20 – 24
…
95 - 99
100+

These groupings are termed cohorts, which have a measurement scale designated as ordinal.

Step 7: Update the Matrix to reflect the Measurement Scales listed in Step 6

		N	R	N	N	N	O	N	N	O	O
		1	2	3	4	5	6	7	8	9	10
N	1					X					
R	2								X		
N	3										
N	4										
N	5										
O	6									X	
N	7										
N	8										
O	9										
O	10										

Step 8: Convert the Interval Scale to Ordinal Scale as per Step 6

		N	**O**	N	N	N	O	N	N	O	O
		1	2	3	4	5	6	7	8	9	10
N	1					**X**					
O	2								**X**		
N	3										
N	4										
N	5										
O	6									**X**	
N	7										
N	8										
O	9										
O	10										

Step 9: Colour-Code the Cells marked with an X

This step ensures that once the data inputting stage is complete, the analysis phase is carried out in an efficient time-saving manner. The colour-coded cells can be run together by row number which reduces the time taken to run single queries individually.

Should Q1 (where Q1 is a Nominal - N) need to be analysed against another Nominal (Q3), then the statistical test to use is that called Chi Squared (covered in Chapter 11).

If both the row and the column are identified as an Ordinal (O), then the relative test would be based on a correlation (Chapter 11) (such as Q6 vs. Q9). Should the Row be an Ordinal (Q2) and one of the attributes (Column Question) be a Nominal (Q8), the latter enforces the test to the same types as Nominal vs. Nominal i.e. Chi square.

		N	O	N	N	N	O	N	N	O	O
		1	2	3	4	5	6	7	8	9	10
N	1					**X**					
O	2								**X**		
N	3										
N	4										
N	5										
O	6									**X**	
N	7										
N	8										
O	9										
O	10										

Legend

	Nominal vs. Nominal and Nominal vs. Ordinal	Chi squared test	NxN NxO
	Ordinal vs. Ordinal	Correlation: Spearman's test	OxO
	Descriptive Statistics	Frequencies	

The Matrix can be amended as more analysis is carried out and new queries may be required based on new findings. The ready-made MATRIX will allow for such additional work without the requirement for starting from scratch.

Data gathering

Once the Matrix has been prepared, the basic groundwork would have been laid. The next phase would initiate the process for eventual data collection. This is easier said than done, since one has to start by deciding the best way to start one's work.

Remember, the first interview is always heart-throbbing and tests one's nerves, even of those few whose nerves were primed from steel! Once the first hurdle has been overcome, the next research steps should follow smoothly.

If the first step or steps prove problematic, you might wish to confer with a supervisor or a manager on either reviewing the *modus operandi* or even restructuring the process. This can be carried out by first tempting the process with what is called a Pilot Study.

What is a Pilot Study?

Consider this as a testing phase, a launching pad and a necessary evil!

The magic ingredient required to carry out this initial step is called 'human targeting': grab a few friends or colleagues and choose between five and ten if taking the quantitative route and one or two individuals if the qualitative one is chosen.

Distribute the material to the nice persons in front of you and ask then the necessary questions or review them while filling it up: both in terms of time and issues reflecting the understandability/flow of the questionnaire.

If using an interview mode ensure that this simulation will record all the steps that will be replicated in the actual sessions. These include availability of recording materials, notepads, ensuring that posture leads to legible hand-writing (*hen-writing* makes it very difficult to go back and revert to the interviewee with an illegible scribbling)…. That's very embarrassing!

It is always surprising how many small items are discovered during this phase, so it must **not** be approached lightly.

Once the devils have been confronted, then the findings can be dissected in a post-mortem:

i) What could have been carried out better?
ii) What did not make sense in such a process?
iii) What needs to be weeded out?
iv) What needs to be included?
v) Are the numbers targeted realistic?
vi) Was the time projected realistic and appropriate or should I restructure the process?
vii) How will it affect the data collection period identified?

Verify that all these steps have been solved and then start the process to go about your real study.

Start the sampling process and then choose the process to elicit the group of persons who will serve as your target group. Chapter 2 covers the sampling process and how best to choose a sample from available sources.

Different ways to go about doing research

Once the target group has been chosen, one needs to identify the on-the-ground mode of operation…. the basic activity which will allow one to gather data. Unless already decided upon through the literature review process and unless pre-prepared data is available, one has to identify the appropriate mode. This is not a pick-out-of-the-Easter-bunny-hat choice but one that must be stemmed from background studies and review of case-studies.

Once chosen, the method must be adhered to as some can take ages before the necessary information can be gleaned. Refer to Chapter 2 for an explanation of the different ways to go about this process.

Analysis

Having completed the data gathering exercise and inputted the data into an analytical tools (refer to Chapter 9 for details on the diverse tools), it is best to understand what steps should be taken in order to analyse that same data.

There are various options one needs to contemplate prior to the actual analysis:

- Has a decision been taken on what tool will be used?
    - Ensure that the tool is well understood and if one has programming skills, such would help to carry out simple queries. If using such skills it is best to record the process used.

- Has the matrix been completed?
    - Ensure that the matrix described above is complete and allows for the analysis process to flow smoothly.

- Have you drafted a series of keywords if using a qualitative approach?
    - Write down a series of keywords that you will look for in the transcribed text. These will serve to elicit commonalities between the different replies.
        - Examples of keywords for a study on youth and drug use:
            - Youth
            - Recreation
            - Stress
            - Drugs
            - Alcohol…

      the keywords list will expand as more interviews and transcriptions are carried out, though it would be ideal to compile the list during the literature review stage.

- Does your method include creating a series of catalogue cards which will allow you to remember the keywords (categories) within which they fall?
    - This is a very interesting issue which was used primarily in pre-computer research. A series of cards would be created which would contain themes or keywords. Any information and data would be described under each theme or through a sub-set series of cards. When one requested information for a specific query the cards would be extracted by theme and the contents clashed.
    - Though an interesting way to carry out cross analysis, this method has been superseded by technology, and computer software carry out this role. However, hardcopies should still not be abandoned as they serve both as a backup and as a visual reality: try reading a 100-card graph on a monitor as against all those cards spread on a floor! In reality both can be used, however the researcher's choice is based on ease of use of digital technology or more traditional reliance on solid material options.

- Have you chosen the statistical measures that will be used if you are taking the quantitative route?

- The statistical measures used depend on the outputs of the NOIR Matrix and on the type of relations one is analysing. Chapter 11 gives a description of some of these measures.

- Can your qualitative results be analysed also through the quantitative approach?
  - This is a very challenging issue for those who do not want to consider using quantitative tools! However, the scope of such a question serves the inherent reality that qualitative and quantitative approaches are divided by a very fine line. What is the scope of keywords if not to elicit how many times that same word surfaced in a set number of interviews? Simplistic as it may seem the interviews pile up, so do the specificities that bind them together.

Having come up with answers to the above, one's next step is to look at how best to carry out the analysis. There are important processes to do this based on a simple rule:

Record Every Step You Take!

One of the problems faced in carrying out analysis concerns the issue of back-tracking. Very few researchers record every step they take to carry out an analysis, and when a discrepancy crops up, one is at a loss to backtrack and find out where the whole thing went awry!
To avoid this problem, all one needs to do is to record the steps taken in every query through what is called a Lineage. The Lineage allows one to follow the steps taken and also records what files were generated, how they were stored, the problems encountered and other relevant steps.

The following Lineage examples give overviews of some of the steps taken in order to (i) create a dwelling zone map, and (ii) find how many persons are aged over 70. The actual lineage is over 15 pages long!

Lineage 1

Project: Creating a dwelling zone map

User: Saviour Formosa

Date: 23 June 2005

Source File Name: lineage creating a dwelling zone map.doc

Destination Directory: C:\data\landuse\descriptive links\lineage

Project Description: a number of dwelling by type maps were created to eventually form a single dwelling zone map

Abbreviations used:

xls	Excel file
doc	Word file
tab	MapInfo file
mif/mid	MapInfo Export file
dbf	Database format file

Lineage Steps

1) Creating a buffer map for each zoning type

Each layer was gathered from the different data sources such as:

i) land availability database
ii) planning permits
iii) address point database

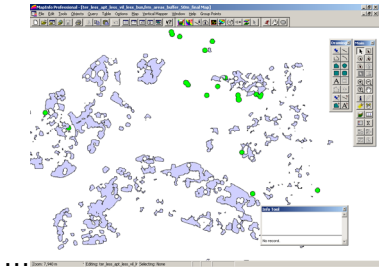
a buffer averaging 30m was given for each area

2) each layer was then overlaid on the respective layer in order to remove any overlaps. The base importance used was as follows:

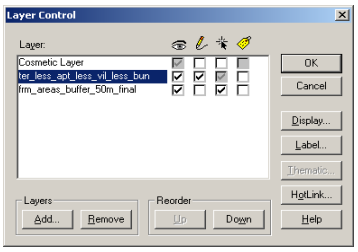
i) TER - terraced (base)
ii) APT – apartment (one up)
iii) VIL – villa (next up)
iv) BUN – bungalow (next up)
v) FRM – farmhouse (top)

3) an example in the integration was kept as follows:

i) Open both layers to be overlaid



ii) Make the bottom layer editable



Comments on difficulties

a. *If buffers were not all at 30m, this may generate a type of inconsistency that should not be a problem since the streets would be covered and any accidental overreach would only be in the region of 3 dwelling plots.*

Lineage 1

Project: ***Lineage for listing of elderly over 70 years of age***

User: Saviour Formosa

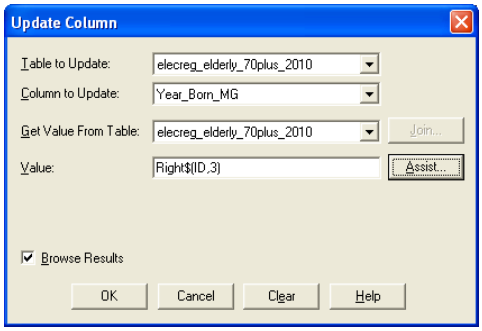
Date: 18 July 2010

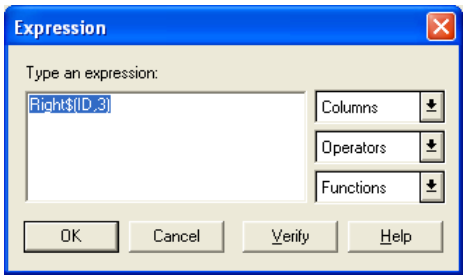
Source File Name: *elecreg elderly_70plus_2010.TAB*

Destination Directory: C:\data\elderly living alone

Steps: Phase I: Eliciting the age of the persons

i) created a file from elecreg Oct2009
ii) added a column named YearBorn MG to extract the last 3 digits found in the ID column
a. Right$(ID,3)





VERY IMPORTANT NOTES

i) The age resultant in Phase 1 is indicative and can generate some errors as follows:
a. The A, P, L categories can generate persons aged over 100. Review each item individually since they may be children: error caused by the ***Step iv****. Those who have an M or G category would more likely be Maltese elderly since the M and G categories were discontinued in end 1999/beginning 2000*

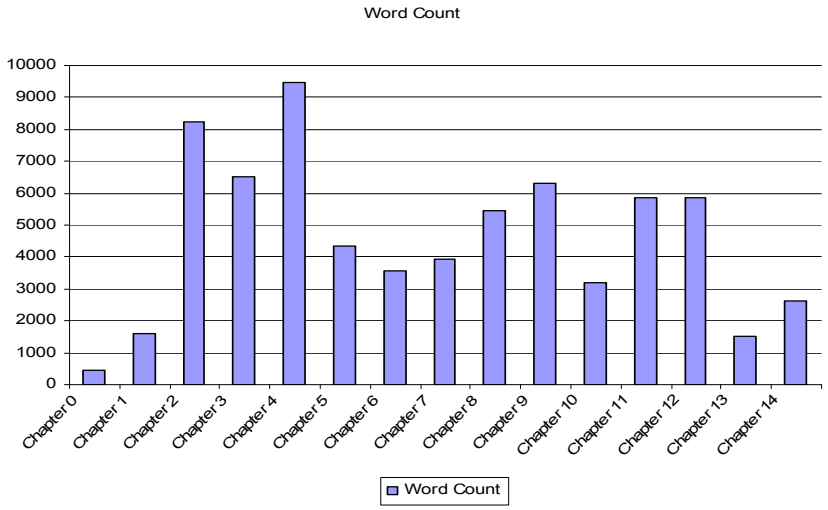
One might be tempted to ignore the lineage process as it appears too time consuming and demanding. It is, but serves its purpose both for cross-checking of data where the results appear too outlying or also to enable re-running that study at a later stage. Do not depend too much on your memory for such processes! Memory deteriorates with time and, once a number of similar but distinct lineages have been drafted, it is difficult to recall which was which.

Analysis

Some rules must exist! Here they come. Keep them in mind at all stages of the analysis.

- Choose the right variables to compare
    - Make sure that you do not complicate matters only to find that the variable has nothing to do with the research question. Do not choose a pollution-related variable to analyse the behaviour of juvenile delinquents.
- Choose simple relationships
    - Do not choose relationships not backed by literature and eventually discover that the relationship between the variables is too complex and difficult to describe.
- Divide complex relationships into smaller simple ones
    - A smaller problem is easier solved. Divide a complex problem into a number of problems that can be solved individually. If still complex, divide some into even smaller ones. Once solved build up the answers towards the whole again.
- Compare different sections together e.g.: demography with transport, etc
    - Ensure that the variables being cross-analysed cover all the themes discussed in the mind map and in the matrix.
- Design graphs which describe the actual data under discussions - do not get bogged down in numbers
    - Ensure that the analysis outputs do not bother the reader to death! They have to follow simple rules (Chapter 7 covers the options):
        - The simple-is-better rule – make your outputs simple (Figure 4.1)
            - Less text
        - Tables should not be very large, containing volumes of data, as no one will read those. It is easier to explain them in graphical or map mode.
        - More graphics and visual tools help the reader understand the context. Visuals include:
            - Graphs
            - Photos
            - Maps
            - Graphics

Figure 4.1 – A Simple Chart



Reporting

As detailed earlier in the book, the results of a study can take various forms; mainly reports, articles in journals, books, computer presentations and other multimedia outputs. The main output has to be descriptive of the findings. The results must reflect the literature review, the methodology used and the analysis process as it is related to the literature and the uniqueness of the study.

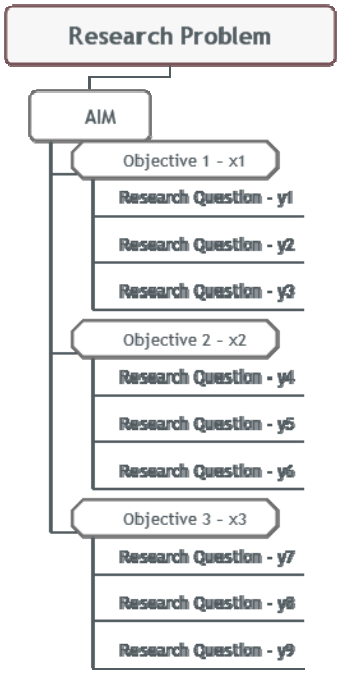
One has to draft a convincing report that mirrors the findings but does not replicate the analysis of the findings chapters. Be concise and ensure that the report includes a series of recommendations as built over a number of scenarios. Do produce an executive summary as it will be the primary document that will be read. It has to be interesting enough to drive the readers to go through the whole report and implement those actions that close the DIKA loop.

Aim, Objectives and Research Questions

Chapter 3 was dedicated to the description of the research phases with one particular step calling for the identification of the research problem. The logical approach identified a 3x3 structure which outlined the aims, objectives and research questions structure. The whole issue revolved around coming up with a series of identifiable research questions which should allow the researcher to test for relationships at the smallest and most detailed level.

Let us review them and see how each section can be aided by the structure described in this chapter, again revising the summary depicted earlier in Figure 3.1, replicated here as Figure 4.2:

Figure 4.2: The Aims, Objectives and Research Questions



i) **A.** - Create an **AIM** which should be expressed as a statement that shows the topic of the study and the direction you wish your research to take.

The aim, though at high abstract level, has to identify the theme under study and has to be clear in terms of what the study should achieve. So if it is aimed at investigating the issue of the phenomenon of cart ruts in Malta, the aim has to specify the term 'cart ruts' and not deviate the reader to such issues as mysticism and prehistoric cults! This is a scientific study so the facts are what should be studied; thus cart ruts ARE an undeniable phenomenon and the researcher aims to understand what their origin was. If s/he aims to investigate who the potential builders were then the title has to state this, however the task to prove such a statement is more difficult, if not impossible as a study, since the facts pertaining to their use have not as yet been ascertained and one might fall in a cyclical trap of the proverbial "chicken and egg", where the establishment of use can lead to the understanding of the culture that built them. Vice-versa, the understanding of the culture can lead to the understanding of the usage. However the former has a fact to base itself on (the ruts), while the latter has yet to provide evidence of the composition of such a culture.

ii) **B**. - The first of the 3x – create 3 **OBJECTIVES** based on the AIM.

The objectives bring the researcher closer to the variables that can be depicted in the matrix and eventually studied at the hypothesis level. Taking the cart ruts' study one can develop the research objectives as follows.

a. The first objective (x1) should describe that you wish to understand a specific topic through the literature review

Objective 1: To seek an understanding of the literature on the phenomenon of cart ruts.

b. The second objective (x2) should state what you want to achieve

Objective 2: To achieve an understanding of what cart ruts have been listed, those that have been lost, the related ancient structures identified close to the ruts, whether they were actually ruts, their purpose and structure.

c. The third objective (x3) should state how you aim to achieve those results

Objective 3: To record all the cart ruts on the islands, to measure the structures and study them using spatial tools and try to identify relationships with meteorology, terrain, topology using digital terrain modelling systems. Also, to recreate the structures in a virtual world for scenario testing.

iii) **C**. – For each of the objectives identify 3 **research questions.** This section closes the 3x3 loop as it allows for the creation of up to nine research questions.

The research questions bring the researcher into an arena where the actual literature themes and variables are investigated at the individual level.

a. For Bx1 – the Research Questions (y1 to y3) could state that the researcher will investigate whether a relationship exists between, for example, the theoretical exponents and the realities on the ground.

Research questions for Bx1 could be as follows:

i) What scientific literature exists on the phenomenon of cart ruts?
ii) What has been written in folklore literature?
iii) Is there evidence of the phenomenon in other countries apart from the Maltese Islands?

b. For Bx2 – the Research Questions (y4 to y6) could state that the study will aim to identify the linkages between one theme and another.

Research questions for Bx2 could be as follows:

i) Which ruts have been listed and where are they located, inclusive of those recorded but lost through degradation, development of reclamation?
ii) Have related structures been recorded in the vicinity of the ruts?
iii) What is the purpose of the phenomenon: were they actually the result of cart activity?

c. For Bx3 – the first Research Questions (y7 to y9) could state that the researcher will investigate whether a relationship exists between the spatial data gathered and the literature review findings.

i) Is there a relationship between the spatial terrain model, the climatic/meteorological issues and the location of the phenomenon?
ii) Can the virtual model show the relationship between the phenomenon and other structures?
iii) Have the findings supported the established literature or has it brought up a unique perspective to the literature?

As one can perceive from this process, considerable thought goes into coming up with research questions at the start of a research process. The process has to follow a methodological structure that enables researchers to provide ‘meat’ to their research’s bones.

Once the Aim to Objectives to Research Questions 3x3 model has been described, the final construct to understand concerns that termed “hypothesis testing”, which takes on where we left at the research questions level.

A need for a hypothesis

What is a hypothesis? This single sentence makes or breaks the study of a research theme. Having followed all the above rules and processes, it is essential that the most detailed analysis fits somewhere. Does a relationship between two variables indicate a relationship or has the relationship not been proven? Such is the need for a hypothesis.

A hypothesis aims to explain a phenomenon using scientific means to test it. Scientific studies call for the testing of two hypotheses, called the null hypothesis and the alternative hypothesis respectively.

Note that the difference between the two is as follows:

- The null hypothesis always states that there is **no** relationship between the variables (phenomena) under study. This hypothesis is presumed to be true unless proven otherwise.
- The alternative hypothesis states that there **is** a relationship between the two.

During testing:

H0 refers to the null hypothesis
H1 refers to the alternative hypothesis

As a case study, one can investigate the relationship between village core dilapidation and population loss. The study is trying to find if there is a relationship between the deterioration of the urban cores (village centres) and population loss from those centres.

H0 would be described as:

H0 - Village core deterioration does not cause depopulation

H1 would be described as:

H1 – Village core deterioration leads to depopulation

In summary, should one have created a DIKA flow, with the aim leading to a set of objectives and the objectives leading to the research questions, the latter should be tested for a relationship. If the relationship is not found then the null hypothesis holds; if it is established, then the alternative hypothesis stands.

Questions (refer to Appendix for the answers)

1. Briefly describe the datacycle approach and state why it is important.
2. How can a researcher achieve a clear view of what is required from her/his study?
3. What are the main methodology issues to be considered?
4. What are the main operational issues to be considered?
5. What are the main technical issues to be considered?
6. How/when can a research prove impossible to conduct?
7. What are the steps needed to structure the (research) mining and trawling process?
8. What is a matrix?

9. List the 3 types of measurement scales. Using the model questionnaire provided in this chapter (without looking/ copying) complete the following table:

Question Number	Measurement Scale	Very briefly explain your choice of Measurement Scale
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

10. Without looking/copying from this chapter, try to complete the following table:

Types of Variables to be Compared	Statistical Tests to be Used
Nominal vs. Nominal Variables	
Nominal vs. Ordinal	
Ordinal vs. Ordinal	
Descriptive Statistics	

11. What is a pilot study? Why is it necessary?
12. What are the questions asked post-mortem, after a pilot study?
13. List the main types of data-gathering methods.
14. When a researcher reaches the analysis phase which are the main issues to be considered?
15. What is a lineage and why is it important?
16. When it comes to research analysis there are some rules one must adhere to. List the main ones.
17. Briefly describe the ideal research report.
18. “Aims, Objectives and Research Questions”: List the 3 main steps, adopting the 3x rule.
19. What is a hypothesis?
20. What is the difference between the null hypothesis and the alternative hypothesis?

Chapter 5
From Concept to Tangibility

The techniques have galloped ahead of the concepts. We have moved away from studying the complexity of the organism; from processes and organisation to composition.

[Commenting that growing use of new technologies and techniques, from molecular biology to genomics, has proved a mixed blessing.]

Sir James Black

Quoted in Andrew Jack, "An Acute Talent for Innovation", *Financial Times* (1 February 2009).

As in all studies and processes, for every great work, there are some rules to follow: Newton's Third Law states that: "For every action there is an equal and opposite reaction". In sociological terms, as every citizen in society has his/her own rights and obligations, one is accorded rights to reside in a country but has the obligations to abide by its rules and regulations. These are commonly referred to as laws. Scientific analysis and statistics also have their rules.

Basic Concepts

Let us take an overview of some basic concepts, before embarking on the process. This process entails a flow from the ideal "idea" to the actual, on-the-ground, tangible aspects (Reeve, 1997).

- **From Ideas and Concepts....**
  - Where does one start from? The idea is to form an abstract ideal image in one's mind, based on readings and experiences. This mental image sets the ground through its summarisation of sets of similar observations or ideas. The process this step takes covers a sequential series of events termed conceptualisation, entitation, quantification and measurement. These are discussed later on in the chapter.

- **To Operations...**
  - Getting those ideas a bit closer to reality
  - One needs to identify the value of the cases in a variable

- **To Transformation...**
  - Initiating the process of linking abstract concepts into empirical indicators that can be analysed through hard and measureable steps
  - Measurement issues need to be brought to the fore through the implementation of the rules

- **To Variables**
  - Identify the characteristics or properties of the variables
    - Variables are entities that **vary**
  - Note that each variable produce two or more different values
  - These variables can be categorised as *(now very discernable through the previous chapters)*:
    - Quantitative – measures a quantity or amount
    - Qualitative – measures a quality or category
- **And finally to Indicators**
  - Indicators are self-descriptive. They indicate how to measure a value.
  - These should also indicate the W6H
  - Indicators also describe the operation (question) used to indicate the value of cases in a variable

Two examples that follow the main steps of the above are the following:

Concepts	Variables	Indicators
Drug taking	Frequency of drug use	"How many times in the last year has the offender taken drugs?"
Air Pollution	Vehicles per length of roads	"How many vehicles are registered per kilometre road length"

Let us now review some data issues, particularly the first rule one must keep in mind!

Data issues – The structures

Before we even start to understand how to get from **A** (ideas) to **B** (the variable and subsequently the indicator), it is best to start from the first basic rule…

The Rubbish Rule!

Or better known as…

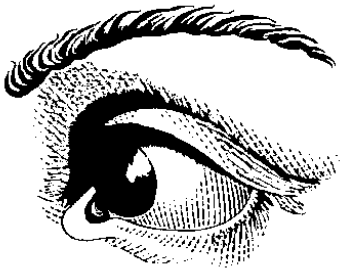
Gallois's Rule

Pierre Gallois described a very realistic rule for computer-based information which can be transposed to any type of data, information or model. The rule states that:

"If you put tomfoolery (rubbish) into a computer, you get tomfoolery (rubbish) out. But the tomfoolery coming out of a very expensive and complicated computer will have attained a sort of respectability and no one will dare criticise it"

THUS:

GARBAGE IN – GARBAGE OUT
(GIGO)



Moral of the story: *Researchers cannot hide behind the technology: but must make an honest evaluation of the validity of the data that is collected.*

As indicated in the previous pages, the process to arrive at a valid variable from a string of ideas is not an easy task. However, taking the process described through database technologies and spatial information systems, one can understand the practical ways described below in order to be able to make sense out of the data created for one's statistical analysis.

To take an abstract to a functional approach, one must start from the concept – the original idea, then structure that idea into something which can be touched in principle, then process that into something tangible on the ground, and finally ensure that it is valid to the item you want to analyse. Once all this is completed, the next step would be to bring all such issues together and enable the modelling function, where the "whole is greater than the sum of the parts".

i. Conceptualisation

Transforming one's idea into a tangible construct requires one to go through a step that allows the researcher to nail down the elusive variable. Think in terms of definition of an idea. If the definition reflects the original idea, then the concept has been established.

This is not always a straightforward task as some social issues are not easily defined. Taking an example from the physical sciences, one can readily identify an orange as having a spherical shape, as being constituted of a series of slices each containing multiple mini-sachets of juicy liquid and that this fruit can be eaten (or drank, if squeezed!).

Now try to define a social construct into a quantifiable and measurable indicator…. Easy? Not really! This is because the boundaries have been defined. These uncertain boundaries are called "fuzzy" concepts – it is not always concretely clear what the definition is. Some examples would include: social segregation, educational need, poverty and areas of special deprivation, among others.

Taking the case of deprivation, one can immediately picture someone suffering from hunger or physical lacking, however… What is deprivation? Is it not a form of poverty? Is there a bias towards one specific type of deprivation? What constitutes deprivation in terms of variables?

The fuzziness here is too real as deprivation means many things and can be analysed using many forms. One has to clarify that deprivation is more complex than poverty, which can be analysed through being short of money (the argument is based on the fact that lack of income results in lack of access to goods and services). Deprivation is based on a model or a group of variables that interlink to help form a tangible variable from the idea.

Take the example of measuring poverty in an apartment block in Valletta. A millionaire might live in an apartment on the tenth storey but is bed-ridden and has no access to a lift. As a result, his social contacts have been severed and he cannot interact with his peers. Though not poor, he is deprived. How would a researcher measure such instances to come up with one definition of deprivation? While building a model is not that easy, some, such as the Index of Deprivation in the UK[1], include hundreds of variables and are very complex. Others may only choose one variable, such as access to the community areas of a town. This process can only be enhanced through review of the literature written on similar circumstances to the ones under study. Then the researcher should review the availability of data for that variable being chosen.

This step requires the rethinking of the process and reworking the choice of ideas until refined enough to elicit the study of a realistic variable.

ii. Entitation

Having taken the mental leap and molded an idea into something that can be measured, the next step in the process is described as entitation. This step describes the way we recognise, understand and define the entities about which we wish to collect data.

What is an entity? An item that can be described and measured in statistical analysis is termed an entity. It spans all the themes under study in research: a single variable to complex models; or single relationships (psychological 1:1 relationships) to complex multiple-organism relationships (sociological multiple-individual interactions).

There are many issues at stake in ensuring the process of entitation. Policy makers may rush to measure things without first pausing to consider whether the same things are worth the trouble. It is essential to ensure that the researcher finds the right things to represent their work with and that they define the boundaries and constitution of the entity as well as the context within which that entity 'lives' (Reeve, 1997).

As an example, one could take spatial development boundaries or the NUTS 5 area boundaries (Refer to Chapter 7 for description of NUTS).

Can anyone tell exactly where Balzan's boundary really is? Where does it end and where does it start? One can equate that to a GPS series of locations on a map, but this might not be so straightforward to someone who has just had the misfortune of being involved in a traffic accident at the traffic lights leading to San Anton Gardens! The boundary is an exact fuzzy location between three NUTS 5 local councils. So, that someone who has just been involved in an accident is perplexed when trying to find out under which police district jurisdiction such an area falls.

---

[1] http://www.neighbourhood.statistics.gov.uk/dissemination/Info.do?page=analysisandguidance/analysisarticles/indices-of-deprivation.htm

The same issue can be experienced if one wished to organise a barbecue at Mistra Bay. Since such an activity requires a permit, the logical solution is to request a permit from the Mellieha Local Council. Fair enough! However, what happens when one has to state on which exact location such an activity would occur? The left-hand side of the bay (facing the sea) is within the Mellieha area, but the right hand side is within San Pawl il-Bahar's jurisdiction. Since one cannot exactly tell where such an activity would occur (due to various factors such as areas taken up by boat carriages, other barbecues, bathers, etc), and one cannot define exactly where the dividing line is drawn, boundaries becomes practically impossible to define. The most logical solution to the dilemma could have nothing to do with the geographic location of the proposed activity, but would have been taken simply to avoid the long drive up to Mellieha (if driving from the southern localities). The organisers could merely state that the activity would occur in San Pawl il-Bahar territory.

Thus, the simple process of moving from the concept to creating an activity (conceptualisation) to mould that activity (entitation) cannot be held in terms of the exact spatial location but has to be concerned more with the activity itself, the barbecue, that can be measured: Did it occur or not? Did it occur in clearly-discernable Mellieha or clearly-discernable San Pawl il-Bahar area? There is no need to be bogged down with extra detail such as the exact spatial location since that is superfluous to the activity. How many party-goers take a map, a GPS or a SatNav with them to find the exact location of the boundary? Now that would be funny! Especially considering that there is a displacement of significant metres in the technology…

iii. Quantification

The next step takes into account those entities that have been given a form. Can these now be given a measuring scale? Can the values that represent those same entities be measured?

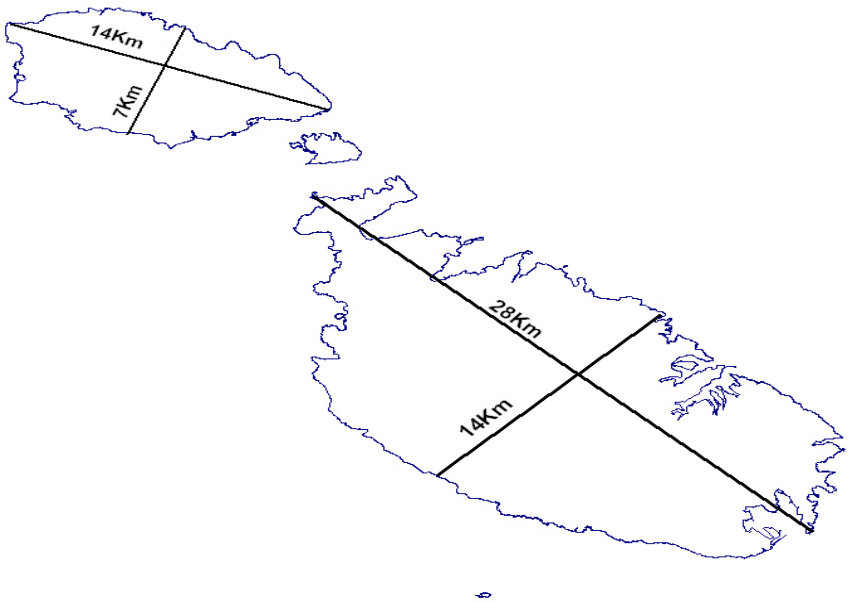
One might say that this is entirely possible.

- How tall is a person? A measuring tape will give the exact length.
- How far is Sicily from Malta? A simple browse through an analogue map or an online browse through a digital map will give you the distance.

But the doubt fiends start creeping in and ensuring that the carpet underneath our feet is not that sturdy any more. The research bugs are attacking one's solidity again. Let us check the following issues based on the above second example.

- Any map will give a distance of 90km from Malta to Sicily, however the researcher has to be more precise: Exactly from where to where has the distance been calculated? From the Valletta Grand Harbour breakwater point to Pozzallo or from any other point to any other point? The exact locations are required. Grab a computer and visit an online map and check this out!
- Another example closer to home concerns that of the geographical construct of the Maltese Islands: a quick review shows that Malta is 28km long and 14km wide, while Gozo is half that – 14km long and 7km wide (Figure 5.1). These figures are easy to remember since they are all divisible by half within the Islands and across the islands – 28/2 equals 14 and 14/2 equals 7! Easy, no? But this just gives one an approximate idea and not an exact measurement since there are no end-point locations given. Quantification requires the exact dimensions pertaining to that measurement.

Figure 5.1: Coastal Distances



Now let us go down research lane. More complex issues are encountered as one moves towards the social domain.

Can these be measured?

- How red is a red car? The eye of the beholder can mislead; red depends on wavelength and perceptual encounters. What is red exactly and how dark or light is dark red or light red? This measurement needs to be clarified before the data is gathered. Whether employing physical, chemical, textual or any other characteristic baseline analysis to ensure that one has a definition of the characteristics of the 'red', the result will hold through as a benchmark for the study and other ones which will be used to compare in the post-study reruns.

Let us take this issue to another level. The same issue discussed in terms of the red factor is encountered in terms of socially-specific data, where data on key variables in the social science research are often missing and one may need a surrogate to measure such.

- Define affluence! What is affluence and how can it be measured? How rich is a rich person? Note that relative poverty may not mean the same thing to two different social groups, so in effect the 'redness' of poverty has to be defined beforehand.

    i. As an example, one can take the case of a farmer who owns a small plot of land in a subsistence-based economy. The farmer can grow his own crop in order to survive. The neighbour, on the other hand has no patch of land and has to beg for his food. The neighbour is definitely poorer in a poor community.
    ii. The second example concerns the case of a yacht owner in Monaco. His yacht is a mere 28 footer and is berthed next to a sleek, gleaming, ultra modern 70 footer. The 28 footer owner is considered a poor fellow compared to his neighbour and could be looked down upon.
    iii. In these two cases the poverty issue is relative to the actors but also across the cases. The 'poor' fellow living in the 28 footer is definitely far richer than the 'rich' farmer working the patch! The latter is utterly poor compared to the rich guys!
    iv. Validity

Once the concept has been transformed into an entity which in turn has been given a measurement structure, then the next step is to ensure that the measurement is valid and that it represents the item under study. The setting up of an indicators list and a lineage process will help the researcher to ensure

that the particular dataset is valid and can be repeatedly gathered, analysed and processed. Indicator lists help the researcher follow the analytical process on a step-by-step basis, which will allow for comparative output analysis.

Ensuring that the data collected is a close representation of the issue is known as an Acid-Test exercise. Follow the rules, and the data processing is considered reliable. However, if your reputation as researcher depends on a dataset of dubious origins bin it!

One item relating to validity regards the use of surrogates.

Surrogate: the substitution of one variable by another corresponding or similar variable.

- Consider the case where a person or a couple would like to conceive but are unable to do so. Cases have been evidenced where a person is contracted to carry the baby for them. The 'carrier' is called a surrogate as she replaces the mother's pregnancy and parturition episodes.
- In social sciences, the use of proxy or surrogate indicators is used in order to help analyse a relationship for which data may not always be available.
- An example would include income data, which question, even in Census takes, has been removed. Such a question may either elicit untruthful or misleading answers. Thus one has to find the best choice that represents that variable, something which is related directly or indirectly, but it must be related. The Census income example would serve as an ideal case through the use of a surrogate variable such as car ownership. The latter may serve as a surrogate for wealth. The car type, cost, number of cars per household, year of production and other related variables serve as a pointer to the level of income.
- Note that any surrogate choice may introduce a measure of doubt as the choice may be deemed subjective by different users.

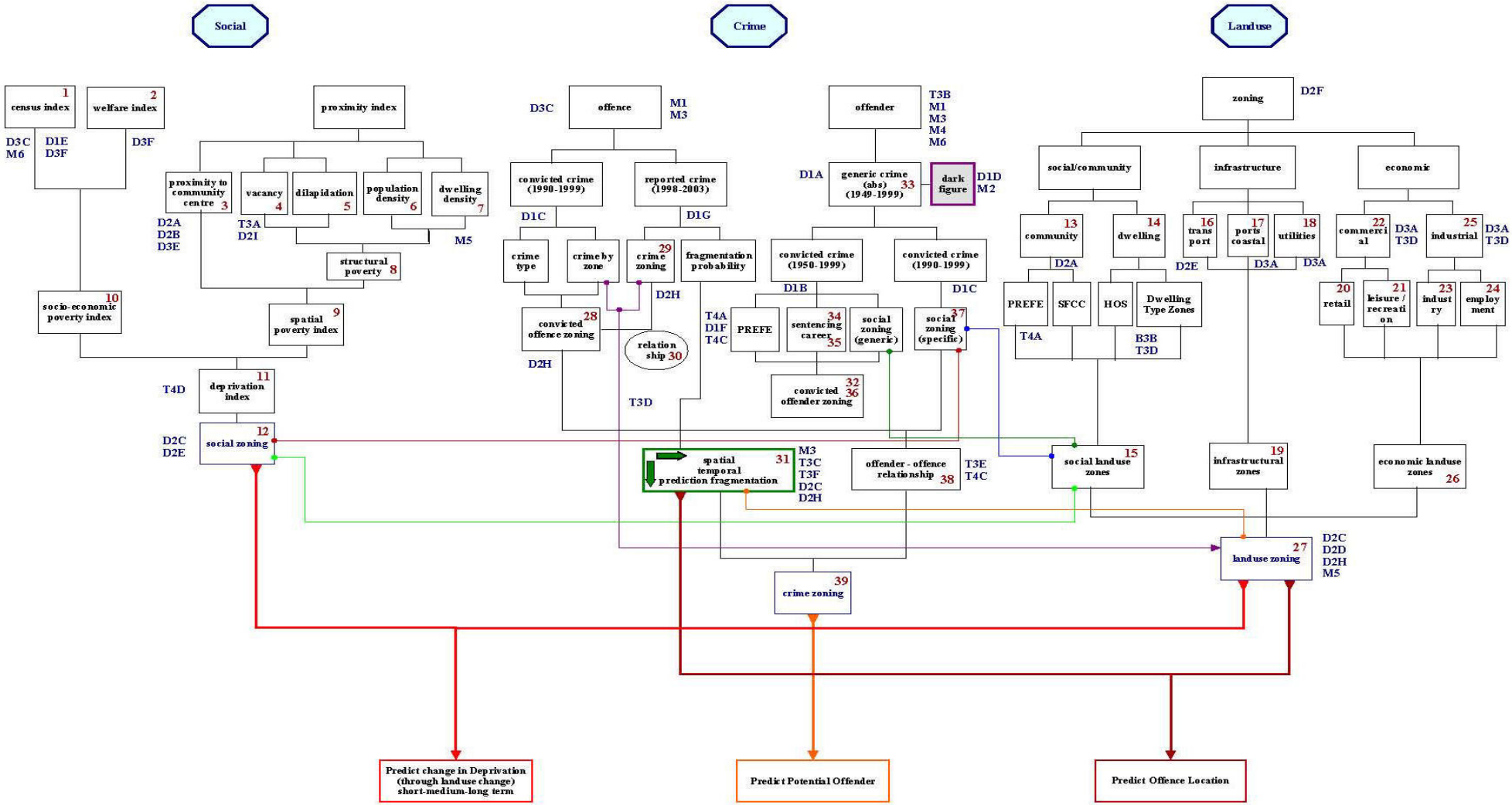
v. Composition

Finally, individual indicators are combined into a composite model and all the data comes together through a process. The composite is a sum of the parts, though in actual fact it is bigger than the sum as information and structures are not emanating from the individual variables but from the linkages between the different variables.

Composition is never an easy process but if approached in the correct way, it easily proves to be a relatively smooth exercise to engage in. This book dedicates a whole chapter (Chapter 8) to the description of how to carry out mind mapping for a research study. The mind map exercise is covered in detail through a model called CRISOLA (Formosa, 2007) which depicts the process described above from conceptualisation to composition (Figure 5.2).

In summary, the composition process allows researchers to develop concepts, visualise variables, create the linkages between the variables and also to identify the statistical measurements required for each linkage, and how they link within themes and across themes.

Figure 5.2: Crisola Model Phase 2 (Formosa, 2007)



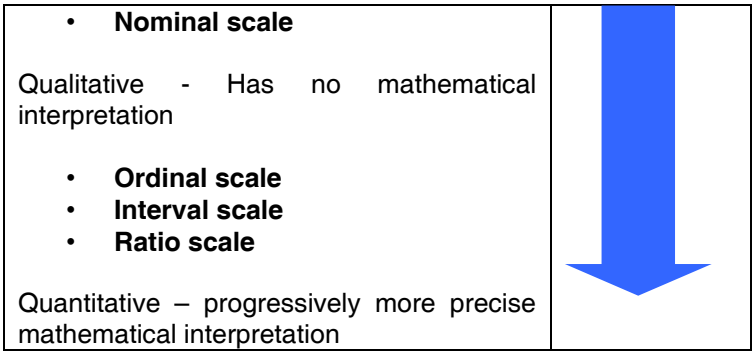
Data Measurement: The scales

Having structured one's thought on the way a research will take place, the researcher is bound to understand which measurement scales to use for each type of query. The previous chapter (Chapter 4) tackled Matrixing and there one could identify the need for the tagging of a measurement scale for every variable.

The Measurement Scales structure follows an ever more detailed and defined process of precision, which is reflected in a mathematical staged process of precision upon which the values of a variable are defined.

There are 4 types of measurement scales, each with the higher level of precision than the previous scale. While Nominal has no mathematical interpretation, Ordinal scale has a higher level of precision, with Interval more precise and Ratio the most precise.

- **Nominal scale**
- **Ordinal scale**
- **Interval scale**
- **Ratio scale**



A description of each would suffice to ensure that one understands which one should be used for the analysis process[2].

Nominal scale

The Nominal scale serves as a pointer to Qualitative variables. There are various characteristics that describe this scale.

- NO quantitative information can be gleaned from this scale
- The scale does not refer to amounts (numbers) but to tags that serve as an identification
- One item in the variable does not signify a greater composition than another item (male is not more important or bigger than female, Attard is not a less important village when compared to Zurrieq)
- The reference to categories is paramount, therefore it refers to a naming structure or a nominal structure
- Frequency distributions are used to measure Nominal scales
- The Mode is the statistic used for Nominal scales

Examples pertaining to the Nominal Scale:

- Sex: male or female
- Local council name: Attard, Valletta, Zurrieq
- Nationality: Maltese, Italian, Greek
- Citizenship: Maltese, Non-Maltese, Foreign
- Country: Malta, China, Antartica

[2] A very good reference on these terms can be found under: http://davidmlane.com/hyperstat/A30028.html)

The Bright Spark Memstore for Nominal:

IMAGINE NOMINAL – IMAGINE NAME

Ordinal scale

The ordinal scale serves the purpose of building up a series of numbers that are ordered in scope, however the number that is larger than another number may not necessarily be followed by another number with the same difference between the first two numbers. Thus an ordinal number that has a designation of 4 may not be bigger than 3 and 2 could be bigger than 1.

There are various characteristics that describe this scale.

- Quantitative information can be employed within this scale
- The scale refers to amounts (numbers) but not as an actual amount but as a representation of an amount
- One item in the variable does not necessarily signify the same difference between a sequence of numbers
    - Aims to discern who is 1st, 2nd, 3rd in recidivist cases
- Indicates Rank Order
    - Score of 1 means the most or the least
    - Score of 2 means the 2nd most or 2nd least
    - Example: first timer, recidivist, twice recidivist...
- The reference to order is paramount, thus one figure must still follow the next as they may signify severity of an issue such as: (1 – not important at all, 2 - low importance, 3 - important, 4 - highly important, 5 – extremely important)
- There is no true 0 (as 0 does not represent a number but a state) and the zero is designated arbitrarily
- The Median is the statistic used for Ordinal scales

Examples pertaining to the Ordinal Scale:

- Affirmation: Yes, No
- Age Cohorts (years): 0-4, 5-9, 10-14
- Density (population per square kilometre): (0, 1-100, 101-1,000, 1,001 – 2,000)
- Mohs Hardness Scale (1, 3, 9, 21, 48. 72, 100, 200, 400, 1,600

As a good indicator, one can use the Mohs Scale for Hardness. The actual scale refers to the differences between the numbers representing the different levels within the scale while each represents the relative actual hardness.

For example, the difference in hardness between scales 8 and 9 and that of 9 and 10 reflects an actual hardness difference of 200x and 400x respectively (compared to the previous number). Thus, though the intervals are the same (8-9 and 9-10 = interval = 1), in reality the hardness difference is major (200-400 and 400-1,600).

(http://www.galleries.com/minerals/hardness.htm)

Mohs Hardness Scale	Actual Relative Hardness	
1. Talc	**1**	Talc
2. Gypsum	**3**	Gypsum
3. Calcite	**9**	Calcite
4. Fluorite	**21**	Fluorite
5. Apatite	**48**	Apatite
6. Orthoclase	**72**	Orthoclase
7. Quartz	**100**	Quartz
8. Topaz	**200**	Topaz
9. Corundum (ruby and sapphire)	**400**	Corundum
10. Diamond	**1,600**	Diamond

The Bright Spark Memstore for Ordinal:

IMAGINE ORDINAL – IMAGINE ORDERED

Interval scale

The Interval scale serves the purpose of building up over the ordinal in more detailed precision levels. A series of numbers is ordered but a number that is larger than another number is followed by another number with the same difference between the previous and following numbers. Thus an Interval number that has a designation of 4 has the same interval difference from 3 as 2 has from 1. However one must keep in mind that 50 is not necessarily twice 25, such as that designated in temperature measurements.

There are various characteristics that describe this scale.

- Quantitative information can be employed within this scale
- The scale refers to amounts (numbers) but not as an actual amount but as a representation of an amount
- Rarely used in social sciences
- Numbers are separated by the same gap or interval
    - Difference between 3 and 4 is the same as between 1 and 2
- There is no true 0 but 0 here represents itself as simply another number
- There can be negative numbers included such as (Negative bank accounts = EUR-1000)
- The Mean is the statistic used for Interval scales

Examples pertaining to the Interval Scale:

- Income: -EUR1,000, -EUR500, -EUR0, EUR500, EUR1,000
- Temperature in Degrees Celsius: -100°C, -50°C, 0°C, 50°C, 100°C

The Bright Spark Memstore for Interval:

IMAGINE INTERVAL – IMAGINE EQUAL INTERVALS – IMAGINE CINEMA INTERVALS

Ratio scale

The Ratio scale has the same structure as the Interval Scale but zero means literally **zero**. The difference between the numbers is the same as in the Interval Scale but there cannot be any negative numbers.

There are various characteristics that describe this scale.

- Quantitative information can be employed within this scale at the highest level of precision
- The scale refer to actual and true amounts (numbers)
- Numbers are separated by the same gap or Ratio
    - Difference between 3 and 4 is the same as between 1 and 2
- There is a true 0 as it represents the base number
- 10 is exactly twice 5, 2 is twice 1
- There cannot be negative numbers (example, one can weigh **30 kilos not -3 kilos)**
- The Mean is the statistic used for Ratio scales

Examples pertaining to the Ratio Scale:

- Temperature in Kelvin: Absolute 0 (0K, 100K, 273K (relative to 0°C), 1,000K
- Age in Years: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

– Richter Scale (Earthquakes Scale where each number signifies a severity that is 10 time as sever than the previous number): 0 (no earth movement), 1, 2, 3, 4, 5, 6, 7, 8, 9,

The Bright Spark Memstore for Ratio:

IMAGINE RATIO – IMAGINE RATIOS

The measurement scales need to be employed within the Matrix as described in Chapter 4. Based on the relative scale, each variable can be given a structure within which to operate. Further description of how to carry out this process will be tackled in the following chapters.

In conclusion, this chapter has enabled the reader to progress through the processes required to make the jump from an idea onto a tangible structure known as a variable. It has also helped the reader to understand how these variables can fit within the various research processes described in previous chapters.

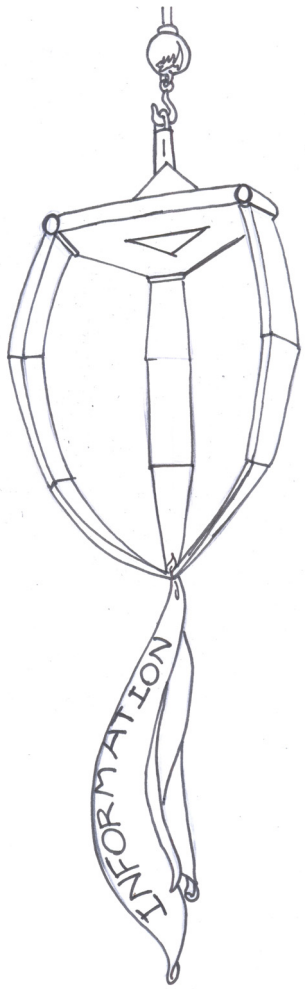
The concept now has something which can be measured. It also has a foundation for measuring that variable through an understanding of the scales that one can use to carry out the analysis. The Matrix can now be completed and the research can be set in motion.

Prior to this launch, however, one must understand issues pertaining to data acquisition and the data quality, to ensure that the type of data is actually 'clean' enough for statistical purposes.

Questions (refer to Appendix for the answers)

1. Very briefly mention the steps between the conception of a research idea and the formulation of an indicators' list.
2. What do you understand by the research rule of Pierre Gallois?
3. What is "conceptualisation" in research?
4. What is a research "entity" and what does "entitation" in research describe?
5. What do you understand by "quantification" in relation to research?
6. After coming up with a research idea/concept, transforming it into an entity and giving it a measurement structure, the researcher still has to do something else. What is it?
7. Why does the researcher need to set up a list of indicators and a lineage process?
8. Why are indicator lists important?
9. Define a surrogate variable.
10. Why is the process of "composition" (in research) important?
11. List the four main types of measurement scales.

Chapter 6
Data Acquisition and Data Quality



No generalising beyond the data, no theory. No theory, no insight. And if no insight, why do research.

Henry Mintzberg

'Developing Theory About the Development of Theory,' in Ken G. Smith and Michael A. Hitt, *Great Minds in Management: the Theory of Process Development* (2005), 361.

One must never forget the dubious origin of the source data. The resultant data sometimes takes up a life of its own. Beware, no magic formula exists! (Reeve. 1997)

The current situation is one where research is suffering from **DRIPS**

Data-Rich-Information-Poor Syndrome[1]

Data is available in vast volumes – whether in book format or in digital real-time format. One must ensure that the data that is being gathered is reliable, has been gathered in the strictest manner and that it pertains to the topic under study and not to an alternative dataset.

One is never sure of the problems and issues pertaining to data gathering, however knowledge of what that data holds is necessary. Data has to be turned into information to avoid a DRIPS situation. Again the issue of context comes up. Information needs context so one needs to know which situational background that information was based on. If in default DRIPS occurs.

The next section gives an overview of which data forms exist and focuses on the topic of METADATA, which is the process of how one ensures that data has a context within which it was created and that it serves as a veritable ID card/Passport for that particular dataset.

Data Categories

As described in the previous chapters there are various types of data ranging from raw data to structured data, there is data that represents the function it was set out to do whilst others serve as surrogates. Then there is the pinnacle of data structuring; the metadata. Each is summarily described below:

i. Raw

The data that is gathered straight from the field and left in its pure form is called raw data. This category provides a stream of data that is either continuously or periodically gathered. It can be remotely gathered or from in-situ technologies. This type of dataset requires investment in cleaning and structuring.

ii. Numerics

This refers to data that has been given a context and can be readily analysed. The data can be structured in a way as to either represent a full population or a sample.

iii. Imagery

A category that is not normally associated with data can be found under the generic term imagery. Imagery is a multi-faceted mode that submits information to the user due to: its raw content, the position it is located in and the context it is located in. This data has been given a spatial context and can be easily understood in today's world of online maps, graphics and information systems.

A short pointer to a photo should do the trick. Take a photo as shown below (Figure 6.1) and look closely at the image.

---

[1] http://www.pharmamanufacturing.com/articles/2005/397.html

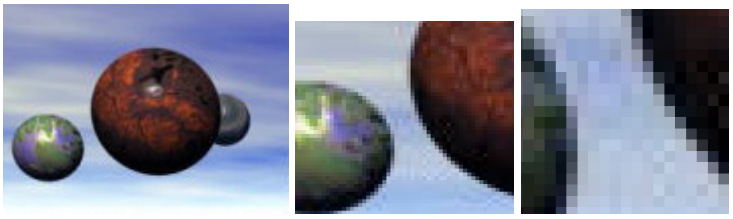
http://jobfunctions.bnet.com/abstract.aspx?docid=158118

Figure 6.1: Main Photo

The image should reveal itself as being composed of tiny dots, something also seen on a TV or monitor. Such structures are called pixels and each pixel depicts a colour, where the colour represents some kind of raw data. Each image can be composed of various collections of such pixels: for example an image of 800x600 is made up of 800 pixels wide by 600 pixels high, whilst a 1280x768 is composed of 1280 pixels wide by 768 pixels high. Quite a lot of pixels, one would say. The former image has 480,000 such pixels and the latter has 983,040 pixels. Best to leave a computing engine to calculate what data each should depict.

Now let us understand what each pixel is saying. Zooming in on a detail of the image shows that the spheres are composed of small pixels (Figure 6.2).

Figure 6.2: Pixels in an Image



Each colour (or greyscale) depicts various messages, amongst them:

- the boundary (shape of the object being depicted) – a round object that can represent a planet, a square boundary that represents a house, a polygon that can represent a field of corn;
- the containment of the pixel: the colour (represents either a direct colour or a false colour (example red would represent height, blue depth) – a blue sphere or a red one can indicate an Earth and a Mars;
- the adjacency (colour of the adjacent pixel which can describe what contained outside of the boundary) – the sky, background, the closest sphere.

Each pixel is given meaning from the above raw data through its interaction with the surroundings, again described by its context.

iv. Metadata

What is a Metadata?

One can immediately hear groans of frustration: isn't data complicated enough as it is without having to understand and learn about another level of data! Actually, metadata eases the understanding process impinged by data. It is simply data about data. A Metadata provides a description of what a dataset is composed of.

A metadata on an image might state the dimensions of an image: its width and height as well as the date and time it was taken. A metadata on a music file can describe the length of the composition, it's composer and a summary of the composition.

Whilst very detailed metadata documentation is available such as those described by International Standards (ISO 19115 and ISO 19119) under the INSPIRE[2] Directive[3] 2007/2/EC which outlined metadata specifications for spatial data, the scope here is to describe metadata in a simple form that would cover most data categories and not just specifically spatial data.

Basic elements that can be included in the metadata form include: information on the creator or guardian of the data, information on the data source and information on the contents of that dataset. Note that the metadata is not restricted simply to numeric datasets but also to imagery, documentation and any other archived and live forms.

A more detailed list would include the following items:

a. Who is the creator of the data? (the aim is to ensure that there is someone responsible for the maintenance and upgrading of that dataset)
   i. Who owns it at the current date?
   ii. Are contact details available?

b. Where did it come from? (aim is to source the original data)
   i. Are details on the original dataset or document available?
   ii. What scope was it created for?
   iii. Is it updated periodically?
   iv. When was it created in its present form?
   v. When was the source data gathered

c. What does the data contain? (the aim is to help check whether one can repeat the capture as well as identify what is held within that dataset)
   i. Title
   ii. Format
   iii. Attributes
   iv. Medium such as dataset, database, spreadsheet, map, document, image
   v. Spatial data such as scale, projection, coordinate system, bounds

d. Operational Issues (generic and overarching abstract data which would help persons querying the existence of access to a dataset.
   i. Keywords
   ii. Access issues
   iii. Charges, if any
   iv. Metadata on metadata

More detailed structures have been created such as the categories that are specified in the INSPIRE Directive Implementing Rules for Metadata, which inhabit a higher abstract level based on broad categories[4]. One can see that most are related to the spatial aspect, however one must not be derailed by the list in order not to create a metadata. If one identifies the data as non-spatial those related spatial elements can be switched off.

- Identification
- Classification of spatial data an services
- Keyword
- Geographic location
- Temporal reference
- Quality and validity
- Conformity
- Constraints related to access and use

[2] http://inspire.jrc.ec.europa.eu/
[3] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R1205:EN:NOT
[4] http://geostandards.geonovum.nl/index.php/6.6.3_INSPIRE_Metadata_elements

- Organisations responsible for the establishment, management, maintenance and distribution of spatial data sets and services
- Metadata on metadata

An online metadata creator based on INSPIRE can be found at http://www.inspire-geoportal.eu/index.cfm/pageid/342. It is straightforward to use and exports an output in xml format for further transposition to a discovery (search tool) service.

In terms of non-spatial data, the following metadata contains a list based on the INSPIRE output which can be used for non-spatial data. The following describes the broad categories of the metadata specified by the EU JRC INSPIRE Implementing Rules for Metadata[5].

- Identification
- Classification of spatial data an services
- Keyword
- Geographic location
- Temporal reference
- Quality and validity
- Conformity
- Constraints related to access and use
- Organisations responsible for the establishment, management, maintenance and distribution of spatial data sets and services
- Metadata on metadata

Always create a metadata for every datum created as it helps one to source back the relevant information and ascertain whether it is relevant for studies being carried out at considerable time post-creation. Together with a lineage, this tool helps one to ensure that the base data on which to run research is reliable, sourced and helps to ascertain whether one can use its attributes in real or surrogate forms.

Data Sourcing

Sourcing data for one's project is not a ready-made task. As technology is becoming more immersive, most organizations are creating their own datasets, allowing access either through online queries, or through a system of dedicated research units. Thus, theoretically, data sourcing is increasingly becoming transparent and easy to access. This said, one must look at the issues concerned with such sourcing.

Main questions to ask

Note that should any of these successive questions prove in the negative, then the options may be limited to a physical/manual check (where a dataset or part of it may be indicated) or to go back to reviewing the research scope. If the dataset is still required, then an actual collection by the researcher is necessary.

- Existence of a dataset
  - Does a dataset exist to one's knowledge?
  - Are there surrogates for the research/project purpose?

- Metadata
  - Is a full metadata available?
  - Does the metadata indicate that one can use such a dataset for the research/project?

[5] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R1205:EN:NOT

- Sources
    - Can one contact the sources?
    - Are the sources reliable?
    - Has one access to the sources?

- Availability
    - Is the dataset readily available?
    - Are there restrictions on access?
    - Is it protected by the Data Protection Act (of 2001)[6] and the Freedom of Information Act (of 2010)[7]?
    - Has one access to the actual dataset or to partial attributes?

- Costs and Access on time
    - Are there charges linked to the dataset?
    - Are these established on full recovery or are they based on a nominal charge?
    - Can the data be made available on time for the researcher to fit in the research/project?

- Fitness for researcher's needs
    - Once the data has been acquired, does it fit the real needs of the research?
    - Does the dataset reflect the metadata?
    - Can one go back to the originator in case the data does not fit the request?

- Acquisition
    - If the data is not made available has the researcher implemented a plan B for acquisition?
    - Can the user start the process to gather his/her own data?

Note that the acquisition of data from other sources may result in overkill or an under-representation of that data for the research needs. The data needs and interests of the creator are different from those of the researcher and as such that dataset may contain large volumes of extra data or even contain data that is not of the level required for the researcher's project.

As an example an inter-census (mid-point between the decadal census years) data request on housing stock is not available from statistics offices but can be acquired from such entities as a planning agency (which would have data on permits given but not necessarily on completions), from a utility-billing agency (which would have data on periodic metering but not on actual habitable dwellings) or from the postal office, which may have a list based on postcode but includes all types of other building types.

Each of the above can have missing or data that is not up to date, extra data such as non-dwelling units, amongst others. Therefore, the acquisition of pre-existing data limitations must be kept in mind. These include: the fact that the user has no control over the original format, over the content, over the attributes structure, over the actual data requested for as well as on the volume of data acquired.

One could even end up with very little data or a massively-populated dataset that includes extra data that is not required for the study but that the seller deems too cumbersome to dissect according to the query. Thus overpayment may occur at this stage. The best way forward is to ensure that the data costs reflect the study requirements and not to go overboard on acquisition. As stated earlier, data is very expensive and in certain cases costs more than the actual software and hardware used to analyse it.

Thus it is best to choose the datasets that fit the needs of the study and to ensure that data maximization occurs. The dataset acquired should also allow other users to gather once and use many times. This process ensures that more people can work on the same dataset and that they can add their own attributes, thus creating value-added datasets.

---

[6] http://www.dataprotection.gov.mt/
[7] http://docs.justice.gov.mt/lom/Legislation/English/Leg/VOL_16/chapt496.pdf

Data Capture

Data capture can occur through various modes which ensure that the process from data source to database is streamlined. This so-called data stream may employ manual or automatic methods of data capture. The automatic methods are governed by the rules integrated in the systems. Conversely, the manual ones may require: manual data input, data encoding, scanning, digitizing, and electronic data transfers. Knowledge of the different technologies is required, particularly knowledge of the software used and the availability of tools to help one become more efficient. These include OCR (Optical Character Recognition) technology that recognises text and converts it to digital text. More modern applications also allow dictation (speech to text) and calligraphy (handwriting) recognition.

Quality

Data quality is concerned with the creation of datasets that conform to specified requirements where the main emphasis is placed on prevention of error generation and not on post-creation analysis. To a certain extent, it is useless to analyse datasets for quality 'after' the whole process has been employed as against ensuring that the quality measures are put in place 'before' the capture even starts. An error generated at the start would be replicated throughout the capture and one may not have the chance at a second run. Other quality elements include the facts that nil defects must be generated during the capture, thus no room for error exists.

Error

Registering as hi-level quality issues, errors are nearly always generated during data capture, either through faulty processing, faulty technologies and sensors as well as data input mistakes.

Errors refer to the difference between the captured data and the real data that exists. This is sort of similar to the communication errors generated during interviewing when one person interprets a reply in a subjective way which may not be a true replica of what the interviewee meant it to be.

Errors can be categorised by type, by source, by the medium, by the technology and by the effects generated. Researchers should always be on the lookout for error generation. The following terms identify particular considerations that must be taken into account (Reeve, 1997):

- Accuracy – extent to which an estimated value approaches the true value
- Precision – level of recorded detail
- Scale and resolution – smallest size that can be displayed (for spatial datasets)
- Bias – systematic deviation from a norm or from the truth
- Completeness – extent to which data is supplied for all component parts and time periods
- Temporal consistency – repeated elements of the data handling process
- Logical consistency – suitability of commands, operations and analysis
- Semantic accuracy - quality with which objects are described
- Repeatability – extent to which independent users can produce the same data or output

Primary, Secondary and Tertiary Sourcing

There are 3 main classifications of data sourcing: primary, secondary and tertiary. Since the definition of data covers all data categories, this includes data types as defined in this book, inclusive of documentation.

Primary sources are those sources that point at data gathered first-hand. A thesis or a professional report for a company is considered as a primary source since it is composed of first-hand and unique information. This type of source also includes such first-hand information as: data gathered by other researchers (which is made available) as well as original documentation emanating from research or writing. This publication has covered such a source and termed it as 'raw data'. This data allows one to analyse at first-hand the original and unique data gathered and can run or rerun tests as well as compare to new data created by the researcher.

An example of a Primary Source: notes gathered during a field survey.

Secondary sources are those sources that are based on the findings of others such as those made available in academic journals. Researchers make references to these secondary sources in order to contrast and compare between different sources and then decide how those findings will affect their own study. Thus, secondary sources affect the outcome of how primary sources are captured.

An example of a Secondary Source: article written on notes gathered by a number of researchers during field surveys.

Tertiary sources are those sources that are not directly linked to an author or editor. These normally refer to actual data sources created by experts. A simple example of a tertiary source would be a book listing all the papers and books published on taxonomy. Papers will not be included, except for example abstracts. To a certain aspect, the academic search engines Athens, Science Direct and Ingenta all fit into the tertiary sources directory. The following is an example of a Tertiary Source: a list of titles and authors of all articles written on notes gathered by a number of researchers during field surveys and structured as an index.

The following amended General Classification (Table 6.1), sourced from the Department of Translation Studies, University of Tampere is a very useful tool to use during the differentiation of the type of source one may use.

Table 6.1: General Classifications of Selected Primary, Secondary and Tertiary Sources

Primary	Secondary	Tertiary
• Autobiographies • Correspondence • descriptions of travel • diaries • literary works • interviews • personal narratives • paintings and photographs • data gathered by researcher • data gathered through technological tools	• Biographies • prior books & papers on a topic • literary criticism & interpretation • history & historical criticism • political analyses • reviews of law and legislation • essays on morals and ethics analyses of social policy • study and teaching material • pre-prepared data	• Abstracts • bibliographies • chronologies • classifications • dictionaries & encyclopaedias • directories • guidebooks and manuals • population registers • statistics compendia

Source: Amended from: http://www.uta.fi/FAST/FIN/RESEARCH/sources.html

v. archival and real-time

Having covered quality, capture and sourcing, one other data issue that needs to be covered concerns the method of data capture. Data can be accessed depending on the sourcing. Primary data can be gathered in two main modes, that through archival and that through real-time capture.

Archival data is best referred to as original records that are gathered by the researcher and/or another researcher. This data is in its original state and has not been interpreted by others. The uniqueness issue is paramount and the archival records must be original records and the data gathered by the researcher must be his/her original work. All others fall under secondary or tertiary sourcing.

Capture Modes

One final item that has to be highlighted on data capture refers to the mode of capture. There are various ways through which one can capture data. These cover all the possible modes of capture and are not restricted to human-based capture.

Manual

This is the simplest and most commonly-employed mode. Data is gathered using time-honoured if not time-consuming effort. Researchers gather data manually through: field work, surveys, interviews and other methods as described in previous chapters. One needs to make use of analogue (hardcopies) material and then transpose findings into a storable system.

Semi-manual

This mode allows for the integration of tools together with the manual data sourcing. It includes the use of data capture technologies such as location-based maps that allow one to input data digitally into a coordinate system. This ensures that data is not inputted twice and spatial error is reduced drastically. Other tools, such as recorders (that aid the researcher in identifying keywords), fall into this category. Cameras also serve this purpose as an additional tool for data capture, particularly those enabled with face-recognition technology, voice-recognition and other innovative technologies that have become everyday tools for the researcher.

Automatic: in-situ/remote

These are the most complex and researcher-presence-free technologies. Automatic systems that gather information for the researcher are being employed more regularly in the real world and also in the virtual world. An example of the former would constitute someone who gathers data on air pollution (pm10) from an air monitoring station. In the virtual world case the researcher can employ electronic robots that gather data on users of particular sites such as the popular social networks examples of which are Facebook[8], Twitter[9] and Google Buzz[10].

In-situ automatic systems include such apparata as air monitoring stations, traffic cameras and council CCTVs. Remote apparata would include such items as drones (pilotless planes), satellites and airplanes.

In summary, data acquisition has to follow stringent rules to ensure that the data that is gathered can be transposed to information for eventual knowledge-building and action. Data acquisition depends on a plethora of capture issues such as: sourcing of new data, relevance to secondary sources, metadata structures, modes of capture and the reduction of error. Excluding any of these issues would endanger the outputs of one's study.

Questions (refer to Appendix for the answers)

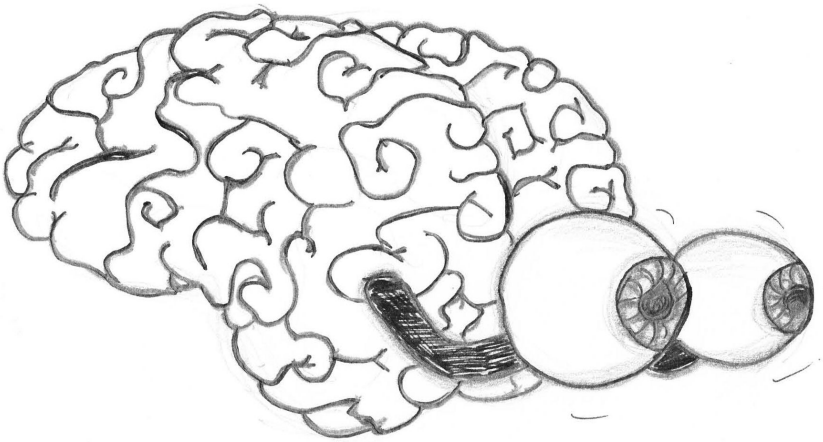
1. Research is suffering from DRIPS. What is this condition and what should be done to avoid a DRIPS situation?
2. Briefly explain what metadata is.
3. List the three main data categories.
4. Why should a researcher always create a metadata for every datum?

---

[8] http://www.facebook.com/
[9] http://twitter.com/
[10] http://www.google.com/buzz

5. List the five main issues associated with the acquisition of pre-existing data.
6. What are research errors? When are they are mostly generated? How can they be categorized?
7. To avoid errors, researchers must consider certain factors. What are they?
8. List the three main classifications of data sourcing.
9. Primary data can be gathered in two main modes. Which are they?
10. What do you understand by “archival data”?
11. List the three main capture modes.

Chapter 7
Visualisation



It was eerie. I saw myself in that machine. I never thought my work would come to this.
Upon seeing a distorted image of his face, reflected on the inside cylindrical surface of the bore while inside an MRI (magnetic-resonance-imaging) machine—a device made possible by his early physical researches on nuclear magnetic resonance (1938).

Isidor Isaac Rabi

Quoted from conversation with the author, John S. Rigden, in *Rabi, Scientist and Citizen* (2000), xxii. Rabi was recalling having an MRI, in late 1987, a few months before his death. He had been awarded the Nobel Prize in 1944, for his discovery of the magnetic resonance method.

Although statisticians consider statistics as a very exciting subject, some could claim that it is one of the most complex to understand. Brainstorm for one minute and write down the first three things that come to mind when stats is mentioned.

Were numbers on the list?
Were tables on the list?
Were questionnaires on the list?

Were graphs on the list?
Were pictures/images on the list?
Were maps on the list?

All of the above should have been mentioned somewhere but we do have our dull moments! Do not worry if only one from the second group was mentioned, since very few people actually tag statistics with imagery. NO, it is not an endless list of figures and tables, it is also transposed into visual realities. In addition, most recent technologies such as the development of interactive content and wider access to spatial information systems has brought imagery to the populous.

This output is called Visualisation. Now there is Visualisation and Visualization!

Visualisation:
Visualisation with an S refers to the actual mental image that one can see within one's mind

Visualization:
Visualization with a Z refers to the process that occurs when one converts data into a graphic representation

Thus the **S** version refers to the image and the **Z** version refers to the process employed to get there.

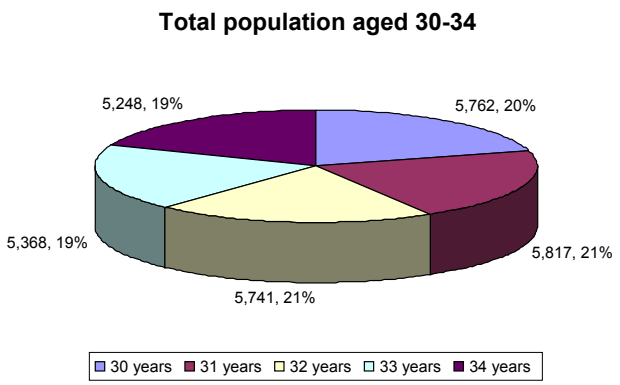
Whilst the final image is important for us to visualise data, the process is to get there is equally important. We shall discuss both in this chapter.

Whichever way one approaches visualization, the outcome is realistic enough to aid researchers in bringing to the fore a tool that renders their research easier to understand.
However this may not be easy to conceptualise for many people. Kindly review the following five visual categories that are linked to the data process. Try to identify them and then check the description below.

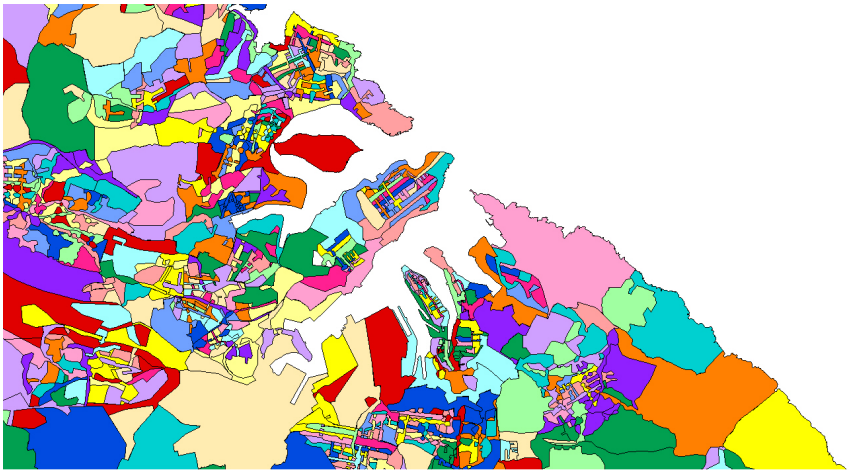
Category 1

Council	Eas_Code	NSO_Code	eas_code_for_route_	Eas_no	M_0_to_4	M_5_to_9	M_10_to_14
1	1	101	11	101	6	8	10
1	2	102	12	102	4	2	4
1	3	103	13	103	5	2	4
1	4	104	14	104	11	7	13
1	5	105	15	105	6	9	19
1	6	106	16	106	10	18	11
1	7	107	17	107	4	8	8
1	8	108	18	108	7	3	11
1	9	109	19	109	6	4	11
1	10	110	110	110	8	5	2
1	11	111	111	111	3	8	17
1	12	112	112	112	5	11	14
1	13	113	113	113	6	3	14
1	14	114	114	114	2	4	3
1	15	115	115	115	5	11	8
1	16	116	116	116	10	9	16
1	17	117	117	117	7	5	13
1	18	118	118	118	8	9	5
1	19	119	119	119	11	15	15
1	20	120	120	120	5	5	12
1	21	121	121	121	9	10	13
1	22	122	122	122	10	14	10
2	1	201	21	201	11	12	10
3	1	301	31	301	10	18	30
3	2	302	32	302	7	10	18
3	3	303	33	303	13	12	14
3	4	304	34	304	14	12	12
3	5	305	35	305	7	9	11
3	6	306	36	306	8	8	10
3	7	307	37	307	8	10	23
3	8	308	38	308	11	8	6
4	1	401	41	401	9	11	15
4	2	402	42	402	13	8	14
4	3	403	43	403	16	9	16

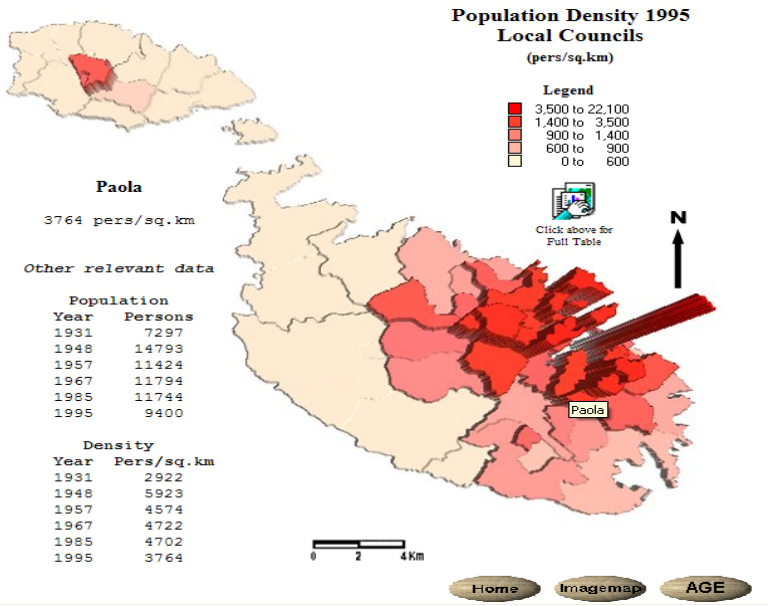
Category 2



Category 3



Category 4



Source: Formosa, (2000)
http://www.mepa.org.mt/Census/archive/age/Pop%20Density/popdens3D.htm

Category 5

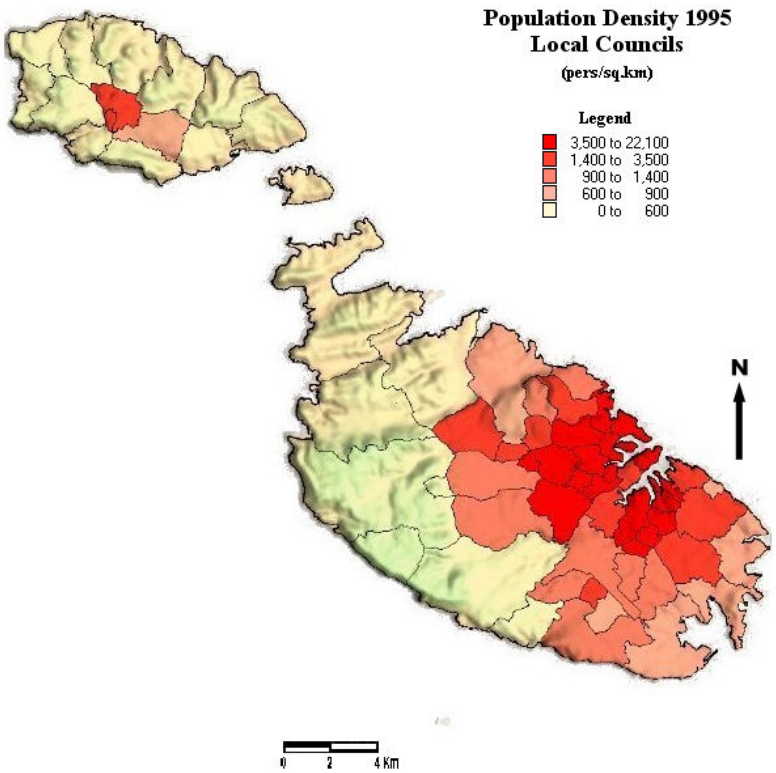
Source: MEPA

The five categories are easily discernible:

Category 1:	Table	Depicts actual data in cells
Category 2:	Chart	Depicts data analysis result in a simple form
Category 3:	Map	Depicts data on a map
Category 4:	3D Map	Interactive map
Category 5:	Photo	Depicts data in purely untainted visual format

Each category serves a specific purpose. One can read data off a table but needs to mentally visualize it in order to understand the relationships between the different attribute contents. The chart can be understood immediately but needs further information which would be contained within a table as there is no description of what the numbers and percentages are actually describing in relation to which population totals. The map is an interesting tool which requires a legend (Figure 7.1) to help readers to understand what those colour represent: is it population, area and if area, is it in kilometres squared, hectares, etc … ? The 3D interactive map takes data visualization to another level. Whilst the map may be difficult to understand, unless one has a tool to visualize it in 3D, the actual data is shown every time the mouse moves over one of the councils. The photo, on the other hand, depicts an image taken from a plane which shows urban development on the ground. It is interesting to note that one can read many items off a map from building types, to roof area, to number of cars to approximate time of days as based on shadowing analysis and a myriad of other data items. Thus, one must not make the mistake of assuming that a photo is a dead image: it speaks volumes.

Figure 7.1: An Example of a Legend



The short introduction given above shows that the eye can read diverse items off a data category. From a simple table to a complex 3D interactive tool, there is a veritable sea of information to be garnered from the use of one of the most vital human senses: sight.

However, visualisation does not stop here. This is only the beginning! The above are just five simple types of visual tools that can be used. An excellent research study by Lengler and Eppler (2007) resulted in the collection of a multitude of visualization methods which they adventurously called the Periodic Table of Visualization Methods. Their work, which is a very interesting case in their comprehensive structuring of visualization processing, was transposed into what is termed the Visual Literacy Project[1].

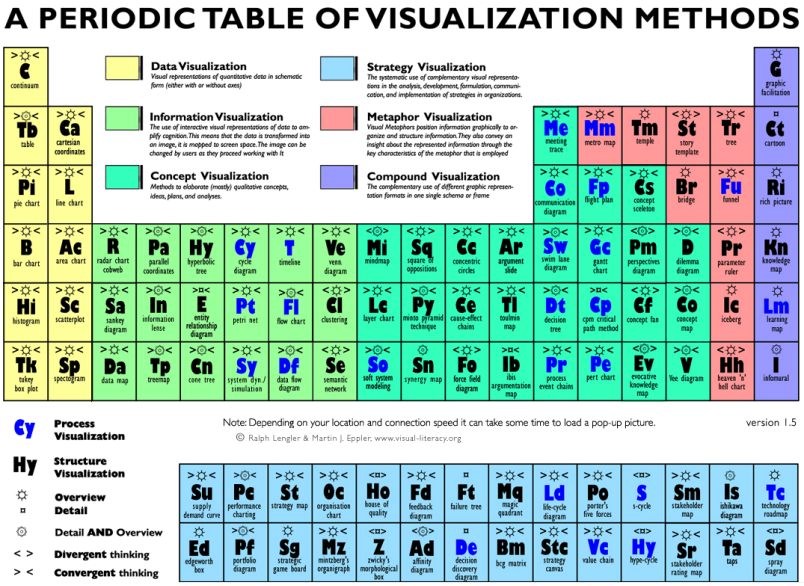
Lengler et al (2007) split the methods into six visualization categories based on what they called the Data, Information, Concept, Strategy, Metaphor and Compound approach (DICSMC).

Data Visualization	Data in schematic form
Information Visualization	Data transformed to an image
Concept Visualization	Qualitative approach
Strategy Visualization	Systematic approach
Metaphor Visualization	Structuring information
Compound Visualization	Combining different methods

This so-called DICSMC approach runs parallel to the DIKA one and structures each of its items towards the formulation of a final outcome; that of usage of the tools for implementation. The table also outlines whether the methods define processes or structure. Each of the 'elements' in the periodic table represents a visual tool which results in one of the most comprehensive tools available totalling a list of 100 methods. The table value-adds to the overall visualization concept as it identifies those 'elements' that focus on convergent or divergent thinking, which is an ideal way of expressing oneself away from traditional methods, especially where new concepts are being researched.

[1] http://www.visual-literacy.org/index.html

Figure 7.2: Periodic Table of Visualization Methods



Source: Lengler et al (2007)
http://www.visual-literacy.org/periodic_table/periodic_table.html

The reader is urged to review the tools identified in the table and partake to those highlighted in their specific area of specialisation. One of the tools covered in this book, (namely, the mind map), is listed as the concept visualization element MI.

Chapter 11 lists a number of books pertaining to the different disciplines, which publications serve as a focus for statistical tests as well as a pointer towards the methodological processing aspects that are endemic to that specific discipline. The rest of the chapter will cover the two categories of visualization most employed in research, those pertaining to graphing and mapping.

Graphing

Graphical representation through shapes is a way to summarise data through a visual medium. Graphs are vital to the understanding of data and how it is represented in non-tabular format.

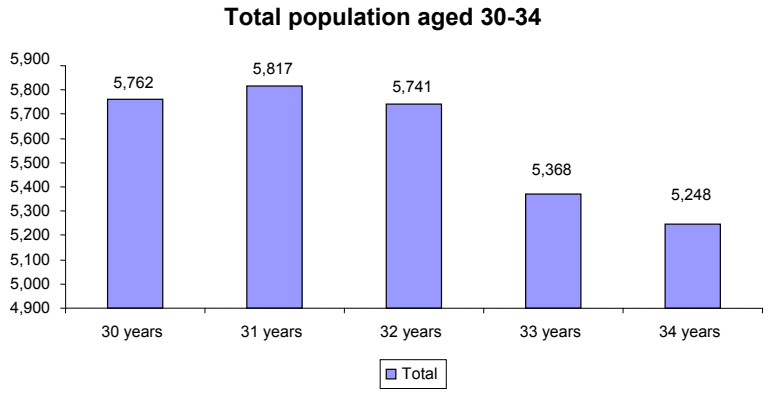
The formats mostly used include the following: Bar Charts, Pie Charts, Line Charts and Histograms.

- **Bar Charts**

Bar Charts are composed of bars separated by spaces

Ideal for displaying the distribution of variables measured at the nominal level.

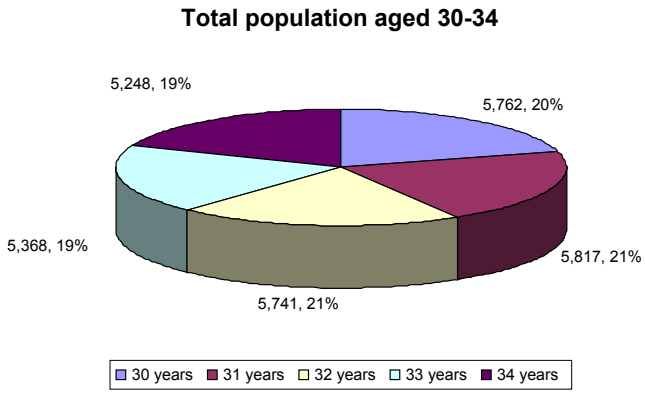
Figure 7.3: Bar Chart



- **Pie Charts**

Circles (Pies), as in the case of bar charts display their data in the form of slices. All the slices make up a cake or a pie!

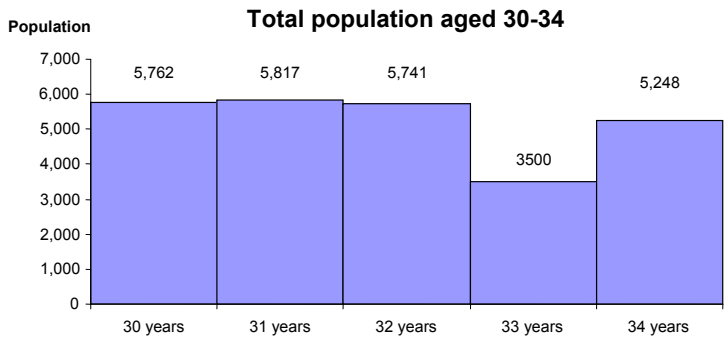
Figure 7.4: Pie Chart



- **Histogram**

Very similar to bar charts but depict a distinct difference. Adjacent bars used to display the distribution of quantitative variables. These variables vary along a continuum with no gaps.

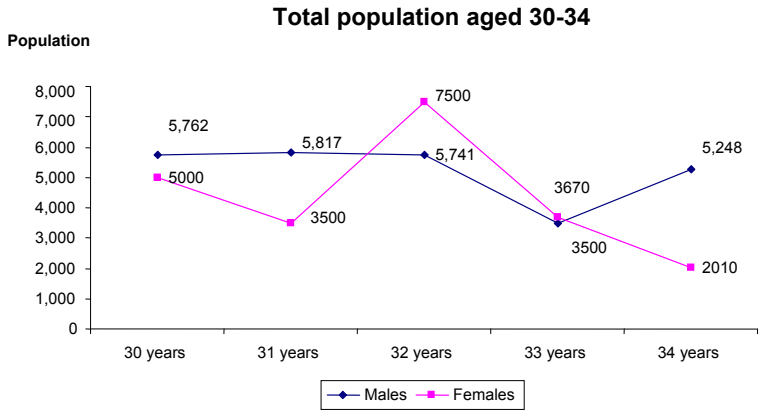
Figure 7.5: Histogram



- **Line Charts**

Line Charts are composed of lines along an axis. This type of chart allows multiple variables to be depicted in the same chart.

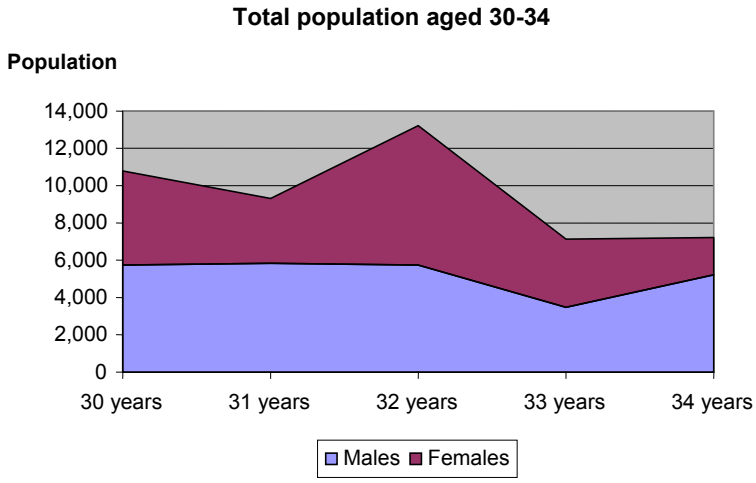
Figure 7.6: Line Chart



- **Area Charts**

Area Charts are ideally used for data that requires depiction of individual variables in relation to a total.

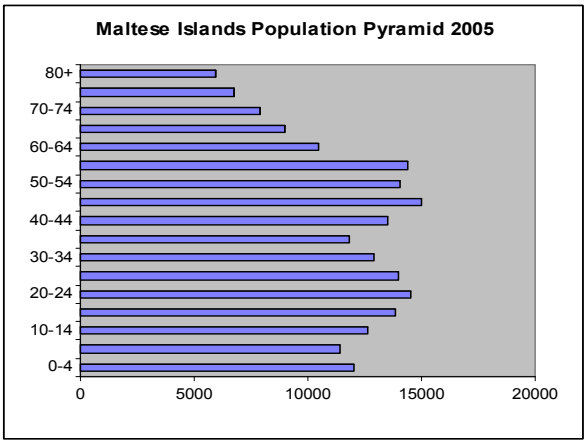
Figure 7.7: Area Chart



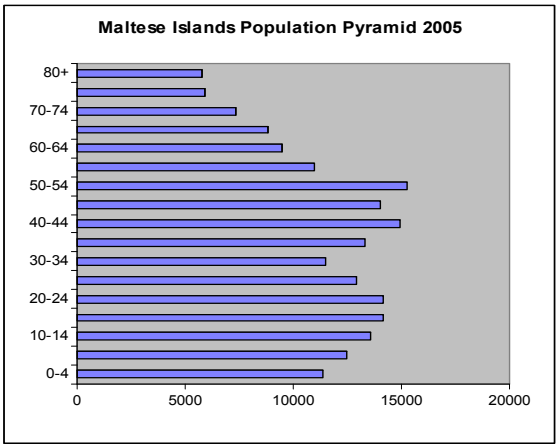
- **Composite Charts**

In statistical research, sometimes composite charts help one to better understand a situation. Let us consider the case of what is termed as a population pyramid. This is essentially a Bar Chart that has been inverted to form horizontal bars.

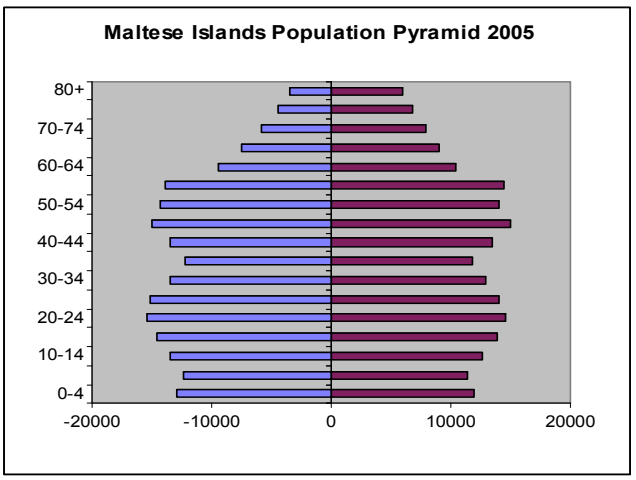
1. The following chart depicts the Maltese Female population in 2005 by age cohort.



2. Another chart depicts the Male population in 2005 by age cohorts.



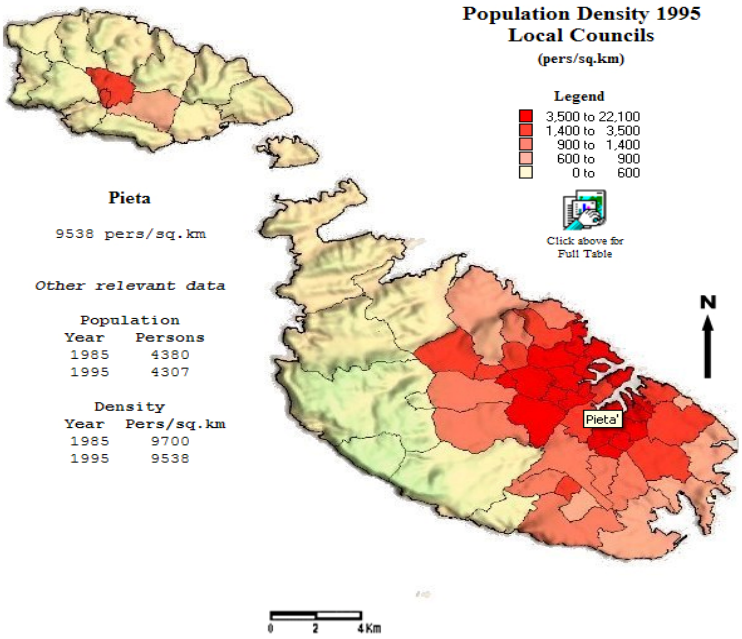
3. It is best if both are combined in order to be compared against each other. The population pyramid always depicts males on the left and females on the right. One can immediately state that the population is growing older as more people move into the higher age groups and that their are more females at the higher age groups compared to the males.



This tool has been developed in static imagemap format (Formosa, 2000) and in interactive format for the Maltese Islands (SAGIS for NSO Malta, 2009).

Static Imagemap Format

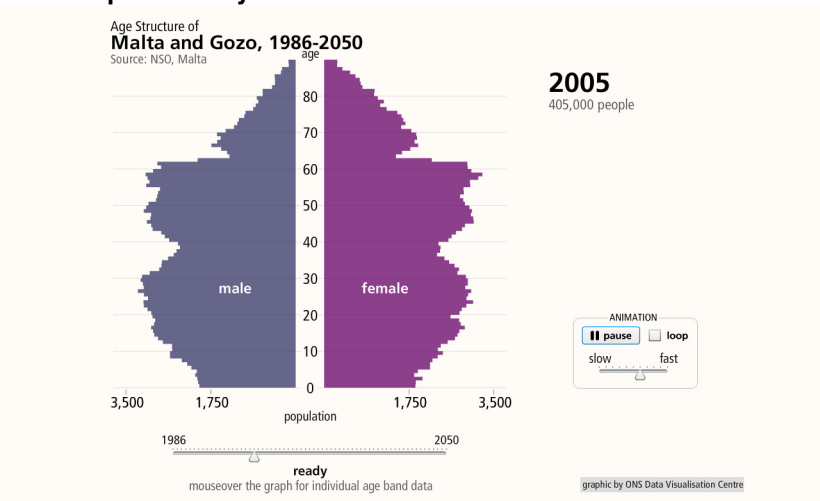
Figure 7.8: Imagemap



Source: Formosa (2000)
http://www.mepa.org.mt/Census/archive/ageimagemap.htm

Interactive Format

Figure 7.9: Interactive Population Pyramid



Source: NSO Malta, 2009
http://www.fernuni-hagen.de/statliteracy/chapter4/Malta_Pyramid/pyramid6_29.html

Charting Tools

Many tools can create charts as are generic Office tools that have spreadsheets integrated in them. However there are specialised tools in the commercial and opensource domains that target specifically the creation of static and interactive charts.

Commercial applications include Microsoft Excel[2], SmartDraw[3], amCharts[4], and the ozgrid[5] collection.

---

[2] http://office.microsoft.com/en-us/excel/
[3] http://www.smartdraw.com/

Opensource tools also exist which allow one to create static and interactive charts which are free to use. Amongst these one can use, OpenOffice Calc[6], FChartsSE[7], The Google Visualization API[8], lovelycharts[9]. A comprehensive list of such tools can be found at the free DOWNLOADScentre[10] and at New Free Downloads[11].

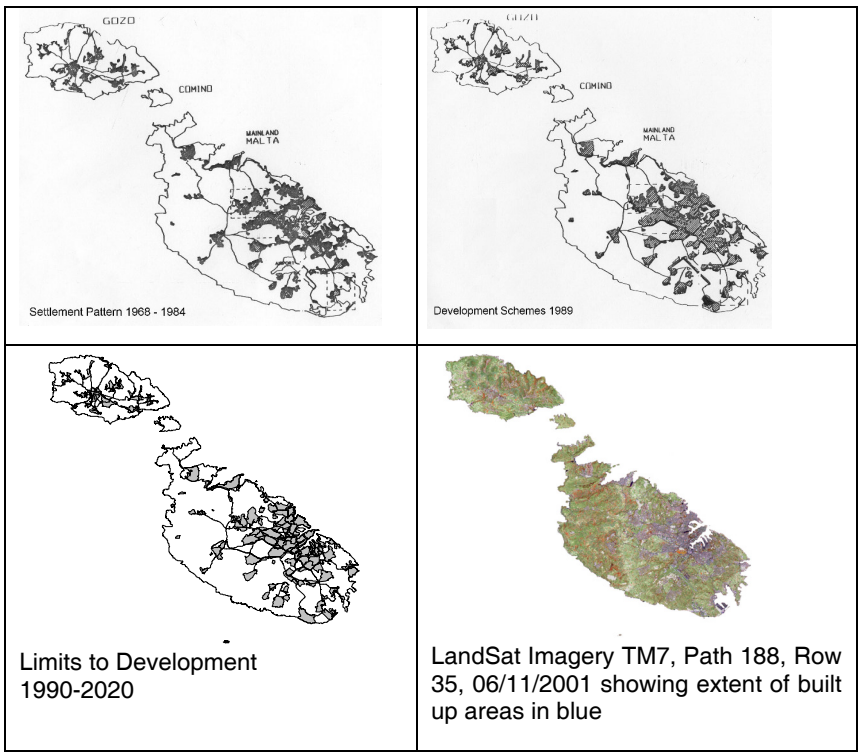
Mapping

Why is mapping so important for statistical analysis? Whilst not necessarily important for calculations and measures related to numerical studies, it is becoming increasingly difficult to carry out statistical analysis without the employment of mapping or spatial techniques.

Let us initiate this part of the study with a review of a map series depicting development in the Maltese Islands. Sourced from various entities and sensors, Formosa (2007; 91) outlined the changes over time in a combination of drawings (1910 – 1989), a map (1990 – 2020) and imagery (2001).

The drawings were originally extracted from old maps and drawn onto a map of the islands. The 1990-2020 map is an extract from GIS data layer (data is stored in a series of layers on top of each other as in the case of a stack of transparencies. This allows analysis across the tables in a spatial format). The satellite image is virtually a photo from space which empowers the researcher with an ability to code each individual colour and relegate it to a particular land cover, in this case the bluish/purple colour represents the urban land cover.

Figure 7.10: Urban Growth 1910 to 2001



Source: Formosa (2007, 91), Structure Plan Report of Survey Vol. 1 (August 1990), LimDev GIS Layer – MEPA, LandSat 2001 Imagery.

---

[4] http://www.amcharts.com/
[5] http://www.ozgrid.com/Services/ExcelChartTools.htm
[6] http://www.openoffice.org/
[7] http://www.spacejock.com/FreechartsSE.html
[8] http://code.google.com/apis/visualization/interactive_charts.html
[9] http://www.lovelycharts.com/
[10] http://www.freedownloadscenter.com/Search/chart.html
[11] http://www.newfreedownloads.com/find/chart.html

On reviewing such a series of maps of images, one can immediately reach the conclusion that the expansion of the urban areas took off after 1968 and expanded rapidly in the 1980s and 1990s forming a very large conurbation.

Tables and charts would give us the exact rate of change but the visual aspect is more direct and to a certain extent highlight the rapid growth that accelerated over the decades.

Today, technologies such as image recognition software help one to carry out this analysis automatically. Software such as eCognition[12] represent a step forward in the analysis of massive volumes of data that such maps generate. To give an idea of the file sizes between a table and a map (grid file format). Whilst the former can register a size of 30Kb, a grid file can take up 2Gb which is equivalent to 2,000,000Kb or 67,000 times larger than the excel file. However the output is immediate and significant.

GIS as a Tool for Scientific Research

The use of mapping tools has been debated for a long time and though used by the military since the 1960s, it really took off in the 1990s due to the physical and environmental sciences and has since been taken up by the social sciences post-2000.

Formosa (2007; 3) in his focus on environmental criminology states that:

> Until recently, most crime investigations concentrated on non-spatial sociological issues whilst some painstaking geographic research looked into specific locations but only in a descriptive way (Campbell, 1993). The advent of high-end information systems and spatial software has changed the direction that these studies are taking. Environmental criminology (as one such a theme delving into GIS) has been brought back as a theoretical issue through the use of Geographical Information Systems (GIS), which has become one of the main means of bringing together previously disparate research analysis (Openshaw, 1993). The use of GIS together with other tools (such as SPSS, Vertical Mapper and CrimeStat), enhance analysis over more than 2 dimensions. It integrates both spatial and temporal crime, whilst linking crime statistics to such information layers as development and urban sprawl, crime hotspots, social and community facilities, locations of policing infrastructure, location of crime near bus stops, amongst others (Hirschfield, 2001; Haining, 1987; Clarke, 1995). In addition, analysis and dissemination tools such as 3-Dimensional mapping, Virtual Reality Modelling Language (VRML), and Web-mapping give access to researchers to carry out comparative spatio-temporal analysis. This said, caution must be taken to understand the limitations of such systems and methodologies (Pease K., 2001)[13].

Whether one is studying crime, population movements, air pollution dispersal, noise dispersal, ecology and a thousand other themes, one needs to be aware that there are tools that one can employ to aid statistical analysis that go beyond tabular and graphing methodology. One has to note that irrespective of the glitz being shown by such maps, there are also limitations and one needs to be specialised in the field to ensure that the tools are used correctly.

What is a GIS?

GIS is known by many names, from Geographical Information Systems / Geographic Information Systems to Spatial Information Systems to Land Information Systems to Automated Mapping/Facilities Management and Geomatics. All these have been integrated into the term Spatial Information Systems.

By definition, GIS is:

> A geographical information system is a group of procedures that provide data input, storage and retrieval, mapping and spatial and attribute data to support the decision-making of the organization. (Grimshaw, 1994)

---

[12] http://www.ecognition.com/

[13] Such a methodological debate is a hot topic in the CrimeMap list (15th August 2006) between the digital-leaning school Dr. Ned Levine (CrimeStat III) creator and Prof. Marcus Felson who promotes the traditional methodologies of crime analysis.

GIS combined spatial data tools, cartographic tools, computer-aided design and remote sensing technologies. This combination, together with its specialist tools, has resulted in a very powerful technology that is only now becoming widely recognised as a very useful tool for researchers in the natural, physical and behavioural sciences. Due to this change, later definitions included the people factor which is now seen as the most important factor!

Very few specialists used to work with GIS as a spatial analytical tool that goes beyond the simple generation of maps. Theoretical and practical issues started spreading beyond mere use to incorporate the hard-scientific physical and earth sciences approach to the more complex fuzzy concepts identified by social-scientific theories. This led to a transformation in the use of maps and spatial tools for research.

In addition, the advent of online maps, mobile gps and real-time mapping services changed all that. Today anyone can go on line and access Google Maps[14], Multimap[15], National Geographic Map Machine[16], MapQuest[17], Bing Maps[18], Yahoo! Local Maps[19], and many others. Thus mapping is no longer the domain of a few specialists but is a tool for all. Also, the tools can be handled by many disciplines and have taken up the work formerly carried out by cartographers and geographers, with the terminology also changing from geographic information systems to spatial information systems. This change has been brought about by the inclusion of context in the various physical, natural and behavioural sciences.

A S.W.O.T. analysis on the USE of GIS for scientific-mapping research

Before one can decide to employ GIS as the main tool in scientific analysis, a SWOT (Strengths, Weaknesses, Opportunities and Threats) exercise was carried out to enable the reader to understand the issues that emerge when implementing such a scientific-mapping system that does away with the rose-tinted glasses perspective of a one-solution product. GIS as used for scientific-mapping has its positive and negative aspects of the technology and its service in a 'scientific' construct. The SWOT analysis helps to clarify these issues.

Each part of the process is analysed for its technical, policy-social-environmental and, marketing-economic functions (Table 7.1).

Table 7.1: SWOT Analysis of the Use of GIS for Scientific Research

SWOT Analysis of the USE of GIS for Research Strengths, Weaknesses, Opportunities, and Threats	
Strengths	***Weaknesses***
Technical • Immediate availability of data to analysts, researchers • Queries are automated and pre-formulated letters sent to decision/policy makers • Attribute data available on one single keyboard stroke linked to a map • Routines automate queries and instructions through cross-referencing • Use of Common Database (CDb) eliminating need for redundancy in databases	**Technical** • Potential bias by employees in favour of older non-technological systems • Confidentiality issues • Inputting, updating and reading rights are not always adequate - wide access • Distribution of data to a large number of people: staff • Incompatibility with older datasets/systems • Prone to rare but possible data theft or sabotage particularly by "rogue

[14] http://maps.google.com/
[15] www.multimap.com/
[16] http://maps.nationalgeographic.com/map-machine#s=r&c=43.74999999999998, -99.71000000000001&z=4
[17] http://www.mapquest.com/
[18] http://www.bing.com/maps/?wip=2&v=2&style=r&rtp=~&&msnurl=home.aspx?%26redirect%3dfalse&msnculture=en-US#
[19] http://maps.yahoo.com/

• Use of buffering analysis and zonal searches can be carried out within single-theme data layers as well as multi-theme layers. Buffering methodology identifies activities within a specific area from the area under study such as offences occurring at a distance of 100m from a school. Zonal searches review population movement-patterns occurring at differing distances from the area under study such as interactions occurring every 100m from a bar	professionals". E.g.: having access to maps of areas of affluence and unprotected areas turns a security map into a treasure-trove of opportunities to an offender.
Policy, Social and Environmental • Faster analysis of alarm calls such as pollution-threshold exceedances • Determination of the effects different parameters types have on different physical and socio-economics variables • Crime analysis results can be utilized by a number of disciplines and activities such as real estate estimation, fraud, security companies and social services suppliers • Integration of data from different sources, leading to improvement in rapid reaction delivery and projections based on trends	**Policy, Social and Environmental** • Specialised data is viewed as being the domain of the managing agency rather than social scientists in general – data is kept at a distance through a series of barriers to access the data • Limited support from management and non-technological-oriented chiefs • Lack of understanding by policy makers of the process to mine and analyse data for analysis • Lack of skills in information technology and information systems by social science students and practitioners • Some data is seen as too dangerous to research as it highlights a nation's weak points or failure in policy making and GI outputs makes it even more dangerous
Marketing and Economic • Real-time mobile input can be easily updated and defaulters acted upon • Incentives for research-related organisations to invest in new technologies • Reduction of data entry errors and overhead costs for field-based and office work	**Marketing and Economic** • Policy and Decision makers take time to realize the utility of GIS • High initial cost of hardware and software plus cost of training, cost of managing the GI system, costs of updating the data and costs of answering queries and requests for information • There is no 'monetary' profit in these activities and hence refusal to see profit against reduction in staff time to analyse data

Source: Adapted from Formosa (2007)

SWOT Analysis of the USE of GIS for Research Strengths, Weaknesses, Opportunities, and Threats	
Opportunities	***Threats***
Technical • Establish a real-time variable identification system • Link to international datasets such as those created by WHO, EUROSTAT • National and International quick analytical function • Need for the setting up of systems compatible with the mapping agency's GIS allowing future exchange of data	**Technical** • Inputted data is not updated regularly • Changes in categorizations can lead to incomparable results • New hardware can make whole systems obsolete • Project stoppage midway through completion
Policy, Social and Environmental • Identification of specific areas of hot-spots • Provide progressive incident-reduction environments • Work towards data sharing schemes • Identification of potential needs by victims of exceedances	**Policy, Social and Environmental** • Data is mishandled or misused by non scientists • Political and economic uncertainty impede investment • Poor timing of decision making • Assumptions can be "mistaken" leading to wrong and costly decisions
Marketing and Economic • More action taken by operational staff, less time in office allowing more efficient and effective outcomes • Wide availability of data to other related agencies for inter-disciplinary data exchange • Spin-offs of the attribute section of data • Allows time-scheduling - third parties are better served	**Marketing and Economic** • Data sold/bought at exorbitant prices that do not reflect reality (includes social data) • Real-time analysis needs real-time updated 3rd-party data exchanges that reflect the ground-truth such as new development, new transport routes

Source: Adapted from Formosa (2007)

The SWOT descriptions given here can be carried out for any of the other methods and processes listed in this book. The choice of GIS as the conveyor for such a description was made on purpose due to its integration of software, hardware and the human component in the inherent DIKA steps within which it operates.

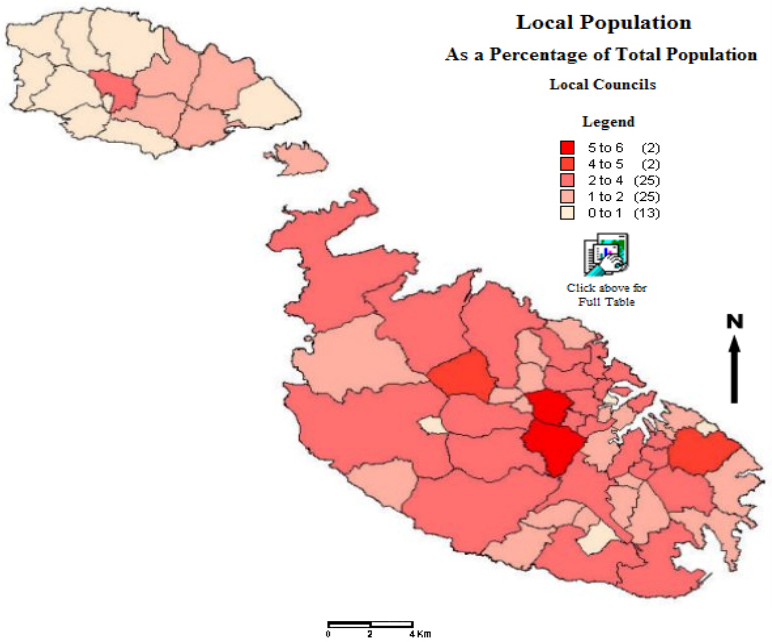
Some mapping outputs:

It is best at this stage to depict some outputs in mapped format which users can use in conjunction with their tables and charts.

Map 1: Choropleth Map

A map that depicts data based on ranges.

Figure 7.11: Choropleth Map

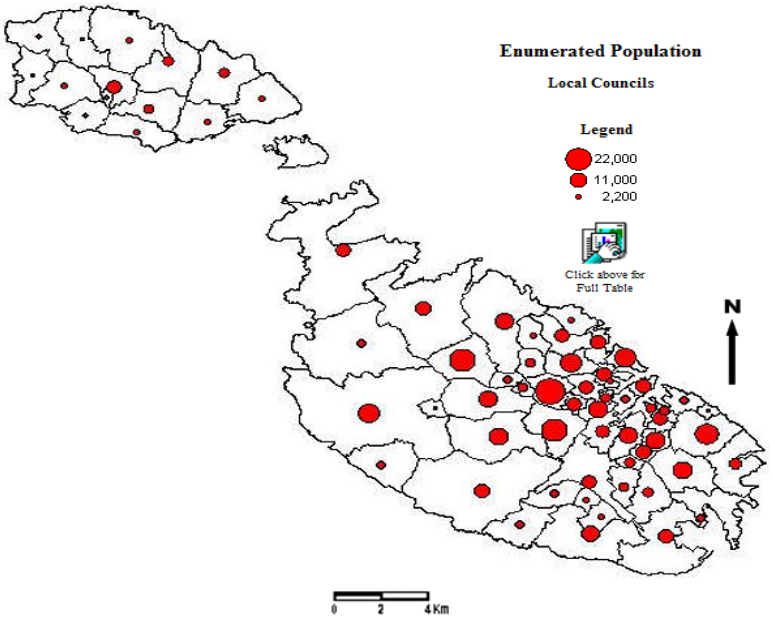


Source: Formosa (2000)
http://www.mepa.org.mt/Census/archive/ageimagemap.htm

Map 2: Graduated Map

A map that depicts data as a series of graduated points (which could also comprise pie-charts).

Figure 7.12: Graduated Map

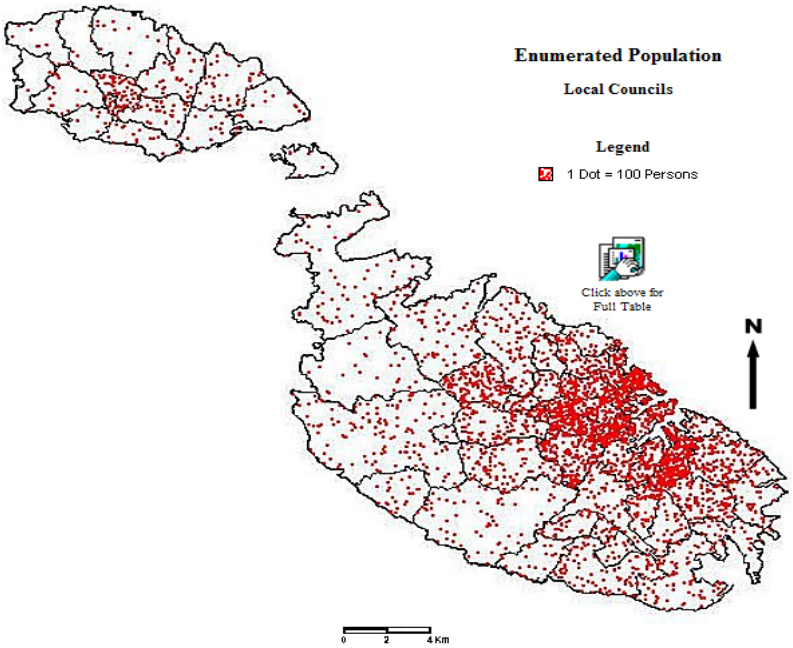


Source: Formosa (2000)
http://www.mepa.org.mt/Census/archive/ageimagemap.htm

Map 3: Dot Density Map

A map that depicts data based on randomly-located dots representing numbers of cases.

Figure 7.13: Dot Density Map

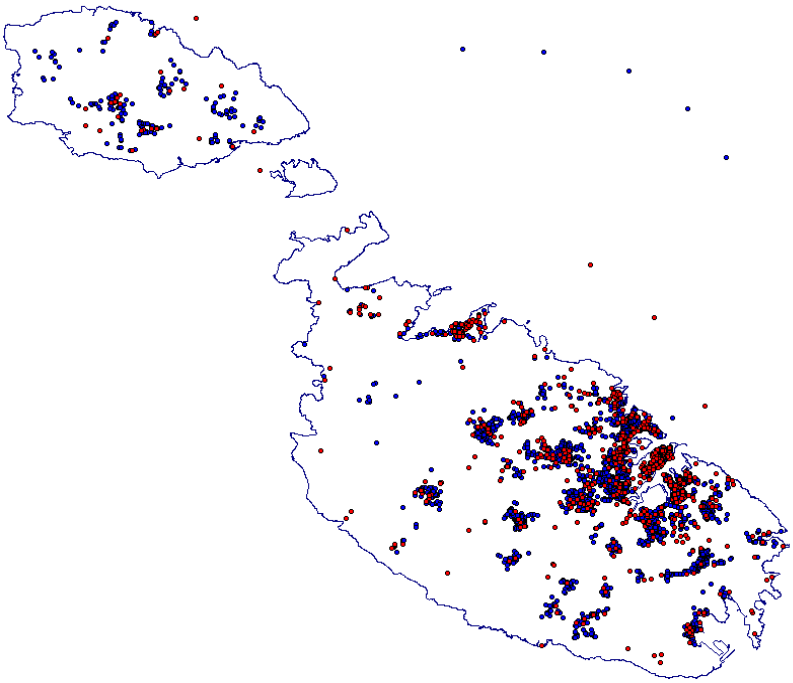


Source: Formosa (2000)
http://www.mepa.org.mt/Census/archive/ageimagemap.htm

Map 4: Point Map: Actual Location of Offences Map

A map that depicts data based on points representing the actual location of an activity

Figure 7.14: Point Map

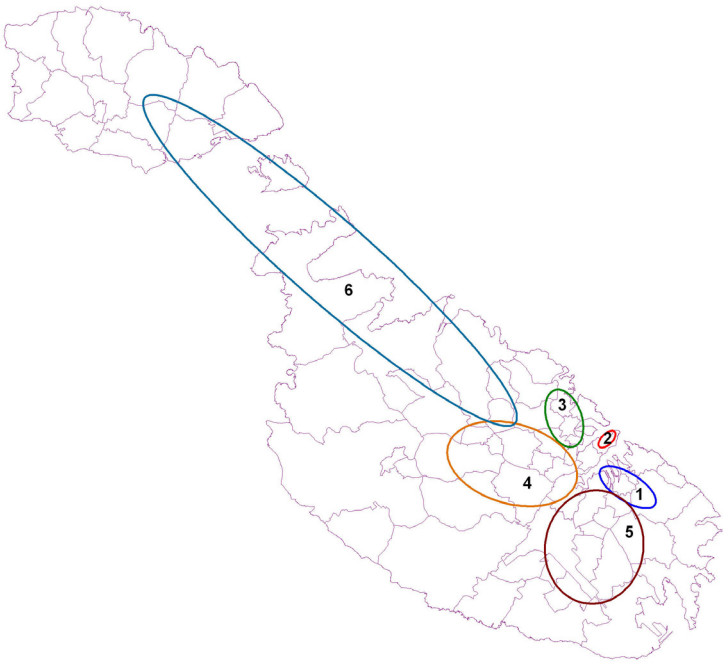


Source: Formosa (2007)

Map 5: K-Means clustering Map

A map that depicts data based on statistical clustering of related data points.

Figure 7.15: K-Means Clustering Map

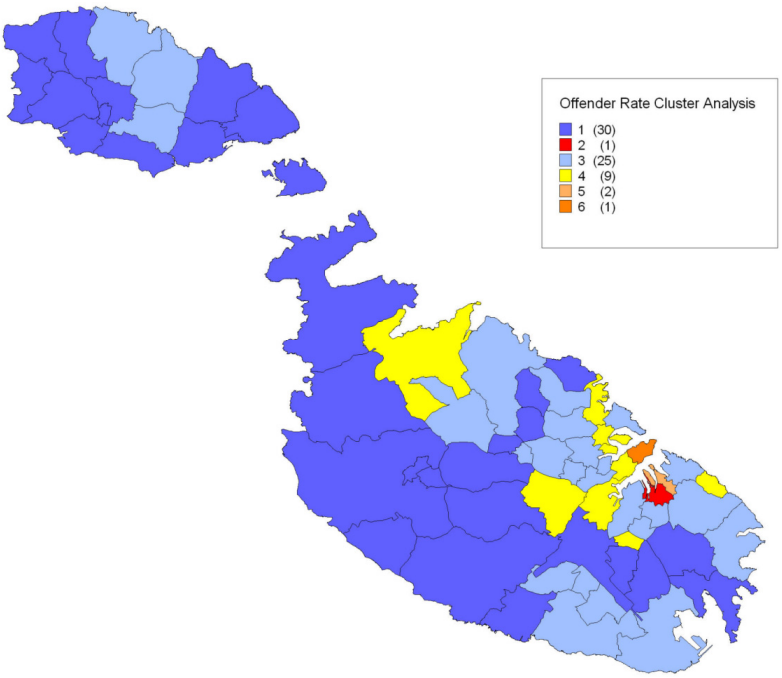


Source: Formosa (2007)

Map 6: Polygon-Based Cluster Analysis

A map that depicts cluster data ranged across polygons (areas).

Figure 7.16: Polygon-Based Cluster Map

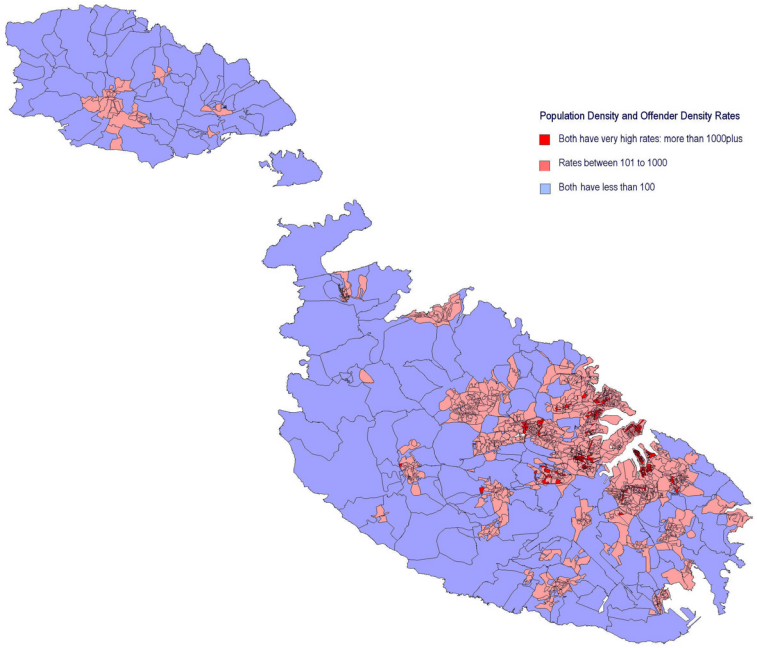


Source: Formosa (2007)

Map 7: Small-Area Choropleth Map

Same as Map 6 but depicted very small polygons (areas) for niche analysis.

Figure 7.17: Small-Area Choropleth Map

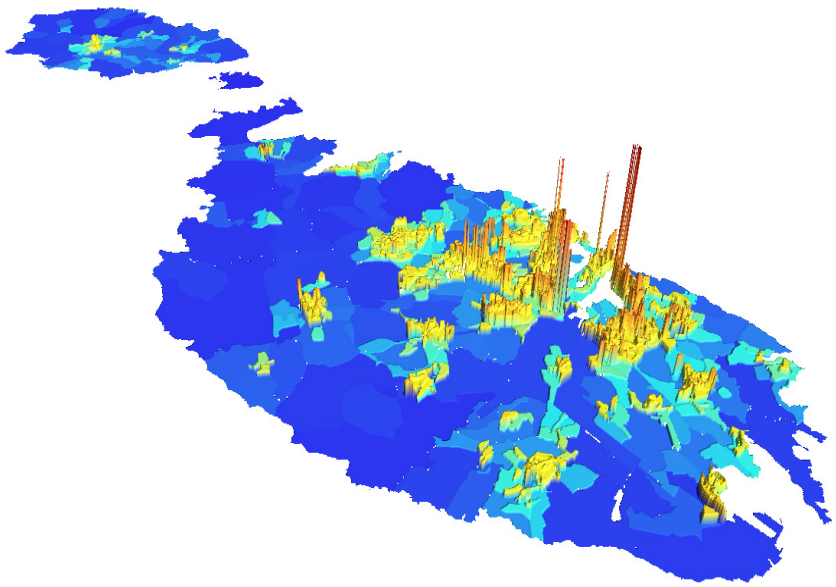


Source: Formosa (2007)

Map 8: 3D Map: Population Density

A map that extrapolates the polygon data of such maps as Map 7 into 3D format. One can immediately tell which areas have the highest population density. This type of map can be overlaid with another map such as that of poverty in order to analyse whether there is a correlation between population density and poverty.

Figure 7.18: 3D Map

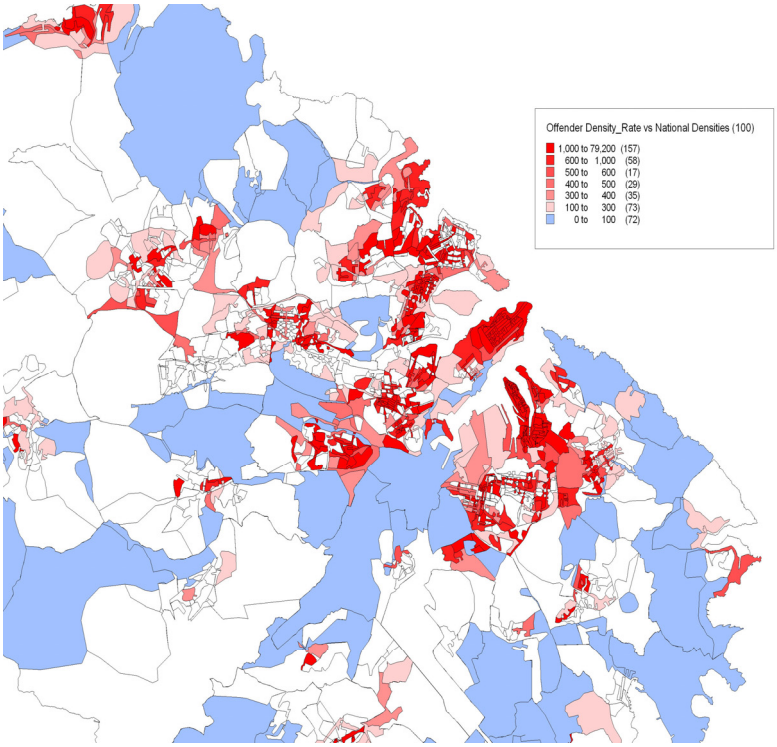


Source: Formosa (2007)

Map 9: Correlation Map: Small Area densities vs. National densities: EAs

A map that depicts correlations between two variables in polygon format.

Figure 7.19: Correlation Map

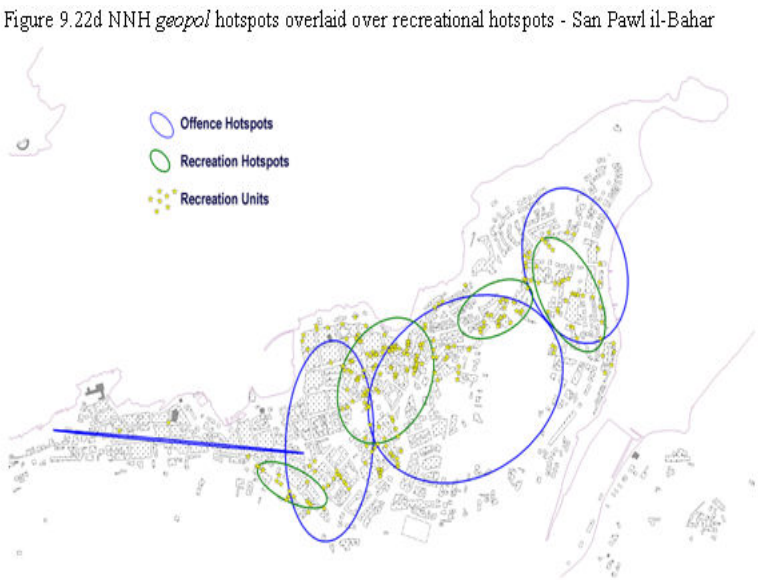


Source: Formosa (2007)

Map 10: Nearest Neighbour Hierarchical Analysis (NNA) Map: Offence Hotspots: Spatial – Retail Crime

A map that depicts data showing hotspots in the form of ellipsoids.

Figure 7.20: Nearest Neighbour Hierarchical Analysis Map

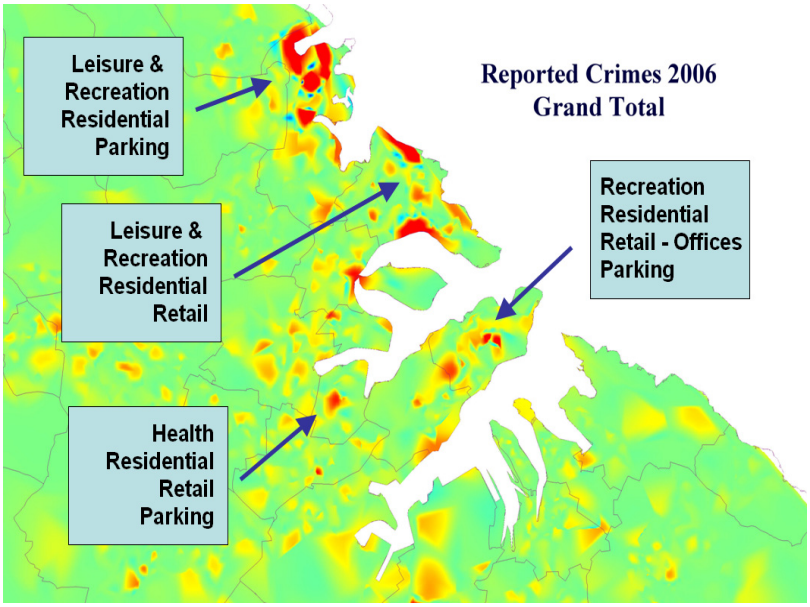


Source: Formosa (2007)

Map 11: Offence NNA: Spatial – Type by spread – Most effected

A map that depicts those areas having similar characteristics and indicating very high levels of activity.

Figure 7.21: Nearest Neighbour Hierarchical Analysis Spread Map

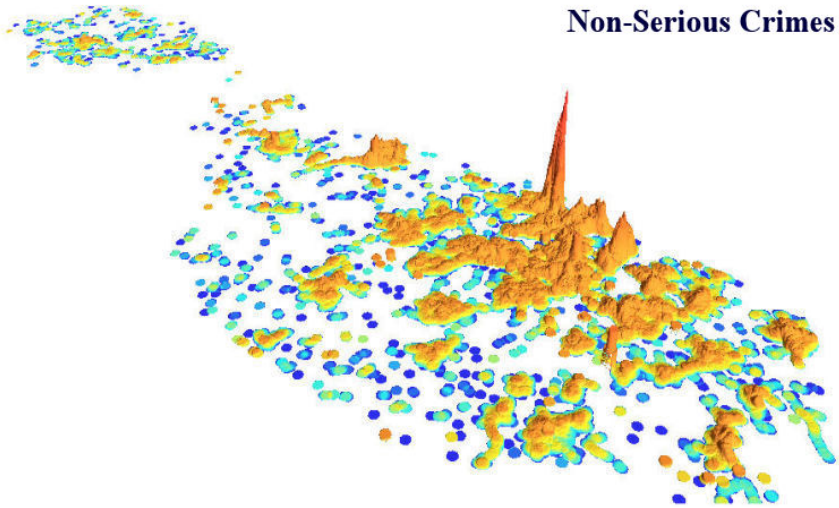


Source: Formosa (2007)

Map 12: 3D extrapolation of activity spread: NNA: Non-Serious Crimes

A map that depicts data developed through the process outlined in Map 11 in a 3D format for ease of visual reference to the hotspots.

Figure 7.22: Nearest Neighbour Hierarchical Analysis 3D Extrapolation Map



Source: Formosa (2007)

GIS Tools

As in the other instances of tool review, GIS has seen a veritable expansion in the availability of tools in both the commercial and opensource domains that target specifically the creation of spatial data and its outputs.

Commercial applications include MapInfo[20], IDRISI[21], ARCGIS[22], and ERDAS Imagine[23].

Opensource tools include GRASS GIS[24], SAGA GIS[25], Quantum GIS[26] and MapWindow GIS[27].

Other free software lists can be found at the GIS Development[28] site and the GeoCommunity site[29]. In addition, comprehensive list of tools and documentation can be found at the Directionsmag[30] site.

In summary, there are various tools that can be used to visualize a research output. To present results, researchers can choose between tabular, graphic or mapped format. There are no hard and fast rules on which one should be preferred, but it is best to refer to the specific discipline to relate to the relative visualization developments.

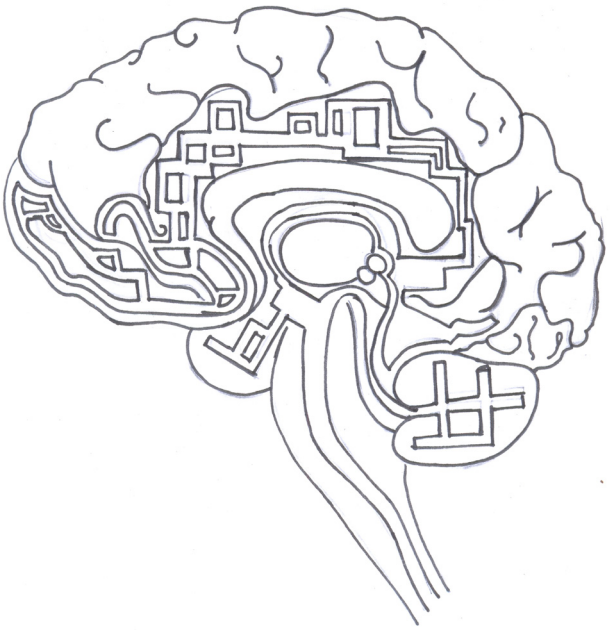
Questions (refer to Appendix for the answers)

1. List five simple types of research visual tools.
2. Graphing is a way to summarise data. List the graphing formats most commonly used.
3. Briefly define the following: (a) Bar Charts; (b) Pie Charts; (c) Histograms; (d) Line Charts; (e) Area Charts; (f) Composite Charts; (g) a Population Pyramid.
4. Why is mapping important?
5. Briefly explain what you understand by GIS.
6. What do the letters "S.W.O.T." stand for and why would one carry out a S.W.O.T analysis on GIS?
7. Briefly DESCRIBE what each of the following maps depicts: (a) Choropleth Map; (b) Graduated Map; (c) Dot Density Map; (d) Point Map: Actual Location of Offences Map; (e) K-Means Clustering Map; (f) Polygon-Based Cluster Analysis; (g) Small-Area Choropleth Map; (h) 3 D Map: Population Density; (i) Correlation Map: Small Area Densities Vs National Densities: EAs; (j) Nearest Neighbour Hierarchical Analysis Map; (k) Offence NNA: Spatial-Type by spread – Most effected; and (l) 3D Extrapolation of Activity Spread: NNA: Non-Serious Crime.

---

[20] http://www.pbinsight.com/welcome/mapinfo/
[21] http://www.clarklabs.org/
[22] http://www.esri.com/
[23] http://www.erdas.com/
[24] http://grass.itc.it/
[25] http://www.saga-gis.org/
[26] http://www.qgis.org/
[27] http://www.mapwindow.org/
[28] http://www.gisdevelopment.net/downloads/gis/
[29] http://software.geocomm.com/
[30] http://www.directionsmag.com/

Chapter 8
Mind Mapping



Always preoccupied with his profound researches, the great Newton showed in the ordinary-affairs of life an absence of mind which has become proverbial. It is related that one day, wishing to find the number of seconds necessary for the boiling of an egg, he perceived, after waiting a minute, that he held the egg in his hand, and had placed his seconds watch (an instrument of great value on account of its mathematical precision) to boil! This absence of mind reminds one of the mathematician Ampere, who one day, as he was going to his course of lectures, noticed a little pebble on the road; he picked it up, and examined with admiration the mottled veins. All at once the lecture which he ought to be attending to returned to his mind; he drew out his watch; perceiving that the hour approached, he hastily doubled his pace, carefully placed the pebble in his pocket, and threw his watch over the parapet of the Pont des Arts.

Camille Flammarion

Popular Astronomy: a General Description of the Heavens (1884), translated by J. Ellard Gore, (1907), 93.

Mind mapping? The word conjures a series of both remote exotic structures and unknown complexities. Imagine your mother screaming at you to get your room clean, your phone ringing, your friends sending you messages on the various social networks, music and TV competing for attention… while you are trying to study. Your mind starts to spin…you need to organize yourself and your thoughts before you start into something that is logical… that is what mind mapping is all about…Mind mapping is a tool to clarify one's mind and help visually draft the process from concept to tangible measuring as described in Chapter 5.

What is a model?

Prior to defining a model, it is best to introduce the issue that mind mapping requires a *sea change* (a veritable transformation) in how researchers visualize information. Thoughts can be listed in a column of data quite easily, however for mind mapping models, one requires an understanding of different dimensions and perspectives as the model allows for an analysis of a phenomenon which can take up both cross-sectional (through space) and longitudinal studies (over time). It allows for the transformation of data into such dimensional structures as 3D. It also allows for dynamics that cannot be replicated in other models since the cell-based variables can be moved within the virtual world and relationship can be better understood through visual effects.

By definition, a model is a representation of a structure which has within it the requirements which allow for the investigation of that same structure. By structure, one can include theories, hypotheses and data. Within a model one can include relationships and statistical elements. Models can also be used to predict future occurrences and also to investigate past events.

A model allows the researcher to study the real world through a series of observational activities. A design of the model is based on the information structured from the observed data. This triggers ideas for further research. The model itself allows for a feedback loop from the ideas to the observed data, since most ideas call for new data requirements and adjustments to the data gathering process.

The mind map exercise elicits the need for the clarification of specific questions that require investigation before one initiates the modelling process. These questions are reviewed in the following section.

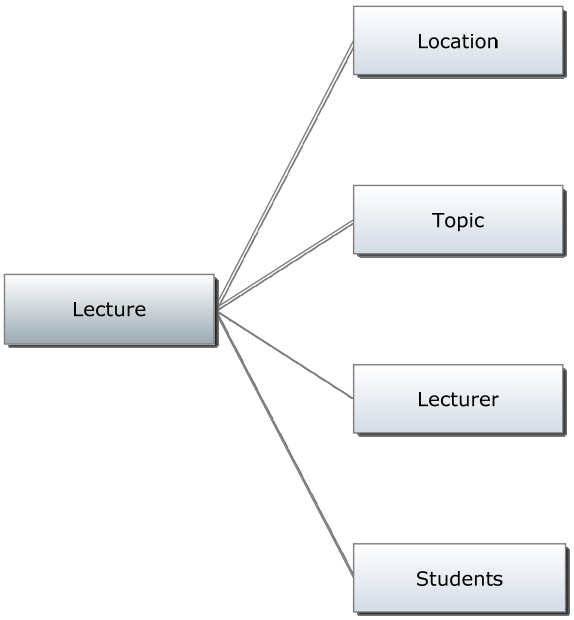
- **Can one translate human activities into a functional model?**

Whilst seemingly impossible to translate human activities onto a map due to the myriad interactions and the intrinsic phenomenology, one can initiate processes to try to understand such interactions or relationships through the component parts. Human activities can be broken down into smaller and smaller issues (that require investigation). Once an activity is disassembled into smaller relationships, these relationships can be identified and then the model can start taking shape.

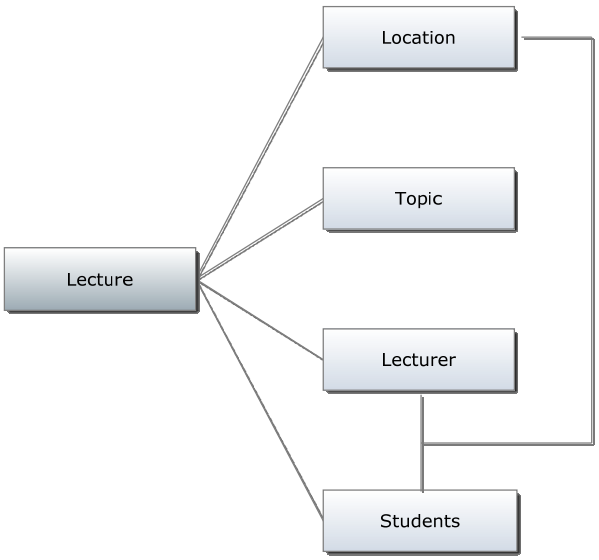
This process (described here) works in the opposite way to mind mapping since it de-structures an issue into the component parts (since the construct is already known as against a mind map where one has to start from scratch). However, this process enables the reader to understand how human activities can be translated into a model.

Let us take the example of a lecture held in a classroom:

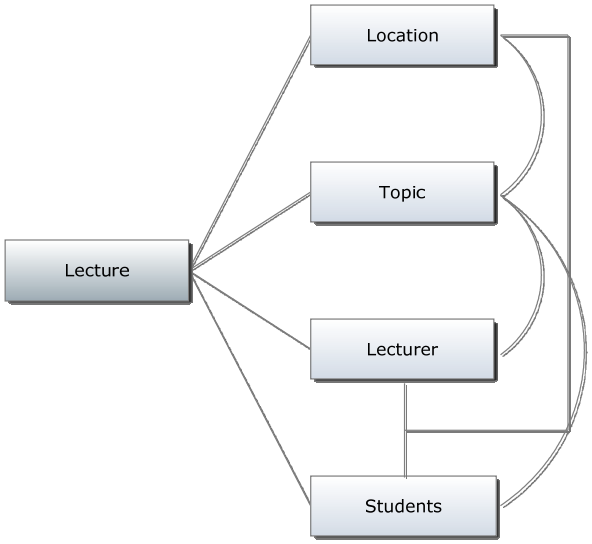
Disassemble the activity into its main components:



Identify the interactions occurring between the different components



There are evident interactions between the lecturer and the students and these occur in the location where the activity is held. Thus the particular interaction can only occur if both lecturer and students are located in the same space. Should the activity be held online, then location is superfluous to the exercise and can be eliminated from the study. In this case, the activity is not absolutely dependent on the topic under discussion, since the lecture is being held in a classroom. Should the topic have been one that necessitated a specific location (such as a dissecting lab), then the links between the entities would have been more diverse and would have included all four components.



Once all the possible relationships are mapped, then a model can be constructed. This model enables one to map human interactions. The move from concept to entitation to quantification and eventually to composition is possible, but only through a thorough understanding of the interactions at play.

- **Can one build a model / Does an application exist?**

One can build a model through the use of both analogue and digital technologies. Whilst the former is drafted on paper/cardboard, the digital version allows for full-range and very large model creation, one that can be given extra layers of data as are required in understanding the
relationships. These include: the theoretical approaches, the datasets being investigated, the sources of data, the actual variable mapping and the linkages between the variables. In addition, one must also allow space for the insertion of future linkages and codes within the model.

Applications that help such model building include Smart Draw,[1] Mind mapper,[2] and Visual Mind,[3] amongst others. The model is based on imagery which establishes links between the different types of information, which links allow for the inclusion of images, lines, arrows, words and bubbles.

The important aspect of such a tool resides in the ability to expand on one's ideas and branch out into the different sub-topics emanating from that topic. The idea of splitting a topic into sub-topics and ever more sub-topics fits our purpose perfectly as it allows problems to be split up into their smallest components.

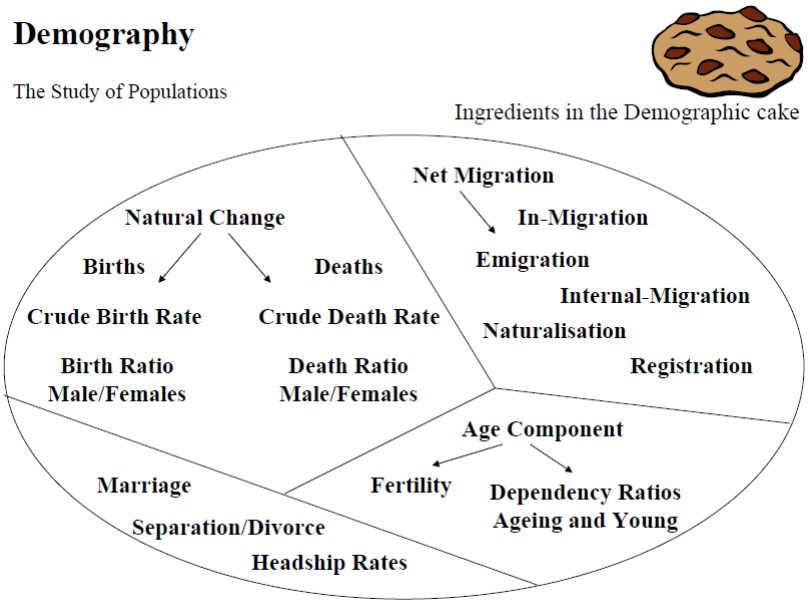
There is no need for a linear (straight-line, direct) link between the topic/subtopics. Links, can be expressed in a variety of ways across and within the levels. The case study example below gives an overview of how such a mind map is created.

Case Study: Creating a mind map for the Study of Population and its links to Housing

Step 1: Start with a rough drawing of what the elements of this mind map constitute (Figure 8.1).

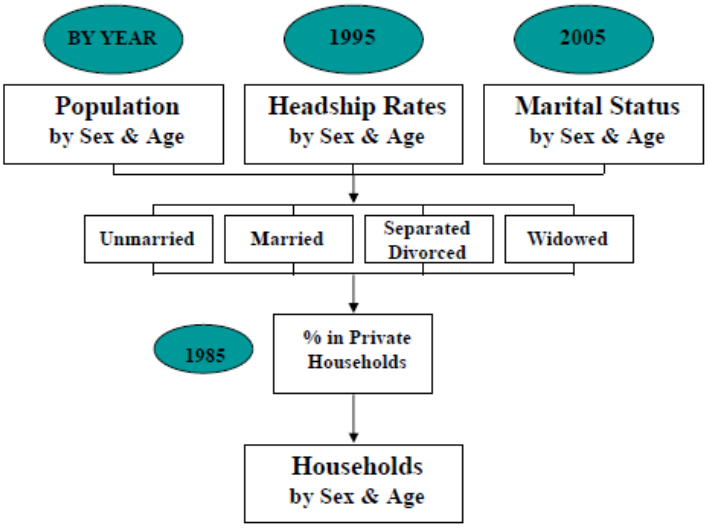
[1] http://www.smartdraw.com/
[2] http://www.mind mapper.com/
[3] http://www.visual-mind.com/

Figure 8.1: The Demographic Cake



There are four main elements that fall within two main themes: demographic structure and household structures (through headship rates). Whilst the image above defines the demographic structure, the figure below shows the steps taken to form an idea of the household structure. These eventually lead to an estimation of the need for housing based on the available infrastructure (Figure 8.2).

Figure 8.2: Households Structure



Step 2: Create the main theme (population)

Step 3: Identify the sub-topics (main tenets)

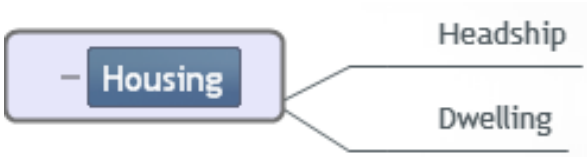
- Housing
- Demographic Structure

Step 4a: Identify the sub-topics for Demographic Structure

- Insert: Natural Growth and Net Migration

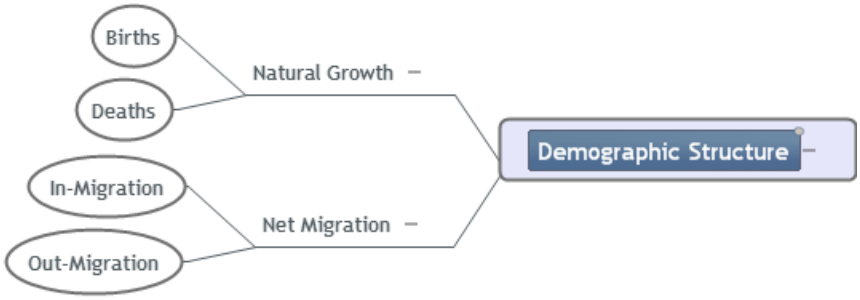
Step 4b: Identify the sub-topics for Housing

- Insert: Headship and Dwelling

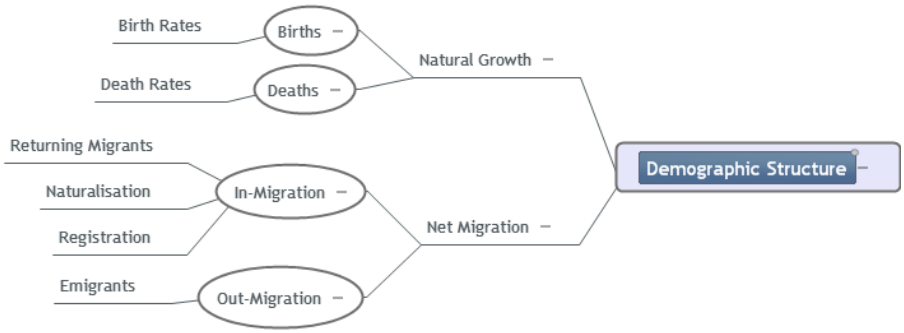


Step 5a: Identify the sub-sub-topics for each of the elements identified in Step 4a

- Insert: Natural Growth and Net Migration sub-elements

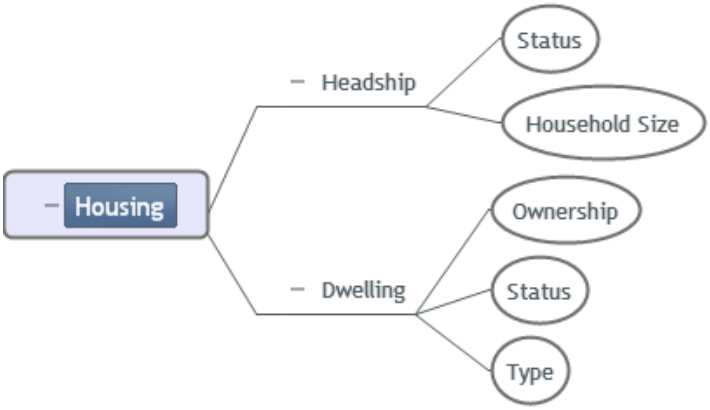


- Insert: the subsequent sub-elements

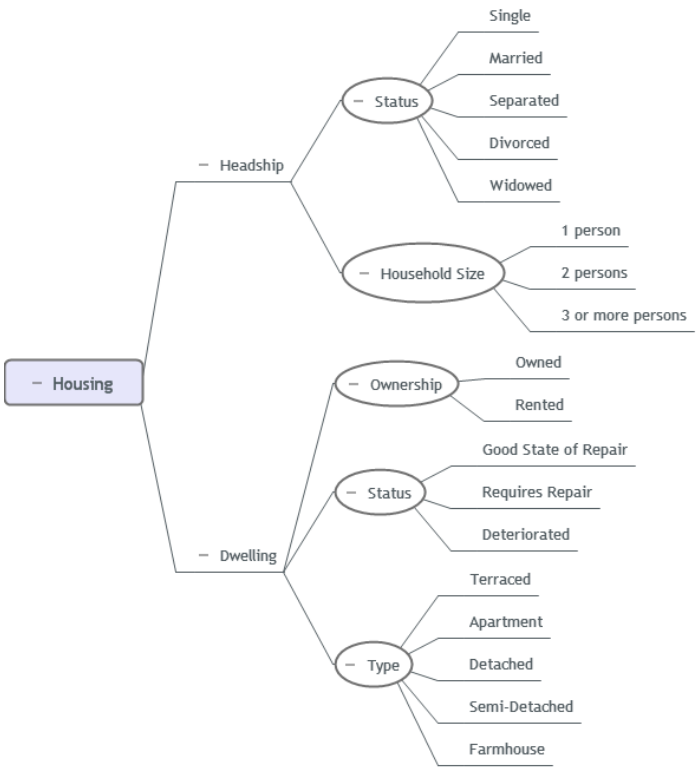


Step 5b: Identify the sub-sub-topics for each of the elements identified in Step 4b

- Insert: Headship and Dwelling Types sub-elements



- Insert: the subsequent sub-elements



Step 6: View the result in its entirety and start thinking about the links between the elements.

The final mind map that depicts all the elements is defined in Figure 8.3.

Step 7: Create the potential links between the different elements

The resultant basic mind map can allow one to acquire an idea of how the elements will interact. The result can then be taken to other levels through the inclusion of more informational issues that are described in the section below and in the CRISOLA example (Figure 8.4).

Figure 8.3: The Mind Map Elements

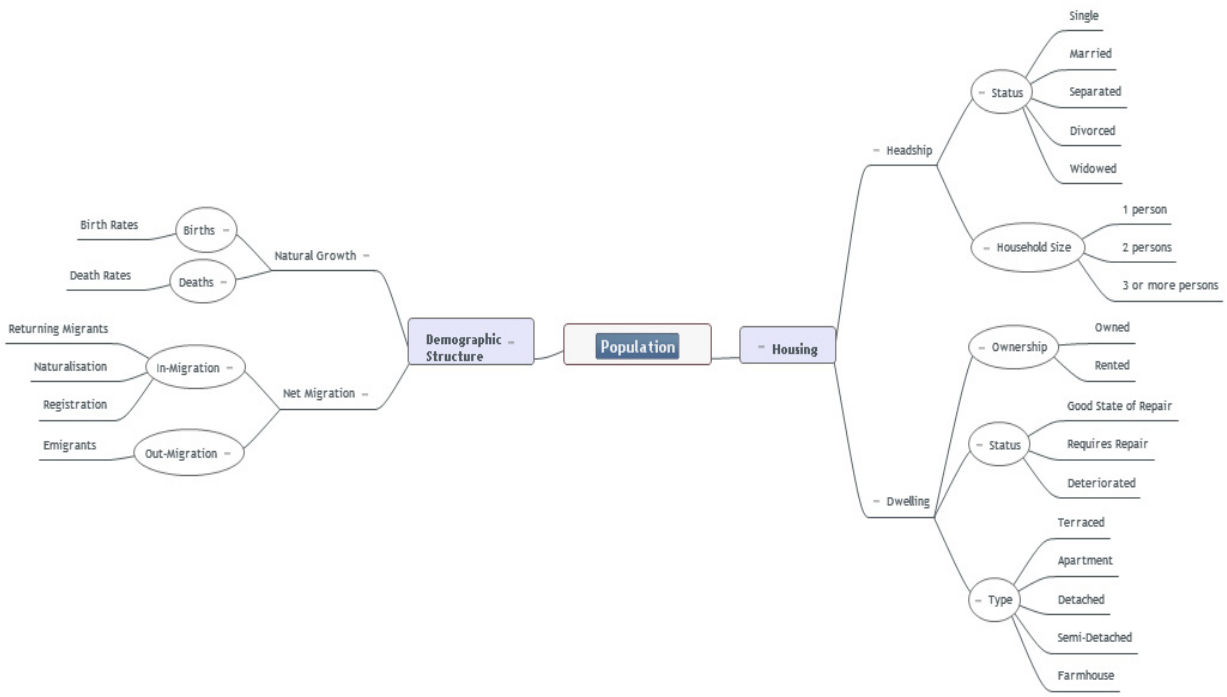
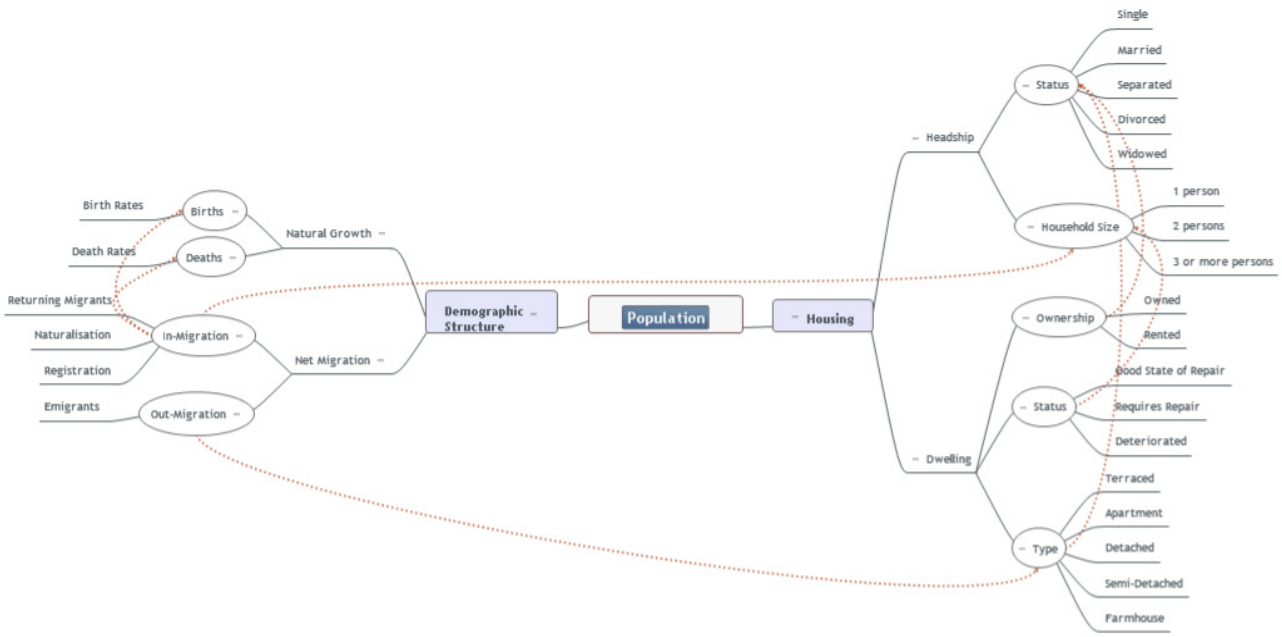


Figure 8.4: Linkages



Who are the players?

The players in a mind map are basically very few: they comprise the following:

a. the main topic
b. the sub-topics
c. the links between the topics
d. the dependencies (direction of dependency)
e. the data sets representing the topics
f. the data sources
g. the measurement scales

Each of these helps to enhance the model into an understandable structure. Although most mind maps only contain the first three items, the rest of the items are necessary to build a model.

- **What infrastructure is needed?**

Very little infrastructure is needed. A reliable computer and the relevant software should do. It is a good idea to prepare a parallel hardcopy version which would allow one to mess around on paper before drafting it digitally.

- **Can researchers afford to stay away from technology?**

A highly interesting question indeed! This book gives an idea of the efforts required in research generated in the social and natural domains. It emphasizes the use of technology, but one must not feel left out if the modus operandi lays preference to a manual approach. There are many ways to carry out research however, over the last few years, effort has centred on the use of technology as an aiding tool. One resultant problem is that faced by technophobic researchers, when they find themselves forced to use technology. However, another problem posed is that of a technology-side effort to always use technology, thereby imposing their wares into the social domains.

The socio-technic approach has proven successful since it has brought technology to the uninitiated – especially those in the social-domain. It has helped research to take off at phenomenal rates and has allowed for phenomenology to be analysed so intensely that we are currently investigating the links between the social sciences, the natural sciences and the physical sciences. Thus, today, it is possible to carry out research aimed at analysing the instance of an offence (criminological - social) occurring in a specific location (spatial - physical) that may have been triggered by a meteorological (atmospheric - natural) effect such as low atmospheric pressure combined with episodes of high pollen dispersion (biological - natural) and high noise levels (environmental - natural). Such links can only be tackled using high-end technology and one must be aware of the possibilities available for such research. This said, such a positive drive has to be counterbalanced by caution: technology must serve the user and not serve as a tool for abuse.

Conceptual Modeling

This above discussion has shown that there are different levels and activities required in creating a model. In effect, creating a model can only take shape through the understanding of the CONTEXT within which that activity occurs. There are various levels that can be described prior to creating a model moving from the outline identified in Chapter 5, from concept to tangibility.

The first step is thus to create a model based on the idea (concept), called a Conceptual Model. As shown in the Population and Housing example above, the different elements must have been sources from somewhere and must mean something to different persons and operations.

Why should one research population in relation to housing? Are researchers studying phenomena for phenomena's sake or are they linked with the operational sector. In order to start identifying the different levels, one must first understand the dynamics that such a model may fit in through a process spanning the visionary (idea) to the planning aspect to the operational and the eventual implementation.

Any model must fit the levels as based on the level of need. For example, a decision-maker does not need to know which house will be occupied by which family, but would require knowledge of how many dwellings are required for the next 20 years to ensure that the supply is there.

The following levels of need are required prior to understanding the type of model to be drafted. One can create the most detailed model and then switch off those levels not deemed necessary for the levels above that.

- Global vision perspective - Visionary:

The philosopher king sits here. The person with a vision is the person who will identify the need for some kind of change. One can identify that there is a need for housing but does not need to know what types as long as the units are available. Neither does a visionary need to identify what the population structure will look like but only the information that it will grow.

This refers to the highest level knowledge of the W6H but at a very abstract level. There IS a requirement to know where one is coming from and where one is going but not necessarily knowing how to implement such. The links between population and housing can be established at this level, but it is up to others to establish the links.

- Strategic Planner

This level requires knowledge of what will happen in the mid to long-term, often basing one's studies on a 10 year (for local or regional levels) and 20 year for strategic levels. The planner needs to be able to overview planning, taking in the ideas generated at the global perspective and moving them towards a more realistic approach. The planner must have knowledge of the on-the-ground levels and the higher level visionary levels.

At this level, one can include the policy maker and the decision maker who have to draft the actual policies and legislate on them. They would need to know the links between the demographic structures and the housing elements in order to effect their proposals for change. Thus, at this level, the model should also show the links between the different elements, though not necessarily those links at the most detailed levels.

- Operational Designer

This level instigates the need for a model that shows how one will implement the requirements of the strategic level. This level would need information on how to link the demographic projections based on different scenarios with the need for specific housing types.

An example would be to link a scenario that predicts an increase of the elderly population and their requirements for specialised housing such as community-based services, the need for smaller housing types, the areas where such dwellings can be situated and the marketing actions to convince elderly persons to move into smaller residences.

- Administrator

This level structure plans on how to operate the required changes on the ground. The model at this level should allow for detailed information on what is required to actuate such a change, in terms of: design, materials and administrative procedures. This is the nitty-gritty level and concerns who does what and when – within the constraints established by the levels above.

- Tactical Planner

At the end of the line one can find the tactical planner – that person who will be working in the field and who requires the deepest level of information to ensure that the job gets done. This level works on the specifics on what needs to be carried out and how best to ensure that it occurs.

The Three Dimensions ensconcing a conceptual model

At any level of research, a conceptual model has to keep in mind three important dimensions within which that model operates.
(Source: www.els.salford.ac.uk).

Models look at…

- the **spatial** dimension: where something is located
    - any policy or decision or research has to occur somewhere, whether it is in real or virtual space, whether in terrestrial, bathymetric or extra-terrestrial terms, whether in a flat digital domain or an integrative virtual digital domain
    - there is always a location to a model

- the **thematic** dimension: the characteristics of that something
    - the character of either the location or the object occupying that location
    - the topic, policy or strategy under study
    - the theme has to be defined by its component categories

- the **temporal** dimension: when something occurs
    - the comparison of data over time
    - why something occurs when it does
    - the span of the research
    - not restricted to the present but also the past and the future. The future can only be visualized at all the levels through a knowledge of the past with the present used as a starting ground for the study of the interim period between the past and the present and the present and the future
    - there is always a time factor, whether when the study occurs or when the activity under study occurs or will occur

Moving towards implementation of the Model

In order to move from a conceptual model to a working model, one must again move towards an unravelling of the complexities built into the reality being studied. Thus, the conceptual model's reincarnations are tools that will be used in a computer to replicate the real world process one is trying to model.

This process is based on the steps identified by Peuquet (1990) in his description of **Levels of Abstraction.** Peuquet's work on GIS can be translated to a mind map which drafts the process from reality to abstract user-oriented information structure to concrete machine-oriented storage structure of the computer. From the latter one can then run the necessary queries and implement to outputs from the model.

The modified Peuquet structure below takes the Peuquet model and transposes it for any data type.

There are 3 stages that need to be taken into account:

- Stage 1: identify those entities one is interested in and decide how to represent them;
- Stage 2: choose a data model that computers are able to display, analyse and store your entity representation;
- Stage 3: draft a "nuts and bolts" stage where one instructs the computer how to recreate the entities identified earlier.

Content Analysis

How does one use a mind map in non-quantifiable studies? Can a mind map be created for information extracted from textual material? Content analysis comes to the rescue.

Content analysis usually refers to the analysis of a written material. This type of analysis is used in political speeches where certain words are analysed to try to envisage the meaning of the writer. When

studying texts these must be studied as products of the society that produced them. For example if we are studying a Second World War speech by Winston Churchill we have to keep in mind the political situation of the time. If we look at his speech "We shall fight …" we see that he uses these words seven times in rapid succession. Content analysis is about the intended content (i.e. the meaning which the author intends to portray) and the received content (the meaning constructed by its audience). It is the difference between the two that makes up the crux of content analysis.

Let us go back to Churchill's speech…:

> We shall go on to the end, **we shall fight** in France,
>
> **we shall fight** on the seas and oceans,
>
> **we shall fight** with growing confidence and growing strength in the air, we shall defend our Island, whatever the cost may be,
>
> **we shall fight** on the beaches**,**
>
> **we shall fight** on the landing grounds,
>
> **we shall fight** in the fields and in the streets,
>
> **we shall fight** in the hills;
>
> **we shall never surrender**, and even if, which I do not for a moment believe, this Island or a large part of it were subjugated and starving, then our Empire beyond the seas, armed and guarded by the British Fleet, would carry on the struggle, until, in God's good time, the New World, with all its power and might, steps forth to the rescue and the liberation of the old. (Churchill, 4th June 1940)

This speech uses repetition to encourage and fire-up the soldiers. If we look at the intended content we will refer to the word "fight" as the most important word. It is obvious that Churchill wants to encourage his soldiers. However there is also the received content from the soldiers. Hearing this speech, soldiers will not only feel encouraged to fight, but there is the hidden message of the greatness of the Empire. The message is not only about the courage to fight, but on the fact that the British will win because of the greatness of the Empire and the British fleet, however, if by some strange happenings they will lose, they will be saved by the New World because God is on their side.

Before you start to conduct content analysis you need to decide on the following:

1. What level of analysis are you going to take i.e. which words/set of words would determine a concept e.g. should we take "we shall fight" or "fight"

2. What is the number of concepts you will code?

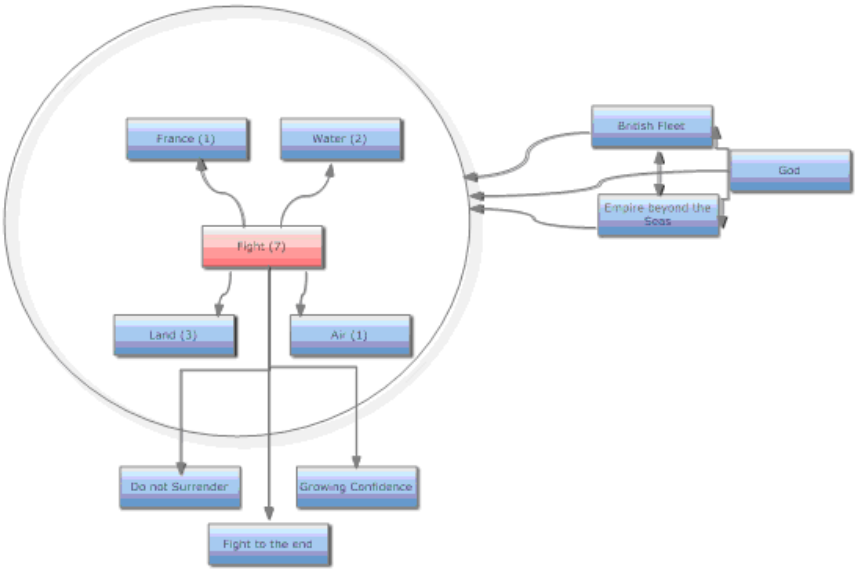
3. Are you going to code for existence or frequency?

4. How are you going to distinguish between concepts? – e.g. fight and defence.

5. You also need to develop a set of rules

6. What are you going to do with the irrelevant information?

At this point you need to start coding the text to construct a mind map of the speech. If we take the above speech a mind map of the speech would look something like this:



The final step would be the analysis of the results. The mind map above shows that fighting is central to the speech. However there is also a mention of God's help, the greatness of the empire and the British fleet. Analysing a number of speeches one could come out with a set of variables that is common in all the war speeches.

Mind maps are also good when you are starting to think about your dissertations/project proposal. It is a way of helping you analyse your thoughts. By drawing out the connections throughout your variables, you will realise where the flaws in logic are as well as which area you will be studying. Constructing a mind map would help you delineate the connections of your study. It would also help you develop your idea. Therefore by using mind maps, you will be able first to expand your topic and later to narrow it down so that it would be manageable.

CRISOLA Model

The conceptual model thus lays the groundwork for a completed model that can allow for predicting what could happen over time and space. The next sections look at this process from the point of view of a real model, created in a study on crime (Formosa, 2007). This model is termed **CRISOLA**. It attempts to create a predictive structure through the analysis of spatio-temporal elements.

It takes the reality of the PRESENT but looks into the FUTURE. This requires not a mere crystal ball viewing but extensive knowledge of the PAST

The main area of study is the interaction between:

- the **crime characteristics**
- the **social characteristics**
- the **physical characteristics**

In a study (Formosa, 2007) carried out over the period of 1997-2007, the topic of environmental criminology elicit some interesting research pathways that emanated from both the issue on data availability and access as well as the methodology used to understand the linkages between the different entities. A model was created to ensure that an understanding of all the parameters was established. This was not an easy enterprise as most datasets are either non-existing or non-available.

Why create a conceptual model?

Such a question lingers through the process of any literature review and in this study there was an amalgamation of the environmental criminology literature, the GIS literature and the Maltese scenario readings. The reviews, together with an understanding of the complex Maltese data availability situation, highlighted the need to bring together each aspect. It also helped to build a mind map that helped set out a process to depict a basic and generic model on how crime, social and landuse issues interact together. The review process also identified techniques and datasets that can be used in the identification and understanding of crime. The use of these datasets is best explained through a conceptual model that is relevant to **CRI**me and to the **SO**cial and **LA**nduse aspects, embedded as the acronym **CRISOLA**.

The model took shape through a tiered 3-phase process, with each iterative phase building up from an abstract level (Phase 1) through the identification of the main datasets (Phase 2) to a final individual attribute listing (Phase 3). The model is not exhaustive as it covers potential datasets that yet need to be created/surveyed, statistical measures identified as well as the inclusion of other crime-relevant theories. The model can be evolved in future studies as it attempts to highlight areas of study that will not be tackled in this research and which may/may not be found to be significant, entailing further change.

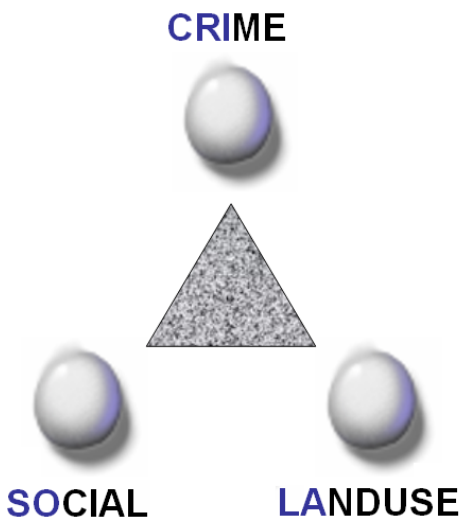
The three CRISOLA radials: Crime, Social and Landuse

Initially the conceptual model catered for the crime aspect in isolation, but crime does not stand alone: it interacts within a wider and more complex environment. The mind map exercise soon sought the inclusion of social and landuse parameters within the model aimed at streamlining the process to facilitate the analysis. The result brings together the three CRISOLA disciplines and attempts to identify theoretical links between the different datasets.

The decision to model crime together with the sociological and landuse disciplines is based on an understanding of the interactivity between the three as identified in the literature. The model attempts to understand criminal activity within the social and physical structures it operates in.

The main area of study is the interaction between:

i) the **crime characteristics** through an analysis of offender and offence composition and the interactivity between them;
ii) the **social characteristics** of an area through an analysis of its poverty/deprivation;
iii) the **physical characteristics** of an area, particularly its landuse, structural and zoning parameters.



The social characteristics of a human society are linked to the physical surroundings it operates in, which two characteristics are directly caused by or affect crime. Offender analysis requires an understanding of the social construct that the offender operates in, such as affluence and poverty. Offence analysis requires an understanding of the landuse structure crime occurs in; the opportunities offered, the mode of travel, and the activities that may lead to the occurrence of crime, amongst others. Theories covered in the study were inclusive of Environmental Criminology (or Urban ecology).

Phase 1 – The Abstract Level

Table 8.1 outlines the Phase 1 thought-process needed to reach an initial structure within which to analyse any relationships between the three disciplines. It is a high-level abstract model that attempts to look at parallel processes between the three disciplines and how an understanding of the processes can be achieved. It develops the concept through a series of five linear steps that can be tackled in order to facilitate later cross-thematic crime studies. It is aimed at an analysis of the thematic structure, focusing on the main parameter in the themes that affect change, identifying the spatial construct within the theme, highlighting the impact on capital and cohesion and finally leading to a change phase.

The latter phase can only be tackled through longitudinal studies that would draw a better long-term picture of what constitutes change. Although the current study looks at crime over a period of time, this model needs to be revisited with longer-term data if one needs to analyse sturdier change processes. This is needed particularly in the final phase that covers change for each of the CRISOLA themes.

Table 8.1: Phase 1 - Conceptual Model Logical Matrix 1

Social		Crime		Urban
Analysis of the Social structure of the area under study		Analysis of crime in the area under study through offences and the behaviour of offenders		Analysis of spatial constructs through a study of landuse zoning, spatial aggregates and physical structures
↓		↓		↓
Focuses on socio-economic and socio-cultural parameters towards an understanding of poverty and deprivation as a surrogate for social and community health		Focuses on offences as a measure of attractiveness of an area and focuses on offender data as a measure of social disorganization		Focuses on landuse zoning as a measure of affluence, leading to an understanding of opportunity structures
↓		↓		↓
Identifies the social-spatial constitution of the areas, leading to a social-zoning structure		Identifies the criminal-spatial constitution of the areas leading to a crime-zoning structure		Identifies the physical constitution of the areas leading to a landuse-zoning structure
↓		↓		↓
Impact on social capital – social cohesion		Impact on security and safety		Impact on spatial capital
↓		↓		↓
Social change		Crime change		Landuse change

Phase 2 – Identifying the linkages

Whilst, the high-level Phase 1 Model enables a generic focus on the study in question, a more detailed second level model was required which helped point at and identify the interactivity between the three parameters. This is accomplished preferably through the identification of datasets that may be used for analysis. Being a mind map model, Phase 2 (Figure 8.5) sought to identify those literature-related issues and integrate them within the model. It also sought to bring together the different: theories, datasets, spatio-temporal aspects, predictors and the main tenets that can be used in such a study on crime. These include such parameters as are age and density.

The deeper one moves into the model (towards the bottom part of each section and where the predictors are highlighted) the more research is needed to identify the real relationships and how each parameter can be predicted. The model does not attempt to solve these issues in this study but depicts the potential future studies that can be attempted.

The following walkthrough of the model in Figure 3.1 shows the three distinct social, crime and landuse sections. Each section has a series of data-boxes each depicting a specific theme, index or concept. The following section describes one such data-box.

A Social section walkthrough: Taking the proximity data-box as an example

Refer to the Phase 2 data model and identify the proximity index data-box within the Social section.

The proximity index attempts to elicit an understanding of each area in Malta through its location in relation to proximity to a number of factors. These are split in two:

i) the proximity to the community centre (identified by the number 3, which number also refers to the relative Phase 3 data-box) and
ii) structures identifiers split into four themes,
   a. two related to building state such as vacancy (4) and dilapidation (5) (indicates broken windows-tipping) and
   b. the other two related to densities – population (6) and dwelling (7).

The latter four would together be developed into a structural poverty index (8) that would be integrated with the proximity to the community centre theme. These two constructs would enable the creation of a spatial poverty index (9) that introduces a concept which identifies that poverty is not essentially an economic construct but is also related to access to the community construct. Taking the model further, integrating the socio-economic poverty index (10) created through a separate integration process, with the spatial poverty index (9) would result in a deprivation index (11). This process is followed by a statistical measure that would eventually result in the identification of a categorisation of different social zones (12).

It is at this stage that the first cross discipline links are highlighted: those of the identification of a possible link between social zones (12) as identified through the process described above and the potential relationship (brown link) to the offender location (37) that looks at the social zoning pertaining to convicted offenders. This link can be further analysed through statistical measures. Other potential cross-discipline relations are identified through the link between the social (poverty) zones (12) and the landuse social and community-related zones (15). This link could better describe the relationship between the 'poor' areas and their location in the landuse designated for social use as against industrial and recreational use. It may identify 'poor' areas that are situated outside of the social zones as well as concentrations within specific areas of the social zones. Other lower-level links between the different themes would relate to the linkages between the final level of each theme and the potential impact on each resulting in a change in the other. The social zoning (12) to landuse (27) link is such a potential link (red line) where one could predict changes in deprivation through changes in the landuse construct and vice versa.

The other sections follow the same logical process and each successive branch highlights its particular theme, theory base and dataset pertaining to it. The best way to follow this is within the model is to once again look at the proximity index example in Figure 8.5. The level 2 model in Figure 8.5 is accompanied by a description and spatial levels key (Figure 8.6). The key describes the different spatial data

aggregates available from national to regional to enumeration areas, which data layers can be employed for most datasets listed. The description section, however lists the different datasets available (D), the theories (T), the main data tenets (M) as well as other relevant information.

Once again, taking the proximity index as an example, the proximity-to-centre data-box (3) is tagged with 3 codes, amongst them D2A. The D2A refers to the key: Data (D) is available at (2A) Address-point spatial detail. Similarly the vacancy (4) data-box is tagged with T3A and D2I, where as an example T3A refers to social disorganization theory and potential to analyse the data based on concentric rings and broken windows concepts.

Other model issues include the identification of a potential to integrate a dark figure of crime, once this is carried out. To date, this has not been covered in Maltese crime studies, except for a crime victimization study in the 1990s, which was never published and another study carried out by Formosa in 2007 where the sample return was too small to prove reliable.

The coloured data-boxes indicate some kind of major studies that were not found in the literature review but are deemed essential to understanding crime, such as the analysis of spatial-temporal-prediction-fragmentation (31) which attempts to understand the spatial aggregate (ex: council, enumeration area, street) at which predictability starts to deteriorate over time and which would allow researchers to know how far to predict at each level in order to remain statistically significant. Such a model would help crime understanding for operational and tactical levels.

Phase 3 – Identifying the datasets and attributes

Taking the model one step further to Level 3 (Figure 8.7), a series of statistical measures are listed for the variables within each dataset identified for model integration. This level is theoretical as each link needs to have a theoretical construct attached to it with the relevant research studies carried out which would validate that such a model can work.

The Phase 3 is highly detailed where it looks at each data-box, identifies the relative dataset as indicated in Phase 2, lists the attributes within that dataset and then attempts to identify statistical measures for each level within the process. In most cases, the statistical measures call for further research into the potential measures to be employed. Also, at this stage new indexes were inputted such as insurance, sentencing practice and recidivism, each of which was identified as vital to a particular complex index.

As in the Phase 2 case, the best way to understand Phase 3 would be through an example, that pertaining to the proximity-to-community-centre data-box (3). In Phase 3, a statistical measure is listed as distance-to-centre which is further explained through the use of a distance ranking index based on GI buffering techniques employing 100m intervals.

New indexes are also identified in Phase 3, which indexes help to clarify how a more complex index is created. The following example is based on the welfare index (2) that is split into two component indexes (persons-at-risk and structural-dependency). Each of these is composed of three data complexes (ex: pensions, social assistance, widows survivors), where each complex is composed of the sum ($\sum$) of a number of welfare benefits pertaining to that category (attributes within the welfare index dataset). For example, Widows survivors is composed of Widows pensions (NM and NMWP), Survivors pension (SRP and ESRP). The results are then integrated with other categories as in the Phase 2 process described earlier.

Conceptual Model Summary

In summary, the main aim of producing these three Phases was primarily targeted at understanding the potential relationships between the CRISOLA constructs. These relationships operate within a human environment that is intrinsically dynamic, where any change in one sector would affect the other two, positively or negatively. The model will be used post-research to further refine the theories and carry out in-depth studies in each of the sectors and linkages.

The conceptual model was drafted to enable the author to focus the direction this study would take though the identification of some of these areas that can be analysed, whether data exist to support such studies and also to identify further areas of research. It also helped to list the relevant theories, the data availability, the spatial and temporal aspects and the potential relations between the different CRISOLA constructs.

Once the model has been drafted, the next steps of the research development process would entail the running of the relevant queries, studies and analysis as described in the earlier chapters. The next chapters will entail a study of the tools available for such study purposes.

Questions (refer to Appendix for the answers)

1. Briefly explain what mind mapping is.
2. What is a “model” (with reference to research and mind mapping)?
3. List the six main steps when it comes to creating a mind map.
4. List the main players in a mind map.
5. Different research stakeholders have different level of needs. Mind maps are designed, keeping in mind the requirements (levels of need) of people in different roles with their different perspectives. List these roles/perspectives.
6. A conceptual model has to keep in mind three important dimensions within which that model operates. List these dimensions.
7. Building a model – moving from a conceptual model to a working model – requires a process based on Peuquet’s (1990) three stages. List these three stages.
8. Briefly explain what you understand by “content analysis”.
9. What does CRISOLA stands for? What is CRISOLA’s main area of study?

Figure 8.5: Conceptual Model Phase 2 – Linkages – Themes - Key 1

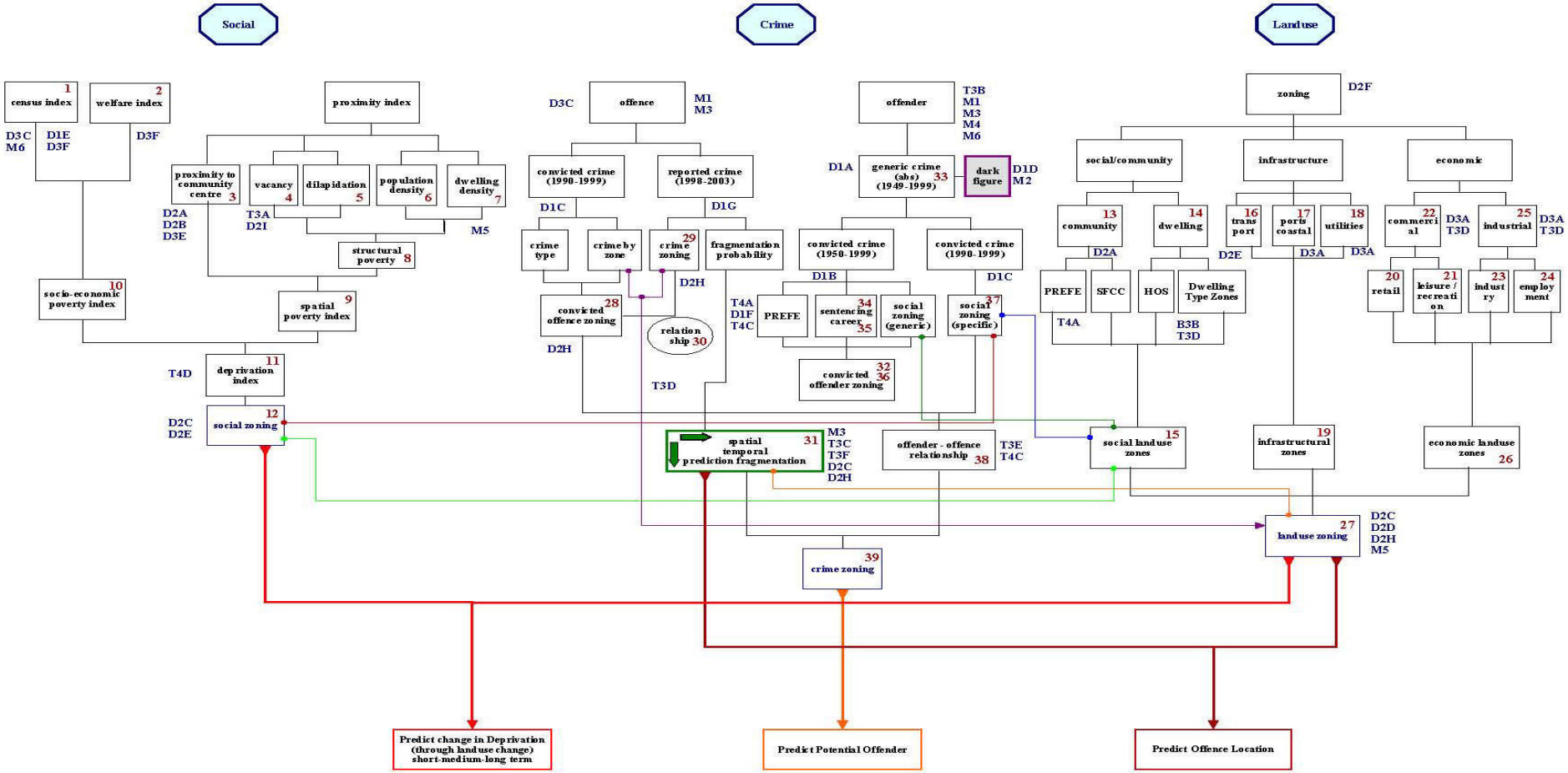


Figure 8.6: Conceptual Model Phase 2 – Linkages – Themes - Key 2

Data (D)

1) Crime Data

a) 1949 - 1999 (Abstract of Statistics - Historic)
b) 1950 - 1999 (CCF Offender)
c) 1990 - 1999 (CCF Offence)
d) 2000 - 2005 (dark figure
e) 2001 (education absences police)
f) 1993 - 2004 (Enforcement development cases)
g) 1998 - 2003 (police reported crime)

2) Landuse Data

a) Address point
b) Basemap extractions
c) Boundaries
- streets
- enumeration areas (EAs)
- councils
- regional
  - local plan areas (LPAs)
  - police
  - health
  - census
  - NUTS (5-1)
  - conurbation

d) Planning applications
e) Transport
- roads
- streets - Link Tables
- bus routes

f) Urban - rural
- development zones
- zoning

g) Urban Conservation Areas (UCAs) / village cores
h) Development PLAN (base map extraction)
i) Maltamal (utilities - vacancies - migration)

3) Social Data

a) Economic
- town centers
- recreation
- retail
- industry
- employment

b) Housing zoning - HOS
c) NSO - socioeconomic Census 1995
d) Welfare - 1995 - 2003
e) Education
f) Health

Theories (T)

1) Rule relativist - social construct
- space
- time

2) Radical conflict - human rights
- law
- (power)

3) Environmental criminology

a) Social disorganisation
- concentric rings
- broken windows

b) Offender
- individual
- family
- community
- country
  - careers
  - networking

c) Offence
- zoning
  - urban / rural
  - housing market
  - commercial
- victimisation
  - networking
- defensible space

d) Opportunity
e) Routine activity
f) Predication fragmentation probability
- space
- time
- thematic

4) Sociological

a) PREFE
- Politics
- Religion
- Education
- Family
- Economy

b) Networking
c) Structuration (3b-3c)
d) Deprivation
- Poverty - Absolute / Relative

Main Tenets (M)

1) Age

2) Dark figure

3) Seriousness

4) Recidivism

5) Density

6) Unemployment

Spatio-Temporal

1) Thematic (H)

2) Spatial (S)

3) Temporal (P)

All datasets in model are:

1) Spatial
2) Thematic across
- social
- landuse
- crime zones

3) Temporal

a) 1950 - 1999 (base data - offender / crim)
b) 1990 - 1999 (specific period - offence)
c) 1995 - 2003 (reported crime - geopol)
- (census)
- (social data)

d) Landuse data
- current
- 1990 - 2004

Predictors

1) Potential offender - TURF

2) Potential Offence - SHIFT

3) Potential Deprivation - SHIFT

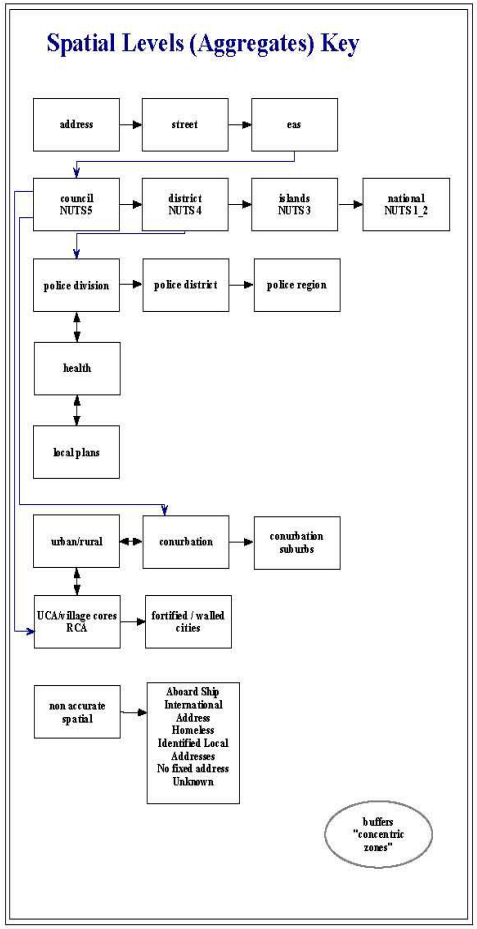
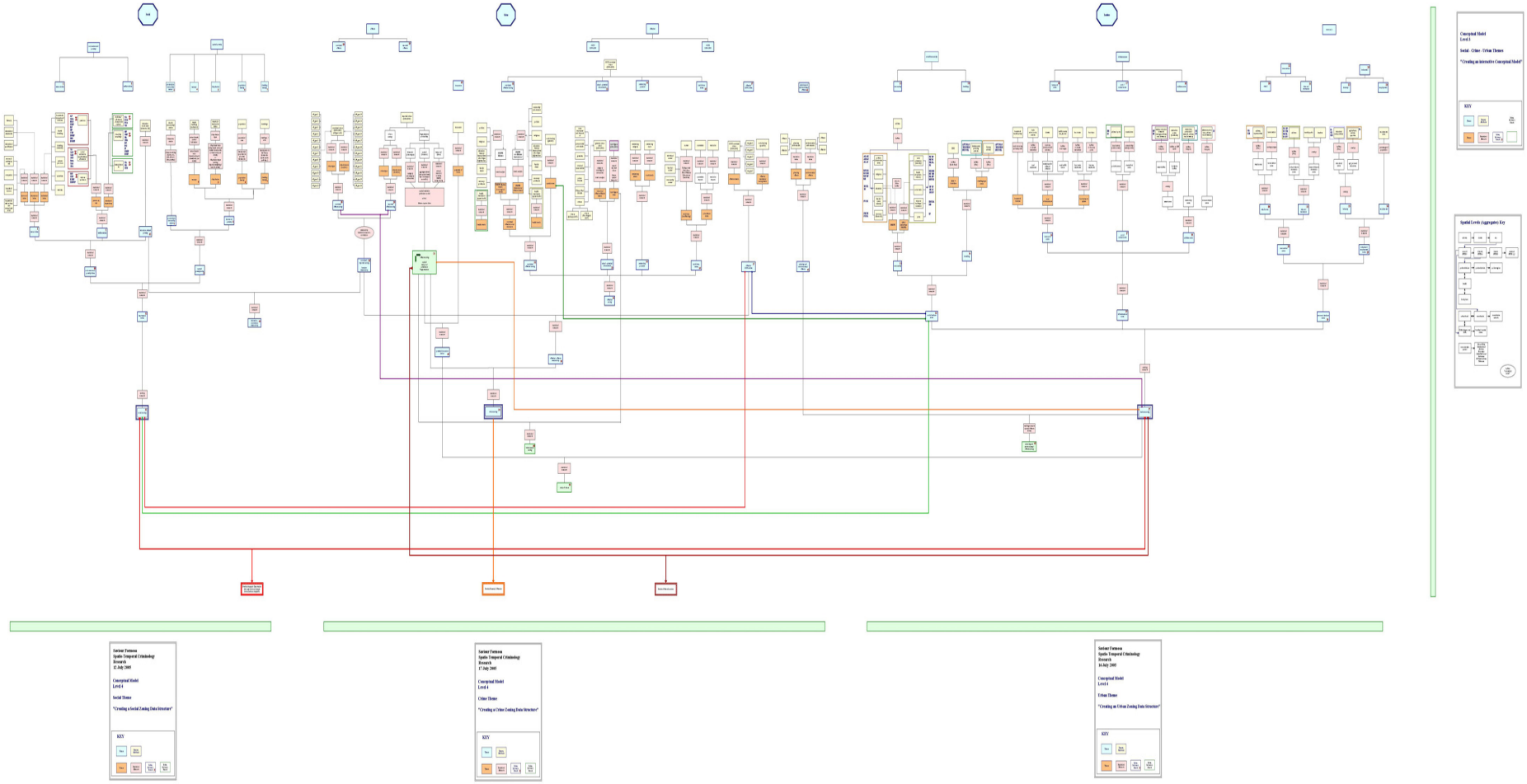
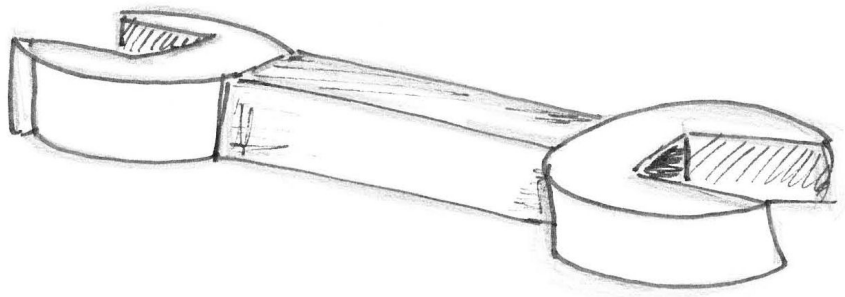


Figure 8.7: Conceptual Model Phase 3 – Datasets, variables and Statistical measures model



Chapter 9
Tools



Most of us who become experimental physicists do so for two reasons; we love the tools of physics because to us they have intrinsic beauty, and we dream of finding new secrets of nature as important and as exciting as those uncovered by our scientific heroes. But we walk a narrow path with pitfalls on either side. If we spend all our time developing equipment, we risk the appellation of 'plumber', and if we merely use the tools developed by others, we risk the censure of our peers for being parasitic.

Luis W. Alvarez

'Recent Developments in Particle Physics', Nobel Lecture, December 1th, 1968. In *Nobel Lectures: Physics 1963-1970* (1972), 24.

There are various tools which one can employ to help in statistical research. These range from simple tools that produce basic calculations (and are readily available in proprietary office packages) to specialized statistics tools that are developed solely for the purpose of analysing data. Another more interesting group of tools is available freeware and has the same functionality as the commercial packages. This Chapter will give an overview of the tools available and will give examples from some of the tools for ease of reference.

Prior to venturing out and buying or downloading/using online software one has to decide on what **Depth of Use** one is going to employ the tools.

A. Does one need basic tools that identify keywords and help to find links between the different inputted data? This is particular to qualitative methodology.
B. Does one need simple tools to aid in quantitative work such as summing, adding, subtracting, sorting by minimum or maximum and averaging as well as exporting in the forms of tables and graphs?
C. Does one need tools that employ the above but also allow for basic statistics, such a measures of central tendency, etc. (refer to Chapter 11)?
D. Does one need a full statistical package that employs the above as well as a whole range of statistical measures (refer to Chapter 11)?
E. Does one need to employ spatial statistical tools that deliver maps as outputs and use specialised statistical measures, inclusive of Moran's I, Standard Deviational Ellipse, Nearest Neighbour Hierarchical Analysis, etc. (refer to Chapter 11)?

An early decision will ensure that one is not sidetracked into wading through forest-equivalent paper trails or wrestling with a software tool that is as complex to manoeuvre through as it is to pilot the space shuttle. Choose wisely and the earlier the decision is taken the quicker the choice is made and the less tears to be shed!

Let us take 4 examples. Read them and try to establish which of the above would fit each example:

i) Analysing correlation of offender residential areas by poverty levels;
ii) Identifying the most common links highlighted during interviews with the elderly in a retirement age study;
iii) Attempting to establish if there is a relationship between youth leisure activities and drug use;
iv) A demographic project that aims to establish the central tendency for a number of attributes such as sex, age, citizenship and residential location; and
v) Identifying the oldest person in the electoral register.

The answers would appear as follows:

Example	Depth of Use
i	E
ii	A
iii	D
iv	C
v	B

Chose well as each choice will imply the acquisition and learning of new tools where not already acquired. This may take some time to establish so it is best to run the above check first.

Which Tools are Available?

There are many tools which one can use in statistical analysis. They are varied: some are linked to traditional approaches and do not require highly-detailed technology, others are based on multi-mode methods employing both manual and digital means, whilst others are based on high-end technology that

basically takes over an analysis function. There is no longer the need to think about how to go about running a statistical test as this is done through a simple drop-down menu command.

This progress in technological innovation is tantamount to a literal JANUS! It is both a blessing and a curse as one now can literally abuse of the test to be used. It is really easy to run a correlation test for a nominal-ordinal structure as it is easy to sum two numbers. The issue is no longer one of access to the statistical tool and it is to abuse of the tests that are relative to the variable type. Great care must be taken to ensure that the relative tests are chosen.

Once one has decided to go for a specific method of employment, one has to choose the best which fits the purpose. There are both manual and fully automatic tools.

Manual

A simple calculator or grey matter should do the trick! In terms of basic mathematical calculations there is no need for hi-end technology, A few soft taps on the simple stand-alone keyboard that is even available in mobile phones will help one to reach an answer. Or else do it the old fashioned way and count on one's fingers, sketch on a piece of paper or use an abacus. Simple – yet very effective!

One can also use a spreadsheet and input all the numbers in successive cells. The statistical measure is inputted manually.

Town	Population
Valletta	1000
Floriana	200
Qormi	300
Qala	500
Zebbug	800
Total	**2800**

Semi-Automated

Taking the spreadsheet concept a bit further, one can create a tool that carries out a number of functions in order to automate the process. These tools are called macros, where the researcher inputs the data in the individual cells and the macro runs simple statistical measures automatically.

Automated

Automated tools serve as the cherry on the statistical cake. The researcher structures the dataset and the relative variable, input or imports the data from his/her survey and runs the relative statistical test. Today's tools are a breeze compared to those employed a few years ago when one had to programme the tool and input the data within the tool, severely limiting the number of variables one could use as well as the number of interviews one could carry out.

Imagine a questionnaire with 10 questions and the researcher had 100 respondents. That equates to a cool 1,000 inputs! Imagine the probability of error generation during the input stage as a series, of 1, 4, 3, 2, 5, 6, 3, 9, 12, 1, 3, 2, 4, 5….a veritable nightmare! At a rate of 1 input every two seconds (allowing for reading from the questionnaire and inputting) and stopping for 5 minutes every 20 minutes for an eye-rest that would easily result in a 45-minute run. Run this for every query and one runs out of research time…

Researchers nowadays get a cool deal: they create the variable structure, then input the data either manually or import it from another tool such as a spreadsheet and *viola* their work is done in as little as a few seconds. The authors have run statistical tests on data sets comprised of 221 attributes and 153,080 records. That would result in inputting time of 392 solid 24-hour days!

Distributed

There are different types of tools, including the automated ones that do not reside in one's computer! Users do not even need to know where they are. The internet phenomenon has helped users to run statistical tests from their desktops using technologies that are automated but are located anywhere on the Earth (or even elsewhere, should a particular technology be located in orbit!). The user inputs the data locally and can run statistical tests across the attributes created by the researcher and even against other attributes created by other parties which again are stored somewhere else. These distributed datasets could again be anywhere and as long as they conform to the same research rules (covered in Chapter 2), they can be used. One can for example analyse Malta's epidemiological data against the WHO's (World Health Organisation) data using such tools, where such options are available.

The following sections describe the different types of tools that exist giving examples of simple outputs.

Spreadsheets

What are spreadsheets? The simple electronic tools remind us of graph paper where one could add or subtract numbers, draw or play simple games. Today's electronic version can allow for the same modes of operation but also allow for some basic statistical analysis. The tools are composed of multiple cells in what are described as rows (records) and columns (attributes).

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
Record 1					
Record 2					
Record 3					
Record 4					
Record 5					

Spreadsheet cells allow for the inclusion of numbers, formulas and alphanumeric text. Formulas range from simple to very advanced and complex examples such as the following:

Example 1: Summing a set of values

Town	Population
Valletta	1000
Floriana	200
Qormi	300
Qala	500
Zebbug	800
Total	**=SUM(F2:F6)**

Town	Population
Valletta	1000
Floriana	200
Qormi	300
Qala	500
Zebbug	800
Total	**2800**

Example 2: Calculating how many persons live in a household aged over 40 years or less than 30 years

Age (Years)	Persons
46	Over 40 years
49	Over 40 years
29	Less than 30 years
38	Not Applicable
25	Less than 30 years
67	Over 40 years
12	Less than 30 years
33	Not Applicable
34	Not Applicable
67	Over 40 years
56	Over 40 years
4	Less than 30 years
78	Over 40 years
100	Over 40 years
10	Less than 30 years

=IF (E4>40,"Over 40 years", IF(E4<30,"Less than 30 years", "Not Applicable"))

Spreadsheets allow various basic statistical tools to be run and some modules also exist to expand on the tools and turn a spreadsheet into an advanced statistical tool. There are also online and standalone spreadsheets that in themselves offer very powerful features. The choice is nearly endless and the best advice one can give is, to use the one with the simplest interface and which has a modest set of commands.

The available tools

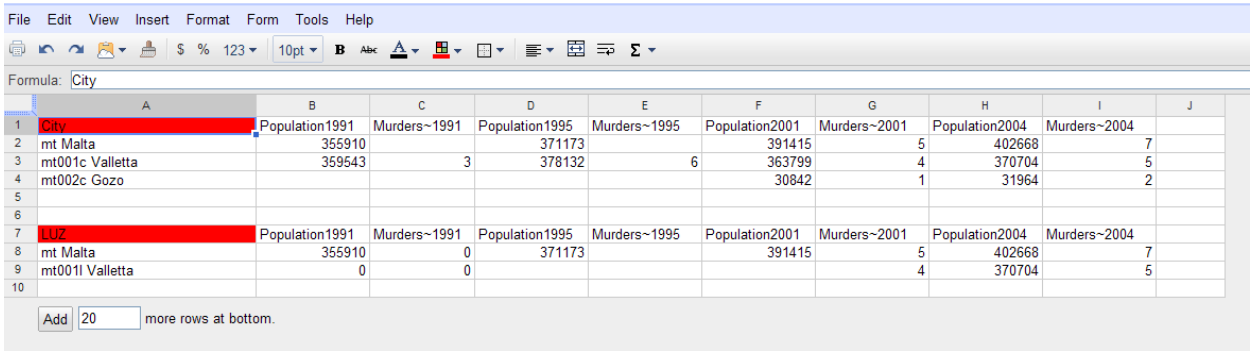
The most popular commercial Spreadsheets are IBM Lotus 123[1] and Microsoft Excel[2]. The former was originally created for DOS but has since been overtaken by Excel. Free/Opensource Desktop Spreadsheets are available that include OpenOffice Calc[3], Gnumeric[4], GNU Oleo[5], Bean Sheet[6], KSpread[7], SIAG[8], and Resolver One[9] (the latter free for personal use).

Free Online Spreadsheets include Simple Spreadsheet[10], wikiCalc[11], Google Spreadsheets[12] (Figure 9.1) and ThinkFree Online Calc[13].

---

[1] http://www-01.ibm.com/software/lotus/products/123/
[2] http://office.microsoft.com/en-us/excel/
[3] http://www.openoffice.org/
[4] http://projects.gnome.org/gnumeric/
[5] http://www.gnu.org/software/oleo/oleo.html
[6] http://bsheet.sourceforge.net/
[7] http://www.koffice.org/kspread/
[8] http://siag.nu/siag/
[9] http://www.resolversystems.com/
[10] http://www.simple-groupware.de/cms/Spreadsheet/Home
[11] http://www.softwaregarden.com/products/wikicalc/
[12] http://spreadsheets.google.com/
[13] http://member.thinkfree.com/member/goLandingPage.action#

Figure 9.1: Google Spreadsheets



	A	B	C	D	E	F	G	H	I	J
1	City	Population1991	Murders~1991	Population1995	Murders~1995	Population2001	Murders~2001	Population2004	Murders~2004	
2	mt Malta	355910		371173		391415	5	402668	7	
3	mt001c Valletta	359543	3	378132	6	363799	4	370704	5	
4	mt002c Gozo					30842	1	31964	2	
5										
6										
7	LUZ	Population1991	Murders~1991	Population1995	Murders~1995	Population2001	Murders~2001	Population2004	Murders~2004	
8	mt Malta	355910	0	371173		391415	5	402668	7	
9	mt001l Valletta	0	0				4	370704	5	
10										

Source: http://spreadsheets.google.com/

A very good site to access free opensource software is the Junauza.com[14] collection. There are many more such tools and an exhaustive list exists in the Wikipedia spreadsheet page[15].

Macros

As described earlier, some tools exist that serve as add-ons for spreadsheets called macros. These tools cater for specific requirements and are normally based on a sequence of commands that are enclosed within a 'shell' that can be run over and over as new data is inputted. Macros allow researchers to input their data in specified cells and run the resultant measure accordingly, thus drastically reducing the need for repeated work.

An excellent example of free macros can be found in the page: http://www.ozgrid.com/VBA/. Other related spreadsheets and macros webpage can be found in Matt H. Evans site[16]. Dedicated statistical macros, ranging from freeware to commercial licenses) can be found through the software.informer website[17].

Dedicated Statistical Software

Dedicated Statistical Software is available for researchers in various forms and covers both methodologies: those catering for quantitative and qualitative. In this section we cover both methodologies as well as the specialised quantitative-based spatial statistics tools. The tools are described in summary and a walkthrough has been created based on one tool for the readers serving as an introductory step towards the use of such tools.

Quantitative

A number of highly polished tools exist and are used by researchers in their process of analytical endeavour. The main tools used are PASW known as SPSS, SAS, Stata and MiniTab, with opensource tools gradually coming to the fore such as R-Commander, PSPP and Gretl. These examples are not exhaustive and many more tools are available.

1. SPSS (PASW)

SPSS[18] (Statistical Package for the Social Sciences), which was also called PASW (Predictive Analytics SoftWare) between 2009 and 2010 is a commercial statistical analytical processing tool. It has a base set of statistical features and can also be enhanced by various specialised modules.

---

[14] http://www.junauza.com/2008/04/freeopen-source-spreadsheet-programs.html
[15] http://en.wikipedia.org/wiki/List_of_spreadsheet_software
[16] http://www.exinfm.com/free_spreadsheets.html
[17] http://software.informer.com/getfree-statistical-macro/
[18] http://www.spss.com/

The user-friendly base tool caters for descriptive statistics such as frequencies and cross tabulation, bivariate statistics such as means, anova and correlations as well as nonparametric tests. It also carries out tests for linear regression, factor analysis and cluster analysis. The add-on modules include such tools as those of: tables, trends, advanced models and maps, amongst others.

2. SAS

SAS[19] (or Statistical Analysis System) is a commercial suite of statistical tools that was formed, based on the integration of a number of software tools. Since SAS works on the integration of a number of tools, the most adequate tool for statistical analysis is that called SAS/STAT[20], which together with BASE SAS and SAS GRAPH form the main components required for analysis within a tool called SAS Analysis Pro.

SAS/STAT can be employed for analysis through categorical data analysis, regression, bayesian analysis, multivariate analysis and other specialised functions as survival analysis, psychometric analysis, cluster analysis and nonparametric analysis, amongst others.

3. Stata

Stata[21] is a commercial general-purpose statistical tool, originally using a command-line interface but recent versions have been enhanced with GUI (Graphic User Interface) which makes it easier to use. Stata allows for such tests as: summary statistics, regressions, ANOVA, cluster analysis, survival models and cluster analysis, amongst others.

It is somewhat limited by its inability to load more than one file simultaneously and that it cannot load very large files. However, it has the capability to operate in the same way as opensource through the integration of online material.

4. MiniTab

MiniTab[22] is a commercial tool that together with another tool from the same company called Quality Trainer provides a range of statistical functions that are both wide-ranging and user-friendly. The statistical functions include basic statistics, descriptive statistics, regression, t-tests, variance, correlation, least squares and ANOVA, amongst others.

5. R-Commander

R-Commander[23] is an opensource tool that is deemed to be the most comprehensive, free statistical software available. Whilst the interface is a bit daunting for new users, there is a comprehensive manual and FAQs which guide the user in its usage. A number of plug-ins also enhances the product.

The tool and its plug-ins provide a wide-range of statistical tests, time-series analysis, classification, as well as linear and non-linear modeling. The outputs include very interesting 3D plots that emulate digital elevation models (DEMs) which can be used also for change analysis.

6. PSPP

PSPP[24] is a free opensource tool that was originally called Fiasco. It replicates SPSS functionality and serves as a useful tool for statistical analysis. The description states that “PSPP is a program for statistical analysis of sampled data. It is a free replacement for the proprietary program SPSS. PSPP development is ongoing. It already supports a large subset of SPSS's syntax. Its statistical procedure support is currently limited, but growing” (PSPP Installation File).

On first reviewing the tool, one would be forgiven to mistake it for SPSS, so mirrored is the whole concept. Whilst the number of statistical measures are limited, this tool serves as a very good tool for

---

[19] http://www.sas.com/technologies/analytics/statistics/index.html
[20] http://www.sas.com/technologies/analytics/statistics/stat/index.html
[21] http://www.stata.com/
[22] http://www.minitab.com/
[23] http://www.r-project.org/
[24] http://www.gnu.org/software/pspp/

most organisations since it covers the most used measures, inclusive of descriptive statistics, means, correlations, factor analysis, linear regression and non-parametric statistics, amongst others. The walkthrough uses this tool to aid researchers get an overview of how to use such tools.

7. Gretl

Gretl[25] is a free opensource asset that provides various statistical tools for econometrical analysis. Gretl stands for Gnu Regression, Econometrics and Time-series Library. Whilst not fully comparable to the other commercial tools as in the case of PSPP, it offers various time-series, maximum likelihood methods, least-squares based statistical estimators and econometric tests (Baiocchi et al, 2003).

A simple Walkthrough using PSPP

Part 1: Using PSPP for the first time - Values and Variables

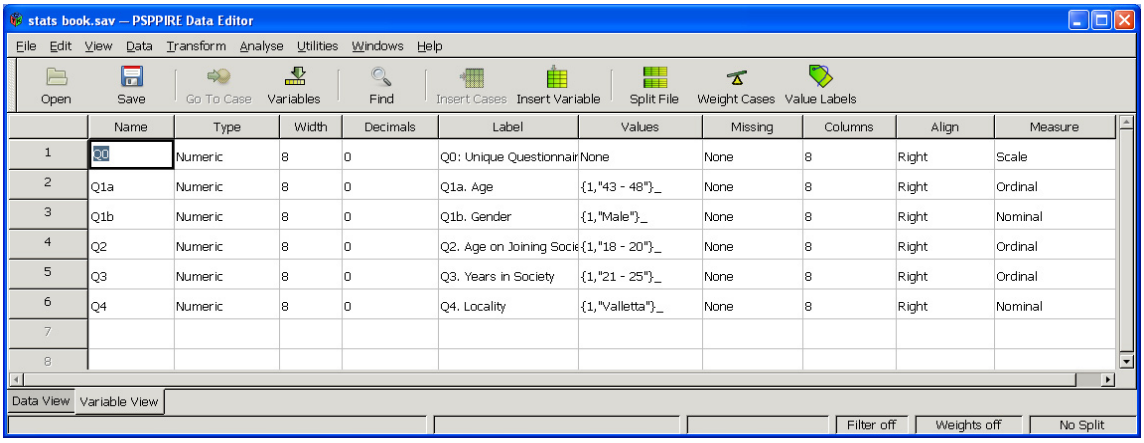
1) Open PSPP

2) Using PSPP for the first time – Inputting Data

There are two Views in PSPP, the Variable View and the Data View.

Variable View – input the variables from your survey here and list the different parameters being analysed, e.g.: for a question on sex: input 1 for male and 2 for female. This will do away with retyping each description for every questionnaire as against typing only one number (Figure 9.2).

Figure 9.2: Data View

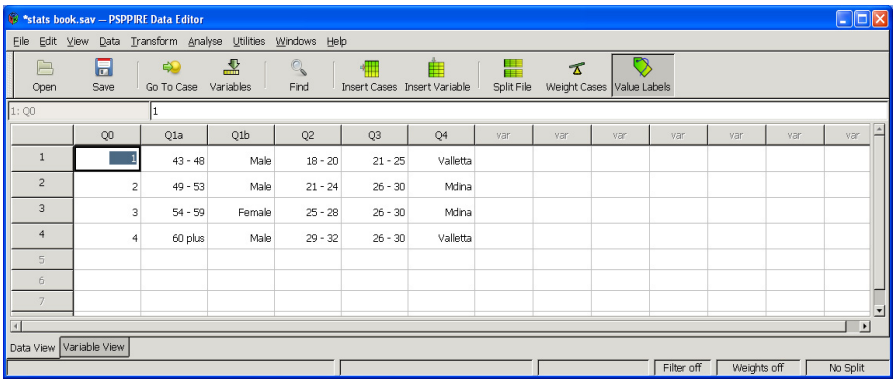


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Q0	Numeric	8	0	Q0: Unique Questionnair	None	None	8	Right	Scale
2	Q1a	Numeric	8	0	Q1a. Age	{1,"43 - 48"}_	None	8	Right	Ordinal
3	Q1b	Numeric	8	0	Q1b. Gender	{1,"Male"}_	None	8	Right	Nominal
4	Q2	Numeric	8	0	Q2. Age on Joining Socie	{1,"18 - 20"}_	None	8	Right	Ordinal
5	Q3	Numeric	8	0	Q3. Years in Society	{1,"21 - 25"}_	None	8	Right	Ordinal
6	Q4	Numeric	8	0	Q4. Locality	{1,"Valletta"}_	None	8	Right	Nominal
7										
8										

[25] http://gretl.sourceforge.net/win32/

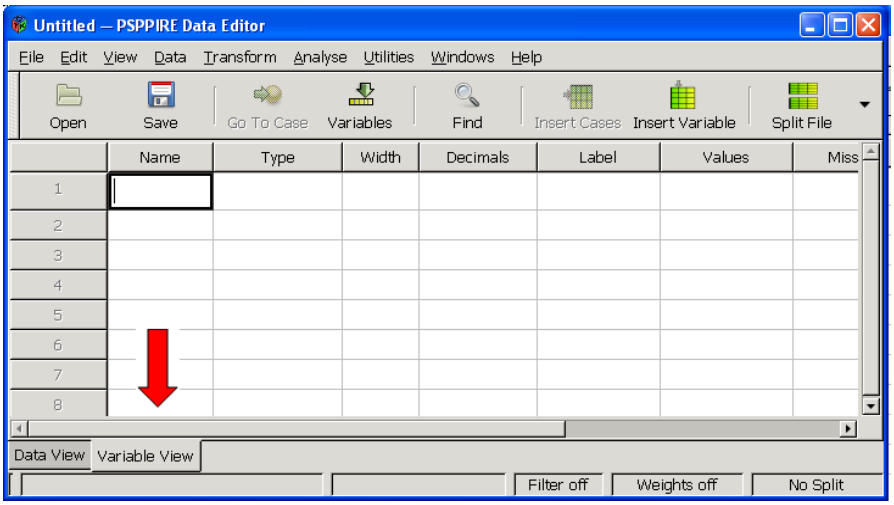
Data View – input the data from your survey in this tab sheet as per codes entered in the variable field above (Figure 9.3).

Figure 9.3: Variable View

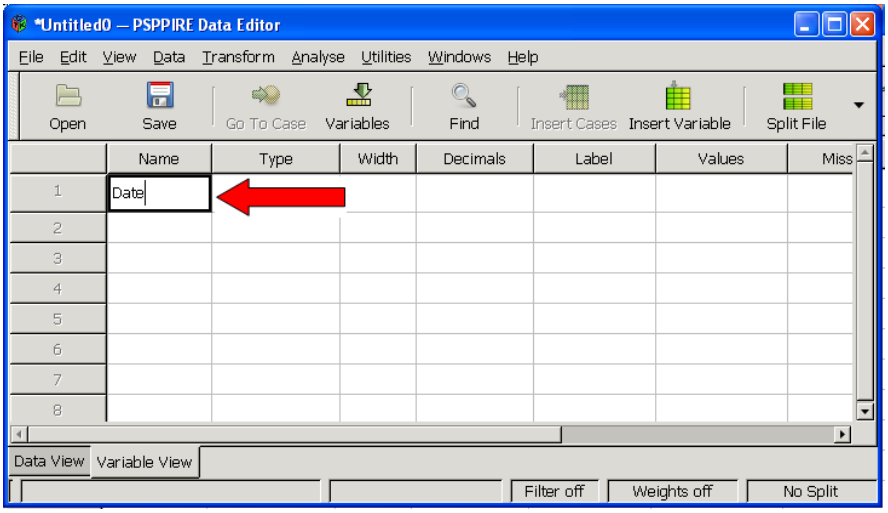


3) Using PSPP for the first time – the first variable.

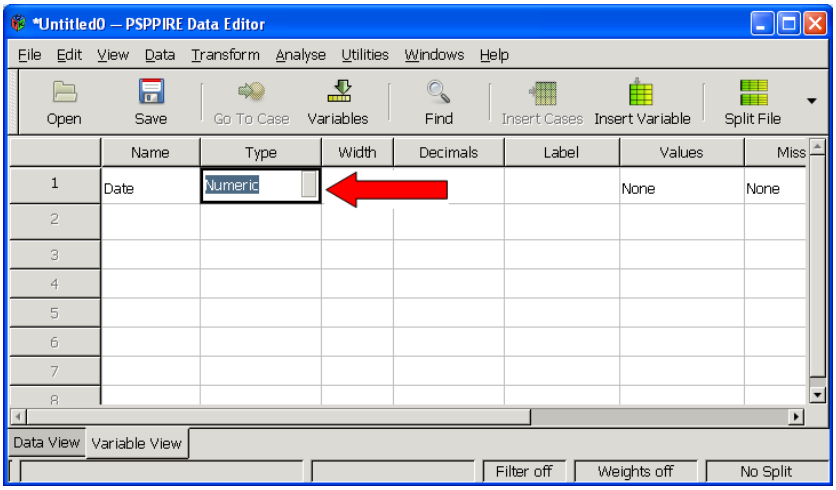
Click on Variable View option

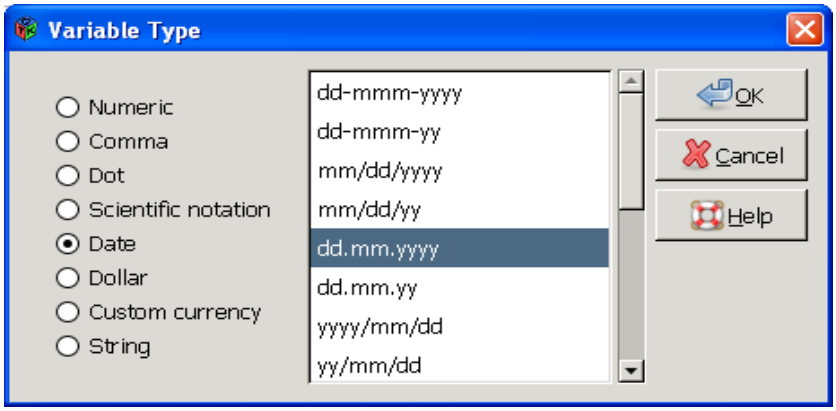


4) Double click on the top left hand cell (under Name) and type in ‘Date’.



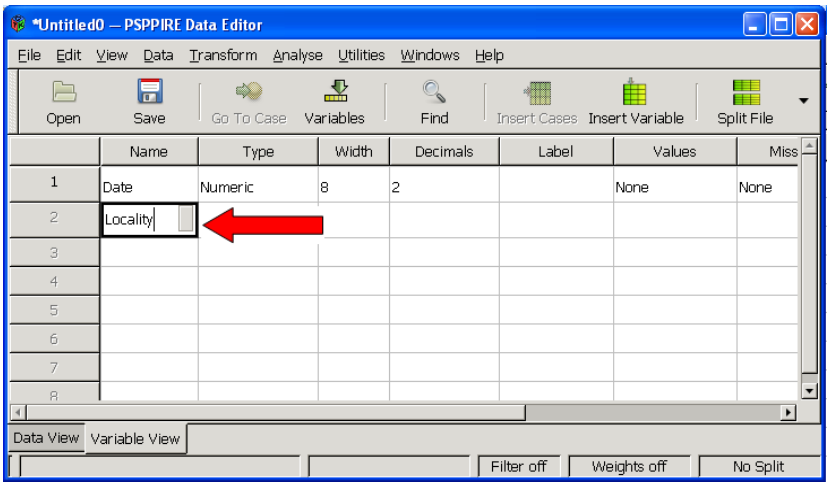
5) Double click on the top cell (under Type) cell. ‘Numeric’ and a choice box appears. Click on the grey box and then choose Date (dd.mm.yyyy) in the popup window.



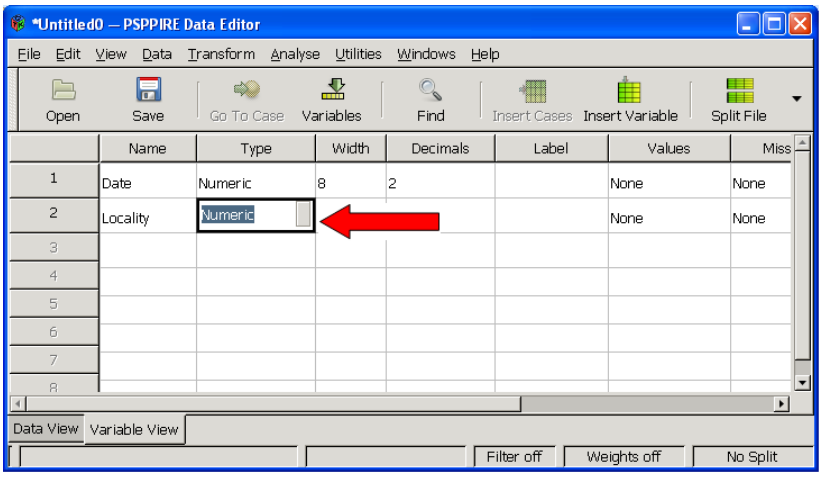


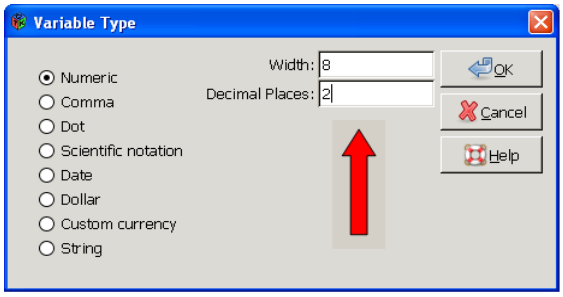
6) Using PSPP for the first time – the Variable Type.

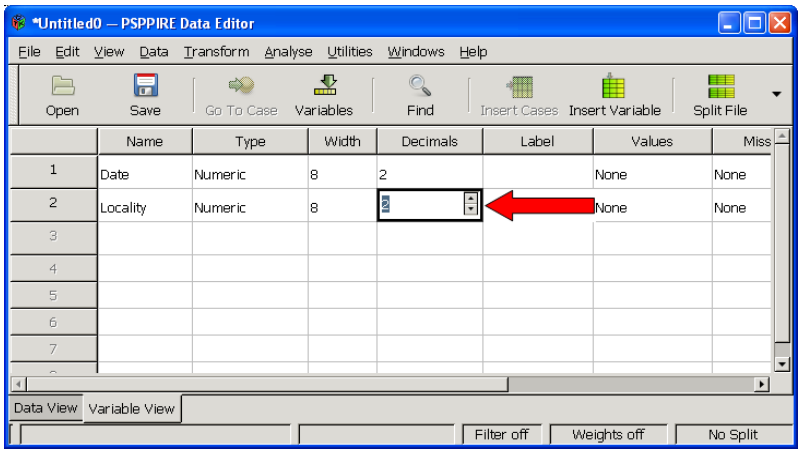
Double click on the second cell (under Name) and type in ‘Locality’



7) Press tab to go to the next parameter (Type). Ignore the number that appears under Width, Decimals, etc. In the popup window (Variable Type) type in 2 for Width and 0 for Decimal Places. Click Ok and you will notice that the numbers in under Width and Decimals change accordingly.

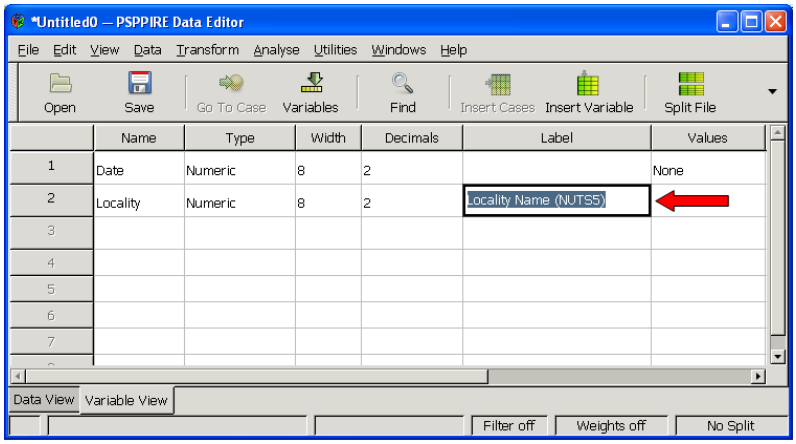




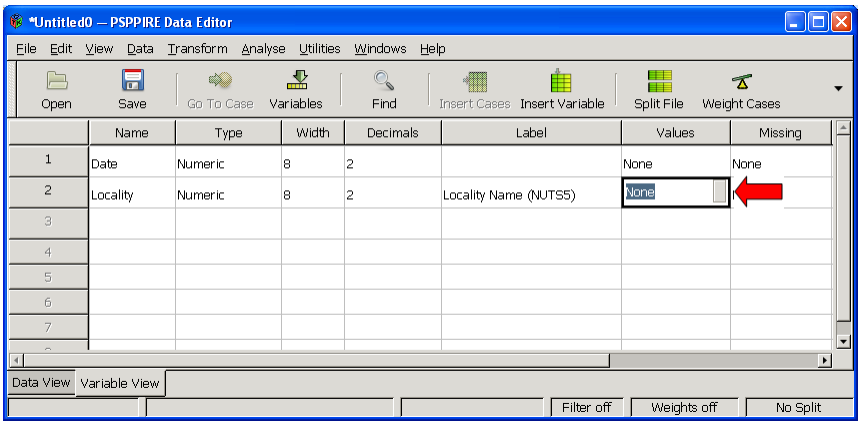


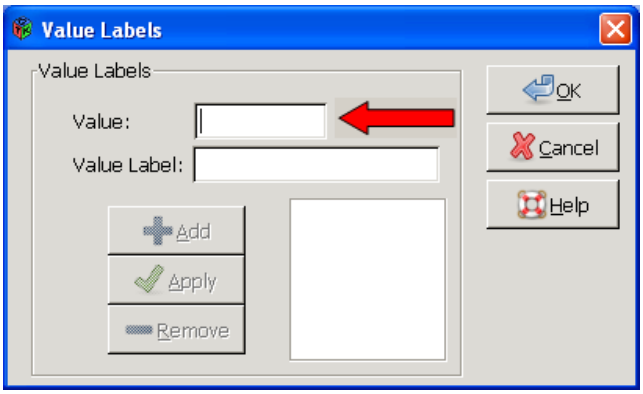
8) Using PSPP for the first time – the Value Label.

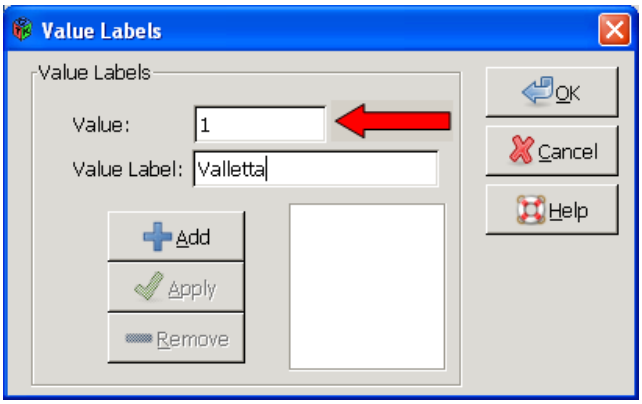
Click on Label and type in ‘Locality Name (NUTS5)’.

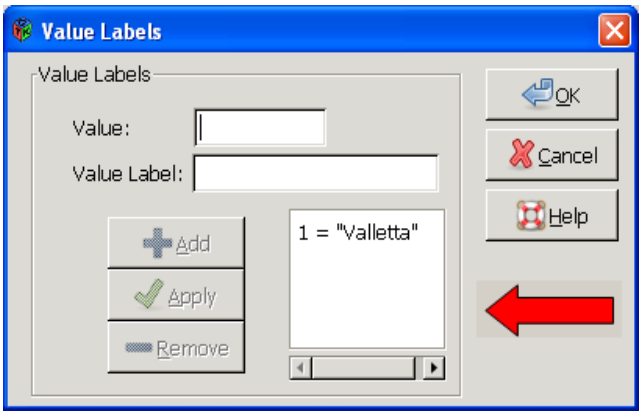


9) Click on Values and press the grey button. A Value Label window will pop up. Click on value and type in 1, then press tab and type in ‘Valletta’ in the Value Label box. Click on the add button. Note that the text ‘1 = “Valletta”’ will appear. Repeat this process for ‘Floriana’ with a Value of 2. Add some more localities.



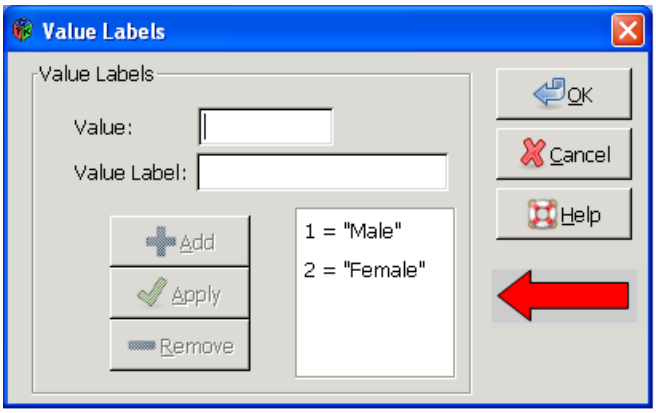






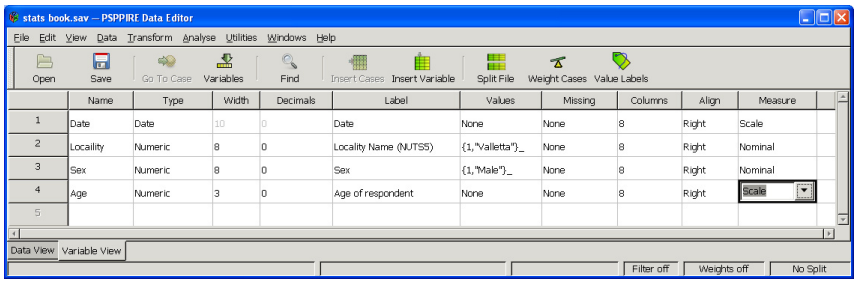
Using PSPP for the first time – some more inputs

10) Add another variable called Sex (under Name), Type (Numeric 2,0), Label (Sex of Respondent) and insert Value Labels (1 for Male and 2 for Female).



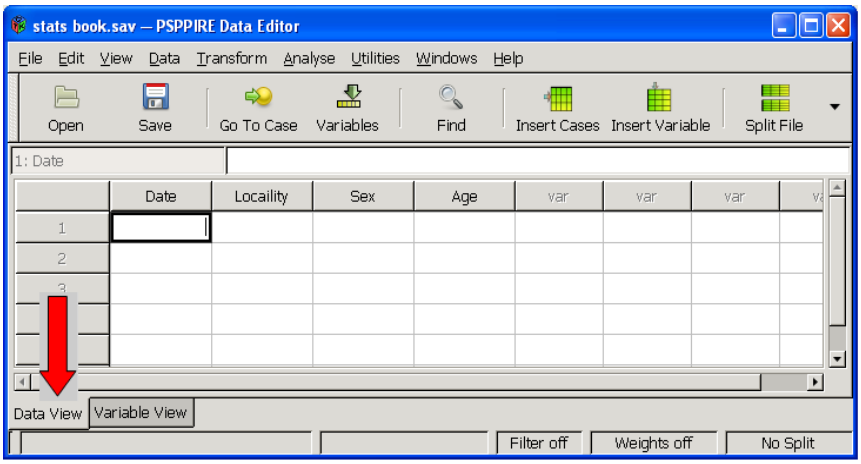
11) Insert another variable named Age, Type (numeric), Width (3) Decimals (0), Label (Age of respondent) and leave Values as none, since this is dependent on each individual's age rather than a specific category.

Note that the Measure has been inserted for the variables accordingly (Locality and Sex are Nominal whilst Age is termed as Scale).

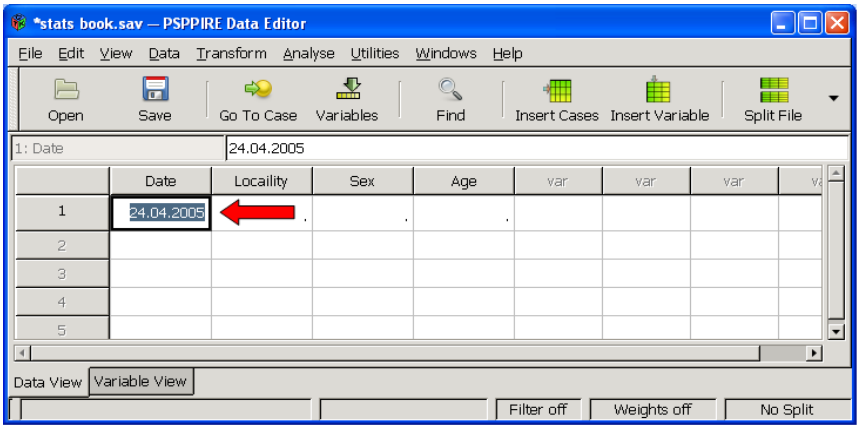


12) Using PSPP – inputting some real data now.

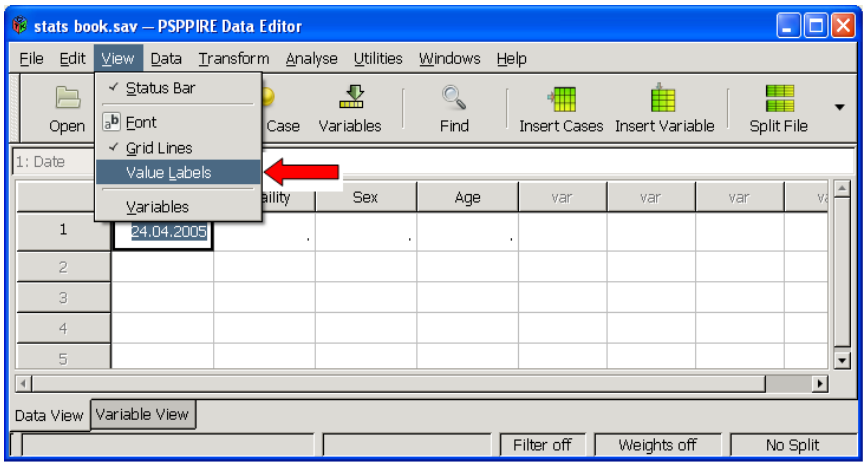
Click on Value View option



13) Type in 24.04.05 (under Date).

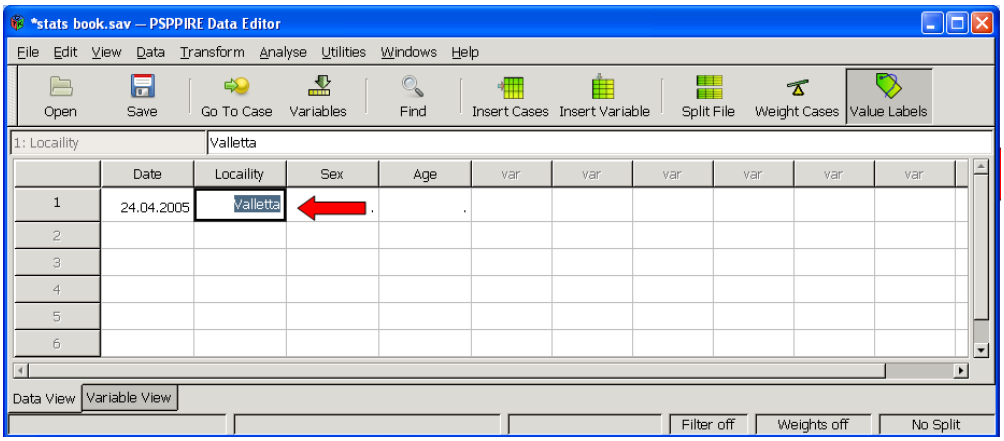


14) Click on the menu button VIEW and click Value Labels. This activates the choice button in the variable fields.

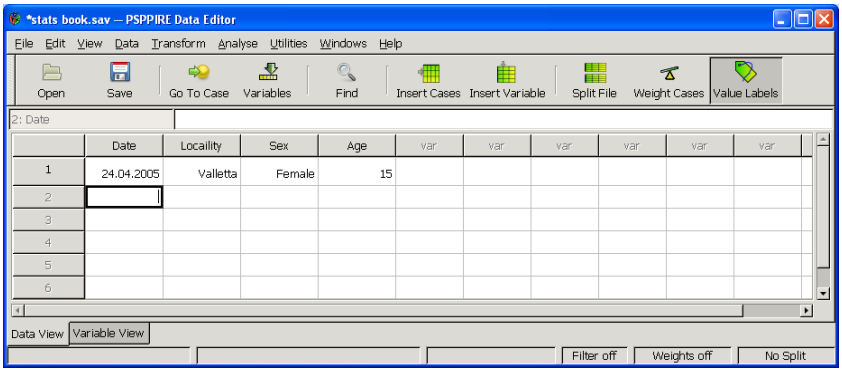


15) Using PSPP: choosing the values.

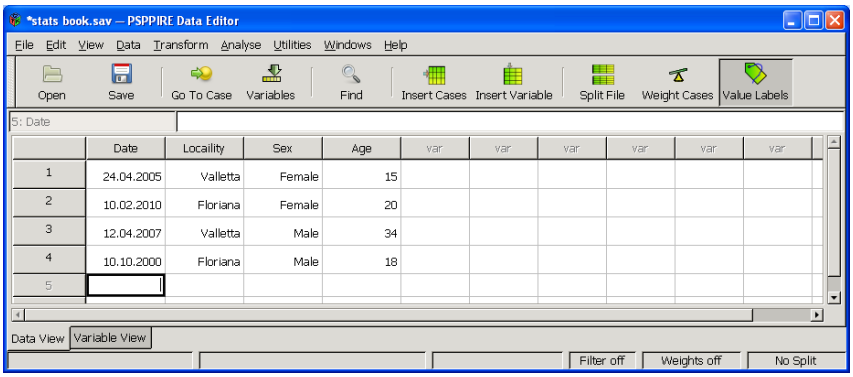
Go back to the input cells and click on Locality and insert your number that relates to the variable – e.g. typing in 1 under locality gives Valletta. Note the activations of the value Label at the Right Hand Side.



16) Repeat for Sex, whilst for Age type in 15.



17) Populate the dataset with a number of records.



Part 2: Using PSPP – The statistical measures: Frequencies

18) Running Frequencies – work instruction.

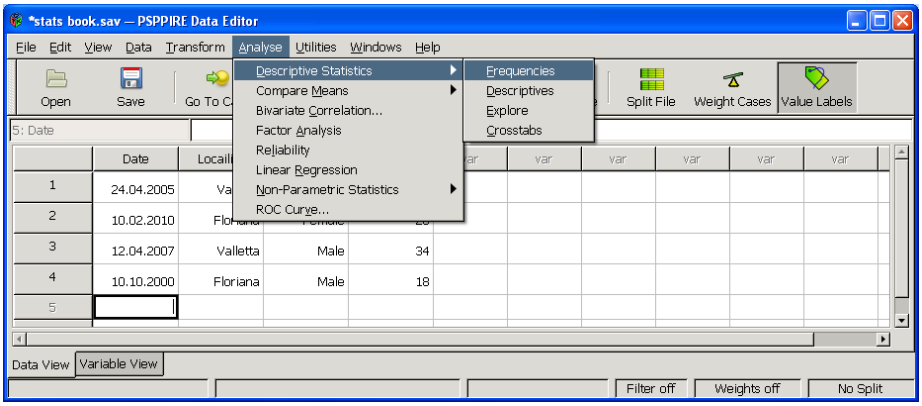
From within PSPP, Run Frequency and Descriptive Statistics

Follow these instructions.

1) Go to Analyse on the menu bar.

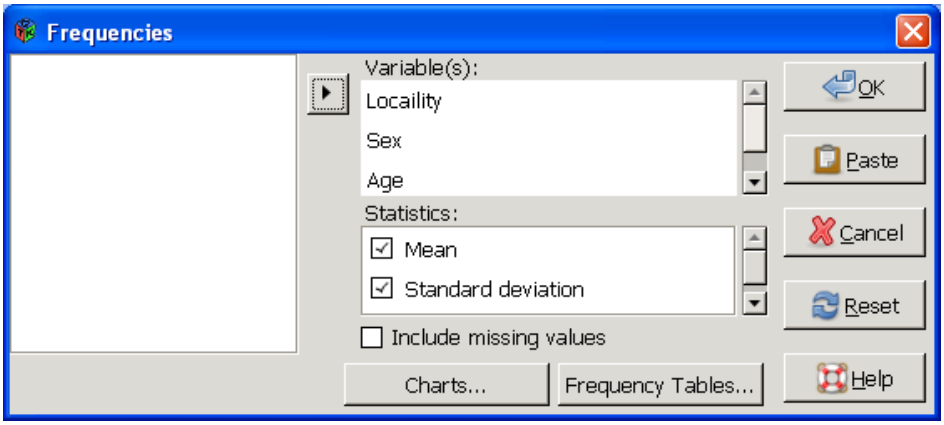
2) Click on Descriptive Statistics and then click on Frequency.

Figure 9.4: Descriptive Statistics

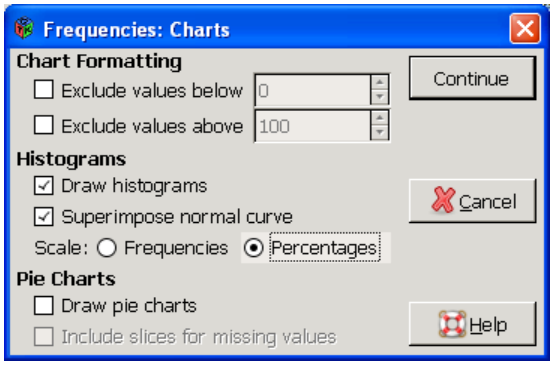


3) Select all variables shown in the left box, send to the right box (using arrow).

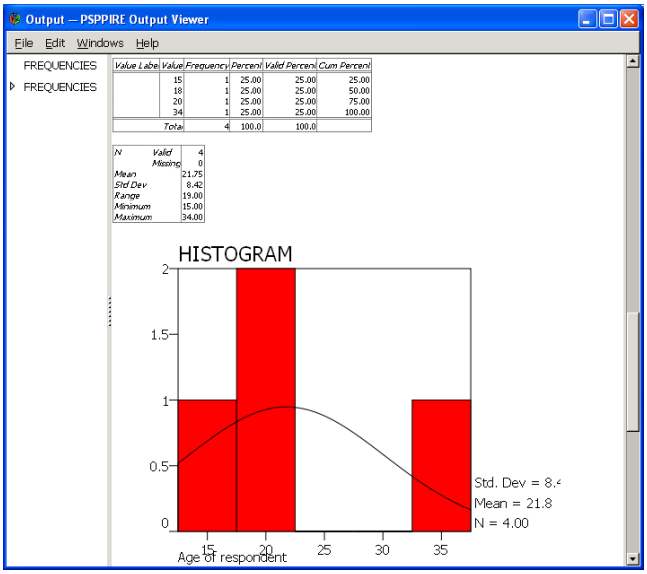
4) Select statistics (e.g., mean, minimum, maximum, standard Deviation, range).



5) Click on Charts, if you want to see charts for your variables.



6) Then click on OK. You will see the results in output file.



The above example gives a brief overview of what can be achieved with statistical tools. There are many statistical functions that can be carried out using PSPP. It is ideal to follow the PSPP tutorial that cover the different functions.

A list of free statistical tools can be downloaded from the Freestatistics website: http://www.freestatistics.info/stat.php

Qualitative tools

The software packages that are used in qualitative research analysis are called Computer-Assisted Qualitative Data Analysis Software (CAQDAS) or Qualitative Data Analysis Software (QDAS or QDA software). This kind of software help to: organize, categorize, annotate textual and visual data. This software aims at building theory while visualizing the relationships between data and/or theoretical constructs. Some of the software are the following: AnSWR[26], ATLAS.ti[27], CDC EZ-Text[28], The Ethnograph[29], Kwalitan[30], MAXqda[31], N6 or NUD*IST (which when superseded was replaced with NVivo[32] and XSight[33]), QDAMiner[34], Qualrus[35], TAMS Analyzer[36], Transana[37] and Weft QDA[38]. As an example of how these software packages function this section will provide a brief walkthrough of ATLAS.ti6 identifying some of functions that are used in qualitative research methods.

CAQDAS packages interpret the data collected through recognition and codification of the various research issues, facilitating the explanation and creation of theories. For instance, some of the approaches include grounded theory and conversation analysis. This does not exclude the possibility that qualitative researchers conjoin quantitative data with their methods. One of the most frequently asked questions in choosing one of the CAQDAS is "which is the 'best' package…?" It is practically impossible to answer this question as all the packages have tools that provide support in various stages of the analytic process and every program has its own advantages and disadvantages.

Some software packages suit certain types of approach more than other. This creates debates as to whether a particular package can be manoeuvred to suit a particular analysis. The researcher should remain in control of the interpretive process and decide if utilizing any software facilitates the chosen analysis approach. Whichever package one chooses, only a selected number of tools will be utilised in data running and analysis. The most sophisticated packages may not suit the envisaged task. Deciding which is the 'best' CAQDAS is a subjective judgment based on a numbers of factors but a more resolute choice of tools determines if a software package will serve over time.

Some similarities between CAQDAS packages

CAQDAS packages have key principles that facilitate the qualitative research process in similar ways. For instance in content analysis, transcribed interviews there is a process called KWIC (Key Words in Context). This tool offers ways to search in the text for singular words, phrases, or a collection of words on a particular theme. This function provides access to those keywords that appear in the analyzed documents.

Some of these software packages integrate code and retrieve functionalities. The user or research here can define key-words and/or theoretical categories (codes) that are embedded in the text. The researcher structures the coding and strategies to be employed. Code is simple and flexible and the researcher can modify and refine the coding as considered necessary. In most of these software packages the coding action rests entirely on the user.

All these packages offer means to control the research project and classify the data according to facts, features and data types. CAQDAS packages significantly facilitate qualitative research and analysis processes enables the researcher to focus on combinations and comparison of singular data. Qualitative data analysis is rarely a linear process. CAQDAS packages include various writing tools that enable the researcher to post comments and annotations of data that could not be reported in any other way; such as the non-verbal communications that one may observer during interviews. Some of the software

---

[26] http://www.cdc.gov/hiv/topics/surveillance/resources/software/answr/index.htm
[27] http://www.atlasti.com/
[28] http://www.cdc.gov/hiv/topics/surveillance/resources/software/ez-text/
[29] http://www.qualisresearch.com/
[30] http://www.kwalitan.nl/engels/
[31] http://www.maxqda.de/
[32] http://www.qsrinternational.com/products_nvivo.aspx
[33] http://www.qsrinternational.com/products.aspx
[34] http://www.provalisresearch.com/QDAMiner/QDAMinerDesc.html
[35] http://www.ideaworks.com/qualrus/
[36] http://tamsys.sourceforge.net/
[37] http://www.transana.org/
[38] http://www.pressure.to/qda/

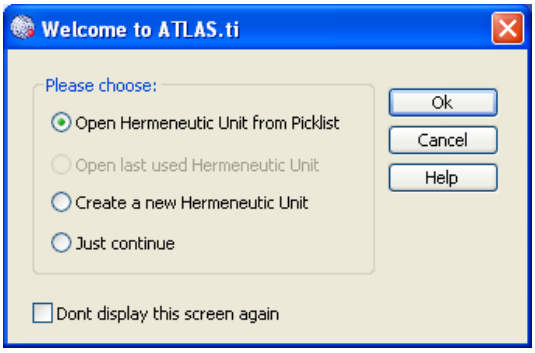
packages allow the researcher to analysis the material in hard copy or to integrate it in other applications such as Microsoft Word[39] and Excel[40]. These similarities indicate that each CAQDAS software package can be an asset that its methodical use of packages assists in continuity, increase precision and thoroughness in qualitative analysis.

A simple Walkthrough in CAQDAS using ATLAS.ti 6

http://www.atlasti.com/

Thomas Muhr from Free University, Berlin created ATLAS.ti and Scientific Software Development GmbH, Berlin is continuing to support the development of this software in the dynamic sphere of research. The features of ATLAS.ti can be applied in various fields of research such as art, social sciences, education and criminology. The last version of this software package is ATLAS.ti6 (Figure 9.5).

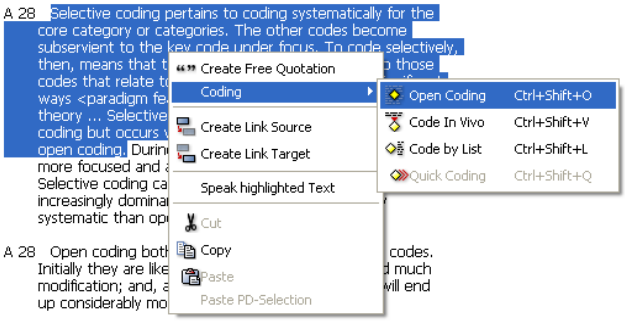
Figure 9.5: Atlas Ti6



The latest version of ATLAS.ti acts as an external database and it is possible to transcribe multimedia data using this software. This software has an electronic "room" called Hermeneutic Unit (HU), which helps text interpretation and prepares the data for each project analysed by ATLAS.ti. Thus the HU is connected between the primary data and any annotations taking in examining the collected data and keeps track of all of your data in ATLAS.ti project file. The data can include text, images audio and visual recordings, pdf files, and data extracted from Google Earth[41].

Functions operate from main menus and drop-down menus that can be access through the Manager windows. Selected quotes, text (Figure 9.6), images (Figure 9.7) or other documents enables flexibility in the use of coding.

Figure 9.6: Coded Text



[39] http://office.microsoft.com/en-gb/word-help/word-help-and-how-to-FX010064925.aspx?CTT=97
[40] http://office.microsoft.com/en-gb/excel-help/excel-help-and-how-to-FX010064695.aspx?CTT=97
[41] http://earth.google.com/

Figure 9.7: Coded Images

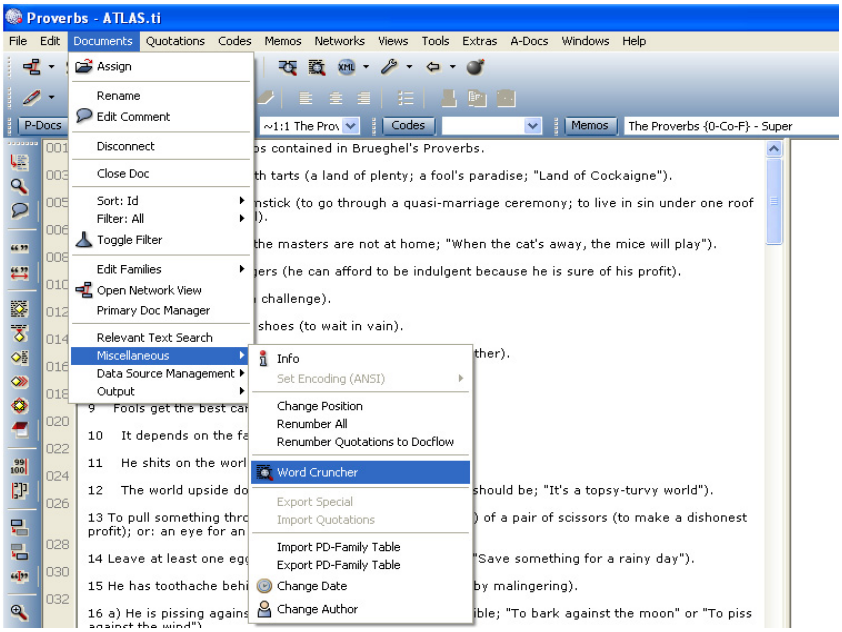
(This image is one of Hitchcock's cameos extracted from the Hitchcock's movie *North by Northwest,* adopted from http://www.filmposters.com/, accessed 9th August, 2010)

However the output results this kind of data are independent from any coding function. Quotations are not dependent on ATLAS.ti software, selected data can be pointed and separately marked without being coded. However this independence from the software does not hinder that quotations are located in the framework. This feature allows one to produce quotations in different formats such as audio and can be play-backed autonomously from the written transcript.

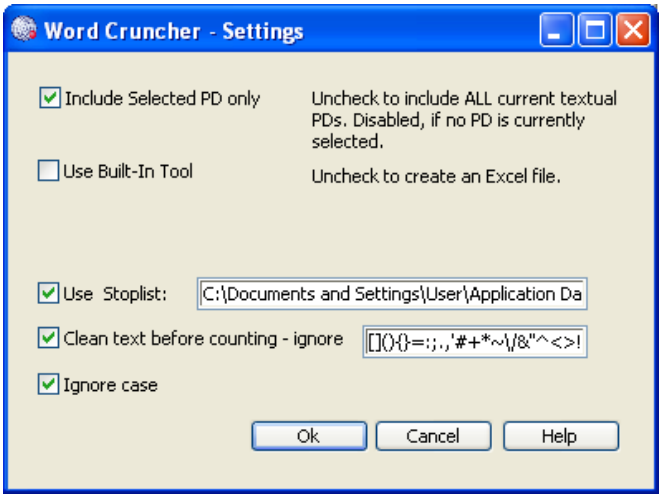
Word Cruncher

The Word Cruncher counts the number of times a word appeared in the whole collected data or a particular document. The results can be escorted in a Microsoft Excel spread sheet or in a form of a memo. After uploading the text in ATLAS.ti, these are the steps taken to use the Word Cruncher:

Step 1 – Click on **Documents**, then **Miscellaneous**, then **Word Cruncher.**

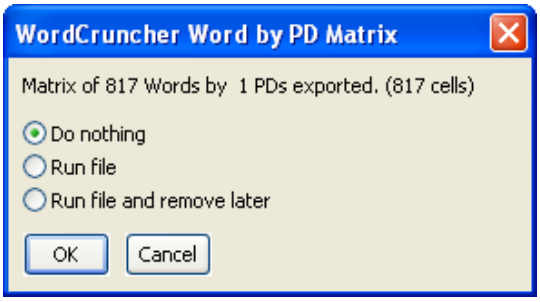


Step 2 – Click on Properties before starting the frequency count.



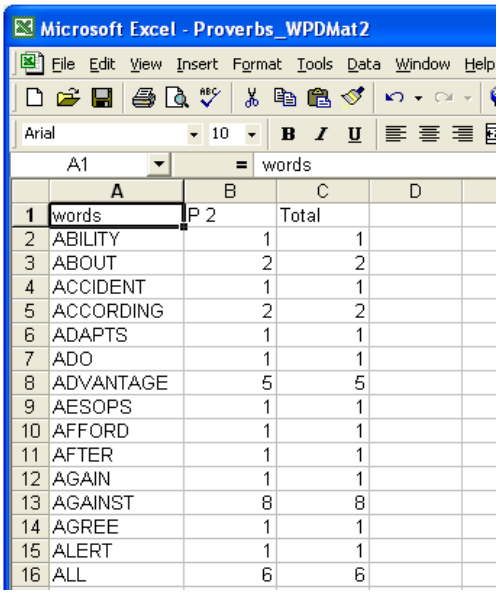
The results can be either saved in Excel or not. For this example the user opted to use Excel as the user wanted to use the data for further analysis.

Step 3 – This is the final window before the user can proceed to the final output. If you click on "**Do nothing**" the frequency count is saved in Excel. If you select "**Run file**" the file is opened in Excel, if you click on "**Run file and remove later**" you may want the file to be removed after viewing the results.



Naturally, you need to have Excel installed before being in a position to see the results.

Step 4 – Open Excel where the information is stored.



	A	B	C	D
1	words	P 2	Total	
2	ABILITY	1	1	
3	ABOUT	2	2	
4	ACCIDENT	1	1	
5	ACCORDING	2	2	
6	ADAPTS	1	1	
7	ADO	1	1	
8	ADVANTAGE	5	5	
9	AESOPS	1	1	
10	AFFORD	1	1	
11	AFTER	1	1	
12	AGAIN	1	1	
13	AGAINST	8	8	
14	AGREE	1	1	
15	ALERT	1	1	
16	ALL	6	6	

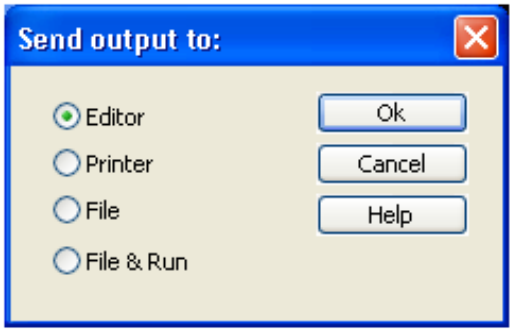
ATLAS.ti6: Remarks

Holistically ATLAS.ti is a flexible system to work with and can be engage in various qualitative projects. This software does not need codes to function but yet one can easily filter the data collected. Thus the researcher has all the liberty to manage the different linkages in the data. The latest version of this software (i.e. ATLAS.ti6) has a number of upgrades when compared to the earlier versions. ATLAS.ti6 supports pdf files and creates a very precise representation of the pdf including the layout, images and other contents. Another important feature that can be accessed through ATLAS.ti6 is Google Earth (GE). This package offers the possibility to include geographical references and navigation through GE facilitating the creation of snap-shots of the interested spots around the world.

However there are some issues that the user of ATLAS.ti6 has to be aware of. This software has no KWIC and the word frequency tool is quite basic. Besides ATLAS.ti does not have any function that integrates quantitative data with filtered qualitative data. Another function that may raise some eyebrows is that the code list does not have a function to create connections in a hierarchical structure to systematically clean up the collected data. However for some other users this is an advantage, as it does not restrict the researcher in anyway. Nevertheless ATLAS.ti6, like any other software, is continuously being updated and new tools are constantly being added.

Output generated by ATLAS.ti

Among the output of the information generated, ATLAS.ti usually creates textual information though exists the possibility of producing a graphical result. Textual reports usually include lists of codes, annotations and citations. In order to create textual output the following window is displayed providing four options (as found in the diagram) of how the user wants to extract the data.



Geo-statistical tools

A third set of tools which are available for researchers are those related to spatial statistics. Though forming a specialised methodology, they form part of the quantitative approach, but bring into context the spatial element. This element as discussed in the Visualization Chapter looks at the analysis of statistics through the spatial dimension. It employs GIS and other specialised tools.

There are various tools covering geostatistics; some commercial such as ArcGIS Geostatistical Analyst and MapInfo Vertical Mapper, whilst others were produced through grants such as CrimeStat, SAGE and STAC. Others have been developed as opensource which include Gstat and SGeMS.

1. ArcGIS Geostatistical Analyst

ArcGIS Geostatistical Analyst[42] is a commercial tool that serves as an extension to ArcGIS Desktop. It consists of a suite of geostatistical tools focusing on spatial data exploration and surface generation. The tools allows for data sampling, interpolation surfacing and prediction modelling. The tool requires the main ArcGIS software to operate.

---

[42] http://www.esri.com/software/arcgis/extensions/geostatistical/index.html

2. MapInfo Vertical Mapper

A commercial tool which is also an add-on to another GIS software, MapInfo Vertical Mapper[43] sits on the GIS tool MapInfo and provide a set of tools that produce: trend analysis, gridding algorithms, prediction modeling, gravity modeling, risk modelling and large dataset correlations.

Though not strictly fully statistical tools, both MapInfo Vertical Mapper and ArcGIS Geostatistical Analyst serve to provide specific geostatistical function not available in mainstream applications. The outputs are in visualised formats such as 2D and 3D maps as well as in VRML and other virtual environments.

3. CrimeStat[44] (Levine, 2002)

'*CrimeStat®* is a free spatial statistics program for the analysis of crime incident locations, developed by Ned Levine & Associates under grants from the National Institute of Justice (grants 1997-IJ-CX-0040 and 1999-IJ-CX-0044). CrimeStat allows the analysis of: standard deviation maps, attribute analysis, journey to crime, hotspot analysis and a series of spatial statistical measures (Formosa, 2007). Though the software was created to analyse crime statistics, it is a robust tool and is a veritable multi-thematic / multi-discipline tools as it can be used for both social and natural scientific analysis.

As an example of types of spatial statistics used in crime, listed below are the CrimeStat categories clustered in four-groups[45]: Spatial distribution, Distance statistics, 'Hot spot' analysis routines and, Interpolation statistics (refer to Chapter 11).

The software works as a standalone and exports its data in GIS format for further mapping through the dedicated GIS software.

4. SAGE

SAGE[46] (Spatial Analysis in a Geographical Environment) was produced under ESRC (Economic and Social Research Council) research grant R000234471 'Developing spatial statistical software for the analysis of area-based health data linked to a GIS' (Craglia M., Haining R., and Wiles P., 2000).

The idea behind the project is described by Wise, Haining; and Ma (2001). It was structured through the creation of a software for statistical spatial data analysis (SSDA) which was concatenated with ARC/INFO. This product eventually produced SAGE.

5. STAC

STAC[47] (Space and Temporal Analysis of Crime) software was developed by the Illinois Criminal Justice Information Authority (ICJIA). A Users Manual is available: Users Manual and Technical Manual, (1996), Chicago, IL: ICJIA. It is a free tool that helps spatial analysts in their statistical analysis. It achieves this through cluster mapping, employing standard deviational ellipse creation. STAC was eventually integrated into the CrimeStat II tool.

6. Gstat

Gstat[48] is a free opensource tool that war developed to enable multivariable geostatistical modeling. It also predicts and simulates modeling scenarios. The tool can now be used in conjunction with R-Commander earlier described in the quantitative section. It is capable of calculating variograms, kriging, and allows unlimited variables to be cross-correlated.

---

[43] http://www.pbinsight.com/products/location-intelligence/applications/mapping-analytical/vertical-mapper/
[44] http://www.icpsr.umich.edu/NACJD/crimestat.html
[45] http://comm-org.utoledo.edu/pipermail/announce/1999-December/000025.html
[46] http://www.informaworld.com/smpp/content~db=all~content=a713811641
[47] http://www.icjia.state.il.us/public/index.cfm?metasection=Data&metapage=StacFacts
[48] http://www.gstat.org/

7. Stanford Geostatistical Modeling Software (SGeMS)

A free opensource tool, SGeMS[49] hosts a plethora of algorithms catering for extensive multiple-point statistics simulation and 3D visualization. Apart from standard data analysis tools such as histogram, QQ-plots and variograms, the tool also provides for kriging, multi-variate kriging (co-kriging), sequential gaussian simulation, sequential indicator simulation, multi-variate sequential gaussian and indicator simulation

In addition to the above, a list of geostatistical tools is available from the http://www.brynmawr.edu/geology/GIS/geostats.html website. The GWR[50] (Geographically Wieghted Regression) website also highlights the importance of specialised tools for geostatistical analysis.

Online Tools

There are two types of online tools: those that cater for the analysis of data and those that help the researcher to create an online survey for world-wide respondent input.

Some basic online analysis tools

There is quite a number of statistical tools that run quick tests for users who do not wish to acquire or use large complex software.

This section lists some of these tools in brief.

i) GraphPad Software

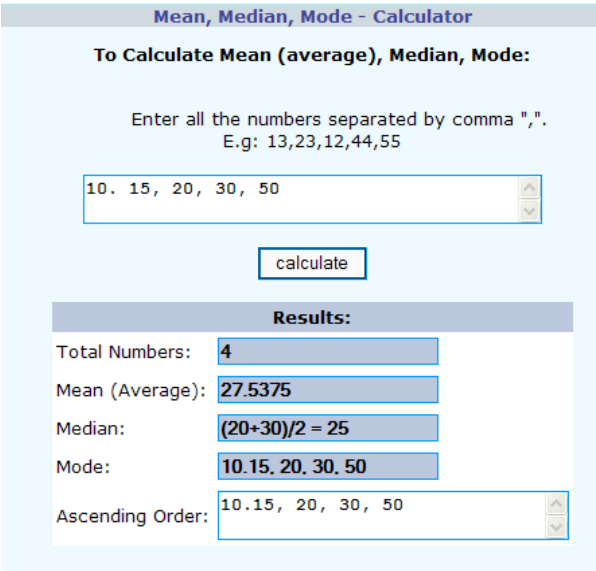
Website: http://www.graphpad.com/quickcalcs/index.cfm

This site provides tools for the following categories: categorical data, continuous data, statistical distributions and interpreting P values, random numbers and chemical and radiochemical data.

ii) Online Measures of Central Tendency Calculator

Website: http://easycalculation.com/statistics/mean-median-mode.php

This simple but very effective tool calculates the Mean, Median and Mode of a set of variables (refer to Chapter 11).



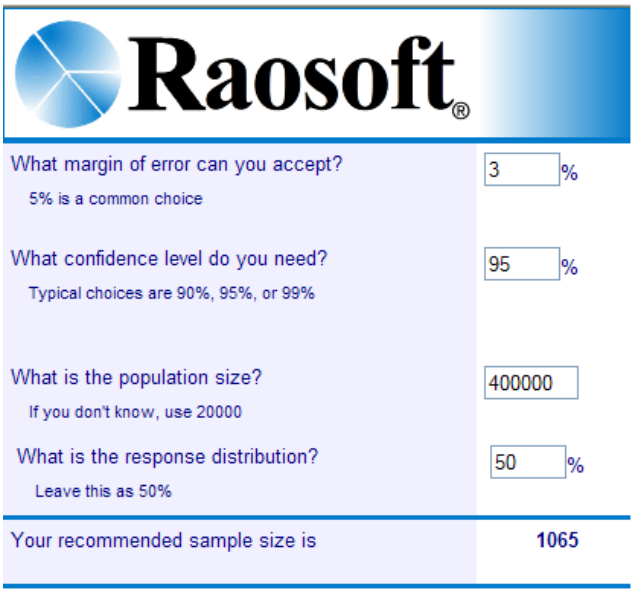
[49] http://sgems.sourceforge.net/
[50] http://ncg.nuim.ie/ncg/GWR/whatis.htm

iii) Sample Size Calculator

Website: http://www.raosoft.com/samplesize.html

This effective sampling size tool by RaoSoft is highly sought and has gone through an incarnation which tool calculates the sample size required for a population. The tool gives an example of what the Maltese islands sample population at 3% margin of error, a confidence level of 95% and a response distribution of 50%.

Interestingly the sample size stands at 1,065. Half the margin of error and the sample size quadruples to 4,224, half it again and the sample size quadruples again to a figure of 16,375.



Other online statistical tools can be found in the StaPages.org website: http://statpages.org/. One particularly interesting tool that can be investigated is StatCrunch[51] which allows online data analysis and even hosts and online mapping tool.

ii. Surveying tools – online

This final section on tools covers those ready-made online packages that help researchers, either free or at low cost to carry out their own online surveys with the tool even delivering either base statistical analysis or raw data for importation into statistical packages.

Other tools do not use ready-made online services but employ other desktop-based tools such as spreadsheets to export to an interactive online format, with replies being returned through email. A sample list of these services is covered below.

1. Survey Monkey

Survey Monkey[52] offers a reliable service, with a free basic service and a low-priced advanced service. The service offers pre-prepared templates hosting a considerable number of question types ranging from multiple choice to text boxes to demographics. Interestingly, Survey Monkey allows randomization and sorting of answers to ensure that each time the survey is run the question appears differently structures.

[51] http://www.statcrunch.com/
[52] http://www.surveymonkey.com/

2. Question Pro

Question Pro[53] provides a free service with an add-on priced account. The free version has quite a large number of features inclusive of multiple choice, Likert scale, open ended and essay open text, rank order, and template library amongst others.

3. FreeonlineSurveys

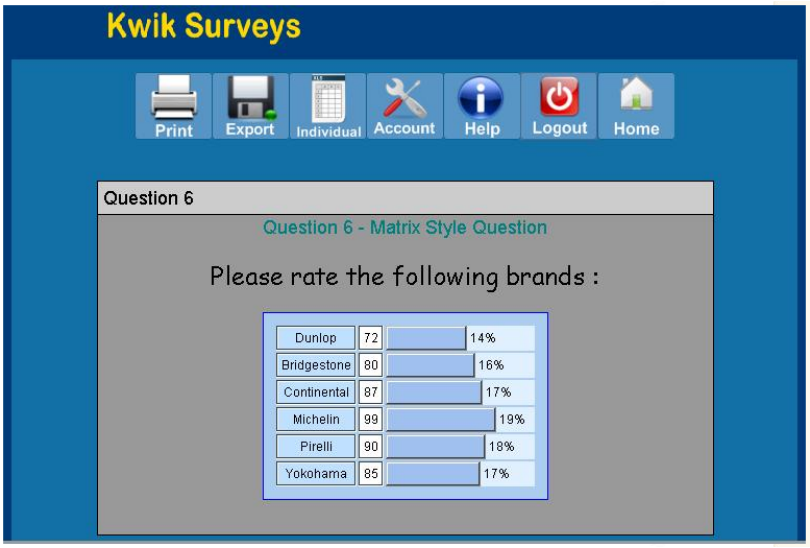
Freeonline Surveys[54] is another online survey which has the limited functionality as that of the other free services in addition to its priced services, but has an interesting output: that of the data appearing at the same time that it is inputted.

4. eSurveysPro

eSurveysPro[55] has a free online version as well as commercial options. It allows one to create free online unlimited surveys, questionnaires and responses as well as providing a survey editor and professional reporting.

5. Kwik Surveys

Advertised as a truly free online survey tool, Kwik Surveys[56] was created to serve as an opensource version of online surveys. It provides a veritable variety of options inclusive of different question types, (such as multiple choice, matrix, star rating and text input), randomization, logo upload, page skipping and other structuring features.



6. Creating one's own online survey

Tools exist that allow researchers to work through their data using proprietary software and then export to an online service for eventual data reporting and analysis. An example of such a tool is the Excel-Based Spreadsheet Converter[57] that takes an Excel survey or questionnaire and converts the relevant input cells into code for online input. The results of this tool are sent through email for eventual inputting. An example of such a survey is shown below. The highlighted colour indicates those cells that the respondent is asked to choose from.

---

[53] http://www.questionpro.com/
[54] http://freeonlinesurveys.com/
[55] http://www.esurveyspro.com/
[56] http://www.kwiksurveys.com/
[57] http://www.exceleverywhere.com/

Microsoft Excel Desktop Version

	A	B	C
1		**Survey: Investigating the use of the Internet**	
2			
3	1	Age:	
4			
5	2	What kinds of research studies would be worth funding in order to formulate internet policy?	
6			
7	3	Sex:	
8			
9	4	Occupation:	
10			
11	5	What is your general interest in Social Networks?	
12			
13	6	Do you have full open access to a PC with internet connection at home?	
14			
15	7	Did you make use of any of the government's Information technology internet incentives?	
16			
17	8	How many hours on average do you spend on the Internet daily?	
18			

Spreadsheet Converter Online Version

	Survey: Investigating the use of the Internet	
1	Age:	
2	What kinds of research studies would be worth funding in order to formulate internet policy?	18 yrs – 20yrs 21 yrs - 23yrs 24 yrs - 26 yrs 27 yrs – 29yrs 30 yrs +
3	Sex:	
4	Occupation:	
5	What is your general interest in Social Networks?	
6	Do you have full open access to a PC with internet connection at home?	
7	Did you make use of any of the government's Information technology internet incentives?	
8	How many hours on average do you spend on the Internet daily?	

In summary, this Chapter has given an overview of the tools available to researchers that are specifically related to statistics. Databases have not been considered as they will be covered in a separate chapter.

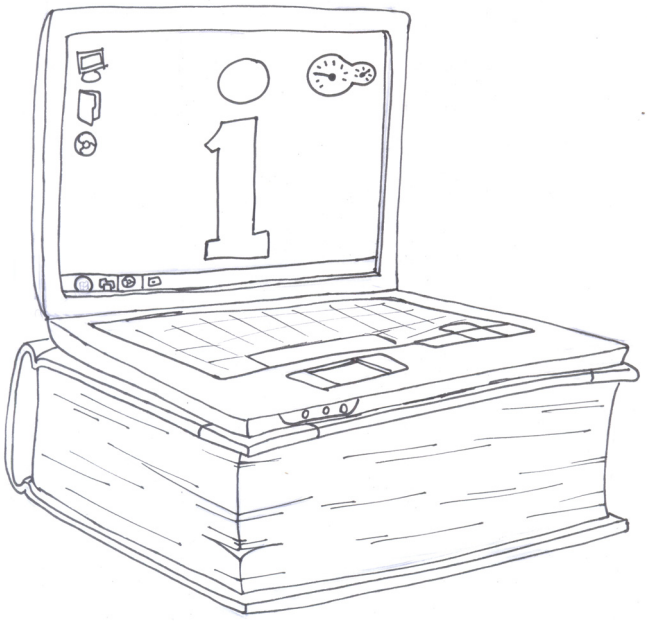
From spreadsheets to quantitative to qualitative and geostatistical tools, the researchers has a veritable sea of choices. The main issue is to keep to one or two tools in order to become deeply conversant with their functions. Also consider whether it is best to use desktop or online versions as these may have a relative impact on process, access and most importantly functionality.

Questions (refer to Appendix for the answers)

1. List the three main different categories of statistical tools.
2. What are “spreadsheets”?
3. What do “Macros” do?
4. What do the letters “SPSS” stand for and what is this?
5. What do the letters “SAS” stand for and what is this?
6. What is “Stata”? What does it do?
7. What is “MiniTab”? What does it do?
8. How would you briefly describe “R-Commander”?
9. What is “PSPP”?

10. What is “Gretl”?
11. What do the letters “CAQDAS”, “QDAS” and “QDA” stand for? What does this software do?
12. What do the letters “KWIC” stand for? What does this tool do?
13. What does the “Word Cruncher” do?
14. What is the “ArcGIS Geostatistical Analyst” and what does it do?
15. What does the “MapInfo Vertical Mapper” do?
16. What is “CrimeStat”? What does it do?
17. List the four main CrimeStat categories.
18. What do the letters “STAC” stand for? What is this and what does it do?
19. There are two types of online tools. Which are they?

Chapter 10
IT/IS and Databases



Research serves to make building stones out of stumbling blocks.

Arthur D. Little

Quoted in Massachusetts Institute of Technology, *Technology Review* (1932), **34**, 4.

Information Technology and Information Systems are two aspects that we have become used to and feel saturated with, in today's world. Add to that Information Communications Technology, Information Resources and other terms and we have given ourselves one nice mesh to entangle ourselves in!

Let us introduce the general organizational management entities before we move on to an important tool used for advanced analysis. Any organisation has a number of experts who carry out analysis across the different information streams. The terms generally used are IT, IS, ICT and IR. These terms are described below:

Information Technology (IT): This is the generic term for the process employed in developing software and managing hardware issues. Most people immediately link information technology with an IT department that runs an organisation's systems and network. This, it does but the term also includes the development, installation and maintenance of computer applications and systems (software and hardware).

Information Systems (IS): These are the conveyors or transporters of information in whatever mode, analogue (written – hardcopy) or digital. The main issue here concerns the systematic approach to managing that information, inclusive of the whole data cycle: design, gathering, cleaning, analysing, storing and disseminating. Information Technology has become integrated within Information Systems.

Information Communication Technology (ICT): The same as Information Technology but includes the integration of different networks such as telephony and computing into one system. More recent integration with video technology is also part of the process. ICT also refers to the strategies companies establish in order to set out their plans for IT investment and maintenance in an organisation.

Information Resources (IR): These comprise those assets or resources that an organisation holds in terms of data and information. These resources also include the human capital and skills gained in different information sciences such as geographical or spatial information systems (GIS), and the implementation of information technology within a socio-technic construct. Thus, IR takes into account the effects and impacts that the information exerts both within the organisation and in society.

To a certain extent, all the above cater for all the data requirements of an organisation and within it, the requirements for the DIKA structure. The DI is mainly catered for by the IT/ICT and IS units while the K and to a certain extent the A are catered for by the IR.

The scope of this chapter is not to describe what work is carried out by the above units. This chapter tries to describe one tool that is employed by the above in order to carry out detailed queries. This tool enables the researcher to do away with basic research tools (such as spreadsheets) and to concentrate on tools that allow linkages across the different research themes.

Moving from basic tools to databases

It is imperative to understand that apart from spreadsheets and specialised statistical software there are other applications which help us carry out research, particularly those that are based on the management of information and focuses on databases.

The intention is not to give an exhaustive discussion on databases but a brief outline of what they are and how they can be used to help researchers.

Databases

What is a database?

SearchSQLServer[1] define databases as "A database is a collection of information that is organized so that it can easily be accessed, managed, and updated. In one view, databases can be classified according to types of content: bibliographic, full-text, numeric, and images."

[1] http://searchsqlserver.techtarget.com/definition/database

The Computer Society[2] describes the process one should employ to manage data through mastering the power of computers; one that use the so-called 'database approach'.
Now that we have covered data and metadata in Chapter 6, it is easy to understand that a database can hold both data and metadata. It is also ideal for researchers in that the actual data is independent from the application using it, it can be shared, controls for error, integrates security issues and ensure that the data rules are maintained.

Should one have used a system that is dependent on a particular application, then one has to create data to fit the programme. Quite a laborious task and one that is not encouraged since the data attributes have to fit a programme and not the other way round. Imagine having to gather data only for persons' height and not for birth as that is the only attribute accepted by the programme! That was the old process employed in decades past and when data still formed a trickle. Now we have a data deluge and the data is paramount, not the tool: that can fit the purpose to accommodate the data!

However, one should not claim that there are no problems emanating from the current system, with researchers experiencing episodes of data redundancy (many copies), unknown versions being recorded with the consequence that researchers risk working on earlier versions instead of the latest versions – without realizing. Issues of data standards, consistency and security are some of the other problems envisaged in the current system. However, databases have major advantages in helping us understand what to do with the data at hand.

Databases are devices which facilitate our search for information within a collection of information, something that is not so easy to do with the so-called "flat tables" as are spreadsheets.

A very apt description but one can understand how it works if a comparison to a library is sought. Databases are accessed through what are termed Database Management Systems (DBMS) which are general purpose computer programmes aimed at making a database work.

If one compares a database to a library system, especially those card indexes, it is easy to understand how much work a database facilitates for the librarian. An analogue card index is composed of a series of drawers each sorted by a different attribute: one sorted by author surname, one by topic, one by accession number, and others. Now, that is one momentous task, especially for the larger libraries. A database holds only ONE structure and presents the user with a query system that can be based on any one of the drawers' resident index. The trick is in the linkages and in the electronic indexing.

One can state that a spreadsheet can do this, but the spreadsheet cannot link to any number of external datasets, whilst a database can. One can state that databases have access to a number of different tables (sort of spreadsheets) that can be compared with each other and retrieved separately while spreadsheets are more static.

What are the functions of a Database?

Primarily a database should be able to carry out a number of functions which include:

- Create, maintain and delete data structures inclusive of data definitions and file structures;
- Data importation;
- Edit data structures such as adding and deleting records;
- Allows searching for and extracting information from data;
- Establishes security protocols in terms of data security and maintenance and access management; and
- Includes a programming language.

Since the purpose of creating a database is very similar to the conceptual modeling to variable structuring process, there are very similar processes that must be followed. The main issue in database creation for an organisation or a project, concerns the establishment of a sturdy design process.

---

[2] http://www.bcs.org/

The process should ensure that the organisational requirements are listed, and the relationships between the different entities (variables or multiples of) and attributes (columns of data) are drafted into a conceptual data-model. The next step requires the move from the conceptual design into a physical structure. Following this one needs to test the theory and if it works implemented. It is important to maintain any system especially due to the requirements of the ever-changing structures in organizations and society.

There are different types of DBMS (Database Management Systems), amongst them the following:

- **RDBMS** – Relational Database Management System

The RDBMS is the most popular model used by commercial and open source systems. It was first discussed by Edgar Frank Codd (1923-2003). It is based on a relational model that links the different elements in a study through the relationships between the attributes in each element. Both the data and the relationships are stored in the database.

The Relational Model which allows for linkages across different datasets and tables and which would fit ideally in our study of social data structures.

- **OODBMS** – Object-Oriented Database Management System

The OODBMS is a niche-market database that builds its model around objects, each of which has its own set of attributes and behaviours that have complex information built into them. An example of an object would be a 12-storey lift with its own commands that state that a lift can only climb up to the 12th storey and no more. Thus, in an object oriented world, the lift would not fit into a 4 storey building as the object is too large to fit within that '4 storey world'. If it is fitted in a 14 storey building then it only reaches to the 12th and has limited functionality. These objects need specific software for their inclusion in a system.

- **ORDBMS** – Object-Relational Database Management System

This DBMS is similar to the relational database but incorporates an object-oriented database model. It is the so-called middle ground between RDBMS and OODBMS since the latter was not really taken up by mainstream society. The inherent OO structures have been integrated and can be accessed through a query language as in a relational database.

Terminologies

It is best at this stage to outline the differences between the different terms relegated to the different formats. Table 10.1 describes the names given to the different elements comprising a table. The purpose of the table serves to ease a researcher's knowledge of terms once a study outlines information based on the different terminologies is used.

As an example a Row on paper is termed a Record/Case in a dataset file and a Tuple in a relational database.

Table 10.1: Terminologies

Paper	File	Relational Database
Table	File	Relation
Row	Record/Case	Tuple
Column	Field	Attribute
No. of Columns	No. of Fields	Degree
No. of Rows	No. of Cases	Cardinality
	Unique Identifier	Primary Key
	Possible Values	Domain

Source: (Reeve, 1997)

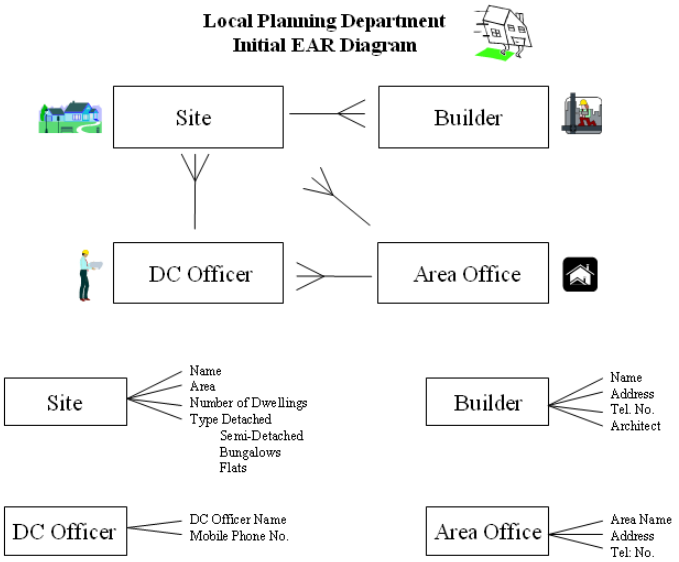
From a Mind to an EAR

As a short example of how a relational database works, visualize a situation which 'employs' a development case officer in an urban development organisation (local planning department) who is required to manage a number of sites.

If one brainstorms the situation in the form described earlier in the mind mapping chapter, then it is fairly easy to come up with a series of elements. In our example, the case officer's world is composed of the site/s, the builder, the area office and our friend the officer. Each of these categories is called an element.

Each element would have a number of components that comprise that same element, for example the site element, is comprised of name, area, number of dwellings and dwelling type (such as terraced, semi-detached, bungalows and flats). The same is done for all the other elements as is shown in Figure 10.1 below.

Figure 10.1: Initial EAR Diagram



Source: (Reeve, 1997)

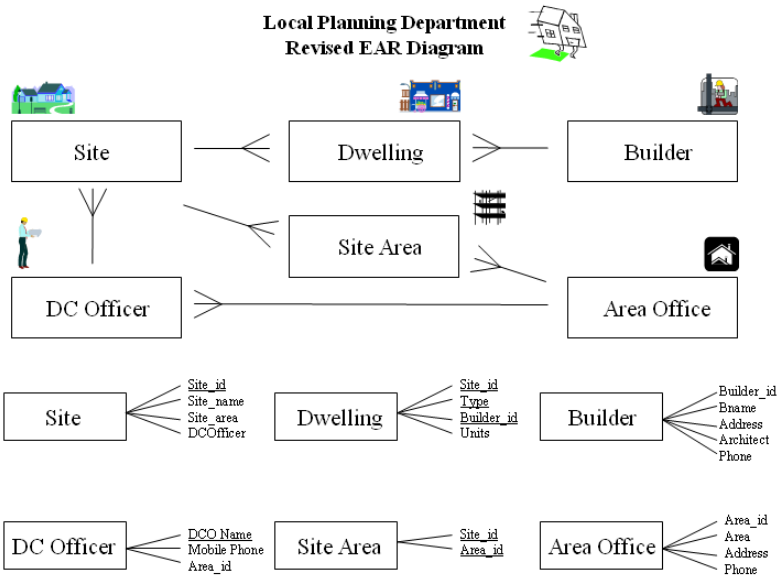
Isn't this very similar to the mind map? It is, and though there is little space for 'fuzziness' in database design as each relationship must be established prior to the designing stage, the concept is similar. The only difference is that the mind map here is called an **EAR Diagram**: an **E**ntity **A**ttribute **R**elationship Diagram! Very easy to remember too from Mind to Ear!

The next step would be to identify the relationships between the elements. There are various ways to do this. Let us take the chicken footprints (i.e. the lines that resemble chicken footprints!) that have already been depicted in the image above. Those lines do mean something and immediately one can tell that the single point at one end and three at another refer to some kind of code. The points refer to the number of relationships: a straight line with a single point at each end means a 1:1 relationship (example a case officer can only have one mobile number), a single point at one end and a multiple point at the other means that that is a 1:Many relationship (one person can have many building sites to manage) and a multiple point to multiple point refers to a Many:Many relationship (example many architects could be constructing many buildings).

The figure above depicts a situation of 1: Many. One can read it as follows: 1 case officer can manage many sites, each site can host many dwellings, each site can host many site areas (plots of land), whilst each builder can be building many dwellings. These and many other relationships can be identified at the first run of the EAR diagram.

Now, let us take this a step further and identify which are those attributes that are common in the different elements. As an example the attribute Site_id is found in both the Site element and the Dwelling element, thus a relationship is established through this attribute which is now called a Primary Key. The attributes are best highlighted using an underline. The Revised EAR Diagram below (Figure 10.2) thus allows the researcher to identify the linkages better.

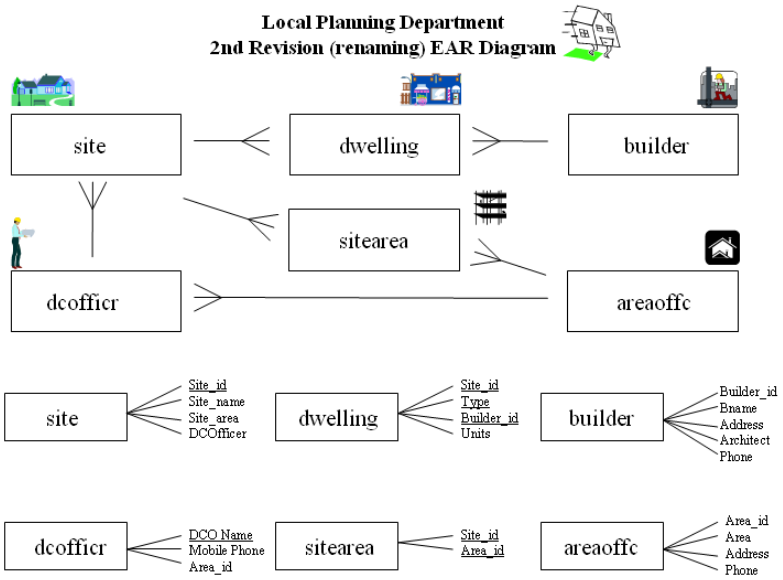
Figure 10.2: Revised EAR Diagram



Source: (Reeve, 1997)

The final step is to rename the elements into words that are readable to a database system through the removal of capitals, removal of spaces and shortening of names as per Figure 10.3.

Figure 10.3: 2nd Revised EAR Diagram



Source: (Reeve, 1997)

Queries can then be run once the database has been populated by the raw data as described in earlier chapters.

Distributed Databases

One of the advantages of using databases is that the same database could have its components distributed over a number of computers. In turn, these components could be distributed anywhere on the globe. In effect, this brings about issues of access and security.

Let us consider a situation where a database on migration has been created to monitor population movements and the relative health, economic and transport parameters. A dataset created by a researcher can bring up the following themes during the initial conceptualization process:

i) Population by country;
ii) Transport routes;
iii) Human development index;
iv) Health;
v) Famine and drought; and
vi) Conflict.

This initial brainstorming session immediately brings to mind the issue of access. Probably all the above data is held in different data bases. Therefore the first question to ask is "What databases exist?". This is quickly followed by "How can I access them?" If you can gain access to them your next query would be if it is possible to gather ALL the databases and rebuild them into one structure? However it might be easier to identify the common attributes within each database and attempt to link those two together.

The task of integrating all data together is a massive one and not highly recommended for various reasons. A data can become obsolete if not updated regularly to reflect changes in the data. Secondly, organisations may not be able to deliver the complete data.

One solution would be to link the databases through a distributed database approach where the data are linked through one or more variables . This allows for data linkages on only those variables that are of relevance to the researcher. The issue to consider at this stage concerns that of access. Access to the attributes may require the resolution of security issues but such is handled by the Information Technology and Systems personnel.

Let us take a look at how the data linkages could be set up:

i. First structure the links into the main theme, the database topic, (the variable) one is going to use and the source as per table below;
ii. Then acquire access through a series of protocols and agreements between organisations;
iii. Set up the new database and ensure that the linkages work;
iv. Create a query tool based on the mind map created as part of the process;
v. Run the relevant queries.

Theme	**Database Topic**	**Variable**	**Source**
Population by country	National Population	Total population	Population Division of the Department of Economic and Social Affairs of the United Nations http://esa.un.org/UNPP/
Transport routes	Main Shipping Routes	Vessel movements	AIS live vessel tracking and movements http://www.marinetraffic.com/ais/
Human development index	Human Development Reports	Human development index	UNDP http://hdr.undp.org/en/statistics/data/

Theme	Database Topic	Variable	Source
Health	WHO Statistical Information System (WHOSIS)	Disease and de-population	WHO http://www.who.int/whosis/en/
Famine and drought	FEWS	Migration due to food insecurity	UN/FAO http://www.fews.net/Pages/default.aspx
Conflict	UCDP/PRIO Armed Conflict Dataset	Persons displaced by armed conflict	Armed conflict database http://www.prio.no/CSCW/Datasets/Armed-Conflict/UCDP-PRIO

Distributed databases are seen as the future for research processing, though there is still a lot of work to be done, which would ensure that all security issues have been covered and that there is no risk of dataset contamination or data theft.

In summary, a distributed database, whilst still having a central database management system, is composed of many structures such as different computers in different physical locations. However, an extensive system of controlled data management (which is based on the principles of replication and duplication), ensures that there is no data loss. New items are updated in each attribute and duplication ensures that the master database is copied and that records of the structures are kept.

Querying language: SQL

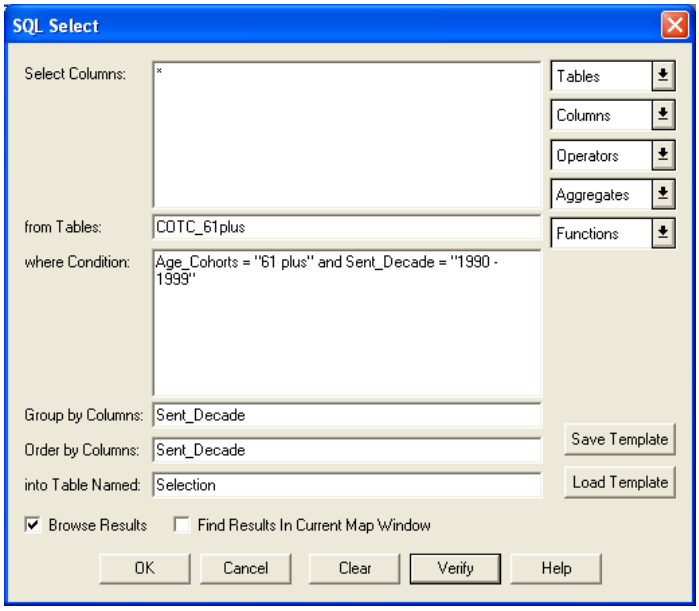
How does one query a database? There are many languages ensuring such a process but the most reliable tool is known as Structured Query Language, better known as (SQL and pronounced also as Sequel). Though not part of the relational model it has become embedded in it as a tool to manipulate and query the data within a database. SQL was the language introduced by Codd (1970) in his relational database launch.

We use it to analyse data across different variables. For example: how many elderly women in care, were transferred to a cure establishment in the 1990s. This language is very simple to use and has less than 30 commands, mostly written in pseudo English, such as ‘create’, ‘order by’ and ‘select’.

SQL allows researchers to carry out most queries based on the W6H described in earlier chapters as it filters the attributes for data that falls within the respective structures as outlined in the commands it was given. In fact, SQL queries also contain spatial commands to help researchers carry out such commands as “how many dwellings can be found within 1km of the centre of a town”.

Figure 10.4 depicts a SQL query that is requesting information on the number of elderly aged 61 plus (attribute name = Age_Cohorts) who were given access to a service (Sent_Decade) between 1990 and 1999.

Figure 10.4: SQL Query



a. Database Tools

There are various database tools available for researchers in both the commercial and open source domain. Some commercial, such as Microsoft Office Access, and others have been developed as open source such as PostGreSQL.

1. Microsoft Office Access

Microsoft Office Access[3] is a commercial tool. It is based on a relational structure and forms part of a suite of applications targeted for office use. It is based on a particular data storage system employing the Access Jet Database Engine and can import or access data in other databases and applications.

2. PostGreSQL

PostGreSQL[4] is a freeware SQL database which has been developed on the ORDBMS model. It is a multiplatform solution. The system is supported by many third-party GUI tools as can be found under the site: http://wiki.postgresql.org/wiki/Community_Guide_to_PostgreSQL_GUI_Tools

Comprehensive lists of commercial and free database tools can be found at Wikipedia's http://en.wikipedia.org/wiki/Comparison_of_database_tools and at Freebyte's http://www.freebyte.com/programming/database/#opensourcedatabases respectively.

Questions (refer to Appendix for the answers)

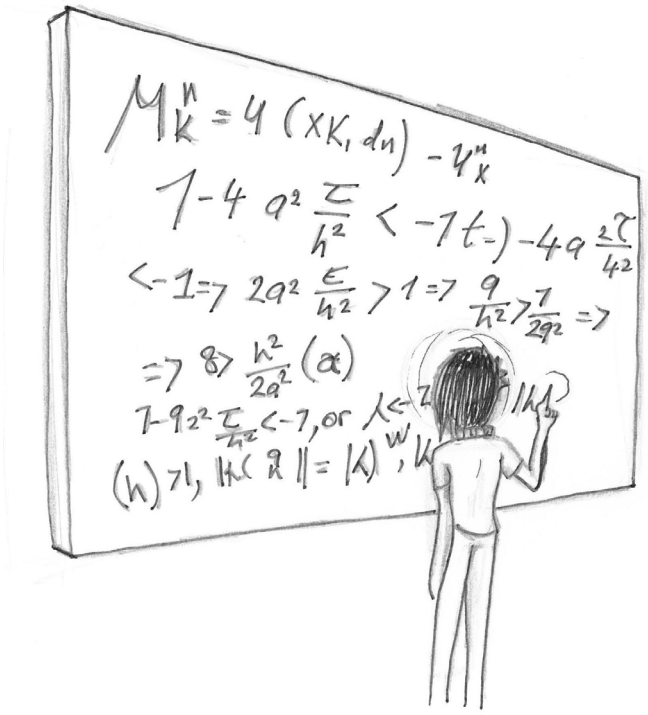
1. List the four main general organizational management entities and very briefly describe them.
2. What is a database and why is it ideal for researchers?
3. What are the problems faced today by researchers?

---

[3] http://office.microsoft.com/en-us/access/
[4] http://www.postgresql.org/

4. What are Database Management Systems (DBMS)?
5. Mention one major way in which a database differs from a spreadsheet.
6. List the main functions of a database.
7. What is the main issue in database creation?
8. Briefly describe how one could establish a sturdy design process.
9. Mention three different types of DBMS.
10. What is an EAR Diagram?
11. List the main steps one needs to take when using an EAR Diagram.
12. Mention one major advantage of using databases as well as the two major issues that it gives rise to.
13. Why is it not a good idea to integrate many datasets together?
14. How can you avoid integrating many datasets together? What could constitute an acceptable solution?
15. List the steps one needs to take to set up the dataset linkages.
16. What is SQL?
17. What is Microsoft Office Access?
18. What is PostGreSQL?

Chapter 11
Statistical Testing



The resolution of revolutions is selection by conflict within the scientific community of the fittest way to practice future science. The net result of a sequence of such revolutionary selections, separated by periods of normal research, is the wonderfully adapted set of instruments we call modern scientific knowledge.

Thomas S. Kuhn

The Structure of Scientific Revolutions (1962), 171.

Any publication on research and statistics cannot exist without a review of the most commonly used statistical tests. The scope here is not to exhaust all statistical testing. There are more than enough books for that dealing with specific disciplines. The scope is to give readers an idea of the basic tests that can be used in their research studies.

One can use this publication in conjunction with specialised publications as per lists below. Always check for new publications as each tackles new methods and case studies. There are two categories of book listed here. The first cover Generic Statistics Publications sorted by Publication Date whilst the second is more targeted. It is thematic in scope and lists publications that are sorted by theme such as: behavioural sciences, criminology, business, and others.

E-Books are also available from the following websites:

HyperStat
http://davidmlane.com/hyperstat/

Probability & Statistics
http://www.e-booksdirectory.com/listing.php?category=15

FreeBookCentre.Net
http://www.freebookcentre.net/SpecialCat/Free-Statistics-Books-Download.html

Generic Statistics Publications sorted by Publication Date

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Introductory Statistics, 7th edition	Prem S. Mann	2010	Wiley	0470444665	978-0470444665
Statistics Unplugged, 3rd edition	Sally Caldwell	2009	Wadsworth Publishing	0495602183	978-0495602187
Statistics for People Who (Think They) Hate Statistics: Excel 2007 Edition, 2nd edition	Dr. Neil J. Salkind (Editor)	2009	Sage Publications , Inc	1412971020	978-1412971027
Elementary Statistics: Picturing the World, 4th edition	Ron Larson and Betsy Farber	2008	Prentice Hall	0132424339	978-0132424332
Statistics: The Art and Science of Learning from Data, 2nd Edition	Alan Agresti and Christine Franklin	2008	Prentice Hall	0135131995	978-0135131992
Elementary Statistics: A Step By Step Approach, 7th edition	Allan Bluman	2008	McGraw-Hill Science/ Engineering /Math	0077302354	978-0077302351
Intro Stats, 3rd Edition	Richard D. De Veaux, Paul F. Velleman and David E. Bock	2008	Addison Wesley	0321500458	978-0321500458
Introduction to Statistics and Data Analysis, 3rd edition	Roxy Peck, Chris Olsen and Jay L. Devore	2008	Duxbury Press	0495557838	978-0495557838
Mathematical Statistics with Applications, 7th edition	Dennis Wackerly , William Mendenhall and Richard L. Scheaffer	2007	Duxbury Press	0495110817	978-0495110811

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Statistics, 4th edition	David Freedman, Robert Pisani and Roger Purves	2007	W. W. Norton & Company	0393929728	978-0393929720
Elementary Statistics (10th Edition) (MyStatLab Series), 10th edition	Mario F. Triola	2007	Addison Wesley	0321331834	978-0321331830
Statistics For The Terrified, 4th edition	Gerald Kranzler, Janet Moursund and John H. Kranzler	2006	Prentice Hall	0131930117	978-0131930117
Discovering Statistics Using SPSS (Introducing Statistical Methods S.), 2nd edition	Andy Field	2005	Sage Publications Ltd	0761944524	978-0761944522
Fundamentals of Statistics	Michael III Sullivan	2004	Prentice Hall	0131464493	978-0131464490
The Basic Practice of Statistics, 3rd edition	David S. Moore	2003	W.H. Freeman & Company	0716796236	978-0716796237
Statistics for Dummies	Deborah Rumsey	2003	For Dummies	0764554239	978-0764554230
The Craft of Research, 2nd edition (Chicago Guides to Writing, Editing, and Publishing)	Wayne C. Booth, Joseph M. Williams and Gregory G. Colomb	2003	University Of Chicago Press	0226065685	978-0226065687
Statistics Without Tears: A Primer for Non-Mathematicians (Allyn & Bacon Classics Edition)	Derek Rowntree	2003	Allyn & Bacon	0205395090	978-0205395095
The Visual Display of Quantitative Information, 2nd edition	Edward R. Tufte	2001	Graphics Press	0961392142	978-0961392147
Your Statistical Consultant: Answers to Your Data Analysis Questions	Dr. Rae R. Newton and Dr. Kjell E. (Erik) Rudestam	1999	Sage Publications , Inc	0803958234	978-0803958234
How to Lie with Statistics	Darrell Huff and Irving Geis	1993	W. W. Norton & Company	0393310728	978-0393310726
Cartoon Guide to Statistics	Larry Gonick and Woollcott Smith	1993	Collins Reference	0062731025	978-0062731029

Thematic Publications sorted by Theme

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Essentials of Statistics for the **Behavioral Sciences**, 7th edition	Frederick J Gravetter, Larry B. Wallnau and Jon-David Hague	2010	Wadsworth Publishing	049581220X	978-0495812203
Statistics for the Behavioral Sciences, 8th edition	Frederick J Gravetter and Larry B. Wallnau	2008	Wadsworth Publishing	0495602205	978-0495602200
Comprehending Behavioral Statistics (with CD-ROM), 4th edition	Russell T. Hurlburt	2005	Wadsworth Publishing	053460627X	978-0534606275
Applied Statistics for the Behavioral Sciences, 5th edition	Dennis E. Hinkle, William Wiersma and Stephen G. Jurs	2002	Wadsworth Publishing	0618124055	978-0618124053
Statistics for **Criminal Justice and Criminology**, 3rd edition	Dean J. Champion and Richard D. Hartley	2009	Prentice Hall	0136135854	978-0136135852
Simple Statistics: Applications in Criminology and Criminal Justice	Terance D. Miethe	2006	Oxford University Press, USA	0195330714	978-0195330717
Research Methods for Criminology and Criminal Justice: A Primer (Criminal Justice Illuminated), 2nd edition	Dantzker, M.L., and Hunter, R.D.	2005	Jones & Bartlett Pub	0763736155	978-0763736156
Statistics for **Business and Economics** (with Bind-In Card), 11th edition	David R. Anderson , Dennis J. Sweeney and Thomas A. Williams	2010	South-Western College Pub	0324783248	978-0324783247
The Practice of Business Statistics w/CD, 2nd edition	David S. Moore, George P. McCabe, William M. Duckworth and Layth Alwan	2008	W. H. Freeman	142922150X	978-1429221504
Statistical Techniques in Business and Economics with Student CD, 13th edition	Douglas Lind , William Marchal and Samuel Wathen	2006	McGraw-Hill/Irwin	0073272965	978-0073272962
Research Methods in **Public Administration** and Nonprofit Management: Quantitative and Qualitative Approaches, 2nd edition	David E. McNabb	2008	M.E. Sharpe	0765617676	978-0765617675
Essential Statistics For Public Managers and Policy Analysts, 2nd edition	Evan M Berman	2006	CQ Press	0872893014	978-0872893016
An SPSS Companion to **Political** Analysis, 3rd edition	Philip H. Pollock III	2008	CQ Press	0872896072	978-0872896079
Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists	Joel Best	2001	University of California Press	0520219783	978-0520219786
Statistics: A Tool for **Social** Research, 8th edition	Joseph F. Healey	2008	Wadsworth Publishing	0495096555	978-0495096559
Statistics Explained: A Guide for Social Science Students, 2nd edition	Perry Hinton	2004	Routledge	0415332850	978-0415332859
Statistics for Social Data Analysis, 4th edition	David Knoke, George W. Bohrnstedt and Alisa Potter Mee	2002	Wadsworth Publishing	0875814484	978-0875814483

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Basic Statistics for **Social Workers**, revised edition	Robert A. Schneider	2010	University Press of America	0761849327	978-0761849322
Statistics for Social Workers, 8th edition	Robert W. Weinbach and Richard M. Grinnell	2009	Prentice Hall	0205739873	978-0205739875
Handbook of Research on Civic Engagement in **Youth**	Lonnie R. Sherrod, Judith Torney-Purta and Constance A. Flanagan	2010	Wiley	0470522747	978-0470522745
Quantitative Research in **Education**: A Primer	Wayne K. (Kolter) Hoy	2009	Sage Publications, Inc	1412973260	978-1412973267
Study Guide for Essentials of **Nursing** Research: Appraising Evidence for Nursing Practice, 7th edition	Denise F Polit and Cheryl Tatano Beck	2009	Lippincott Williams & Wilkins	0781785812	978-0781785815
Applied Spatial Statistics for Public Health Data	Lance A. Waller and Carol A. Gotway	2004	Wiley-Interscience	0471387711	978-0471387718
Statistics for **Psychology**, 4th edition	Arthur Aron, Elaine N. Aron and Elliot Coups	2005	Prentice Hall	0131931679	978-0131931671
Understanding Research Methods and Statistics: An Integrated Introduction for Psychology, 2nd edition	Gary Heiman	2000	Wadsworth Publishing	0618043047	978-0618043040
Planning, **Construction**, and Statistical Analysis of Comparative Experiments (Wiley Series in Probability and Statistics)	Francis G. Giesbrecht and Marcia L. Gumpertz	2004	Wiley-Interscience	0471213950	978-0471213956
Elementary Statistics for **Geographers**, 3rd Edition	James E. Burt, Gerald M. Barber and David L. Rigby	2009	The Guilford Press	1572304847	978-1572304840
Practical Statistics for **Environmental** and **Biological** Scientists	John Townend	2002	Wiley	0471496650	978-0471496656
Using Statistics to Understand the Environment (Routledge Introductions to Environment)	Penny A. Cook and C. Phillip Wheater	2000	Routledge	0415198887	978-0415198882
Statistics for the Environment, Pollution Assessment and Control, Volume 3, 3rd edition	Vic Barnett and K. Feridun Turkman (Editors)	1997	Wiley	0471964352	978-0471964353
Environmental Statistics and Data Analysis	Wayne R. Ott	1995	CRC-Press	0873718488	978-0873718486
Environmental Statistics, Assessment, and Forecasting	C. Richard Cothern and N. Phillip Ross	1993	Lewis Publishers	0873719360	978-0873719360
Introduction to **Engineering** Statistics and Six Sigma: Statistical Quality Control and Design of Experiments and Systems	Theodore T. Allen	2006	Springer	1852339551	978-1852339555
Applied Statistics for **Marine** Affairs Professionals	Niels West	1996	Praeger	0275951723	978-0275951726

Title	Author/s	Publication Date	Publisher	ISBN-10	ISBN-13
Handbook of **Spatial** Statistics (Chapman & Hall/CRC Handbooks of Modern Statistical Methods)	Alan E. Gelfand, Peter Diggle, Peter Guttorp and Montserrat Fuentes	2010	CRC Press	1420072870	978-1420072877
Spatial Statistics and Modeling (Springer Series in Statistics)	Carlo Gaetan and Xavier Guyon	2009	Springer	0387922563	978-0387922560
Applied Spatial Data Analysis with R (Use R)	Roger S. Bivand , Edzer J. Pebesma and Virgilio Gómez-Rubio	2008	Springer	0387781706	978-0387781709
Statistical Methods for Spatial Data Analysis (Chapman & Hall/CRC Texts in Statistical Science)	Oliver Schabenberger and Carol A. Gotway	2004	Chapman and Hall/CRC	1584883227	978-1584883227
Spatial Statistics through Applications (Advances in Ecological Sciences)	J. Mateu and F. Montes (Editors)	2002	WIT Press / Computational Mechanics	1853126497	978-1853126499
Statistics for Spatial Data (Wiley Series in Probability and Statistics), revised sub edition	Noel A. C. Cressie	1993	Wiley-Interscience	0471002550	978-0471002550

Statistical testing helps researchers to control and validate the analysis carried out in their studies. These tests ensure that errors are not committed during the course of an analytical process. In addition, one should also be able to identify the quantity of errors generated.

There are many tools available for research, some are simple whilst other are quite complex and require increasing levels of tests to ensure precision and accuracy. This chapter will cover a few of the simplest tests ranging from measures of central tendency to regression analysis.

Before reviewing some basic statistics, it is best to define four words that are mostly used in statistical analysis: Descriptive and Inferential Statistics as well as Independent and Dependent Variables.

Descriptive Statistics are used to describe a dataset quantitatively through summarization rather than through the usage of probability analysis. Examples of descriptive statistics include the measures of central tendency (Mean, Median and Mode), standard deviation and variance.

Inferential Statistics, also called inductive statistics, on the other hand, employ probability tests through comparative tests that allow one to infer on a population. Inferential tests include the Z-Score, the T-Tests, the ANOVA and the Chi squared. When researchers present their data employing mainly the inferential test, the submission of descriptive statistics is still deemed necessary as they enhance any study and aid in the understanding of the results.

Basic Statistics

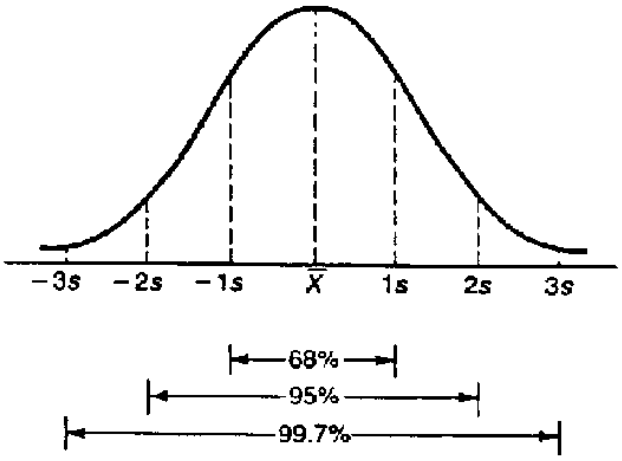
In statistics a basic assumption is made, which revolves around the issue that a set of data has a **Normal** or **Gaussian** distribution. Behavioral sciences' statistical tests assume a normal distribution. These tests can be used even where the distribution is only approximately normal.

Normal distributions are consonant with values that are more concentrated in the middle of the distribution curve than in the tails. They are defined by two parameters: the mean (μ) and the standard deviation (σ).

The Normal assumes that the data has very specific ranges within which that data falls, abiding by the **Empirical Rule** which states that in a **normal distribution:**

- about 68% of the scores are within one standard deviation of the mean ($\mu \pm \sigma$) and

- about 95% of the scores are within two standard deviations of the mean ($\mu \pm 2\sigma$) and
- about 99% are within three standard deviations of the mean ($\mu \pm 3\sigma$)



1. Measures of Central Tendency

As you would recall in Chapter 9 we mention the measures of central tendency. To recapitulate these measures refer to the values that are either at the middle point of a set of data or are typical of that type of data. There are three measures of Central Tendency: the Mean, Median and Mode.

1. **Mean** or "**average**" value: This value computes the central tendency of a frequency distribution ex **Interval** / **Ratio** data;

2. **Median** or **middle** value: Appropriate measure of central tendency for **Ordinal** level data;

3. **Mode** or **most frequent value**: Providing the least precise information about central tendency for **Nominal** (categorical) data.

Now let us go on to working out the Measures of Central Tendency.

1. The Mean

The Mean is the score located at the mathematical centre of a distribution and represents the arithmetic mean which is also called the average: The mean is calculated as the sum of all the scores divided by the number of scores.

The Greek letter Σ (a capital sigma) is used to designate summation.

The Mean Formula

Mean = sum of elements / number of elements
= a1+a2+a3+.....+an/n

Note that sometimes it is difficult to calculate the mean of a whole population as that would take forever, thus sometimes it is best to use a sample and calculate the mean for that.

The table below gives the formulas for both the whole population and a sample population. In the following examples we are assuming that the sample population is being analysed.

	Sample	Population
Mean	$\bar{X}$ (X-BAR)	μ (mu)
Variable	X	X
Add up all the Scores	ΣX	ΣX
No of Scores	N	N
Formula	$\bar{x} = \frac{\sum x_i}{n}$	$\mu = \frac{\sum x}{N}$

Each section contains a few working examples. Kindly attempt to analyse the figures based on the formula given for each measure.

- Mean: Calculations
    - Calculate the Mean for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000

Worked Example: Q1 - Mean

Formula =

$$\bar{x} = \frac{\sum x_i}{n}$$

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Add all the scores

Σ^x = 15 + 3 + 48 + 23 + 8 + 18 + 6 + 19 + 54

Σ^x = 194

Step 2: Count the number of Scores (**N**) = **9**

Step 3: Divide the Sum by the Count

$\bar{x} = \Sigma^x / N$

$\bar{x}$ = 194 / 9

$\bar{x}$ = 21.56

Mean = 21.56

- Answers

Q1 21.56
Q2 31.90
Q3 90914.10 – should we use this or another measure since the figure 1000000 is such a large outlier?

Note that should one outcome be registered as far from the rest of the data, this number is called an outlier as in Q3 above, which would strongly affect the data. One can use an alternate measure called the median.

2. The Median

The Median refers to the score located at the 50th percentile. The median allows researchers to identify that middle value which serves as a divider between the two halves of a dataset. Thus, Median is the middle score.

Symbol is M or Mdn

The Median Formula

Median = position of the value
= number of elements / 2

= (N + 1)/2

There are various sequential steps to calculate the Median:

1. Sort the observations smallest to largest;
2. Compute (n + 1)/2. This gives the *position* of the median (not the median itself) in the ordered data set;
3. Then find the corresponding number in the ordered set;
4. Median = number of units plus 1/2
   = (n + 1)/2 = (5 + 1)/2 = 6/2 = 3;
5. What happens when there are two elements in the middle? (occurs in even number of elements) – both middle ones are chosen.

- Median: Calculations

  - Calculate the Median for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000
Q4 1, 2, 3, 4, 5, 6

Worked Example: Q1 - Median

Formula: Mdn = (n + 1)/2

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Sort the scores

3, 6, 8, 15, 18, 19, 23, 48, 54

Step 2: Add the number of scores to 1 and divide by 2 (where **n** = **9**)

Mdn = (n + 1)/2

Mdn = (9 + 1)/2

Mdn = (10)/2

Mdn = 5 (the fifth score)

Step 3: Check which score is in the fifth position

3, 6, 8, 15, 18, 19, 23, 48, 54

3, 6, 8, 15, 18, 19, 23, 48, 54

Median = 18

- Median: Calculations
    - Answers

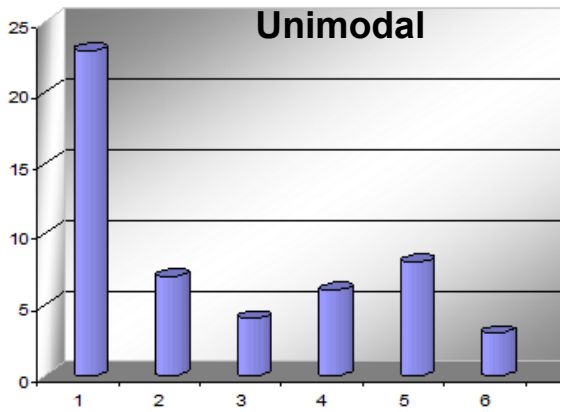
Q1 18
Q2 20
Q3 6
Q4 (3+4)/2 = 3.5

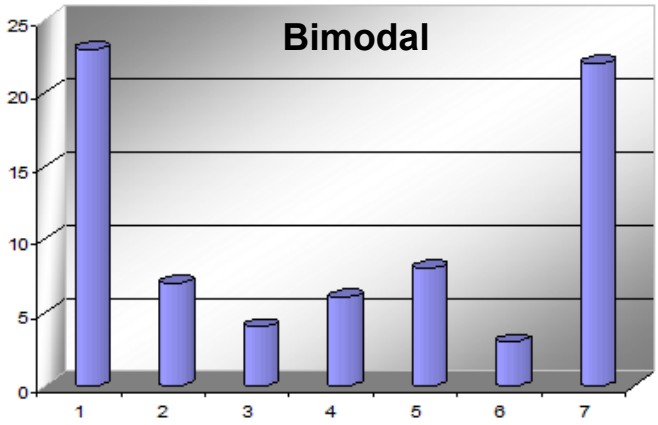
(actually 3rd and 4th Place – the mean of those can be calculated)

3. Mode

The Mode refers to the score that occurs most frequently: example there are more females than males in the elderly age cohorts. There are two types of Mode: the Unimodal and the Bimodal.

- The Unimodal type has one peak (highest point in a distribution – indicating the most frequent score)
- The Bimodal type has two peaks (highest point in a distribution – indicating the most frequent score)





There are various sequential steps to calculate the Median:

- Run a test in Microsoft Excel or another spreadsheet;
- Then run a pivot table
- The element with the highest Count renders the Mode

- Mode: Calculations

  - Calculate the Mode for the following:

Q1 15, 3, 3, 23, 8, 8, 6, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 1, 1, 15, 15, 1
Q3 14, 18, 14, 1, 1, 15, 15, 1, 15

Worked Example: Q1 - Mode

Formula: The score with the largest number of instances

Q1 15, 3, 3, 23, 8, 8, 6, 48, 23, 8, 18, 6, 19, 54

Step 1: Sort the scores

3, 3, 6, 6, 8, 8, 8, 15, 18, 19, 48, 54

Step 2: Check how many instances there are for each score

3	**6**	**8**	**15**	**18**	**19**	**48**	**54**
2	**2**	**3**	**1**	**1**	**1**	**1**	**1**

Step 3: Check which score has the largest number of instances

3	**6**	**8**	**15**	**18**	**19**	**48**	**54**
2	**2**	**3**	**1**	**1**	**1**	**1**	**1**

8 has 3 instances

Mode = 8

(since the mode falls on a score and not between scores, it is termed unimodal)

- Mode : Calculations

    - Answers

Q1 8 – unimodal
Q2 1 – unimodal
Q3 1, 15 – bimodal

Proportions and Percentages

These measures are heavily dependent on the issue of proportion. Such refers to the degree that an attribute is found within a population. One can calculate this through defining whether one needs to depict the degree as a fraction or as a percentage.

- **Proportions**

Proportions refer to fractions of the total. As an example one can state that the fraction of Maltese who have brown hair (300,000 of 400,000) equates to a proportion of ¾ or 0.75.

300,000/400,000 = ¾ or 0.75

- **Percentages**

Percentages refer to the same method as proportions but expressed as a figure out of 100. In effect:
Percentage = proportion * 100

Therefore the example above results in a figure of 75 percent:

0.75 * 100 = 75%

2. Measures of Variability

The next step to understand concerns the fact that numbers are rarely found aggregated around a single figure such as the improbable state where all the population of Valletta is aged 35. Since we study real life populations our study group rarely falls under the same number. In fact, all populations range from 0 to 120 in extreme cases. All start at Day 1 and have a somewhat different end Date!

How does one calculate for such a variation in numbers? There must be hundreds of hundreds of thousands individuals in a population which we cannot calculate individually! This is where measures of Variability or of Dispersion come in. There are three parameters that help in understanding such variability: the **Range**, the **Standard Deviation** and the **Variance**.

The next steps analyse each of the parameters in depth.

Each section contains a few working examples. Kindly attempt to analyse the figures based on the formula given for each measure.

1. **Min – Max – Range**

Whilst the first parameters refer to the Range, it is best to understand Range through its components. The Range is defined as the difference between the two extremes in the data range: the **Minimum** and **Maximum**.

Range = Maximum - Minimum

A. The Minimum (Min)

The Minimum (Min) refers to the smallest number in the dataset.

- Min Calculations

- Calculate the Min for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000
Q4 1, 2, 3, 4, 5, 6

- Min Calculations

  - Answers

Q1 3
Q2 16.5
Q3 1
Q4 1

B. The Maximum (Max)

The Maximum (Max) refers to the largest number in the dataset.

- Max Calculations

  - Calculate the Max for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000
Q4 1, 2, 3, 4, 5, 6

- Max Calculations

  - Answers

Q1 54
Q2 88.88
Q3 1000000
Q4 6

The Range

The Range as already stated refers to the difference between the Minimum and the Maximum values.

Range = Maximum - Minimum

- Range Calculations

  – Calculate the Range for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1000000
Q4 1, 2, 3, 4, 5, 6

Worked Example: Q1 - Range

Formula: Range = Max - Min

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Sort the scores

3, 6, 8, 15, 18, 19, 23, 48, 54

Step 2: identify the smallest and the largest number

(smallest) **3, 6, 8, 15, 18, 19, 23, 48, 54** (largest)

Step 2: Deduct the smallest from the largest

Range = **54 – 3**

Range = 51

- Range Calculations
  - Answers

Q1 51
Q2 72.38
Q3 9 (remove the outlier)
Q4 5

2. Standard Deviation

Standard Deviation is a widely used measure to calculate the deviation (dispersion) of the data around the mean. It helps researchers to understand the structure of their data in terms of how the individual observations deviate from or vary around the mean of that variable.

Thus, standard deviation allows for variation and no variation can exist where the standard deviation is marked as 0. The larger the spread of the data, the larger the standard deviation.

Standard Deviation is designated as σ (sigma)

As indicated in the opening section the following table depicts how many values normally fall within each standard deviation.

1 Standard Deviation	**68% of cases within a normal distribution would fall within one standard deviation of the mean**
2 Standard Deviations	**95% of the cases would be catered for**
3 Standard Deviations	**99% of the cases would be catered for**

It is best to understand how one can calculate deviation. The following simple examples depict the differences from the Mean.

18	**18**	**18**	**18**	**19**	**19**	**19**	**20**	**20**	**20**	**20**
-1	**-1**	**-1**	**-1**	**0**	**0**	**0**	**+1**	**+1**	**+1**	**+1**

- 18 deviates -1 from the Mean
- 20 deviates +1 from the Mean

1	1	1	1	1	10	19	19	19	19	19
-9	-9	-9	-9	-9	0	+9	+9	+9	+9	+9

- 1 deviates -9 from the Mean
- 19 deviates +9 from the Mean

Standard Deviation is calculated as follows:

> Subtract the mean from all of the numbers, square the differences, find the average of all of these squared-differences and finally take the square-root; in short one calculates the 'root-mean-square-deviation' about the mean.

The Formulae

Note that the standard deviation for a population and that for a sample are slightly different.

Population	Sample
$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n - 1}}$	$\sigma = \sqrt{\frac{\Sigma\ [x - \bar{x}]^2}{n}}$

This section contains a few working examples. Kindly attempt to analyse the figures based on the process given above.

- Standard Deviation Calculations
  - Calculate the Standard Deviation for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Q4 1, 2, 3, 4, 5, 6

Worked Example: Q1 – Standard Deviation

Formula:

$$\sigma = \sqrt{\frac{\Sigma\ [x - \bar{x}]^2}{n}}$$

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Find the Mean (or the average of the scores) – Refer to workings for Mean above (Section 1)

15, 3, 48, 23, 8, 18, 6, 19, 54

$\bar{x}$ = **21.56**

Step 2: Find the Deviation of each of the scores from the Mean
($\mathbf{x - \bar{x}}$) (score minus mean)

15 - **21.56 = - 6.56**
3 - **21.56 = - 18.56**
48 - **21.56 = 26.44**
23 - **21.56 = 1.44**
8 - **21.56 = - 13.56**
18 - **21.56 = - 3.56**
6 - **21.56 = - 15.56**
19 - **21.56 = - 2.56**
54 - **21.56 = 32.44**

Step 3: Square the deviations found in Step 2. This step amplifies the positive numbers and changes the negative results to positive results $(\mathbf{x - \bar{x}})^2$

-6.56^2	=	43.03
-18.56^2	=	344.47
26.44^2	=	699.07
1.44^2	=	2.07
-13.56^2	=	183.87
-3.56^2		12.67
-15.56^2	=	242.11
-2.56^2	=	6.55
32.44^2	=	1052.35

Step 4: Sum the obtained squares (as a first step to obtaining an average) $\Sigma(\mathbf{x - \bar{x}})^2$

= 43.03 + 344.47 + 699.07 + 2.07 + 183.87 + 12.67 + 242.11 + 6.55 + 1052.35

= 2586.22

Step 5: Divide the Sum the obtained squares by the number of scores $\Sigma(\mathbf{x - \bar{x}})^2$ **/ n**

= 2586.22 / 9

= 287.36

Step 6: To find the Standard Deviation, run a square root of then result of Step 5 $\sqrt{\Sigma(\mathbf{x - \bar{x}})^2/\mathbf{n}}$

$$\sigma = \sqrt{\frac{\Sigma\ (x - \bar{x})^2}{n}}$$

= $\sqrt{287.36}$

= 16.95

Standard Deviation = 16.95

- Standard Deviation Calculations
    - Answers

Q1 16.95
Q2 24.41
Q3 2.87
Q4 1.71

3. Variance

The variance is defined as the sum of the squared deviations from the mean, divided by n-1. It is computed as the average squared deviation of each number from its mean.

The Variance is designated as σ^2 (sigma squared) (S^2 for a sample)

In other words, the variance is the square of the standard deviation. Thus, vice versa, the standard deviation formula is very simple: it is the square root of the variance. Both variance and standard deviation provide the same information; one can always be obtained from the other.

The Formulae

Note that the standard deviation for a population and that for a sample are slightly different.

Population	Sample
$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$	$S^2 = \frac{\Sigma (X-\bar{X})^2}{n-1}$

- Variance Calculations
    - Calculate the Variance for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Q4 1, 2, 3, 4, 5, 6

Worked Example: Q1 – Variance

Formula:

$$S^2 = \frac{\Sigma (X-\bar{X})^2}{n-1}$$

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54

Step 1: Find the Mean (or the average of the scores) – Refer to workings for Mean above (Section 1)

15, 3, 48, 23, 8, 18, 6, 19, 54

$\bar{x}$ = **21.56**

Step 2: Find the Deviation of each of the scores from the Mean
($\mathbf{x - \bar{x}}$) (score minus mean)

15 - **21.56 = - 6.56**
3 - **21.56 = - 18.56**
48 - **21.56 = 26.44**
23 - **21.56 = 1.44**
8 - **21.56 = - 13.56**
18 - **21.56 = - 3.56**
6 - **21.56 = - 15.56**
19 - **21.56 = - 2.56**
54 - **21.56 = 32.44**

Step 3: Square the deviations found in Step 2. This step amplifies the positive numbers and changes the negative results to positive results $(\mathbf{x - \bar{x}})^2$

-6.56^2	=	43.03
-18.56^2	=	344.47
26.44^2	=	699.07
1.44^2	=	2.07
-13.56^2	=	183.87
-3.56^2	=	12.67
-15.56^2	=	242.11
-2.56^2	=	6.55
32.44^2	=	1052.35

Step 4: Sum the obtained squares (as a first step to obtaining an average) $\Sigma(\mathbf{x - \bar{x}})^2$

= 43.03 + 344.47 + 699.07 + 2.07 + 183.87 + 12.67 + 242.11 + 6.55 + 1052.35

= 2586.22

Step 5: To find the Variance, divide the Sum the obtained squares by the number of scores
$\Sigma(\mathbf{x - \bar{x}})^2$ **/ n**

$$S^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$$

= 2586.22 / 9

= 287.36

Variance = 287.36

OR SIMPLER

Standard Deviation Squared

σ^2

$= 16.95^2$

= 287.36

Variance = 287.36

- Variance Calculations
    - Answers

Q1 287.36
Q2 595.69
Q3 8.25
Q4 2.92

4. The Z-Score

All the above begs the question: How does one calculate where a particular value falls within a standard deviation? Does a 30-year old male fall within 1 standard deviation (that is within 68% percent of the population) or within the other deviations?

This is carried out using the Z-Score test. The test calculates the position where a number on the x axis resides in terms of the standard deviation. In summary, the z-score defines the distance the sample value is from the mean – always in terms of standard deviations.

The larger the values, the further away from the Mean that value resides and into the higher standard deviations. If a Z score is negative, then the value (X) is below the mean. If it is positive, X is above the mean.

The Formula

$$z = \frac{x - \bar{x}}{s}$$

The Z-Score is calculated as follows:

Z = (a given value - mean) / standard deviation

For example, for a young population that is normally distributed with a mean(μ) of 20 and a standard deviation (σ) of 5, you want to find out the Z score for a value of 30 (*x*). This value (X = 30) is 10 units above the mean, with a Z value of:

Z = (30 - 20)/(5) = (10)/(5) = +2

The point is within 2 Standard Deviations of the Mean

This section contains a few working examples. Kindly attempt to analyse the figures based on the process given above. Use the following example as a guide:

- Z-Score
    - Calculate the Z-Score for the following number in brackets: **15, 3, 48, 23, 8, 18, 6, 19, 54 (23)**

i) Firstly, calculate the Mean (μ) of 15, 3, 48, 23, 8, 18, 6, 19, 54. This results in a mean of 21.6
ii) Secondly, calculate the standard deviation(σ) as per relative guide above. This gives a standard deviation of 17.0
iii) Thirdly, deduct the mean from your value - 23 (*x*)and divide by the standard deviation.
iv) The result is that of 0.1 which fall within 1 standard deviation (between 0 and 1 and is a positive number) - (+0.1)

- Z-Score Calculations

- Calculate the Z-Score for the following:

Q1 15, 3, 48, 23, 8, 18, 6, 19, 54 (23)
Q2 16.5, 18, 63.2, 88.88, 19, 20, 21, 22, 18.5 (80)
Q3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (2)
Q4 1, 2, 3, 4, 5, 6 (5.5)

Worked Example: Q1 – Z- Score

Formula:

$$Z = x - \mu / \sigma$$

Therefore:

$$\frac{23 - 21.6}{17.0} = 0.1$$

Note that 0.1 fall within the 1st Standard Deviation (less that 1), thus the results can be said to be within 1 standard deviation (+0.1).

- Z-Score Calculations
  - Answers

Q1 within 1 standard deviation (+0.1)
Q2 within 2 standard deviations (+2)
Q3 over -1 standard deviation (-1.2)
Q4 over 1 standard deviation (+1.2)

Statistical Tests

This section will outline some statistical tests that are used to help researchers understand their data as well as carry out comparative studies. The tests include the F-test, the T-tests, Regression Analysis, ANOVA and Chi Squared. The first two employ both the mean and the standard deviation in order to test if two sets of normally distributed data are similar or otherwise. If they are similar then they can be attributed to the same population. Regression analyzes how the independent variables are related to the dependent variable. The ANOVA is used where more than one factor can exert an influence on a value. Chi-Squared is used to check whether a categorical sample represents the population.

The next steps describe each of the tests in summary. Refer to one of the indicated statistical books for a full description and step by step process of how to employ these tests. If one is referring to online tutorial, ensure that they are produced by reliable sources such as Stat Trek[1].

1. **F-Test**

One of the first inferential tests that one can use in order to establish whether the variances between two populations are equal is the F-Test. This test compares the ratio of the two variances which, if equal, should result in a value of 1.

2. **T-tests**

The T-tests are employed for testing standard deviations when the population is normally distributed. It is a random interval or ratio sample, where the standard deviation is computed from the sample data.

There are different tests which are given names according to the type of similarity between the datasets.

[1] http://stattrek.com/

Independent datasets that have very similar Standard Deviations: employ the **Student's t-test**. This is applied for small datasets having N (number of data) less than 30 and where the F-test shows that they are similar.

Independent datasets that have significantly differing Standard Deviations: employ the **Cochran t-test.** This is applied for small datasets having N (number of data) less than 30 and where the F-test shows that they are dissimilar.

Highly Dependent datasets employ the **paired t-test**. This can be employed when the same samples are used for two different tests.

3. Regression Analysis

Normally described as the Line of Best Fit, regression is used to establish the existence of a linear relationship.

Regression analysis assumes that a change in dataset X brings about a definite change in dataset Y. In correlation analysis, a change in X brings about a change in Y, which could be anything from an increase to a decrease or even no change at all. Regression is not so 'easy' on the relationship as a change in X must bring about a change in Y.

4. ANOVA

The Analysis of Variance, also known as the ANOVA, determines the existence of differences in datasets that contain two or more sample means. A two-way ANOVA is tested for when two independent variables are chosen.

5. Chi Squared

Chi Squared ($\chi 2$) is a critical test that investigates, looking for the frequencies of category (Nominal) presence in a sample and analyzes whether they represent the predicted frequencies in the total population.

This short summary does not do justice to the beauty that is statistical testing and readers are encouraged to acquire a statistics book that pertains to their particular theme as listed in the Thematic Publication Table.

Spatial Statistics

This section outlines some specialised spatial statistical tests that are used to help researchers understand their data both in normal statistics and in a higher-level mode where statistics are also depicted in visual modes.

Spatial statistical books need to be referenced for a full description and step by step process of how to employ these tests. There are different types of spatial statistics, best clustered in four-groups: Spatial distribution, Distance statistics, 'Hot spot' analysis routines and Interpolation statistics:

1. Spatial Distribution

Spatial distribution refers to the spread of values around a spatial mean. These include the mean centre, centre of minimum distance, standard deviational ellipse and Moran's I spatial autocorrelation index, or angular mean.

2. Distance Statistics

Distance statistics calculate the values based on proximity and statistical tests include the nearest neighbour analysis, linear nearest neighbour analysis, and Ripley's K statistic.

3. Hotspot Analysis Routines

'Hot spot' analysis routines are some of the most interesting and used spatial statistics as they depict data based on the concentration of values are a spatial location. Tests include the hierarchical nearest neighbour clustering, K-means clustering and local Moran statistics. There are alternative measures of hotspot analysis such as Kernel Density Estimate, Getis-Ord GI* and also multiple regression as an alternative to bi-variate analysis.

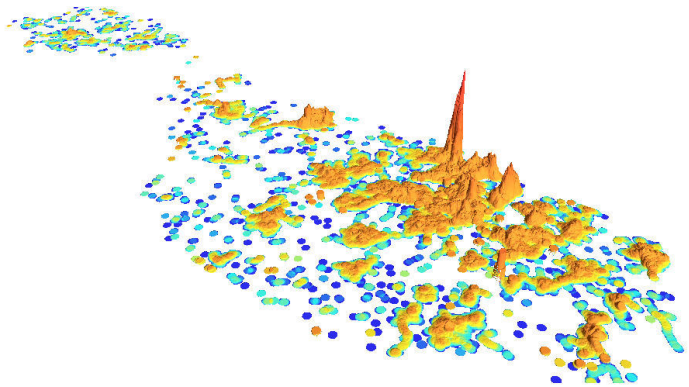
4. Interpolation Statistics

Interpolation statistics use the primary data and interpolate it to predict the probable value of areas within a boundary relative the location of the primary data. These tests include single-variable kernel density estimation resulting in an output of incident density (e.g. burglaries) and a dual-variable kernel density estimation that compares once variable to another baseline variable (e.g., burglaries analysed in relation to dwelling units in the area).

Of all the above, one of the best methods of analysing behavioural patterns is to use clustering methodology. Formosa (2007, Pg149-152) describes that due to the large number of behavioural patterns (such as crime, recreation) occurring in a particular area, analysis may concentrate on the aggregation of these data into specific areas rather than spread them all over the town/city. Clustering helps in identifying areas that are hotspots for specific behaviour types.

ii) Another method that can be employed is the Nearest Neighbour Analysis (NNA) which helps to aggregate data based on the proximity of a crime to the nearest location of another crime (Craglia *et al*, 2000). If an activity occurs within a specific parameter of say '20m' from that being analysed, then these two activities are aggregated, before searching for other activities within the next specific boundary[2]. Once there are no activities left within the recurrent buffers then the hotspot intensity dies out and stops. Where a large number of activities occur in a small area the hotspot is very pronounced and cluster densities can be calculated. Figure 11.1 depicts an example of such an NNA interpolation based on non-serious offences in Malta between 1998-2003 transposed in 3D (Formosa 2007, 150). The shape of the Maltese Islands is easily discernable, particularly the conurbation area. High offence counts are depicted as with red peaks in the main leisure and recreation areas and very few if anything in the rural and rural-urban boundary areas (blue and white respectively). The same methodology can be used to elicit statistical results as well as for visualization purposes.

Figure 11.1: Interpolation of Non-Serious Offences – 1998-2003



Source: Formosa, 2007, Pg 150

[2] Note that variance in the boundary width can produce different results.

Note that each of these methods necessitates knowledge of the limitations in using that specific method which limitations are dependent on a number of factors. These include the sample size taken, the number of minimal points set as the threshold for identifying the least hotspot size, amongst others. The limitations of the methodologies used such as the Nearest Neighbour Hierarchical Analysis Method (NNH) include differing hotspot locations for different spatial aggregations employed, such as a minimal 25-point hotspot cut-off, which signifies where an ellipse boundary should be drawn once no more points falling within those thresholds are encountered. Consistency in the results is ensured as the analysis in this study employ the same threshold limits. Another limitation relates to the issue of cross-comparison of two data-layers that may have widely-differing counts, such as a 10,000 point offender data layer and a 1,000 point poverty layer. Using the same standard-deviation levels and thresholds, error generation can be reduced to a minimum.

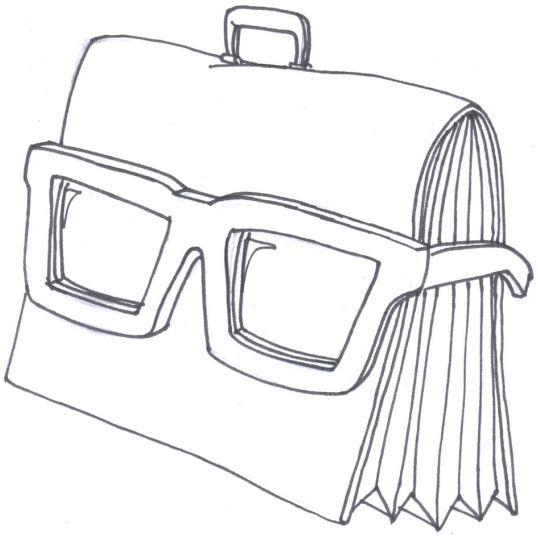
In addition, NNH as well-as K-Means employed in the study show their results through ellipsoids, which in effect can cover areas that may not be prone to high incidences being investigated but still fall within the ellipsoid since such a tool cannot eliminate areas within its boundary without compromising the ellipsoid integrity. Also, some ellipsoids might show areas that have high concentrations of incidences when the base data might show few data points, which result is mainly due to a multiplicity of overlapping points found within the base data layer and weighted for in the ellipsoid. Knowledge of the base data layer is required in order to interpret the results of such methodologies.

In summary statistical tests are many and varied, they cover simple descriptive statistics to inferential statistical tests to spatial statistics. This chapter sought to introduce the new initiate to the idea that, whilst appearing 'scary', statistical tools are easy to understand. However, one really needs to consult specialist books in the field which employ theme-specific examples in order to understand how each of the tests is carried out. The initial section outlining basic statistics gave a walkthrough with examples of how to carry out the calculation required, whilst the other two sections gave a summary of the types of statistical tests that exist for inferential statistics and for spatial statistics respectively.

Questions (refer to Appendix for the answers)

1. Why is statistical testing important?
2. Briefly explain what descriptive statistics are, providing examples.
3. Briefly explain what inferential statistics are, providing examples.
4. What do you understand by independent variables?
5. Why are dependent variables also known as criterion variables?
6. List the three measures of central tendency.
7. There are two types of Mode. Name them.
8. How would you define "the range" in statistics?
9. Briefly describe standard deviation, explaining its function in statistics.
10. What is the variance (in statistics)?
11. What does the Z-score do?
12. Mention five statistical tests and very briefly describe each of them.

Chapter 12
Case Studies



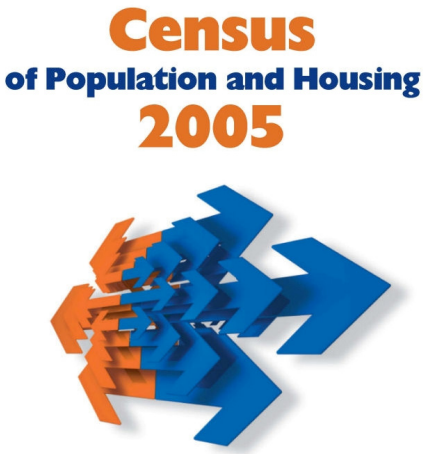
Do not do anything that anyone else can do readily.

Edwin Herbert Land

In Alan R. Earls and Nasrin Rohani, *Polaroid* (2005), 16.

This chapter describes two case studies which help the researcher to review how some studies are carried out and what use is made of the data. The first study describes how Census data is used to develop a series of online datasets that allow users to interact with the data. The second case study takes readers on a tour to the National Archives of Malta.

A taste for working with Census data



http://www.nso.gov.mt/site/page.aspx?pageid=351

The Census of Population and Housing of the Maltese Islands is carried out every 10 years and entails a major enterprise that is carried out by the National Statistics Office (NSO). Works starts early in the year or the year before the Census where various phases are launched in order to enable a smooth running of the exercise. These phases include: the drafting of a list of enumerators (these are officials who establish contact with the members of the public and collect information from the household members), the identification and mapping of the routes they have to follow and which households they have to interview (over 160 thousand), logistical issues like interviewing of enumerators, monitoring their progress, inputting and double checking the questionnaire replies, chasing those persons who were not at home and a thousand other activities. Not a simple task but managed expertly by the Census Office and its Officers (Source: NSO presentation, Zammit S. & Mizzi R., (16 August 2010).

The Management

The Census is managed by an extensive team of professionals who take up the day-to-day running of the Census and ensure that the project runs smoothly. This management team is composed of the following persons:

- **Census Officer -** Person responsible for the Census and empowered to hold the Census in terms of the Census Act of 1948;
- **Chief Coordinator -** Person who runs the Census Office and is responsible for field operations;
- **Census Officers -** Group of persons carrying out the backend process of the project;
- **District Managers -** Responsible for field operations and the work being carried out by the Supervisors;
- **Supervisors -** Responsible for the guidance and direction of the Enumerators; and
- **Enumerators -** Officials who establish contact with the members of the public and collect information from the household members.

The Pre-Collection Process

Approximately 160K Census questionnaires were mailed to all private dwellings in Malta some weeks prior to Census Day (27 November 2005). Households were encouraged to fill in the questionnaire themselves. Enumerators were requested to assist the other households. The way the enumerators were managed was based on a NUTS5 structure further separated through a streets route mapping exercise

carried out by MEPA, which also delivered a map of the area each enumerator had to visit. There were around a 1000 maps generated along with the documentation listing the streets that can be found within each enumeration area.

In the images below, the first map (Figure 12.1) shows the streets within Attard where each colour represents the area that each enumerator had to visit. The second map (Figure 12.2) depicts the boundary inclusive of the streets within each boundary.

Figure 12.1: Attard Enumeration Areas - Streets

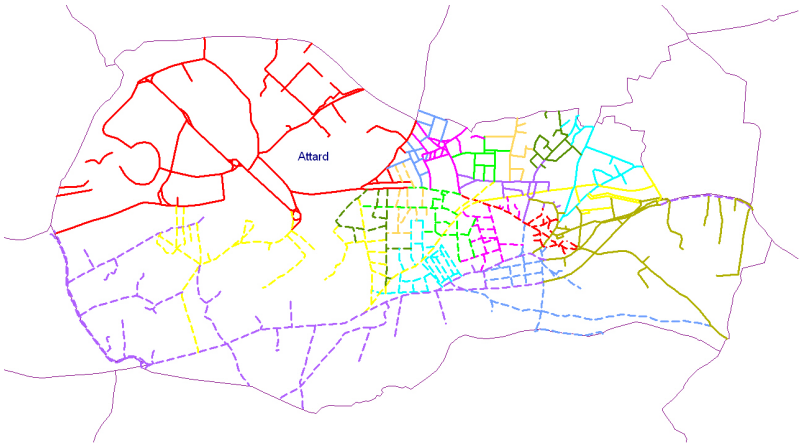
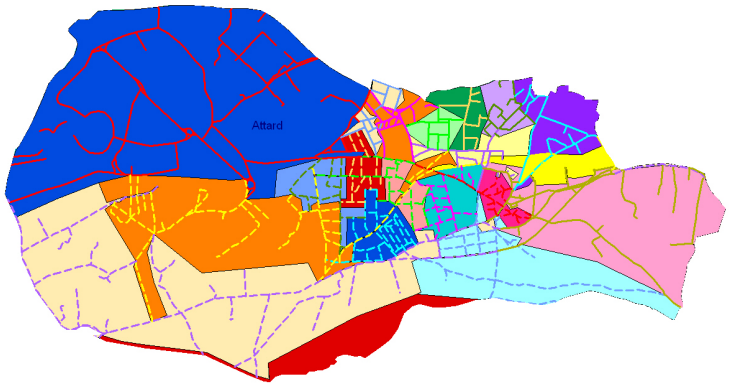


Figure 12.2: Attard Enumeration Areas - Boundaries



Data Collection Process

The collection process was carried out through a system where all households in Malta and Gozo were contacted by the NSO appointed enumerators during a three week period starting 21 November 2005 till the 11th December 2005. This process required the services of 87 supervisors who were appointed to act as a focal point in most localities and 6 district managers who were appointed to assist the supervisors.

Data Processing and Analysing

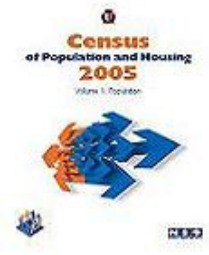
As the process necessitated that collected questionnaires were delivered to the Census Office on a regular basis, the actual data processing started immediately. The inputting phase lasted 2 months, including the inputting of those questionnaires that were collected after the reference period. The data analysis and production of statistical tables lasted about 12 months.

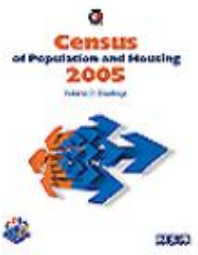
Publishing

There were 3 analogue publications and a digital publication

- Preliminary Report was published in April 2006

- Volume 1: Population was published in August 2007
- Volume 2: Dwellings was published in October 2007
- CD containing Volume I and 2 as well Interactive Maps pertaining to the tables published in the Volumes.





The extract from the Census 2005 website[1] summarises all this work through an overview by the Census Officer.

Census 2005

After a Census Order was issued in 2005, a Census of Population and Housing was undertaken between 21st November and 11th December 2005, with 27th November 2005 being established as Census Day. This was the sixteenth census to be carried out since the first one was undertaken in 1842. On my part, it was the second Census in which I served as Census Officer.

Preliminary findings, detailing the population count on a locality basis by age and sex, were published in April 2006. While that report presented a preliminary population count, this report presents the final count of the population, as on 27th November 2005, together with other socio-demographic results as well as information on our housing stock. A chapter outlining census methodology and a commentary on key census results are also being presented.

This Census presents a snapshot of the socio-demographic profile of our population in the early years of the 21st century. It presents a wide array of statistical reports, including a range of reports with internationally comparable indicators that are important for policy-making and research purposes. We feel confident that the results of this Census will be actively and profitably used for the benefit of our people.

In carrying out this project, particular attention was continuously paid to the quality aspects of our work. The enumeration process was rigorously monitored and evaluated with a view to ensuring coverage and the taking of remedial action whenever problems arose. This was then followed up by an extensive follow-up exercise in order to ensure the exhaustiveness of the enumeration process. This, together with the results of the post-enumeration survey, facilitated the imputation of any missing responses and the coding of a sample of occupation and economic activity codes. Finally, all the results were appropriately validated against other independent data sources in order to ensure their reliability.

A project of this magnitude cannot be carried out successfully without the full co-operation of the general public. Therefore, I would like to express my deep appreciation to all those who co-operated fully with the Census Office during and after the taking of the Census. Lastly, I would also like to express my heart-felt thanks to all the staff at the National Statistics Office and the Census Office for their hard work and dedication in implementing this project successfully.

Alfred Camilleri
Census Officer

25 July 2007

[1] http://www.nso.gov.mt/site/page.aspx?pageid=351

Using the Census for Research

A series of post-census analytical steps that researchers should employ is included below. The target is to map population data for comparison across the different NUTS5 areas.

Step 1: Getting to grips with the terminology

Once all the data has been inputted, the same data is made available at diverse levels of scale which protects information pertaining to individual respondents as it produces data at the various NUTS levels and at a detailed Enumeration Area level. This case study reviews the publication of interactive Census maps for researchers at the diverse NUTS, LAU and the EAS levels. However, once we introduced NUTS and EAS it is best to define them.

NUTS, LAU and EAS

The Nomenclature of Territorial Units, also known as NUTS, was developed by EUROSTAT[2] as far back as the 1970s but become a legal instrument through Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003. The regulation targeted the establishment of a common classification of territorial units for statistics (NUTS). It was aimed at ensuring the classification of territorial units into comparable levels which would allow researchers to compare data on the same level, such that an area in Gozo is compared to a similar area in France or The Netherlands where one would be assured that the base data refers to the same territorial category as based on population thresholds. The table below defines the thresholds for each of the NUTS 1 to 3 levels.

LEVEL	MINIMUM	MAXIMUM
NUTS 1	3 million	7 million
NUTS 2	800 000	3 million
NUTS 3	150 000	800 000

Source: EUROSTAT
http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/principles_characteristics

It is interesting to note that Malta would fall under a NUTS 3 as a state but the categorisations cater for the inclusion of each level for every country, thus Malta has classifications from NUTS 1 to 5.

Recently, levels 4 and 5 were termed as Local Area Units or LAUs and they cater for the smaller administrative units such as small districts and municipalities or local councils.

At this stage, readers are surely scratching their heads in bewilderment. All these terminologies could sound like the jargon we hear on science fiction films. No one will blame readers for feeling they have gone nuts!

In tabular format one can see the categorisation and the number of units in Malta (Table 12.1).

[2] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

Table 12.1: NUTS/LAU Levels

NUTS Level	LAU Level	Designation	Maltese equivalent	Number of Units in Malta
1_2[3]		National (1) Regional (2)	National	1
3		Sub-Regional	Island	2 (Malta) (Gozo)
4	1	District	Districts	6
5	2	Least Administrative Units	Localities (Local Councils)	68 (54 in Malta) (14 in Gozo)

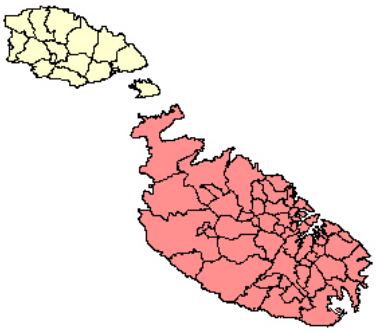
Another spatial layer that is used for research concerns that called the Enumeration Area layer. This level of data splits up the Maltese Islands in over 1,000 areas comprising on average 135-130 households. This level of data is very detailed and helps in the analysis of economic, social and behavioural issues at very detailed levels.

If one had to review these levels in map format (Figure 12.3), one would easily understand which areas the different NUTS levels are comprised of. Each map has the local councils layer overlaid for ease of reference.

Figure 12.3: Map Data Aggregations
(Note: colours represent individual areas under the different categorisations)

NUTS 1_2 – National

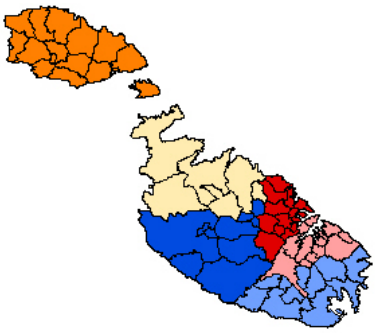
Description:
1 main map representing the Maltese Islands as a single unit: cyan colour with the magenta representing local council boundaries.



NUTS 3 – Islands

Description:
2 island-based spatial groupings: namely mainland Malta as one entity and Gozo and Comino as the other entity.

[3] Note that Malta's NUTS 1 and 2 categories have been designated as the same due to the size of the state. In larger countries, such as Germany, the Landers would be given a NUTS 2 and the Federated State a NUTS 1 category.

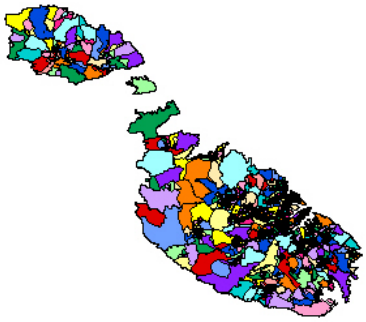


NUTS 4 – Districts

Description:
6 Census Districts. No real administrative powers exist at this level.

NUTS 5 – Local Councils

Description:
68 councils serving as the least administrative unit as defined by the NUTS nomenclature. Each area has its own elected administrative council.



Enumeration Areas

Description:
1,157 Census-based enumeration areas based on the boundaries set for each enumerator nominally representing 150 households.

Once the base knowledge on what is required for this case study was gathered, the next step was to source the data required for the generation of a population map for the Maltese Islands with data going back as far as possible at NUTS4 (LAU1) and NUTS5 (LAU2) levels.

Step 2: Sourcing the Data

The table was sourced from the National Statistics Office, which population data was produced as from 1901, a total of 9 Censuses. Note that some localities did not exist prior to 2005 and one needs to take note that the current locality boundaries reflect the latest change that occurred in 1998 when Mtarfa was extracted from the locality of Rabat. For publishing clarity, the 2005 map is used, but one should note that the data referring to earlier Censuses should technically reflect the maps pertaining to the relative period, something which is difficult to employ since the boundary location is not always defined.

Table 12.2 was acquired and converted to spreadsheet and spatial format.

Table 12.2: Overview of total population by locality: Censuses since 1901

	1901	1921	1931	1948	1957	1967	1985	1995	2005
MALTA	**184,742**	**212,258**	**241,621**	**305,991**	**319,620**	**314,216**	**345,418**	**378,132**	**404,962**
Malta	**164,952**	**189,697**	**217,784**	**278,311**	**292,019**	**288,238**	**319,736**	**349,106**	**373,955**
Gozo and Comino	**19,790**	**22,561**	**23,837**	**27,680**	**27,601**	**25,978**	**25,682**	**29,026**	**31,007**
Southern Harbour	**70,244**	**79,001**	**87,811**	**84,206**	**90,705**	**87,879**	**86,843**	**83,234**	**81,047**
Birgu	6,093	5,887	6,573	3,816	4,242	4,017	3,572	3,069	2,701
Bormla	12,148	11,536	12,163	4,822	9,095	9,123	7,731	6,085	5,657
Fgura	-	-	-	-	-	2,737	8,254	11,042	11,258
Floriana	5,687	5,907	6,241	5,074	5,811	4,944	3,327	2,701	2,240
Isla	8,093	7,741	7,683	2,756	5,065	4,749	4,158	3,528	3,074
Kalkara	1,158	1,698	1,899	2,068	2,101	1,945	2,086	2,833	2,882
Luqa	3,670	3,607	4,059	4,318	5,382	5,413	5,585	6,150	6,072
Marsa	-	4,838	7,867	11,560	10,672	9,722	7,953	5,324	5,344
Paola	2,812	5,475	7,297	14,793	11,424	11,794	11,744	9,400	8,822
Santa Luċija	-	-	-	-	-	-	3,208	3,605	3,186
Tarxien	2,065	2,876	3,247	4,607	7,706	7,989	7,016	7,412	7,597
Valletta	22,768	22,392	22,779	18,666	18,202	15,279	9,340	7,262	6,300
Xgħajra	-	-	-	-	-	-	-	685	1,243
Żabbar	5,750	7,044	8,003	11,726	11,005	10,167	12,869	14,138	14,671
Northern Harbour	**42,774**	**52,347**	**63,941**	**101,526**	**104,889**	**102,938**	**113,730**	**118,409**	**119,332**
Birkirkara	8,417	8,565	10,345	16,070	16,987	17,213	20,385	21,281	21,858
Gżira	-	-	-	6,295	8,545	9,575	8,471	7,872	7,090
Ħamrun	10,393	10,434	11,580	17,124	16,895	14,787	13,682	11,195	9,541
Msida	3,826	5,196	6,334	9,690	10,663	11,437	6,219	6,942	7,629
Pembroke	-	-	-	-	-	-	-	2,213	2,935
Pieta'	-	-	-	-	-	-	4,380	4,307	3,846
Qormi	8,187	9,286	10,165	14,396	14,869	15,398	18,256	17,694	16,559
San Ġiljan	1,444	2,594	3,998	9,122	8,285	7,394	10,239	7,352	7,752
San Ġwann	-	-	-	-	-	-	8,179	12,011	12,737
Santa Venera	-	1,910	2,639	4,535	5,246	6,134	7,827	6,183	6,075
Sliema	10,507	14,362	18,880	24,294	23,399	21,000	14,137	12,906	13,242
Swieqi	-	-	-	-	-	-	-	6,721	8,208
Ta' Xbiex	-	-	-	-	-	-	1,955	1,732	1,860
South Eastern	**17,546**	**20,090**	**23,052**	**34,208**	**36,854**	**35,224**	**42,475**	**50,650**	**59,371**
Birżebbuġa	-	1,219	1,724	5,339	5,297	4,876	5,668	7,307	8,564
Għaxaq	1,518	1,629	1,896	2,448	2,830	2,866	3,655	4,126	4,405
Gudja	1,133	1,167	1,283	1,486	1,712	1,729	2,156	2,882	2,923
Kirkop	633	707	805	1,016	1,204	1,225	1,559	1,957	2,185
Marsaskala	-	-	-	-	888	876	1,936	4,770	9,346
Marsaxlokk	446	791	829	1,431	1,469	1,462	2,405	2,857	3,222
Mqabba	1,228	1,282	1,468	1,965	2,088	2,120	2,269	2,613	3,021
Qrendi	1,333	1,526	1,611	2,144	2,155	2,094	2,199	2,344	2,535
Safi	367	459	448	1,040	709	784	1,323	1,731	1,979
Żejtun	7,234	7,701	8,731	11,980	11,665	10,440	11,321	11,379	11,410
Żurrieq	3,654	3,609	4,257	5,359	6,837	6,752	7,984	8,684	9,781

	1901	1921	1931	1948	1957	1967	1985	1995	2005
Western	**21,666**	**23,587**	**26,393**	**34,899**	**36,196**	**36,142**	**44,580**	**51,961**	**57,038**
Attard	1,837	2,058	2,354	2,480	2,663	2,570	5,681	9,214	10,405
Balzan	1,096	1,313	1,661	2,637	2,734	3,301	4,781	3,560	3,869
Dingli	807	1,087	1,258	1,869	2,041	1,795	2,047	2,725	3,347
Iklin	-	-	-	-	-	-	-	3,098	3,220
Lija	1,692	1,612	1,795	1,950	2,119	2,143	3,078	2,497	2,797
Mdina	304	816	982	1,384	823	988	421	377	278
Mtarfa	-	-	-	-	-	-	-	-	2,426
Rabat	7,211	7,985	9,050	12,503	12,792	12,243	12,920	12,995	11,473
Siġġiewi	3,265	3,355	3,537	4,583	5,055	4,971	5,864	7,097	7,931
Żebbuġ	5,454	5,361	5,756	7,493	7,969	8,131	9,788	10,398	11,292
Northern	**12,722**	**14,672**	**16,587**	**23,472**	**23,375**	**23,933**	**32,108**	**44,852**	**57,167**
Għargħur	1,377	1,327	1,483	1,690	1,813	1,774	2,321	1,991	2,352
Mellieħa	2,357	2,637	3,198	4,549	4,290	4,279	4,525	6,221	7,676
Mġarr	745	1,271	1,627	2,218	2,167	2,115	2,188	2,672	3,014
Mosta	4,629	4,866	5,251	7,186	7,377	8,334	12,148	16,754	18,735
Naxxar	3,429	2,886	3,249	4,389	4,688	4,643	6,461	9,822	11,978
San Pawl Il-Baħar	185	1,685	1,779	3,440	3,040	2,788	4,465	7,392	13,412
Gozo and Comino	**19,790**	**22,561**	**23,837**	**27,680**	**27,601**	**25,978**	**25,682**	**29,026**	**31,007**
Fontana	-	-	-	-	-	893	836	817	850
Għajnsielem and Comino	1,121	1,250	1,449	1,878	1,860	1,755	1,809	2,176	2,570
Għarb	1,091	1,402	1,398	1,555	1,269	1,117	983	1,030	1,146
Għasri	467	409	467	594	471	374	335	369	418
Kerċem	1,037	1,143	1,212	1,307	1,272	1,251	1,411	1,557	1,665
Munxar	-	-	-	-	-	420	507	780	1,052
Nadur	2,948	3,460	3,354	3,465	4,136	3,694	3,482	3,882	4,192
Qala	1,219	1,340	1,601	1,569	1,616	1,522	1,369	1,492	1,616
Rabat	5,057	5,219	5,531	6,175	6,357	5,462	5,968	6,524	6,395
San Lawrenz	643	528	499	413	428	511	517	552	598
Sannat	1,116	1,228	1,324	1,625	1,656	1,297	1,309	1,604	1,725
Xagħra	2,562	3,262	3,522	4,759	4,056	3,517	3,202	3,669	3,934
Xewkija	1,762	2,314	2,470	3,079	3,281	2,999	2,772	3,128	3,111
Żebbuġ	767	1,006	1,010	1,261	1,199	1,166	1,182	1,446	1,735

Data from various censuses

Notes:
A Gżira shown as a separate locality since 1948.
B New locality of Msieraħ (San Ġwann) constituted from parts of Birkirkara and San Ġiljan and shown as separate locality in 1967.
C New locality of Fgura constituted from parts of Paola, Tarxien and Żabbar in 1967.
D Marsaskala shown as a separate locality since 1957.
E New locality of Munxar constituted from parts of Sannat and Fontana.
F New locality of Fontana shown as separate locality in 1967.
G Gwardamangia formed part of Ħamrun in 1967.
H Pieta' formed part of Msida in 1967.
I St Luċija formed part of Tarxien and Paola in 1967.
J Ta' Xbiex formed part of Msida and Gżira in 1967.
K Pembroke formed part of San Ġiljan in 1985.
L Swieqi formed part of San Ġiljan in 1985.
M Xgħajra formed part of Żabbar in 1985.
N Iklin formed part of Lija, Birkirkara, Naxxar and San Ġwann in 1985.
O Mtarfa formed part of Rabat (Malta) in 1995.
P The boundaries of some localities were changed between 1995 and 2005.

Source: National Statistics Office, (2007), Census of Population and Housing: Volume 1 – Population, Valletta, ISBN: 978-9909-73-51-8.

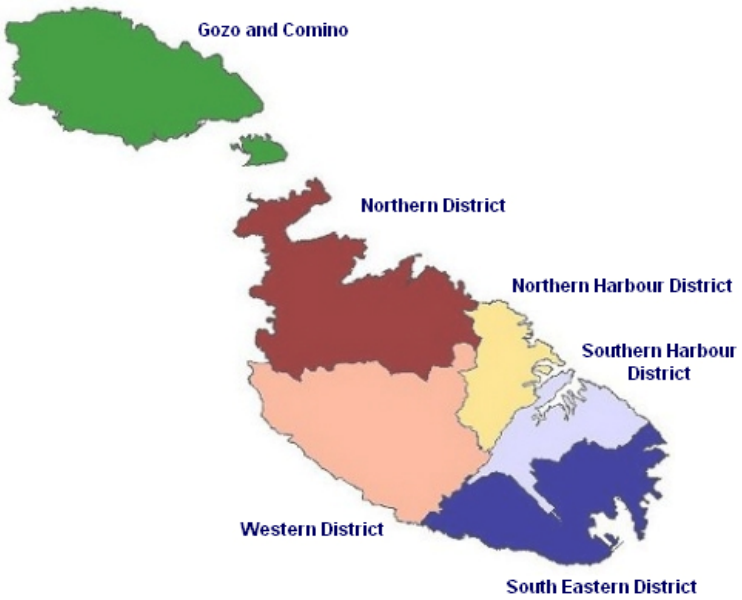
Step 3: Analysing the Data

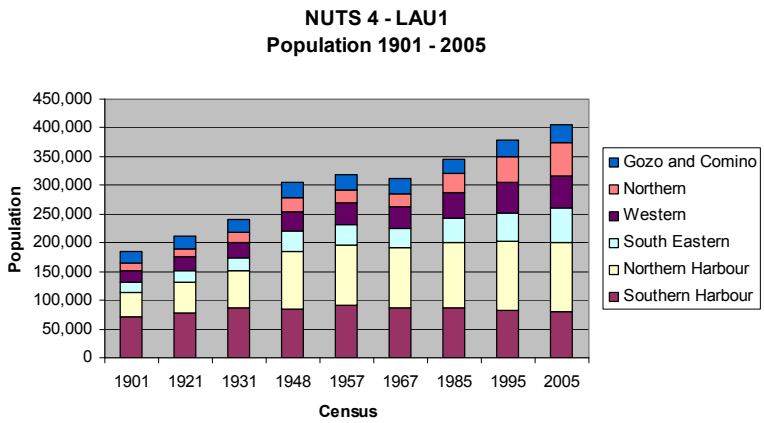
The table shows that two levels of NUTS/LAU can be used as the data produced is a district (NUTS4 – LAU1) and at locality (NUTS 5 – LAU2) levels. This situation enables the researcher to gather comparable data from other themes at those levels.

In our study we shall carry out simple exercises to analyse which immediate information one can extract from the dataset, such as the district (or locality) with the highest population, those that have seen declines, those that have seen a surge in population and other such queries.

Taking the NUTS 4 data and creating a graph would result in an image as depicted below in Figure 12.4.

Figure 12.4: NUTS 4 - Malta





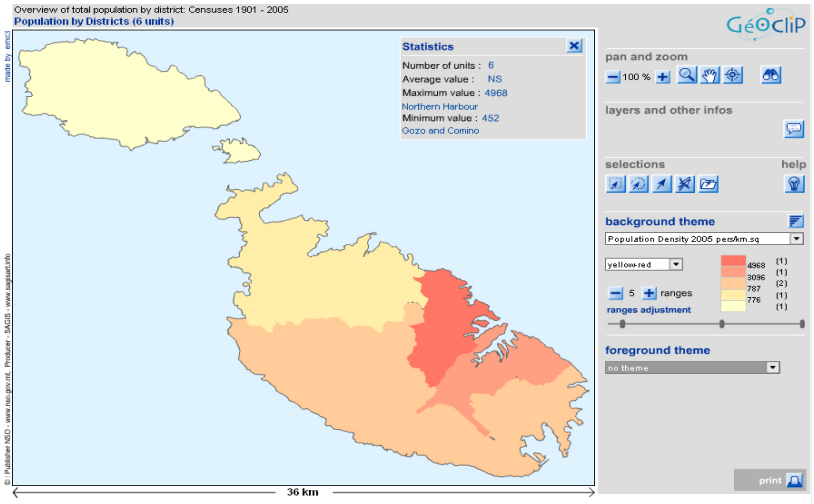
After reading the graph, one can reach a conclusion that the NUTS 4 district which experienced most growth was the Northern Harbour region. Interestingly, the emigration experience of the 1960s shows a dent (decline) which picks up again and increases rapidly after that as shown by Censuses 1985, 1995 and 2005.

Step 4: Visualising the Results

Spatial queries also allow us to carry out such analysis on spatial-related data such as the area which gives us population density, since that is a more realistic variable than absolute numbers.

The following map depicts the output in digital format produced from Census 2005 as part of the publications. The maps have been generated through a GIS product embedded in a multimedia application. The interface contains a left-hand-side menu for document/application launching, a central map area which one can click on to view data and a right-hand-side interactive manipulation system for map and data generation as choropleths (Figure 12.5), and graduated formats. The source for the following figures (12.5-12.12) is http://www.sagisart.info/nso/census2005/

Figure 12.5: Choropleth map depicting population density at NUTS4 – LAU1 in 2005



The next maps show two image employing graduated maps which depict the populations between 1901 (Figure 12.6) and 2005 (Figure 12.7) respectively.

Figure 12.6: Population Graduated map – NUTS4 – LAU1: 1901

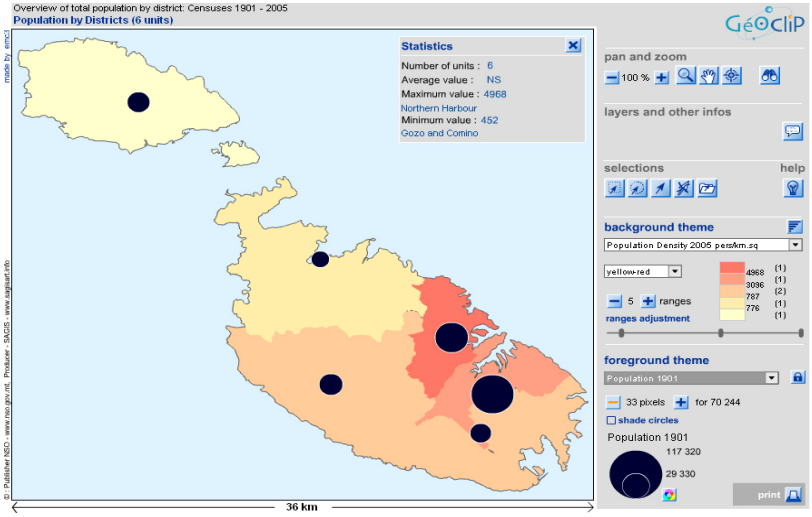
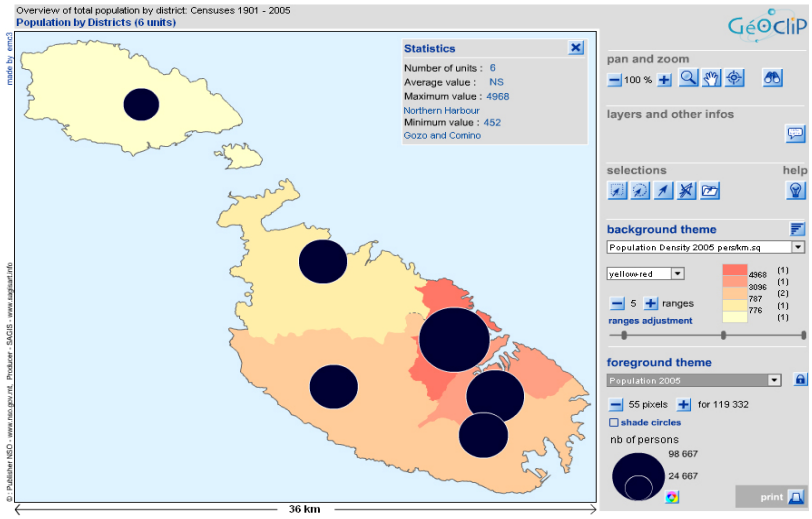
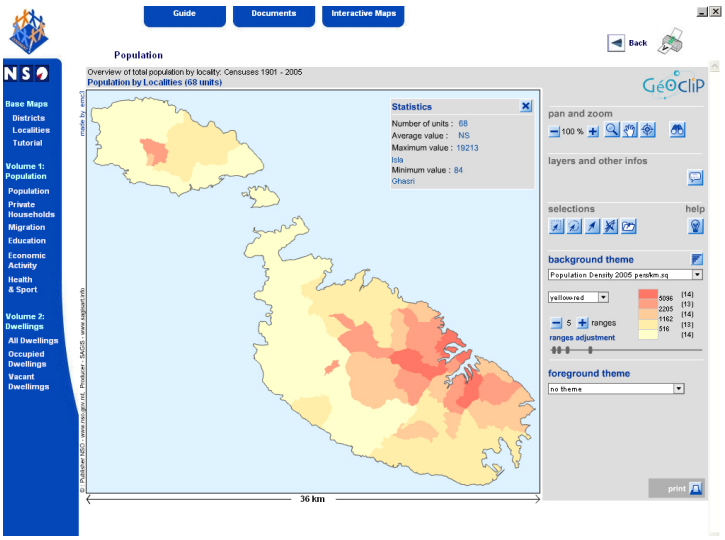


Figure 12.7: Population Graduated map – NUTS4 – LAU1: 2005



The next set of outputs depict the data at NUTS 5 (LAU2) level with Figure 12.8 depicting the data for 2005.

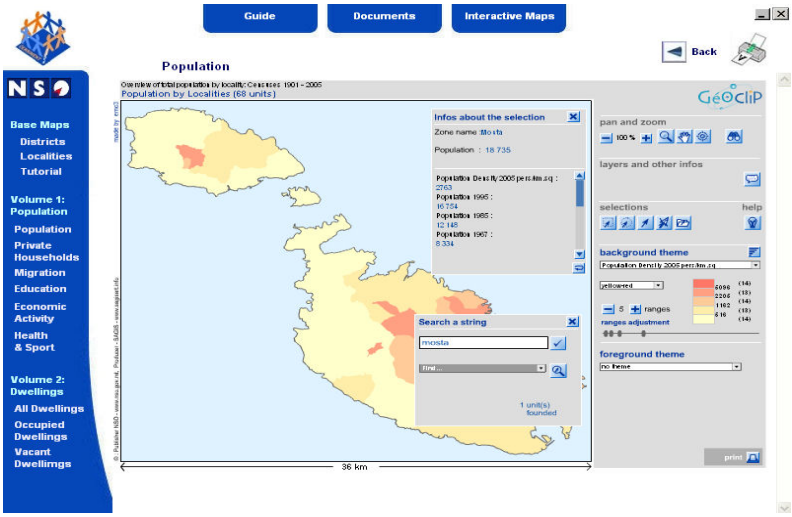
Figure 12.8: Choropleth map depicting population density at NUTS5 – LAU2 in 2005



The map has been generated through a GIS product embedded in a multimedia application.

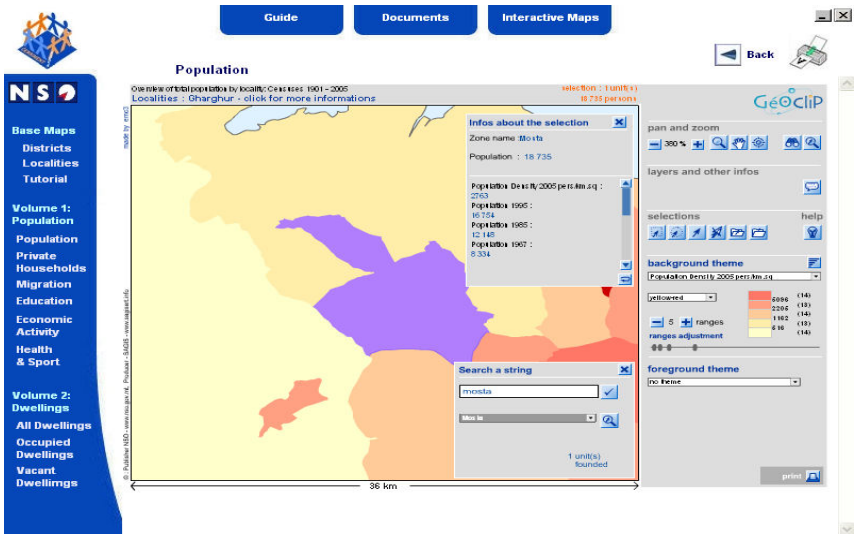
One can search through the data through string searches (Figure 12.9).

Figure 12.9: Search Facility



There is also the option to zoom in to the search results (Figure 12.10).

Figure 12.10: Zooming Facility



The next maps show two image employing graduated maps which depict the populations between 1901 (Figure 12.11) and 2005 (Figure 12.12) respectively.

Figure 12.11: Population Graduated map – NUTS 5 – LAU 2: 1901

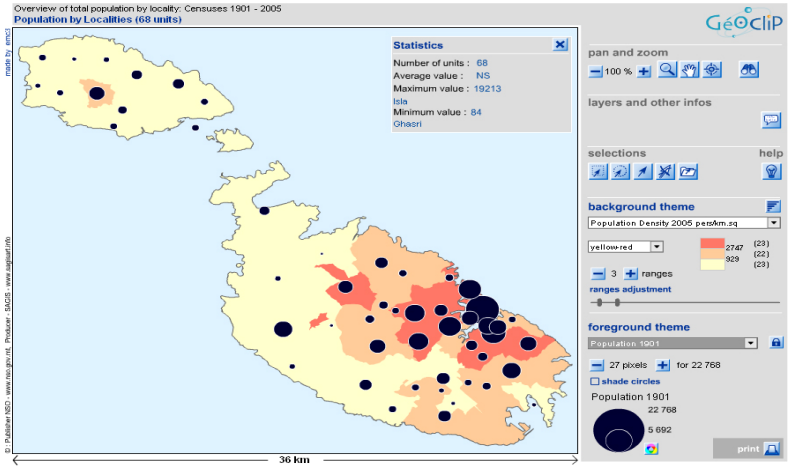
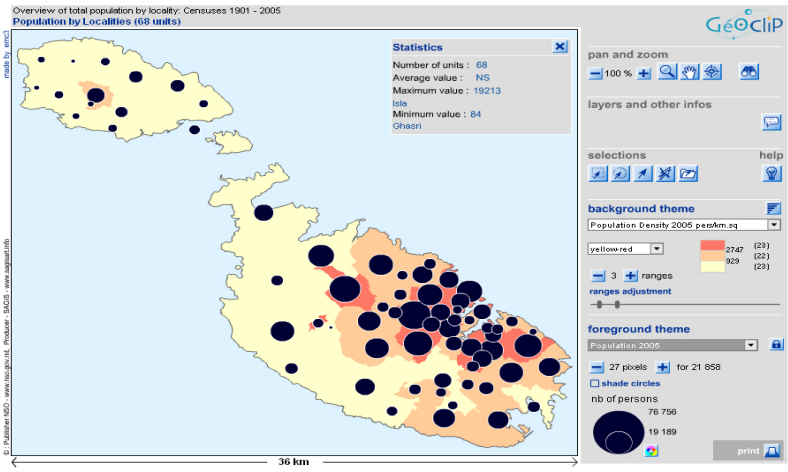


Figure 12.12: Population Graduated map – NUTS 5 – LAU 2: 2005



This case study gave an overview of the process one can employ in order to source data such as the Census 2005 case and to depict the analytical outputs in either chart of spatial format. The technology used to develop the mapped output required knowledge of GIS and of spatial tools which serve as add-ons to the base software. The result is relatively easy to use and allows users who have no knowledge of mapping to actually create their own maps and to save that data for their particular reports.

In an Archive

The following section will give a brief summary of a PhD research (Scicluna, 2004) conducted between 1998 – 2004 utilising the archives of the Malta prison. This research analysed data between 1850 and 1851. Access is not always easy, and in this case, due to the sensitive data that might be stored in the archive access was problematic. The national archives were worried about giving permission to the researcher to consult the prison ledgers because they were afraid that there was some information that could not be published or that should not be made available to the general public. Consequently, they asked to researcher to ask the prison authorities for their permission.

Even here the authorities were not at ease granting permission, mainly because they were not sure what was in the archives. At the end it was the permanent secretary for Home Affairs who granted permission with the condition that names and information could not be used in a way that would identify or incriminate people. Therefore, the researcher could only identify individuals who were previously mentioned in public documents.

When the idea of conducting a research utilizing the public archive was conceived, the researched had assumed that the prison archive was well organised. This was not so and the first thing was the actual organisation of the archive before any research work could start. Although this was hard work it enabled the researcher to really know what was kept in the archive. It is not usual that researchers are allowed to organize archives as this is usually done by people working in the archive itself. Undoubtedly, the researcher was known as a trust-worthy person since the archive authorities are very protective of their documents. Naturally, they expect no papers to be damaged or found missing. Maintaining an archive is indeed a very sensitive job.

It is important that the researcher knows that the records that are being analysed are the real records and not some imitation. In this case, the passage of the prison records could easily be traced. They had been kept in a room in prison until 1982, when the then Director of Prisons, Mr Ronald Theuma, asked a prisoner to organise them. This prisoner was neither a librarian nor an archivist therefore he did his best and categorised the material by subject. After two prison riots in 1992, all the ledgers were packed quickly in boxes and sent to Fort St. Elmo for storage. Storage here was not ideal. They were kept in two closed rooms, elevated from the floor by wooden planks, but humidity in the rooms was very high. The rooms overlooked the Grand Harbour in Valletta resulting in a combination of humidity and sea spray that could damage and corrupt the ledgers.

DePew (1991:45) points to the hazards of storing documents in high humidity or areas exposed to the sun. For most of their existence, these documents have been kept in these conditions. High humidity is to be found in both the prison and Fort St. Elmo. Sauna-like temperatures and high humidity cause chemical reactions in the paper that quicken deterioration. Due to the type of material used and the conditions under which they were kept, some ledgers proved illegible. After six years of inadequate storage, on the 15th April 1998 they were finally transferred to the National Archives of Malta (NAM), still packed in boxes.

Storage in the NAM was better. The material was raised on iron planks and dehumidifiers were in use, but staff shortages meant that the records were being conserved but not processed. The researcher first saw the prison archival material in this state. The researcher decided to wait for six months in order to enable the finalization of the NAM categorisation of the records. This did not take place and in December 1999, the researcher took the decision to open the boxes and sort the documentation. This was an unexpected and unwarranted additional research task. It was complex, time consuming and laborious – and for the first time 'hard labour' took on a new meaning!

The volume of archival material was staggering. A sizable room was packed with 303 large boxes (Figure 12.13). The names of the ledgers alone amounted to 58 A-4 pages. With a span of almost 150 years ledgers tended to be of various shapes and sizes. The easiest were the admission ledgers. These had kept the same size and almost the same shape. They are big ledgers (A3 size or bigger) and thick, therefore heavy. There are 214 such ledgers. The ledgers were haphazardly placed, and there was no continuity concerning relevant year(s), categories of information and even ledger titles.

Figure 12.13: Prison Archives in Boxes

Under the direction of the officer in charge of the NAM the researcher started the process of sorting the archive. The first process was to divide the ledgers into various types of categories. Once this was done, a process, which took more than a hundred hours with the help from employees of the archive and some friends, the researcher could embark on the second phase. This involved organising each category according to the way it was produced, including tagging the ledgers with their date of origin and placing them in the appropriate order (Figure 12.14). A parallel process involved the recording of ledgers on separate sheets of paper, so that at the end of the process one had all the information ordered.

After a year, working an average of 15 to 20 hours per week, the researcher could actually begin data collection. The total absence of archival classification at the start of the project was a major research problematic. Each prison ledger is given a code, always beginning with CCP (Corradino Civil Prison) followed by a slash and a number representing the category of the ledger it is referring to, and then another number referring to the ledger itself. For example the 1850 admission register would be referred to as CCP/01/01. As not all the categories have been numbered in certain cases the researcher would not be able to give the number given by the archive of a certain document. In these cases the name and the year of the ledger would be given.

Figure 12.14: The Sorted Archive

The documents being analysed were the prison documents as movement from one site to the other was always carefully monitored. For more than a century, they had piled up in prison and when they were moved, they were moved to safety. Therefore, they were surely authentic. They were also credible in that it was possible to compare some rough or draft ledgers with the final version, and see corrections appropriately made. Another example includes an admissions ledger where pages had been bound incorrectly; giving the misleading impression that offenders had been imprisoned before having been convicted. These documents were also representative of the era as the information found in the ledgers was substantiated in other documents found in the archives from the same era. The documents were also meaningful in that what was written made sense. Therefore the researcher could conclude that these documents were authentic, credible, representative and meaningful.

Safe storage is always a priority. When the ledgers were kept in prison, they were housed in a central area with limited access. Once they were moved a list was drawn up of the individual ledgers. When the researcher opened the boxes, all the ledgers were checked against this index. They all matched. The documents were clearly authentic but storage was not always ideal. Safety was threatened by prison riots in 1992. There was some fire damage. The records were removed to Fort St. Elmo, Valletta, but high humidity led to physical deterioration, although they could still be readily consulted. There are very few gaps in the sequence of journals. Most ledgers survived and there is internal

consistency in entries from one journal to another. This confirms representativeness (Scott, 1990:106), but there are some puzzling anomalies. For example the prison regulations of 1850 refer to the chaplain's journal but no record of this was found prior to 1887. This means either that in the first 37 years no material was produced or that if it did it was lost or destroyed.

Not all the documents contained important data. Some documents were produced to meet simple organisational needs (Scott, 1990:11) and others were so heavily predicated or taken-for-granted assumptions about every day routines that they tend to be banal (Scott 1990:123). An example of the latter includes 'Nothing to report' entries in the superintendent's journal. Conversely, other entries report daily routine in painstaking detail. Human idiosyncrasies add to data richness (which supports authenticity), although not necessarily ease of interpretation.

The admission ledgers offered near uniform data collected over 100 years. There were omissions, the most frequent being about school attendance. For data analysis purpose, certain data categories were collapsed for example length of imprisonment into 30-day periods (because most prisoners were sentenced to short terms) and age into five-year bands. Analysis by type of crime was problematic because over one hundred offences were recorded. Only the most common offences were used for the analysis of data. The decision to collapse categories was dictated mostly by analysis and data presentation problems.

The Maltese Islands were divided into six regions (in line with the division of the Maltese islands used in the 1985 census) for the purpose of analysing prisoners by area of residence. There was also a provision for prisoners coming from a military or naval base, a commercial ship and persons coming from abroad. Some place names in Malta and Gozo have either undergone a change or have disappeared. Some examples are Macabiba which is today's Mqabba, Garbo and Caccia two villages in Gozo that do not exist anymore. Spelling of Maltese words varied considerably, therefore deciphering the hand written script was sometimes problematic.

In the 19th century recidivism by receipt of prison sentence was recorded on each occasion until the tenth prison term, when the term 'several' is routinely used. From 1931 onwards a distinction was made between first-time and second-time offenders, whilst all others were classed as 'several'. The professions of inmates were classified as professional (including merchants), skilled (such as bakers and farmers), semi-skilled (such as servants and bus drivers), unskilled (such as hawkers and street sweepers), housewives, unemployed, beggars and school children. Some job names, such as carter (someone who constructed or repaired carts) are no longer in use.

The prison archive was indeed very rich in information. However other information was to be found at NAM, which continued to give the global picture of what was happening in the prisons during the period under study. The office of the chief secretary of state was the administrative office of the Civil Government. All departmental, consular, ecclesiastical and individual correspondences were channelled through it. This was the fulcrum where all orders from the central government to the various departments originated. A central filing system kept records of these letters. This office originated with the first government on 5th October 1813 and continued to function until October 1921 when the mandate of the self-government gave responsibility for administrative papers to the Maltese government. Destruction of documents was common, as with the destruction of letters in 1870 to make room for later records; and others were lost through ignorance, accidents or bad storage (Scott, 1990:25).

The dispatches from and to the secretary of state cover a period from 1800 to 1901. These registers are copies of the correspondence between the governors and the secretary of state or the civil commissioner. Therefore they give a good idea of the problems being faced by the prison authorities. Copied documents pose problems as they could have been mistranscribed or some entries left out (Scott, 1990:102). The researcher tried to solve this problem by seeing various copies of the same documents where possible and assessing that the instructions received made logical sense when compared with the knowledge gained. Some misprints were identified in this manner, for example: the entry in the admission register gave a two year old boy, imprisoned for theft. This was surely a misprint. Probably a distracted clerk wrote 2 instead of 22.

Other document consulted at NAM were, the lieutenant government office files, which cover a period between 1835 and 1847. They document the administrative work of the lieutenant government's office. These files are the most substantial having an average of twelve ledgers of about 150 letters each for each year. Most of these files do not have indexes, therefore when looking for correspondence about the Corradino prison one had to laboriously leaf through all the files.

The blue book, a collection of yearly reports written by Maltese institutions (e.g. health, education, prison and so on) was also found to be useful. It covers a period from 1800 to 1939. After this period the annual departmental reports perform the same function. The blue book was of particular importance as it contained a collection of information on the prison and about the year's activity, giving the researcher the information needed for the construction of the initial framework for the analysis of the year. Annual or quarterly reports were produced by the Board of Visitors to be submitted to the lieutenant government office, and by the inspector of prisons, the superintendent/director of prisons, by the chaplain and by the chief medical officer to be submitted to the Board of Visitors.

In any archive there are numerous ledgers, some are important for research others are routine documents that do not give much information. The researcher has to decide which ledgers are the most important and what to analyse. Looking at every bit of paper will prove impossible, laborious and not very fruitful. Searching within an archive is like conducting an investigation. You will never know what you will find, some days will be laborious and unfruitful in other days you will strike a pot of gold. Be prepared to work hard and spend hours going through ledgers but at the end the results are worth it.

Conclusion

The two case studies presented here depict drastically different modes of data collection, contrasting a real-life collective exercise such as the Census with the solitary collection in an Archive alongside 100 year-old documentation.

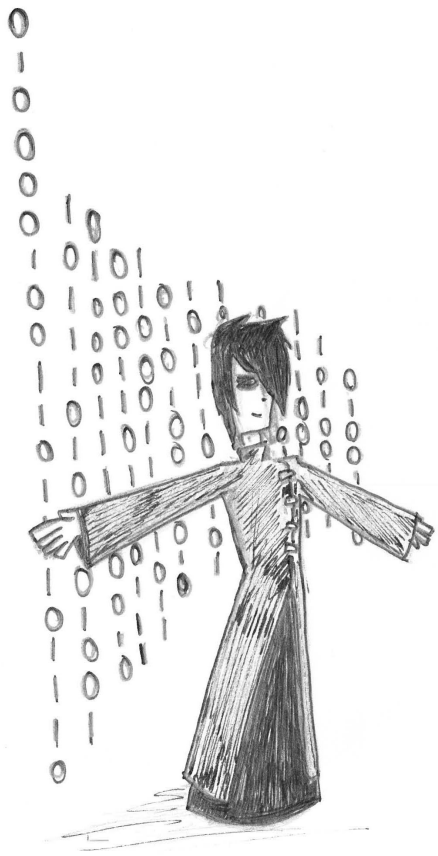
Both had the ultimate aim of carrying out research within the rules and structures that such studies should be undertaken and both reached their targets. Also both fall under strict data protection rules which must be observed such as the protection of the individual household and the protection of the individual incarcerated person.

The data gathering process is a laborious task as seen above and barriers can come up at any moment, however with a determined approach the targets will be reached.

Questions (refer to Appendix for the answers)

1. How frequently is the Census of Population and Housing of the Maltese Islands carried out?
2. Who conducts this laborious survey (the Census of Population and Housing of the Maltese Islands)?
3. List the six main steps done by the Malta National Statistics Office before the actual census starts.
4. Who are the people (mention just their roles/official nomenclature) that comprise the Malta Census Management Team?
5. List the four main steps a researcher would take when using the Census for research.
6. Briefly describe the main problems, an archival researcher might encounter.

Chapter 13
Data Sources



'Science in itself' is nothing, for it exists only in the human beings who are its bearers. 'Science for its own sake' usually means nothing more than science for the sake of the people who happen to be pursuing it.

Rudolf Virchow

'Standpoints in Scientific Medicine', *Disease, Life, and Man: Selected Essays* (1958), 42.

This Chapter lists data sources and identifies those organizations that gather this data. Whilst it attempts to list as many organisations and links as possible, the list is not exhaustive and requires the user to update it through online searches and other data source.

Analogue – a library/archive approach

The existence of different archives in Malta

In this section we will give a brief explanation of the different archives that exist in Malta. The information has been sourced from Charles Farrugia's book "L-Arkivji ta' Malta" (2006).

The National Archives of Malta holds documents from 1530 till today. In the former Santu Spirito Hospital in Rabat, Malta we mainly find document pertaining to the British Period, while the Banca Giuratale in Mdina holds documents from 1530 until 1899 and the archives in Gozo, situated in Victoria hold documents from 1560 till the present. All these fall under the National Archive holdings. There is also the Bibjoteka in Valletta which holds documents from 1107 till about 1800s.

The Department of Information also holds a number of important documents. It has media releases starting from 1957, the Government Gazette holdings start in 1813, an archive of films starting from 1959 and a photo archive starting from 1970.

Other public archives which might be of interest are:

- The Archives of the Notaries that cover the period from the 15th Century up to the present;
- The Public Registry Archives expanding the period from 1863 till today;
- The Archives of the Courts starting from 1900;
- The Archives of the Lands Department;
- The Records and Archives of the Department of Works that start from 1800;
- The Parliamentary Records starting from 1849;
- The University of Malta Archives which starts from the 11th Century;
- The Public Broadcasting Services Ltd Archives which commences from the 1970s;
- The Central Bank Archive which begins in 1964;
- The Medical Records starting from 1978;
- The National Museum of Arts beginning in 1798;
- The centre of documentation for Teachers starting from 1851;
- The Archive of the Administration of Burials;
- The Enemalta Archives commencing from 1853; and
- The Maltacom Archives starting from 1943.

The Church also holds a number of archives. The Archive of the Cathedral of Malta holds documents from the 11th Century onwards while the Archbishop's Curia holds documents from the 16th Century to the present, the Bishop's Curia in Gozo commenced in 1554 and the Gozo Cathedral archives starts from 1623. There are other archives held by the church such as the Archive in the Wignacourt museum and the various archives held by the diverse religious orders such as the Dominican Archive which holds documents starting from the 15th Century and the Archives of the Sisters of Charity which holds documents from 1868 to the present.

There are also a number of private archives on the Maltese islands such as the Archives of Dr Albert Ganado which hold material from 1296 till the present, the archives of "The Times" which start in 1930 and the archives of the Lanfranco family which starts in 1540.

There are a number of archives held abroad which are important to the researcher venturing in archival research about Malta. If you are interested in researching the British period the most important foreign archive is The National Archives situated near Kew Gardens in London, England. Other archives, found in England that might be useful for the researcher are: the Family Records Centre, The Historical Manuscripts Commission and the British library to mention a few. In Italy one finds the Magistral Library

and Archives of the Order of Malta, while in America the National Archives and Records Administration might prove fruitful in some cases. For more information on archival research in Malta consult Farrugia's book.

In addition to archives, one can find other libraries which are managed by the relative institutions such that those held by the National Statistics Office, the Malta Environment and Planning Authority, the United Nations International Institute on Ageing, the National Archives and other agencies. These entities host a vast amount of publications and data which is made available for public use.

Specialised Libraries

National Statistics Office
http://www.nso.gov.mt/site/page.aspx?pageid=29

A repository of over 10,000 statistical publications sourced from international and national statistics organisations which has resulted in one of the largest specialised libraries in Malta. A Collection Development Policy was developed by the NSO to ensure a coherent and sustainable development of the material available for the library users. In addition, a Web OPAC (Online Public Access Catalogue) is one of the tools envisaged in the running of the library services.

Malta Environment and Planning Authority
www.mepa.org.mt/library.html

Houses a collection of national and international books related to planning, the environment and legislation. Also houses publications produced by the same organisation. Also houses plans and old documents in plan format as well as having a legacy of maps and other data.

University of Malta Library
http://www.um.edu.mt/library

Has access to an online OPAC Search facility and e-Resources/e-Journals search facility

Division of Library and Information Studies, University of Malta
http://home.um.edu.mt/lis/

United Nations International Institute on Ageing
http://www.inia.org.mt/library.html
Has over 1000 books and digital databases comprising national and international material.

Malta Library and Information Association
http://www.malia-malta.org/

Various links to information and library services

Malta Public Libraries Online Catalogue
http://opac.library.gov.mt/

MCAST Library
http://www.mcast.edu.mt/llrc_aboutus_generalinformation.asp

Department of Libraries
http://www.libraries-archives.gov.mt/
Contains information on the National Library and the individual Libraries in Malta and Gozo

National Archives
https://secure2.gov.mt/nationalarchives/

National Archives Council
http://www.doi.gov.mt/EN/bodies/councils/national_archives.asp

National Book Council
http://www.ktieb.org.mt/

Schools' Library Service
http://www.education.gov.mt/edu/edu_division/student_services_sls.htm

Heritage Malta
http://www.heritagemalta.org/

Department of Information
http://www.doi.gov.mt/

Istituto Italiano di Cultura - Biblioteca
http://80.22.205.87/bwnet/Frameset.asp?OPAC=ICLV

KOPJAMALT - Copyright Licensing Agency for the Maltese Islands
http://cwebdesign.com/kopjamalt/

Paolo Freire Institute, Malta
http://www.jesuit.org.mt/justice/freire.html

The Malta Chamber of Commerce, Enterprise and Industry
http://www.maltachamber.org.mt/content.aspx?id=186719

Digital – online

Digital data sources are limited as very few national agencies supply their data online, however the list is growing as legislation on Freedom of Information and Access such as the Aarhus Convention are implemented. On an international level, data is made available as well as a variety of tools. The following list highlights some main Maltese and international datasets and links to agencies who might offer the researcher information as well as raw data as available.

Throughout the book, specific theme-related tools were identified. The following lists a number of datasets and tools that researchers would find useful for their studies.

- **Statistics**

StatPages
http://statpages.org/

A comprehensive list of digital statistical tools

National Cancer Institute
http://www.cancer.gov/statistics/tools

Has theme-related databases and includes various statistical tools

StatTrek
http://stattrek.com/Tables/StatTables.aspx

Comprises a variety of tools and online calculators

Rice Virtual Lab in Statistics
http://onlinestatbook.com/rvls.html

A wide range of tools available even for download

- **Data Repositories and Map Servers**

National Statistics Office Links to Organisations
http://www.nso.gov.mt/site/page.aspx?pageid=33

The NSO external links page contains a series of links to national, European and international organisations who hold information and who can be contacted for thematic research, list: 27 national agencies, 31 European agencies and 27 international agencies.

National Statistics Office Statistical Database (StatDB)
http://www.nso.gov.mt/site/page.aspx?pageid=31

The NSO online statistical database generates time-series data across a wide range of themes. The service is free and offers both view and download options.

Census 2005 Interactive Maps
http://www.sagisart.info/nso/census2005/
http://www.nso.gov.mt/site/page.aspx?pageid=351

99 interactive maps based on the 2005 Census of Population and Housing.

MEPA Site and Mapserver
http://www.mepa.org.mt/home?l=1

A map server depicting base maps, aerial imagery, planning and environment data layers, national and international project outputs. Website also includes a vast amount of documents and information for public consumption. The map server can be found on the right-hand side of the site.

Laws of Malta
http://justiceservices.gov.mt/lom.aspx?pageid=24

Site contains all the Laws of Malta. The main MJHA also contains sentencing databases
http://www.justiceservices.gov.mt/courtservices/Judgements/default.aspx

EEA CDR
http://cdr.eionet.europa.eu/
http://cdr.eionet.europa.eu/mt

Contains all the Malta (and EEA members) dataflows to the European Environment Agency, the DG Environment, related EU obligations, Eurostat and OECD material as well as other international flows based on conventions inclusive those pertaining to the United Nations.

EUROSTAT
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home

One of the main sources of information at European level is the EUROSTAT, of which the Malta National Statistics Office is part of. A European Statistical System (ESS) publishes comparable statistics at EU level.

EUROSTAT Statistical Database
http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

EU online database enabling querying within the statistical themes and which also has download services and hosts metadata structures.

Sourcing data from these sources and others is encouraged. Of course, researchers are advised to be specific in their requests to such agencies and to avoid such requests as: “anything on flora and fauna”. This request would be too generic and unrealistic, considering that the databases may hold hundreds, if not thousands, of parameters. Most of which will not be of use to the researcher.

Plan your requests well. These should be based on the literature or project requirements review. Researchers should submit specific requests such as: “the number of new species identified between the years 1950 and 2010” or “the areas under developments between 2005 and 2010”.

Such requests reduce the research time taken up by the agencies and also focus the researcher’s field of research. In addition do check whether such data is already available online at the agency’s website.

Questions (refer to Appendix for the answers)

1. John is compiling a study on the British period in Malta. Which are the best archives for him to visit?
2. Melinda is researching particular events that occurred in Malta and Gozo between 1530 and 1899. Which are the best archives for her to visit?
3. Francesca is researching particular events that occurred in Malta between 1107 and 1800. Which are the best archives for her to visit?
4. What would one find at the Department of Information (Malta)?
5. List at least five other public archives in Malta.
6. List the five church archives (in Malta) mentioned in the text.
7. List the three private archives (in Malta) mentioned in the text.
8. Give at least two examples of English archives that could be of interest to a researcher compiling a study on Malta and Gozo.
9. Give at the two examples (mentioned in the text) of Italian archives that could be of interest to a researcher compiling a study on Malta and Gozo.
10. Give at least one example of American archives that could be of interest to a researcher compiling a study on Malta and Gozo.
11. Give at least four examples of important libraries in Malta that could be of interest to a researcher compiling a study on Malta and Gozo.
12. Briefly describe how a researcher's request for data should be.

Chapter 14
Ethical Issues

Do what you can where you are with what you have.

Theodore Roosevelt

The Homiletic Review, Vol. 83-84 (1922), Vol. 84, 380.

Empirical research generates questions about an area or topic of interest and ultimately whilst conducting research we aim to provide answers. In doing so, researchers aim at gaining scientific knowledge about an area of interest; to link theory with scientific research. For example, researchers may try to understand why people resort to criminal activity after all. The use of applied research turns out to be practical in criminal justice as well as other social sciences. Examples include 1) the relationship between drug use and crime 2) stress and the police 3) interviewing tactics during investigative interrogation 4) memory and eye witnesses 5) sentencing – judge vs. jury amongst other areas of interests. In other words, we are examining the criminal justice system (court, police prison), the individuals who infringe the laws, victims (rape, sexual abuse, domestic violence, assault, murder, fraud etc) and families of respective victims. Whatever the methodology the researcher decides to use (interviews, surveys, case studies, field studies, archival research etc), there are ethical considerations that need to be safeguarded. As researchers we need to adhere to this rule "Inflict No Harm"; the interests of the participants need to be protected throughout as participants have rights and human dignity (Dantzker and Hunter, 2000).

What is Ethics?

The field of ethics, or what is also known as moral philosophy, provides us with recommended guidelines of what is right and what is wrong (International Encyclopedia of Philosophy, 2010) . These guidelines aim at securing the nature of human well-being. As researchers we are creating new knowledge, but in doing so we cannot adhere to the principle *the end justifies the means*. Ethics stands to be one of the most crucial areas of research. Two cases that attest breach of ethics are presented below.

Case 1: The Stanley Milgram experiment

The Stanley Milgram experiment was designed to explain the Nazi camp atrocities of World War 2. Milgram was after a particular research question, that is: to show how subjects obey morally wrong instructions whilst aware that in obeying such instructions (commands to give electric shocks whenever the learner made a mistake) they are inflicting physical harm and emotional distress, following the pain (Milgram, 1963). In the 1960s Milgram's experiment seemed realistic but today critics of the Milgram's experiment claim that this should have not been allowed in the first place. The main ethical concerns for the Milgram's experiment arise from 1) deception – subjects were told that they were participating in a learning experiment 2) physical harm was inflicted 3) some sort of emotional distress could have been inflicted since *teachers* were aware of the harm inflicted to the *learners.*

Case 2: The Stanford Prison Experiment

The Stanford Prison experiment was carried out by the undergraduate Philip Zimbardo in 1971. In this experiment, with the prison guards and convicts as subjects, Zimbardo wanted to show how guards and convicts behave according to the perceived predefined roles whilst disregarding personal judgments and values (Zimbardo, 2010). The findings of this experiment were criticized for validity but also for the ethical issues including 1) adverts in newspapers promising rewards for participants 2) prisoners suffered humiliation and punishment 3) the use of forced exercise and physical punishment by guards increased 4) prisoners were made to sleep on cold floors since mattresses were withdrawn 5) toilet use became a privilege not a basic human need and right 6) prisoners had to clean toilets without the use of gloves 7) sexual humiliation – prisoners were often stripped 8) prisoners manifested signs of emotional distress.

Criteria for Ethical Research

1. **Informed consent**

All researchers need to carry out their studies after obtaining consent from participants. However, participants need to be provided with full information about the research (its aims and outcomes), before giving their consent to the researchers. Also, participants have the right to withdraw from the study at any stage of the research (British Psychological Society, 1997). In the case of children (legally defined as minors); people suffering from a learning disability or an impairment which hinders their ability to make judgments and decisions and the elderly or emotionally frail, researchers need to determine who has the legal capacity to give consent on their behalf. Example: you are interested to study the consumption of alcohol among youths so you opt to use questionnaires with students in form 3 and form 4 as participants in your study. To be able to do so by ethical means, you need to gain consent from Education Division as well as countersigned consent forms from parents of participants.

2. Confidentiality and Anonymity

All information acquired through research needs to be treated as strictly confidential. In addition, the identity of individuals and organizations/institutions needs to be kept anonymous (British Psychological Society, 1997).

Imagine this scenario; we are studying crime patterns in each locality in the Maltese Islands. The researchers need to make sure that the information provided does not deliberately or unintentionally reveal the identity of individuals. In the case of Comino, the results could yield sensitive information about the only few residents who reside on the small island. So, in this case, the researchers could make use of local council boundaries to counteract this ethical problem. This is done so as to protect the privacy of individuals and to avoid exposing subjects to psychological, social and economic harm.

Whilst conducting research, we can create a database, make use of audio visual recordings and photographs. Researchers need to take the necessary steps to ensure that the information remains identifiable as long as it is necessary throughout the course of the investigation and this should be done with discretion. Moreover, this information needs to be rendered anonymous when personal identification of data is deemed no longer essential. In circumstances where anonymity and confidentiality cannot be guaranteed, subjects need to be fully aware of this before consenting to participate (British Psychological Society, 1997).

3. Objectivity

Biases, beliefs and personal values do contaminate and compromise the validity of research. At this stage, one would ask how the social scientist can escape from his/her own mind. According to Weber (Gerth and Mills, 2001), the social scientist should be aware of his/her own values and perspectives and then would be in a position to conduct a value free investigation. The researcher needs to adopt a neutral stance. In other words, if two independent researchers were to conduct the same study separately, they should come up with the same results. Scientific research is interested in facts and unbiased interpretation of the data created, rather than the interpretation of facts (Wimmer and Dominick, 1991).

4. Deception

It is unacceptable to provide participants with misleading information or with partial information about the study they have consented to act as subjects (British Psychological Society, 1997). On the other hand, a researcher can decide to work *undercover* as a volunteer in a drug rehabilitation institute to study the effectiveness of drug rehabilitation programmes. This type of research turns out to be controversial since participants have been deceived and also informed consent was not granted.

Referencing

Scientific research distinguishes itself through the fact that knowledge provided by researchers is made public. The reviewing of existent literature (journal articles and books) and relevant theories is a key factor to producing a valid piece of work. This section in the research report summarises all other research work that has been carried out in the field (Wimmer and Dominick, 1991). All information provided here should be relevant to the area being investigated and also correct. Also understanding how to use and acknowledge the work of others is essential to learn how to avoid plagiarism. Researchers who fail to include a full and adequately presented reference section can be harshly penalised.

Plagiarism

Plagiarism is “the unauthorized use or close imitation of the language and thoughts of another author and the representation of them as one's own original work” (http://dictionary.reference.com/browse/plagiarism).

Plagiarism can occur not only in the use of text but also in the use of tables, illustrations, maps, diagrams … etc … The internet has become a popular and accessible tool for most of us. So you might be tempted to make use of paragraphs read on-line, edit and change a few verbs and present the paragraphs as if they were your own. This is a classical example of plagiarism. So make sure that every text cited is fully referenced and also any captions are fully referenced; remember any ideas that are not your own need to be fully referenced. Example: you are aware that in the Maltese Islands the population stands to be 404, 962. However, this data has been collected by the National Statistics Office in the last census exercise! Thus, you need to acknowledge and adequately reference studies so as to avoid plagiarism. Also, universities provide guidelines for students so, when they still plagiarise, they have no excuse! The University of Malta guidelines can be accessed through this URL addresses: http://www.um.edu.mt/registrar/regulations/general.

How to compile a reference section

In this section, guidelines are provided so as to help researchers compile a reference section that is requested at the end of a scientific report. All references should be arranged alphabetically by the surname of the first author. **The guidelines outlined in this chapter are based on the guidelines provided by the American Psychological Association (2002).** For a free tutorial about the use of APA style please go to:
http://www.apastyle.org/learn/tutorials/basics-tutorial.aspx

Two frequent questions posed by student researchers are: (1) What if the same author has different publications? (2) What if the same author has two or more publications in the same year? In the first case, the sources of evidence need to be arranged according to the year of publication – with the earliest one first. For example:
Wagstaff, G.F. (1981)
Wagstaff, G.F. (1982)

In the second case one needs to use the lower case letters just after the year of publication so as to distinguish between the two different sources.
Gelsthorpe, L. (1985a). *Gender Issues …*
Gelsthorpe, L. (1985b). *Girls Crime and…*

Researchers want their hard work to be recognised, however may encounter sources where the author signs as anonymous or anon. This can still be used as evidence. Also, in consulting documents, such as the Census data mentioned previously, one would come across a document with no identifiable author. In this case, the data has been gathered by the National Statistics Office during the last Census exercise. When consulting documents compiled by institutions and organisations (government and NGOs), this occurs frequently. The following examples will guide you through referencing.

Anonymous, 2010 ….

National Statistics Office, 2007

Federal Bureau of Investigation (1986). *Age- specific Arrest Rates and Race-specific Arrest rates for Selected Offences 1965-1985*. US Department of Justice. Washington DC: Government Printing Press.

Referencing books

The use of books is a valid research tool as it exposes one to theories, knowledge and also scientific findings. Also, these are readily accessible to students (books are provided to students on loan by most University libraries). However, one has to clearly distinguish between an edited book, a revised edition and books that have more than one edition.

One author book:
Ainsworth, P.B. (2000). *Psychology and Crime: Myths and Reality*. Essex: Pearson Education Limited.

A third edition book:
Wimmer, R.D., & Dominick J.R. (1991). Mass *Media Research: An Introduction* (3rd ed.). United States of America: Wadsworth Inc.

A revised edition book:
Bass E., & Davis L. (2003). *Beginning to Heal: A First Book for Men and women Who were Sexually Abused as Children* (Rev. ed). New York: Harpers Collins Publishers.

An Edited book:
Jackson, J.L. & Bekerian, D.A. (Eds.). *Offender Profiling: Theory, research and practice*. Chichester: Wiley.

However, we often make use of a chapter in a book rather than the entire book. Then referencing should be as follows:
Example:

A chapter in a book:
Gudjonsson, G.H. and Copson, G. (1997). The role of the expert in criminal investigation. In J.L. Jackson and D.A. Bekerian (Eds.), Offender *Profiling: Theory, Research and practice* (pp.61-76). Chichester: Wiley.

Sometimes researchers need to make use of legal sources to consolidate arguments. These sources need to be considered as works that have no identifiable author example:

The Probation Act: Chapter 446 (2002). The Laws of Malta.

Referencing Journal articles

Consulting and reading journal articles enriches the literature review and exposes one to findings of other researchers in the field. Referencing should follow this procedure.

Example:
Crombag, H.F.M. (1994). Law as a Branch of Applied Psychology. *Psychology, Crime and Law*, *1*, 1-9.

However if a journal has more than six authors, use et al. The following example will help you.

Foley, D.L., Pickles, A., Simonoff, E., Maes, H.H., Silberg, J.L., Hewitt, J.K., et al. (2001). Parental concordance and Comorbidity of psychiatric disorder and associate risks for current psychiatric symptoms and disorders in a community sample of juvenile twins. *Journal of Child Psychology and Psychiatry*, 42, 381-394.

Another frequent occurrence is the use of abstracts of journal articles and not the entire journal as these are provided free by universities' online libraries. In this case, one has to indicate clearly that the source is an abstract.

Example:
Kovandzic, T.V., Vieraitis, L.M. & Boots, D.P. (2009). Does the death penalty save lives? New Evidence from state panel data, 1977 to 2006 [Abstract]. *Criminology and Public Policy*, 8 (4), 803.

Information from secondary sources

This happens when you read about a study in a book, site or an article but you do not have the original source. Use the 'cited in' rule.
For example:
"In a study conducted by Moffit (1993, as cited in Farrington, 1995) found that.."
"Research shows that … (Moffit, 1993, as cited in Farrington, 1995)"

However in the reference section only the details of the secondary source are to be listed.

Example:

Farrington, D.P. (1995). The Development of Offending and Antisocial Behaviour from Childhood: Key Findings from the Cambridge Study in Delinquent. *Journal of Child Psychology and Psychiatry*, *36*, 929-964.

Internet and Newspaper sources

The use of the internet has surged in the last decade. This information is public and easily accessible. However, all information, data, charts, imagery etc. derived from the net needs to be referenced, otherwise this would be a case of plagiarism. Referencing should be as follows and one needs to distinguish between a document provided by an organisation and online journals.

Example:
International Encyclopaedia of Philosophy (2010, July 14). *Ethics*. Retrieved July 14, 2010 from *http://www.iep.utm.edu/ethics/*

Example of a journal article retrieved online:
Kassin, S.M., Drizin, S.A., Grisso, T., Gudjonsson, G.H., Leo, R.A. & Redlich, A.D. (2010). *Police-Induced Confessions risk factors and Recommendations. Law and Human Behaviour, 34,* 3-38. Retrieved July, 17, 2010, from
http://www.williams.edu/Psychology/Faculty/Kassin/files/White%20Paper%20-20%LHB%20(2010).pdf

One may also consult local newspapers in collecting information about a topic of interest.

Example:
Chetcuti, K. (2010, August 29). Book lures readers like bees to honey. *The Sunday Times,* pp. 41.

Referencing Doctoral Dissertations; Encyclopedias and dictionaries

The use of dictionaries and reference books such as encyclopedias is useful, although one should be careful when to use them. Also the use of doctoral dissertations exposes one to knowledge and research carried out by other experts in the field.
Examples:
Waite, M. & Hawker, S. (Eds.). (2009). *Oxford Paperback Dictionary & Thesaurus* (3rd ed.). Oxford: Oxford University Press.

Formosa, S. (2007). *Spatial analysis of temporal criminality evolution: an environmental criminology study of crime in the Maltese Islands*. Un-published doctoral dissertation, University of Huddersfield, United Kingdom.

Unpublished Paper presented in a Conference

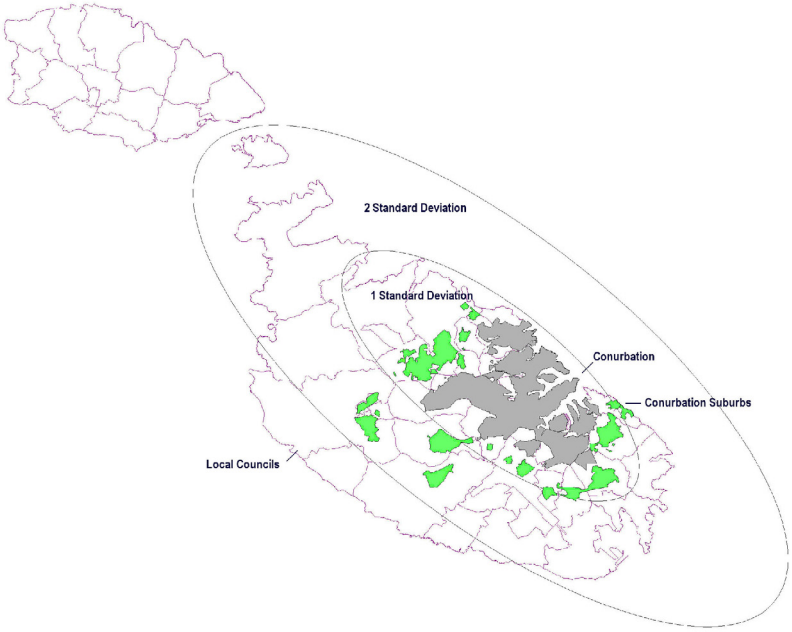
Attending conferences and symposia is also a good way of gaining knowledge and exposure to studies carried out in the field. The following is an example of referencing an unpublished paper presented in a conference.

Scicluna, S. (2010, April). The Development of Imprisonment for Women in Malta. *In Criminality in Island States*. National Crime Conference 2010 of the Malta Criminology Association, Malta.

Referencing drawings

All information that is not your intellectual property has to be referenced. This also applies to tables, figures, diagrams, photographs, maps and graphs. The following is an example of referencing a figure presented in a journal in an edited book.

Figure 1: Offence Standard Deviation Ellipse (SDe) at 1NNH and 2NNH[1]



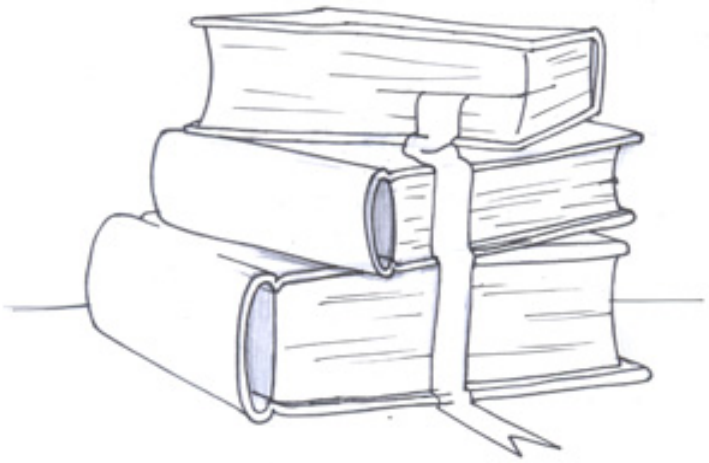
1 From "Maltese Criminological Landscapes – A Spatio-Temporal Case Where Physical and Social Worlds Meet," by S. Formosa, 2010, *Peer Reviewed Proceedings of Digital Landscape Architecture 2010*, p.115. Copyright 2010 by the Anlhalt University of Applied Sciences. Reprinted with permission of the author.

Ethics and referencing form the backbone of any research study. One cannot conduct a study without first checking the ethical issues involved. On the other hand referencing gives your study a professional orientation. It is not acceptable for one to conduct research without keeping in mind these two very important notions. Throughout this book we have seen a variety of methodologies that could be adopted in a research study. These, together with referencing and ethics make a research study complete.

Questions (refer to Appendix for the answers)

1. What does ethics (or moral philosophy) provide us with?
2. Why are ethical considerations important for researchers?
3. Mention two cases of research studies that attest breach of ethics.
4. List the four main criterions for ethical research.
5. What is plagiarism?
6. In Criminology, referencing should be compiled on the guidelines provided by a particular association. Which association is this?
7. With regards to referencing: what needs to be done if the same author has different publications?
8. With regards to referencing: what does one need to do if the same author has two or more publications in the same year?

Bibliography



Every discovery opens a new field for investigation of facts, shows us the imperfection of our theories. It has justly been said, that the greater the circle of light, the greater the boundary of darkness by which it is surrounded.

Sir Humphry Davy

Humphry Davy and John Davy, 'Consolations in Travel--Dialogue V--The Chemical Philosopher', *The Collected Works of Sir Humphry Davy* (1840), Vol. 9, 362.

American Psychological Association, (2002). *Publication Manual of the American Psychological Association* (5th ed.). Washington: APA.

Azzopardi, J., Camilleri Cassar, F. and Scicluna, S. (2006). *A comparative study on domestic violence between the islands of Malta and Sicily*. A projected financed by the RTDI programme.

Azzopardi, J. and Scicluna, S. (2003). *Youths in Malta: The criminal justice system.* Unpublished report.

Azzopardi, J and Scicluna, S. (2009). Criminal justice in Malta. In J. Cutajar and G. Cassar (Eds.). *Social Transitions in Maltese society*. Malta: Agenda.

Azzopardi Cauchi J., Formosa S., and Scicluna S. (2010). Criminal issues. In M. Formosa (Ed.). *Sustainable Development Strategy Dingli 2010*. Malta: Dingli Local Council.

Azzopardi Cauchi J., Formosa S., and Scicluna S. (2010). Technological enhancements to sustainable tourism. In M. Formosa (Ed.). (2010). *Sustainable Development Strategy Dingli 2010*. Malta: Dingli Local Council.

Baiocchi G. and Distaso W. (2003). GRETL: Econometric software for the GNU generation. *Journal of Applied Econometrics*, 18 (1), 105-110.

Boissevain J. (1965). *Saints and fireworks: Religion and politics in rural Malta*. London School of Economics Monographs on Social Anthropology No. 30.

Boissevain J., (1980). *A village in Malta: Fieldwork edition*. New York: Holt, Rinehart & Winston, Inc.

Borg M. and Formosa S. (2008). The Malta NPI project: Developing a fully-accessible information system, In S. Wise and M. Craglia (Eds.). *GIS and evidence-based policy making, innovations.* Florida: Taylor & Francis.

British Psychological Society. (1997). *Code of conduct, ethical principles and guidelines.* Leicester: British Psychological Society.

Burrough P.A. (1996). *Principles for geographical information systems for land resources assessment*. Oxford: Oxford University Press.

Calafato T. and Knepper P. (2009). Criminology and criminal justice in Malta. *European Journal of Criminology*, 6(1), 89-108.

Campbell B. (1993). *Goliath: Britain's dangerous places*. USA: Methuen Press.

Chainey S. (2004). GIS and crime mapping : Going beyond the pretty hotspot map. *Geomatics World*, 11, 24-25.

Chorley R.J. and Haggett P. (Eds.). (1967), Models in Geography. London. Methuen.

Clarke K. (1995). *Analytical and computer cartography*. New Jersey: Prentice Hall.

CMAP. (2002). Retrieved from http://www.nlectc.org/cmap/.

Codd, E.F. (1970). A relational model of data for large shared data banks: Communications of the ACM. *Association for Computing Machinery*, 13 (6), 377-387.

Conchin S., Agius C., Formosa S. and Rizzo Naudi A. (2010). Does visualisation of digital landscapes serve itself? How topographic, planning, environmental and other thematic information is integrated and disseminated via web GIS. In E. Buhmann, M. Pietsch, E. Kretzler (Eds.). *Peer Reviewed Proceedings of Digital Landscape Architecture 2010.* Germany: Anhalt University of Applied Sciences.

Craglia M., Haining R., and Wiles P. (2000). A comparative evaluation of approaches to urban crime pattern analysis. *Urban Studies*, 37(4), 711-729.

Craglia, M., Haining R. and Signoretta P. (2001), Modeling high-intensity crime areas in English cities. *Urban Studies.* 38(11), 1921-1942.

Dantzker, M.L., and Hunter, R.D. (2000). *Research methods for criminology and criminal justice*: *A primer*. USA: Butterworth-Heinemann.

Eley, G. (1980). Some recent trends in social history. In G.G. Iggers and H.T. Parker (Eds.). *International handbook of historical studies: Contemporary research and theory*. (pp. 55-70). London: Methuen and Company Limited.

Ethics (n.d.). In *International Encyclopaedia of Philosophy.* Retrieved from http://www.iep.utm.edu/ethics/.

Farrugia, C. (2006). *L-Arkivji ta' Malta*. Malta: PIN.

Fisher M.M., Nukamp P. and Papageorgiou, Y. (Eds.). (1990), Spatial choices and processes*. Studies in Regional Space and Urban Economics*, Vol 21. (pp. xix +317). Amsterdam: North-Holland.

Formosa S., (2000). Coming of age: Investigating the conception of a census web-mapping service for the Maltese islands. (Unpublished MSc dissertation). University of Huddersfield, United Kingdom.

Formosa S. (2003). Analytical tools for environmental management: Geographical information systems. *Proceedings from the International Conference on Sustainability Indicators*. Malta: Valletta.

Formosa S. (2005). Maltese islands landscape assessment. In S. Formosa, C. Agius and M. Sant (Eds.). *Use of Corine land cover and image 2000: Corine land cover applications in support of national and European policies (Vol 2)*. Copenhagen: European Environment Agency and Joint Research Centre (Ispra).

Formosa S. (2006). Crime in Gozo: A Spatio-temporal Analysis. *The Gozo Observer*, No.15.

Formosa S., (2010). Maltese criminological landscapes: A spatio-temporal case where physical and social worlds meet. In E. Buhmann, M. Pietsch, E. Kretzler (Eds.). *Peer Reviewed Proceedings of Digital Landscape Architecture 2010.* Germany: Anhalt University of Applied Sciences.

Formosa S. (2010). Crime mapping: A Gozitan scenario using the RISC methodology. *The Gozo Observer* No.22.

Formosa S., Aguis M., Grech A., and Pace C. (2007). Hi-end spatial information technologies: A case for mental health. *Psychiatria Danubina*, 19 (1), 27.

Fritz N. (2002). The Growth of a profession: A research means to a public safety end. In *Advances crime mapping techniques, Results of the First Invitational Advanced Crime Mapping Topics Symposium* held in June 2001. Denver Colorado: CMAP.

Gerth, H.H. and Mills, C.W. (2001). *Max Weber: Essays in sociology*. London: Routledge.

Grimshaw D. (1994). *Bringing geographical information systems into business*. USA: John Wiley & Sons.

Hagan, F.E. (1997). *Research methods in criminal justice and criminology.* (4th ed.). USA: Allen and Bacon.

Hakim, C. (1983). Research based on administrative records. *Sociological Review* 31 (3), 489-519.

Harries K.D. (1974). *The geography of crime and justice*. New York: McGraw-Hill.

Harris , M. B. (1998). *Basic statistics for behavioural science research*. (2nd ed.). USA: Allyn and Bacon.

Heywood I., Cornelius S. and Carver S. (1998). *An introduction to geographical information systems*. New York: Longman Ltd.

Hirschfield A. (2001). Decision support in crime prevention. In A. Hirschfield and K. Bowers (Eds.). *Mapping and analysing crime data: Lessons from research and practice*. London: Taylor & Francis.

Howard, J. (2002). How to de-mystify GIS by understanding the role of common-sensed GIS and blue collar GIS in public Safety. In *Advances crime mapping techniques, Results of the First Invitational Advanced Crime Mapping Topics Symposium* held in June 2001. Denver Colorado: CMAP.

Johnson, B. and Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches.* (3rd ed.). USA: Sage Publications.

Josephson J.R. and Josephson S.G. (1994). *Abductive inference: Computation, philosophy and technology*. Cambridge: Cambridge University Press.

Jupp, V. (1989). *Methods of criminological research.* London: Routledge.

Kitzenger, J. (1994). The methodology of focus groups: The importance of interaction between research participants. *Sociology of Health and Illness,* 16 (1), 103-121.

Knepper P. and Calafato T. (2010). Crime and Punishment in Malta. In M Aebi and M. Pina (Eds.). *Crime and punishment around the world. Greenwich*: Praeger.

Koster, A. (1984). *Prelates and politicians in Malta: Changing power-balances between church and state in a Mediterranean island fortress.* (1800-1976). The Netherlands: Van Gorcum & Comp. B.V.

Lengler R. and Eppler M. (2007). Towards a periodic table of visualization methods for management. *Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007).* Florida: Clearwater.

Longley P.A., Goodchild M.F., Maguire D.J. and Rhind D.W., (2001), *Geographic information systems and science*. England: Wiley & Sons Ltd.

Macdonald, K. and Tipton C. (1994). Using documents. In N. Gilbert (Ed.). *Researching social life*. (pp. 187-200). London: Sage Publications.

Martin, P. Y. and Turner, B A. (1986). Grounded theory and organizational research. *The Journal of Applied Behavioral Science*, 22(2), 141.

Malta Census. (1995). Accessed on http://www.mepa.org.mt/Census/index.htm.

Maxfield, M.G. and Babbie, E. (2006). *Basics of research methods for criminal justice and criminology*. USA: Thomson.

McNeill, P. (1994). *Research methods*. (2nd Ed.). London: Routledge.

Milgram, S. (1963). *Behavioural study of obedience*. *Journal of Personality Social Psychology*, 1, 127-134.

Morris D. (2002). *Peoplewatching: The Desmond Morris guide to body language*. Surrey: Vintage.

National Statistics Office. (2007). *Census of population and housing 2005: Volume 1.* Valletta: NSO.

National Statistics Office. (2007). *Census of population and housing 2005: Volume 2.* Valletta: NSO.

Openshaw S. (1993). GIS 'crime' and GIS 'criminality'. *Environment and Planning*, 25, 451-458.

Orwell G. (2008). *Nineteen eighty four*. UK: Penguin.

Parker, H.T. (1980). Concluding observations. In G.G. Iggers and H.T. Parker (Eds.). *International handbook of historical studies: Contemporary research and theory*. (pp. 433-436). London: Methuen and Company Limited.

Pease K. (2001). What to do about it? Let's turn off our minds and GIS. In A. Hirschfield and K. Bowers K (Eds.). *Mapping and analysing crime data: Lessons from research and practice*. London: Taylor & Francis.

Reeve. D. (1997). Data acquisition. UK: MMU.

Scicluna, S. (1997). Types of supervision and 'what works': *Proceeding from the conference* on *Promoting probation internationally*. Rome: UNICRI.

Scicluna, S. (2000). Malta. In A.M. van Kalmthout and J.T.M. Derks (Eds.). *Probation and probation services: A European perspective*. The Netherlands: Wolf Legal Publications.

Scicluna, S. (2001). Community service in Europe: Report of the CEP workshop. The Netherlands: CEP.

Scicluna , S. (2002). Substance abuse and domestic violence. In A. Bell and S. Arpa (Eds.). *Euro med networking in substance abuse: Aetiological and policy perspectives*. Malta: Sedqa.

Scicluna, S. (2004). The prison in Malta: 1850 – 1870 and 1931 – 1951. (Unpublished PhD Thesis). University of Leicester, United Kingdom.

Scicluna, S. and Knepper, P. (2008). Prisoners of the sun: The British Empire and imprisonment in Malta in the early nineteenth century. *The British Journal of Criminology*, 48, 502 – 521.

Scicluna, S. (2008). Malta. In A.M. van Kalmthout and J.T.M. Derks (Eds.). *Probation in Europe*. The Netherlands: Wolf Legal Publications.

Scott, J. (1990). *A matter of record*. Cambridge: Polity Press.

Valentino C. and Formosa S. (2006). Strategies and development of cartographic and GI issues in the Maltese archipelago. *The Cartographic Journal*, 43 (3), 250–255.

Wimmer, R.D. and Dominick J.R. (1991). *Mass media research: An introduction* (3rd ed.). United States of America: Wadsworth Inc.

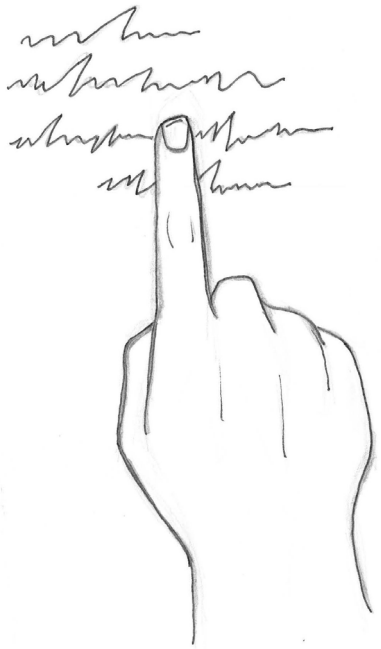
Wise S., Haining R. and Ma J. (2001). Providing spatial statistical data analysis functionality for the GIS user: The SAGE project. *International Journal of Geographical Information Science*, 15 (3), 239-254.

Worboys M.F. (1997). *GIS: A computing perspective*. London: Taylor & Francis Ltd.

Zammit S. and Mizzi R. (2010). The Census Process: From data collection to dissemination. *Delivered as a NSO presentation*. Valletta, Malta: NSO.

Zimbardo, P. (n.d.). *Stanford Prison Experiment*. Retrieved from http://www.prisonexp.org/.

Appendix – Questions and Answers



Keep on going and the chances are you will stumble on something, perhaps when you are least expecting it. I have never heard of anyone stumbling on something sitting down.

Charles F. Kettering

Answers to the questions asked in the chapters

Chapter 2

1. What are the two main points that you should keep in mind before deciding what your research topic should be?
   i) The area should interest you.
   ii) Ensure that you have access to the subjects.

   It is necessary to narrow down one's area of study. Ex: researching young people's behaviour during weekends.
   - Which type of behaviour?
   - Which locality?
   - Which part of the locality?
   - During which weekends?

2. The research question is formulated using two approaches: the deductive method and the inductive method. Briefly describe these two methods.

   i) The deductive method or Top-Bottom approach
   - Uses theoretical interpretations and logically interpretive prepositions to start.
   - Starts with 'Why?' certain behaviours occur to 'Whether?' they will occur.
   - Ex: the Mediterranean climate has a dry summer therefore it will not rain in July.

   ii) The Inductive method or Bottom-Up approach
   - Uses the observation of reality without any theory.
   - Constructs a number of indexes on which theory will be built later.
   - Moves from the 'whether' to the 'why'.
   - Ex: as it has never rained in July in Malta, next July it won't.

3. List the nine main steps of research design.

   Step 1: study the existing theories and works on the topic
   Step 2: define your study
   Step 3: literature review
   Step 4: formulate your hypothesis
   Step 5: research design
   Step 6: data collection
   Step 7: data analysis
   Step 8: drafting a report
   Step 9: presenting results

4. What is the empirical research (social scientific research) method?

   A method applied to understand social reality using LOGIC and OBSERVATION (Hagan, 1997).

5. Good research is based on objectivity. Very briefly explain this.

   Researchers must avoid being, even inherently/emotionally biased.

6. Briefly describe one major difference that exists between the social sciences and the natural sciences.

As human behaviour is based on free and individual choices, Social Science conclusions can never be 100% accurate. Conversely, in the Natural sciences, conclusions are based on hard facts.

In empirical studies, conclusions are based on interpretation. Whilst a subjective analysis of the data may be made, objectivity ensures that the rules of research methodology are adhered to resulting in the establishment of facts. This is a major difference between Social and Natural Sciences and can cause

some researchers to view the process as a stumbling block for the Social Sciences, though strict rules have ensured that the method is based on scientific fact.

7. Why is it very important for the researcher to collect and choose the right data?

Ideally the researcher should provide the readers with all the available data so they can reach their own conclusions/own interpretations. However, since subjectivity must come into play, some form of evaluation is necessary. Thus, the researcher must define which data is correct and necessary. To enable in-depth study of the data chosen (to get a wider view), the Researcher must focus on certain data and ignore other. Therefore, it is important to collect and choose the right data.

8. List the five main rules that researchers must take into account when conducting a research.

   Researchers must take the following rules into account when conducting a research:

   i) reliability
   ii) validity
   iii) credibility
   iv) causality
   v) representation

9. What do you understand by "sampling"?

   Samples are sets of targets (groups, data, persons, entities) that represent the population from which they are drawn.

10. List the five main types of sampling.

    i) Simple random sampling: one in which each person has the same chance of being chosen. Data can come from different sources.
    ii) Purposive or systematic sampling: when you need to target a group.
    iii) Cluster random or stratified sample: when you have to choose a number of people from the same place to minimise costs.
    iv) Disproportionate stratified sample: when you need to make sure that even minority groups are represented in a sample.
    v) Snowball sampling: when access to the sample proves difficult. You ask your contact to put you in touch with a subject and this subject, in turn, puts you in touch with another.

11. What do you understand by "sampling error"?

    There will always be a sampling error (since not everyone was included!). The larger the sample, the smaller the sampling error. A sampling error of 5% is acceptable.

12. What do you understand by "causality"?

    When a change in 'A' brings about a change in 'B'.

13. When does a perfect positive relationship between variables occur?

    When a change in 'A' brings about a clear and direct change in 'B' [A ↑ B ↑].

14. When does a perfect negative relationship between variables occur?

    When a change in 'A' brings about a clear and direct change in 'B' but the change may not be in the same direction [A ↑ B ↓].

15. What conditions enable the researcher to claim that there is a correlation between variables?

    When a change in 'A' brings about a change in 'B' but the relation is not necessarily even.

16. List the four main problems associated with using formal official data.

    i) Possible problems to access data.
    ii) Researcher has to work with what is available.

    iii) May not be (data) what the researcher actually needs (not collected for his/her purpose!).

    iv) May not be comparable with the researcher's other data (collected by her/him)

17. Very briefly describe the two main categories of research: qualitative and quantitative.

    i) Qualitative approach: observations, interviews, documentary analysis. In-depth interviews with a small number (especially with difficult-to-access populations). Qualitative research reports observations.
    ii) Quantitative research: bases research on a large sample. Yields statistical data which is usually analysed through a statistical package tool (SPSS). Quantitative research assigns numbers.

18. Triangulation is of paramount importance for archival research to be valid. List the four main types of triangulation.

    i) Data triangulation
    ii) Investigator triangulation
    iii) Theory triangulation
    iv) Methodology triangulation

19. Very briefly explain what you understand by "adduction" (with reference to archival research).

    Adduction is 'finding the best explanation of a set of data' (Josephson and Josephson, 1994:157).

20. List the three main types of official documents (with reference to archival research).

    i) routine: central in administration (ex: admissions)
    ii) regular: for everyday purpose
    iii) special: for a specific reason

21. Eley (1980) warns about a critical point in archival research. What is it and why does it happen?

    Often, it generates facts about interpretation. Why? Sheer volume of material makes it difficult to integrate data with theory.

22. Scott (1990) claims that the status and standing of the archive material has four sequential dimensions. List them.

    i) Authenticity (verification that the documents are original!).
    ii) Credibility (includes an assessment of potential and actual sources of error and distortion).
    iii) Representativeness (document is typical of another document from the same context).
    iv) Attribution of meaning (ensuring that the documentation reflects existing and relevant information about the subject).

23. List the two main problems associated with archival research.

    - Access and restrictions in viewing data.
    - Not all data is available. Maltese law makes official and personal data only available for researchers after a certain number of years (ex: prison data has a 30-year moratorium; official ledgers and personal data have an 80-year moratorium).

24. Briefly explain what case studies are and state their main problem.

- An in-depth study of a particular site, individual or occurrence to find some common interpretation or principle (Johnson & Christensen, 2008).
- An event or an individual is studied over a period of time.
- Main problem: No set rules but the researcher take notes (usually keeping a format in mind). The main problem revolves around generalisation where one assumes that the case study represents the whole population.

25. Survey research can be divided into two main categories: interviews and questionnaires. Briefly describe these two categories.

i) Interviews: questions asked orally (in person or by phone). Qualitative tool.
ii) Questionnaires: a set of questions respondents are expected to answer in writing. Quantitative tool.

26. List three main advantages of interviewing research participants.

- Useful to get the story behind a person's experience.
- Allows the researcher to dig out hidden or in-depth information about the subject.
- Questions are usually open-ended and elicit most information.

27. List the four main advantages of using questionnaires and the two main disadvantages of using questionnaires.

Advantages:

i) Can be filled in the respondent's free time.
ii) They can reach more people.
iii) They can be filled in privacy.
iv) The respondents remain anonymous.

Disadvantages:

i) Low response rate (30%).
ii) The respondents can't ask for clarifications.

28. Why should questionnaires be piloted (tested)?

Questionnaires should be piloted to make sure that the intended meaning and the way people understand the questions are the same.

29. List the three main types of data and very briefly describe each one, even if by simply providing an example.

i) Nominal: fits distinct categories (ex: male or female); only measures of central tendencies (the mean/median/mode) such as frequencies can be used.
ii) Ordinal: ordered in categories (ex: Likert scale).
iii) Interval: grouping (ex: ages 0-5/6-10/11-15 yrs…).

30. What are "focus groups"?

"Focus groups are group discussions organised to explore a specific set of issues" (Ketzinger, 1994:1).

31. List the four main problems associated with conducting focus groups and list the four main advantages reaped by conducting focus groups.

Main problems

i) Generalisation of findings.
ii) People tend to be reluctant to discuss certain issues in groups.
iii) The non-response rate. Those who didn't participate could've completely changed the outcomes.

Main advantages

i) helps the researcher identify the participants priorities and language.
ii) promotes discussion between participants.
iii) helps identify group norms and the working of the group.
iv) helps people listen and reflect on each other's ideas.

32. Briefly describe ethnography/participant observation.

- Qualitative
- The researcher spends time analysing and observing a group.
- Tool developed by anthropologists.
- Used to describe a cultural group.
- Can be conducted either overtly or covertly. The latter yields truer and richer data but has huge ethical issues.

33. List the main advantages of conducting ethnography/participant observation.

- Enabling the researcher to understand a reality, foreign to his/her culture.
- Subjects may be observed in their natural setting and it enables researchers to conduct a "study of social process" instead of being restricted to a mere "snapshot or series of snapshots" (Mc Neill, 1994:83).

34. List the main disadvantages of conducting ethnography/participant observation. Mc Neill, 1994:83)

- It is difficult for the researcher to remain detached from the situation (especially in covert research).
- Cannot be empirically tested and it is very difficult for the researcher to remain detached and unbiased.
- The presence of a researcher might alter the group's behaviour.
- This research cannot be generalised.

Chapter 3

(1) What 3 major developments changed the process of conducting research? Were research problems solved, or were they merely replaced by new problems? Mention some of these modern research-related problems.

The 3 major upheavals that changed the process of conducting research: (a) the introduction of computers in the 1980s; (b) the introduction of the Internet and the World Wide Web in the 1990s; and (c) the availability of raw, real-time data in the 2000s. Research problems just mutated. Now, we have too much data giving rise to 2 main problems: (a) gaining access to last-version information and (b) lack of know-how when it comes to interpreting data. In fact, data may: (a) not be readable; (b) not be comparable; (c) not be of a reliable format; (d) not be current; and (e) not follow standard research regulations. Therefore, data faces similar problems to those faced by the biblical Babel: (a) too much data; (b) easily-abused choices of statistical measures; and (c) over-reliance on online data and technologies.

(2) What are triangulation studies?

Triangulation studies are researches in which qualitative and quantitative research tools merge.

(3) What is DIKA and what do the letters stand for?

DIKA is a mnemonic. It stands for the research process: Data-Information- Knowledge –Action.

(4) What is the W6H in relation to conducting research? How are the W6H elements helpful?

The W6H represents the questions researchers need to ask/address before embarking on a research, namely: (a) Who? (target group); (b) What? (research question); (c) When? (Indicator); (d) How? (method); (e) Why (linkages); (f)Where? (location analysis); Why not? (controversial/cross thematic). The W6H elements help one to understand which research method should be employed prior to establishing a research process.

(5) What is research and why do we need it?

Researchers shed light on a problem under study (the topic of the research), through empirical research which is scientifically sound.

(6) What are the main questions that need to be asked to a potential researcher?

The questions that need to be asked to a potential researcher are: (a) Does the researcher have the drive to actuate such a study? (b) Is s/he aware of the time it will take up? (c) Is the data available? (accessible?) (d) Is the target group willing? (e) Does the researcher need a particular software? Is it accessible/available? (f) Does the researcher need to go back for clarification after the survey? (g) Is the language to be used, too technical? (h) Do the results reflect the original aims of the study? (i) Does the referencing conform to the establish protocols?

(7) Mention the 3 main forms of research and briefly explain each one.

The 3 main forms of research are: (1) the basic form; (2) the applied form and (3) the multipurpose research. <u>(1) The basic/pure form</u> drives towards an understanding of concepts. It is more theoretical and it does not aim at the immediate provision of tangible results. The aim of such research is the acquisition of new information and the development of the scholarly disciplines. It provides a descriptive approach to research. The results that emanate from the basic form of research are more conceptual in nature. Such a research may not reach the stage where the concepts start taking on a factual form and ending up analyzed through statistical measures. <u>(2) The applied form</u> of research has a more immediate/real-world time frame. It is an inquiry of a scientific nature designed for and conducted with an operational and practical application as its goal. This type of research answers questions dealing with the real world. <u>(3) Multipurpose research</u> finds itself somewhere in between the basic form and the applied form. It is both conceptual and factual (incorporated in one study). This type of research aims at launching scientific enquiries into issues or problems that could be descriptive and evaluative. It is both theoretical and empirical. The main function of multipurpose research is the exploration of operational and applicable results.

(8) List the 4 main types of research.

The 4 main types of research are: (1) descriptive (what something is); (2) explanatory (why something occurs); (3) predictive (to establish future actions) and; (4) intervening knowledge (allows pre-crisis interventions).

(9) The choice of type of research depends on 2 factors. Name them.

The choice of type of research depends on: (1) the availability of information and (2) pre-established knowledge.

(10) When conducting a research, the most important issue to keep in mind is context. What do you understand by "context"?

Any study has to be carried out in a specific time and space. A good research entails a full understanding of the: social, physical, cultural, economic and structural constructs within which the study is occurring. This is the context.

(11) Before starting off on a research project: one needs to be clear on what is going to be studied; one needs to decide whether to adopt a quantitative or a qualitative approach and to ascertain whether the research problem has been identified. This requires a logical approach and the authors recommend the 3x3 structure. What is this 3x3 structure?

The 3x3 structure referred here is the following: A researcher should start by establishing an aim (i.e. a topic and a direction). Then, create 3 objectives based on this aim (i.e. [1] a description of what one wishes to understand/the topic through the literature review; [2] What the researcher wants to achieve and [3] how the researcher aims to achieve those results). Subsequently, the researcher would identify 3 research questions for each of the objectives mentioned above.

(12) What quality should a sequence of Aims-Objectives and Research Questions have and why is it important?

One has to create a sequence of Aims-Objectives-and-Research-Questions that ensure a flow between each section. This would ensure an understanding of the link between the literature review and the research itself.

(13) Why is the literature review important?

The literature review is the glue which provides cohesion to the research, yet allows it to flow. It anchors the theory into a space and time with the context under study.

(14) What are the main research hurdles researchers are prone to encounter?

The main potential research hurdles are: (a) data access issues; (b) access to persons/interviewees issues and; (c) non-completion of data-gathering process.

(15) How would creating a "mind map" assist the researcher?

This helps by enabling the researcher to build an idea of: (a) what exists; (b) what should exist; (c) how best to come up with a method to identify the links/research questions.

(16) What percentage of the total research time does data collection take up?

Data collection takes up more than 80% of all the time available for the research.

(17) What do you understand by "data analysis"?

Data analysis takes into account the post-data collection process. It looks into the diverse ways that one can employ to make sense of the data in conjunction with the findings extracted from the literature review. Data analysis deals with the interpretation of the data within the context under study. It deals with the identification of the lacunae in data availability and how this will affect the analysis. Data analysis requires the running of statistical tests to seek relationships between variables. It leads to the translation of these statistics into readable and understandable text. In summary, data analysis focuses on the employment of statistical and other research tools which aid the researcher to reach more informed and reliable results.

(18) What does reporting of results entail?

This is the final stage. The reporting of results brings together the divers finding from the analysis in line with the findings from the literature review and the context under study.

(19) How did technological advances in research tools affect research?

Instead of remaining only reserved for a few specialists in certain topics, techno-tools have now been made accessible (for example: on-line maps through Googlemaps. Research has become more accessible to more people. However, now more than ever, there is a need for formal training in the use of technological research tools to ensure reliability and professionalism.

(20) Explain the difference between the use and abuse of statistics.

A famous axiom states that there are "lies, damn lies and statistics". Statistics are used correctly when it is used by people to think ahead and be proactive. Statistics are abused when researchers are animated by ulterior motives and have their own agendas. If these so-called-researchers lie repeatedly (backed by state-of-the-tools and data-backing), not only are their claims given credence by the public but they, themselves may even start to believe the lies! In fact, another famous saying claims that "a bad research is worse than no research at all".

(21) Before embarking on a research, the researcher should ask her/himself 4 questions. List these questions.

Before starting a research, researchers should ask themselves: (a) Does the research problem involve questions of value rather than fact? (b) Is the solution to the research question already determined, effectively annulling the findings? (c) Is it impossible to conduct the research effectively and efficiently? (d) Are the research issues vague and ill-defined?

(22) List the 10 major data problems.

The 10 main data problems are: (1) Data can be very expensive; (2) Access can be restricted (for example by the administration or by the law); (3) Data can be jealously hoarded (a commonly-used excuse is: so that the data would not be misinterpreted); (4) Some countries even lack addresses!; (5) Some countries have unreliable zip/post codes; (6) Researchers requesting data from 3rd party agencies should be aware that that data was gathered for the purpose of that agency and not for the researcher – possibly rendering the data partially or wholly irrelevant; (7) Datasets need to be accurate, up-to-date, complete and tagged; (8) One needs to decide earlier on at which level the datasets are required (this is referred to as the NUTS levels); (9) Versioning is very important that is why researchers need to be really sure that the data they are using is the most recent version; (10) Researchers must ensure that a lineage exists (this is a step-by-step record of the process employed to reach the end result).

(23)What are the 3 most important aspects that make data vital for one's study?

The 3 most important aspects that make data vital for one's study are: (a) relevance; (b) timelines and (c) accuracy.

(24) Define "information", in relation to research.

"Information" should not be mistaken for data. Information bridges the gap between coded data and the link that some data has to the reality it appears on. Information refers to the meaning given to data by the way in which it is interpreted (British Computer Society, 1989). Thus, data becomes information when it is given a meaning that ensconces it into a construct.

(25) Define "meaning", in relation to research.

"Meaning" is the second life that the data takes when placed in a context. It is "meaning "that leads to the drafting of policies.

(26) What do researchers mean when they say that the data cycle is suffering from DRIPS?

It means that the data cycle is suffering from Data-Rich-Information-Poor Syndrome. This happens when data ends up as an end in itself rather than a means to an end. It is when data-gathering becomes an obsession and when, as a consequence, research concentrates on description rather than analysis.

(27) Describe geographic information as opposed to spatial information.

Geographic information is information which can be related to specific locations (points) on earth. Conversely, spatial information is information which, unlike merely having an earth tag (a point), has a space-relationship tag (E.g.: on a hill/mountain next to a village where mountaineers initiate their ascent). There is another dimension to it.

(28) Define "knowledge", in relation to research.

"Knowledge" is the jump from "information". It is not an easy step because "knowledge" represents and fits within the social reality under study. Knowledge serves as the tool that extracts the meanings given to the data trawlers and changes that to policy.

(29) Define "action", in relation to research.

"Action" is the implementation of policy. It could involve: training, employment, capacity-building, legislation, setting up bodies that manage the outcomes of legislation and enforcement.

(30) What do researchers mean when they claim that action requires a continuous feedback loop?

When researchers claim that action requires a continuous feedback loop they mean that action needs regular/periodic monitoring surveys.

Chapter 4

(1) Briefly describe the datacycle approach and state why it is important.

A clear design of the research method.
The choice of tools.
Drafting a matrix to help to identify the needed Questions.
Data gathering.
Analysis process (with inherent querying and recording methodologies).

Where issues are deemed problematic to the process are identified, this LOOP is flexible enough to allow the researcher to go back and re-initiate either the whole process or part/s.

(2) How can a researcher achieve a clear view of what is required from her/his study?

Through the drafting of a

a. CLEAR AIM.
b. A set of objectives.
c. A set of respective questions.

(3) What are the main methodology issues to be considered?

a. Is the researcher in a position to initiate studies using a particular methodology (qualitative/quantitative)?
b. Has the literature review brought up very specific data requirements? If yes, is it available? If not, surrogate?
c. Will the researcher need to carry out archival research, interviews, and surveys or use readily accessible distributed data?

- Archival: material physical/hard copy? Deteriorated? Accessible? In same country? If not, costs?
- Surveys: methods understood? On-line emails? Physical mail shots? Return envelopes? Prize? Permits?

(4) What are the main operational issues to be considered?

a. Have costs been factored in?
b. Have the contacts been made? Time to fit them in schedule? Use IT? System works? Backup power to record the sessions?
c. Will the sessions be recorded? Permission? Ethical Issues? Time to transcribe?
d. How to store files? Always keep multiple copies/back-ups of the digital files and at least one copy of the analogue (hard copy) material. Keep digital copies in CD/DVD format. If possible keep one in a secure online location. Ensure the formats are readable in more than one document format in case of software malfunction.

(5) What are the main technical issues to be considered?
If the files are highly sensitive, how will they be stored?

a. Where will the files be kept? Secure place? Which site can't be compromised?
b. How will anonymity be protected? System to convert names to codes.
c. Has the process been recorded in detail to allow for replication?
d. If using imagery, is this available?

(6) How/when can a research prove impossible to conduct?

i. Is a Plan B available if the topic proves to be impossible?
ii. Data is unavailable.
iii. Contacts are uncooperative.
iv. Topic is too sensitive/feedback limited.
v. Topic is superseded by new legislation.
vi. Topic is overtaken by events.
vii. Literature review leads to a deviation from the aim of study.
viii. Researchers may have to change their research topic therefore draft an alternative topic early in the proposal drafting stage, not necessarily a full topic alternative but one that changes part of the topic or the methodology.

(7) What are the steps needed to structure the (research) mining and trawling process?

The steps needed to structure the mining and trawling process:

Step 1: <u>How will the data be gathered?</u>
- Manually? - In-situ?
- Automatically? - Remotely?

Step 2: <u>What forms will be used?</u>
- Pre-prepared forms
- Open-ended – no formal forms

Step 3: <u>Which tools will be used?</u>
- Analogue – paper/clipboard/maps for in-situ study
- Digital- using PDAs/laptop/scanner

Step 4: <u>Will the forms have all the variables inserted</u>?
- Yes and includes all sub-categories.
- Partial- allows for the inclusion of new variables and new types of archival input.

(8) What is a matrix?

A collection of cells that serve as an aid to structure data according to set columns and records in what can best be described as a spreadsheet.

	Column	**Column**	**Column**	**Column**
Record				

(9) List the 3 types of measurement scales. Using the model questionnaire provided in this chapter, (without looking/copying), complete the following table:

The 3 types of measurement scales as used by the matrix (these define the mathematical levels of precision with which the values of a variable are expressed):

- Nominal scale (N)
- Ordinal Scale (O)
- Interval Scale (I)

Question Number	**Measurement Scale**	**Very briefly explain your choice of Measurement Scale**
1	N	Imagine Name
2	I	Imagine Cinema Intervals
3	N	Imagine Name
4	N	Imagine Name
5	N	Imagine Name
6	O	Imagine Ordered
7	N	Imagine Name
8	N	Imagine Name
9	O	Imagine Ordered
10	O	Imagine Ordered

(10) Without looking/copying from this chapter, try to complete the following table:

Types of Variables to be Compared	**Statistical Tests to be Used**
Nominal Vs Nominal Variables	Chi squared
Nominal Vs Ordinal	Chi Squared
Ordinal Vs Ordinal	Correlation Spearman's test
Descriptive Statistics	Frequencies

(11) What is a pilot study? Why is it necessary?

- A testing phase.
- A launch pad.
- A necessary evil.
- The magic ingredient required to carry out the initial step is Human Targeting.
- Quantitative: 5/10 friends/colleagues.
- Qualitative: 1/2 friends/colleagues.
- A pilot study is necessary to check timing and clarity flow.

(12) What are the questions asked post-mortem, after a pilot study?

a. What could've been carried out better?
b. What did not make sense?
c. What needs to be weeded out?
d. What needs to be included?
e. Are the numbers targeted realistic?
f. Was the time projected realistic? Was it enough?
g. Should I restructure the process?
h. How will it affect the data collection period identified?

Make sure you address all the problems and start your research.

(13) List the main types of data-gathering methods.

The method of research should be chosen very carefully and once chosen should be adhered to.

Types of data gathering methods:

a. Uncontrolled/naturalistic Observation.

b. Participant observation.
- Complete participant – ethics? Total immersion.
- Participant as observer – participants know.
- Observer as participant – not total immersion.
- Complete observer- group aware/detached.

c. Surveys
- Location or remote data gathering.
- Gather data from the field or desk-based.

d. Interviews
- Mostly pre-prepared questions yet the trained interviewer can elicit additional relevant info.
- Slow but reliable.
- May take hours!
- Laborious to transcribe.
- Questions understood according to interviewee's life-world. Riddle!

Advantages: body language (non- verbals); tone of voice (verbals); feelings.

Disadvantages: psychological help may be needed by both the interviewer and interviewee (open wounds).

e. Questionnaires:
- reach a large number.
- mainly used for quantitative research.
- highly impersonal.
- absolute non-interaction.

How to carry it out:
- distribution
- post it
- on-line
- farm to an agency
- employ people to visit target groups and help them fill in questionnaire

Issues/Disadvantages of Questionnaires

- participants may not bother to reply.
- questionnaire may be deemed too long.
- participants may leave out data.
- participants may be suspicious of the questionnaire scope.
- participants may lie.
- participants opt to answer questions they feel comfortable.
- interpretation issues- Noise Level.
- bias can contaminate questionnaire.
- low response rate (35% with prize!).
- follow-up; problematic and can raise suspicion.
- ensure sampling based on latest available data.
- entries have to be reflected into an input sheet that has all the variables fleshed out.

(14) When a researcher reaches the analysis phase, which are the main issues to be considered?

a. Has a decision been taken on what tools to be used?
b. Has the matrix been completed?
c. If using a qualitative approach, what are the relevant keywords? (Literature Review)
d. Do you need catalogue cards to remember the key words?
e. Have you chosen the statistical measures (quantitative)?
f. Can your qualitative results also be analysed through quantitative approach?
g. Does this fall within your mind map model?

(15) What is a lineage and why is it important?

Recording STEPS throughout the research process is called keeping a LINEAGE.

A simple but very important rule: Record every step you take! Why? So when a discrepancy crops up, one can back-track and find out when/where the problem started.

- It records the steps in every query.
- Allows one to follow the steps.
- Records what files were generated.
- How files were stored.
- Problems encountered and other relevant steps.

(16) When it comes to research analysis there are some rules one must adhere to. List the main ones.

a. Choose the right variables to compare.
b. Choose simple relationships.
c. Divide complex relationships/problems into smaller simple ones.
d. Rule of normalisation; Codd's Rules – if using databases one should know Codd's rules.
e. Compare different sections together ex: demography with transport – ensure that all variables being cross-analysed cover all the themes discussed in the mind map and in the matrix.
f. Design graphs which describe the actual data under discussion (don't get bogged down in numbers); outputs/graphs/tables should be presented in a SIMPLE and CLEAR MANNER; Less text; Use more graphics and visual tools.

(17) Briefly describe the ideal research report.

Results must reflect the Literature Review, the Methodology and the Analysis Process.

- The report should mirror the findings.
- Be concise.
- Include a series of recommendations.
- Produce an executive summary (serves as a 'film trailer'!).

(18) “Aims, Objectives and Research Questions”: List the 3 main steps, adopting the 3x rule.

(A) Create an AIM which should be expressed as a statement that shows the topic of the study and the direction you wish your research to take.
(B) The first of the 3x: create 3 objectives based on the aim.
(C) For each of the objectives identify 3 research questions.

(19) What is a hypothesis?

- A hypothesis aims to explain a phenomenon using scientific means to test it.
- Scientific studies call for the testing of 2 hypotheses called the null hypothesis and the alternative hypothesis.

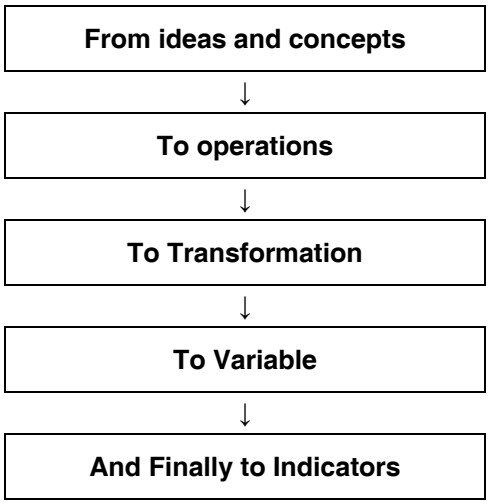
(20) What is the difference between the null hypothesis and the alternative hypothesis?

The null hypothesis always states that there is NO relationship between the variables under study. Presumed to be true unless proven otherwise.

The alternative hypothesis states that there is a relationship between the two.

Chapter 5

1. Very briefly mention the steps between the conception of a research idea and the formulation of an indicators’ list.



2. What do you understand by the research rule of Pierre Gallois?

Garbage In – Garbage Out. “But the tomfoolery (rubbish) coming out of a very expensive and complicated computer will have attained a sort respectability and no one will dare criticise it”.

3. What is “conceptualisation” in research?

Conceptualisation is nailing down the elusive variable – moulding an idea into something that can be measured.
Fuzzy concepts: uncertain boundaries ex: social segregation, educational needs, poverty…

4. What is a research “entity” and what does “entitation” in research describe?

Entity: an item that can be described and measured in statistical analysis.

Entitation: this step describes the way we recognise, understand and define the entities about which we wish to collect data.

5. What do you understand by “quantification” in relation to research?

Can the values that represent the entities be measured / quantified? Yes!

6. After coming up with a research idea/concept, transforming it into an entity and giving it a measurement structure, the researcher still has to do something else. What is it?

Concept is transformed into an Entity built through a measurement structure which ensures that the measurement is valid and can replicate the item under study!

7. Why does the researcher need to set up a list of indicators and a lineage process?

To ensure that a particular data set is valid, and can be repeatedly gathered, analysed and processes, a list of indicators needs to be set up and a lineage process undertaken.

8. Why are indicator lists important?

Indicator lists help the researcher to follow the analytical process on a step-by-step basis which will allow for comparative output analysis.

9. Define a surrogate variable.

Surrogate: the substitution of one variable by another corresponding or similar variable.

10. Why is the process of “composition” (in research) important?

This process allows researchers to develop concepts, visualise variables, create the lineages between the variables and also to identify the statistical measurements required for each linkage and how they link within themes and across themes.

11. List the four main types of measurement scales.

    i) Nominal scale
    ii) Ordinal scale
    iii) Interval scale
    iv) Ratio scale

Chapter 6

1. Research is suffering from DRIPS. What is this condition and what should be done to avoid a DRIPS situation?

DRIPS: Data-Rich-Information-Poor-Syndrome. To avoid a DRIPS situation, data needs to be turned into information. Information needs context.

2. Briefly explain what metadata is.

The process of how one ensures that data has a context within which it was created and that it serves as a veritable ID card/passport for that particular dataset. It is Data about data. Provides a description of what a dataset is composed of.

3. List the three main data categories.

    i) Raw
    ii) Numerics
    iii) Imagery

4. Why should a researcher always create a metadata for every datum?

Always create a metadata for every datum created as it helps one to source back the relevant information and ascertain whether it is relevant from studies being carried out a considerable time

post-creation. Together with lineage, this tool helps one to ensure that the base data on which to run research is reliable, sourced and whether one can use its attributes in real or surrogate forms.

5. List the five main issues associated with the acquisition of pre-existing data.

   i) The user has no control over the original format.
   ii) The user has no control over the content.
   iii) The user has no control over the attributes structure.
   iv) The user has no control over the actual data requested.
   v) The user has no control over the volume of data acquired.

6. What are research errors? When they are mostly generated? How can they be categorised?

   - The difference between the captured data and the real data that exists.
   - Nearly always generated during data capture, either through faulty processing, faulty technologies and sensors as well as data input mistakes.
   - Can be categorised by: Source, the Medium, the Technology and by the Effects generated.

7. To avoid errors, researchers must consider certain factors.

   What are they?

   i) accuracy
   ii) precision
   iii) scale and resolution
   iv) bias
   v) completeness
   vi) temporal consistency
   vii) logical consistency
   viii) semantic accuracy
   ix) repeatability

8. List the three main classifications of data sourcing.

   i) Primary – data gathered first hand (ex: thesis).
   ii) Secondary: based on findings of others (academic journals).
   iii) Tertiary: not directly linked to the author or editor; data sources created by experts ex: book listing all the papers published on Taxonomy (Classification).

9. Primary data can be gathered in two main modes. Which are they?

   - Archival capture
   - Real-time capture

10. What do you understand by “archival data”?

    Original records that are gathered by the researcher or another researcher. This data is in its original state and has not been interpreted by others. Uniqueness issue is very important. The data must be original.

11. List the three main capture modes.

    i) Manual
    ii) Semi-manual
    iii) Automatic: in-situ/remote

Chapter 7

1. List five simple types of visual research tools.

    i) Tables
    ii) Charts
    iii) Maps (1 dimension)
    iv) 3D maps
    v) Photographs

2. Graphing is a way to summarise data. List the graphing formats most commonly used.

    - Bar charts
    - Pie charts
    - Line charts
    - Histograms

3. Briefly define the following: (a) Bar Charts; (b) Pie Charts; (c) Histograms; (d) Line Charts; (e) Area Charts; (f) Composite Charts; (g) a Population Pyramid.

a) Bar charts: composed of bars separated by spaces. Ideal for displaying the distribution of variables measured at the nominal level and other discrete categorical variables.

b) Pie charts: circles (pie) display their data in the form of slices.

c) Histogram: very similar to bar charts but depict a distinct difference that is with no gaps! Adjacent bars used to display the distribution of quantitative variables. These variables vary along a continuum.

d) Line charts: composed of lines along an axis. This chart allows multiple variables to be depicted in the same chart.

e) Area charts: used for data that requires depiction of individual variables in relation to a total.

f) Composite charts: combination of styles that help one to better understand a situation.

g) A population pyramid: a bar chart that has been inverted to form horizontal bars. Always depicts males on the left and females on the right.

4. Why is mapping important?

Mapping is very important because while tables and charts provide the exact rate of changes the visual aspect depicting data on a map is more direct.

5. Briefly explain what you understand by GIS.

A geographical information system is a group of procedures that provide data input, storage and retrieval, mapping and spatial and attribute data to support the decision-making of the organisation (Grimshaw, 1994). Later definitions include the People Factor that is now seen as the most important factor.

6. What do the letters “S.W.O.T.” stand for and why would one carry out a S.W.O.T analysis on GIS?

Strengths Weaknesses Opportunities Threats. S.W.O.T is carried out to help understand the issues that emerge when investigating an entity, a process and/or a system.

7. Briefly describe what each of the following maps depicts: (a) Choropleth Map; (b) Graduated Map; (c) Dot Density Map; (d) Point Map: Actual Location of Offences Map; (e) K-Means Clustering Map; (f) Polygon-Based Cluster Analysis; (g) Small-Area Choropleth Map; (h) 3D Map: Population Density; (i) Correlation Map: Small Area Densities Vs National Densities: EAs;

(j) Nearest Neighbour Hierarchical Analysis Map; (k) Offence NNA: Spatial-Type by spread – Most effected; and (l) 3D Extrapolation of Activity Spread: NNA: Non-Serious Crime.

a) Choropleth map: depicts data based on ranges.
b) Graduated map: depicts data as a series of graduated pie charts.
c) Dot density map: depicts data based on randomly-located dots representing number of cases.
d) Point Map: depicts data based on points representing the actual location of the activity.
e) K-Means clustering Map: depicts data based on statistical clustering of related data points.
f) Polygon-Base Cluster Analysis: depicts cluster data ranged across polygons (areas).
g) Small-Area Choropleth Map: same as Polygon-Base Cluster Analysis but depicts very small polygons (areas) for niche analysis.
h) Population Density (3D Map): extrapolates the polygon data of other maps such as Small-Area Choropleth Map into 3D format.
i) Correlation Map: Small area Densities vs. National Densities: EAs – depicts correlations between two variables in polygon format.
j) Nearest Neighbour Hierarchical Analysis Map: Offence Hotspots: Spatial-Retail Crime – depicts data showing hotspots in the form of ellipsoids.
k) Offence NNA: Spatial-Type by spread- Most affected – depicts areas having similar characteristics and indicating very high levels of activity.
l) 3D extrapolation of activity spread: NNA: Non-serious crime – depicts data developed through the process outlined in a 3D format for ease of visual reference to the hotspots.

Chapter 8

1. Briefly explain what mind mapping is.

A tool to clarify one's mind and helps visually draft the process from concept to tangible measuring.

2. What is a "model" (with reference to research and mind mapping)?

A model is… "either a theory, law, a hypothesis or a structured idea. It can be a role, a relation or an equation. It can be a synthesis of data. Most important, from the geographical viewpoint, it can also include reasoning about the real world by means of translations in space (to give spatial models) or in time (to give historical models)". A model allows the researcher to study the real world through a series of observational activities.

3. List the six main steps when it comes to creating a mind map.

Step 1: a rough drawing of what the elements of this mind map constitute.
Step 2: create the theme.
Step 3: identifying the sub-topics.

Step 4: identify the sub-sub-topics for each of the elements identified in step 3.
Step 5: view the result in its entirety and start thinking about the links between the elements.
Step 6: create the potential links between the different elements.

4. List the main players in a mind map.

i) main topic
ii) the sub topics
iii) the links between the topics
iv) the dependencies (direction of dependency)
v) the data sets representing the topics
vi) the data sources
vii) the measurement scales

5. Different research-stake-holders have different level of needs. Mind maps are designed keeping in mind the requirements (levels of need) of people in different roles with their different perspectives. List these roles/perspectives.

    - Global vision perspective-visionary
    - Strategic planner
    - Operational designer
    - Administrator
    - Tactical planner

6. A conceptual model has to keep in mind three important dimensions within which that model operates. List these dimensions.

    i) the spatial dimension
    ii) the thematic dimension
    iii) the temporal dimension

7. Building a model – moving from a conceptual model to a working model – requires a process based on Peuquet's (1990) three stages. List these three stages.

Stage 1: identify those entities one is interested in and decide how to represent them.
Stage 2: choose a data model that computers are able to display analyse and store your entity representation.
Stage 3: draft a "nuts and bolts" stage where one instructs the computer how to recreate the entities identified earlier.

8. Briefly explain what you understand by "content analysis".

Content analysis usually refers to the analysis of written material ex: a political speech where words are analysed (within historical/political context) in an attempt to envisage the meaning of the writer. Content analysis is about the intended content and the received content; the difference between the two is the crux of content analysis.

9. What does CRISOLA stands for? What is CRISOLA's main area of study?

CRIme SOcial LAnduse. The main area of study is the interaction between the crime characteristics, the social characteristics and the physical characteristics (land use).

Chapter 9

1. List the three main different categories of statistical tools.

    - Manual
    - Semi-automated
    - Automated

2. What are "spreadsheets"?

Spreadsheets are the electronic version of the graph paper. Composed of multiple cells structured in what are described as rows (records) and columns (attributes). Spreadsheet cells allow for the inclusion of numbers, formulas and alphanumeric text. Spreadsheets allow various basic statistical tools to be run and some modules also exist to expand on the tools and turn a spreadsheet into an advanced statistical tool.

3. What do "Macros" do?

Macros are pre-programmed processes that allow researchers to input their data in specified cells in a spreadsheet and run the resultant measure accordingly thus drastically reducing the need for repeated work.

4. What do the letters “SPSS” stand for and what is this?

SPSS stands for Statistical Package for the Social Sciences. It is a commercial statistical analytical processing tool.

5. What do the letters “SAS” stand for and what is this?

SAS stands for Statistical Analysis System. It is a commercial suite of statistical tools that was formed, based on the integration of a number of software tools.

6. What is “Stata”? What does it do?

Stata is a commercial general-purpose statistical tool, originally using a command-line interface but recent versions have been enhanced with GUI (graphic user interface) which makes it easier to use. Stata allows for such tests as: summary statistics, regressions, ANOVA, cluster analysis, survival models, cluster analysis etc. Stata can’t load very large files.

7. What is “MiniTab”? What does it do?

MiniTab is a commercial tool that together with Quality Trainer (another tool/same company) provides a range of statistical functions. It is wide ranging and user friendly.

8. How would you briefly describe “R-Commander”?

R-Commander is an opensource tool deemed to be the most comprehensive, free statistical software available.

9. What is “PSPP”?

It is a free opensource tool. It replicated SPSS functionality and serves as a useful tool for statistical analysis.

10. What is “Gretl”?

Gretl is a free software asset that provides various statistical tools for econometrical analysis.

11. What do the letters “CAQDAS”, “QDAS” and “QDA” stand for? What does this software do?

CAQDAS: Computer-Assisted Qualitative Data Analysis Software.
QDAS: Qualitative Data Analysis Software.
QDA: Qualitative Data Analysis.

This software helps to organise, categorise, and annotate textual and visual data. It aims at building theory while visualising the relationships between data and/or theoretical constructs.

12. What do the letters “KWIC” stand for? What does this tool do?

KWIC stands for Key Words in Context. This tool offers ways to search in the text for singular words, phrases, or a collection of words on a particular theme.

13. What does “The Word Cruncher” do?

Word Cruncher counts the number of times a word appeared in the whole collected data or a particular document.

14. What is the “ArcGIS Geostatistical Analyst” and what does it do?

This is a commercial tool that serves as an extension to ArcGIS Desktop. This Geostatistical tool focuses on spatial data exploration and surface generation.

15. What does the “MapInfo Vertical Mapper” do?

MapInfo Vertical Mapper provides tools that produce trend analysis, gridding algorithms, prediction modelling, gravity modelling, risk modelling and large dataset correlations.

16. What is “CrimeStat”? What does it do?

CrimeStat is a free spatial statistics programme for the analysis of crime incident locations. It allows the analysis of standard deviation maps, attribute analysis, journey to crime, hotspot analysis and a series of spatial statistical measures.

17. List the four main CrimeStat categories.

i) spatial distribution
ii) distance statistics
iii) hotspot analysis routines
iv) interpolation statistics

18. What do the letters “STAC” stand for? What is this and what does it do?

STAC stands for Space and Temporal Analysis of Crime Software. It is a free tool that helps statistical analysts in their statistical analysis achieving this through cluster mapping, employing standard deviational ellipse creation.

19. There are two types of online tools. Which are they?

i) Those that cater for the analysis of data.
ii) Those that help the researcher to create an online survey for worldwide respondent input.

Chapter 10

1. List the four main general organizational management entities and very briefly describe them.

i) Information Technology (IT)
- The process employed in developing software and managing behavioural hardware issues.
- IT runs an organisation’s systems and network.
- IT develops, installs and maintains computer applications and systems (software and hardware).

ii) Information Systems (IS)
- The conveyor/transporters of information.
- Manages the information and the entire datacycle.
- IT has become an integral function for organisational functioning in conjunction with the management of IS.

iii) Information Communication Technology (ICT)
- The same as IT but includes the integration of different networks such as telephony and computing into one system.
- The strategies companies establish to set out their plans for IT investment and maintenance.

iv) Information Resources (IR)
- Assets an organisation holds in terms of data and information including human capital and skills.
- IR takes into account the effects and impacts that the information exerts within the organisation and in society.

2. What is a database and why is it ideal for researchers?

A database is:

- A device that facilitates our search for information within a dataset and across datasets.
- A collection of interrelated data stored together with controlled redundancy to serve one or more applications: the data are stored so that they are independent of programmes which use the data; a common controlled approach is used in adding new data and in modifying and retrieving existing data within the database (Martin, 1983).

The database approach is the process one should employ to manage data through mastering the power of computers. Databases can hold both data and metadata.

Databases are ideal for researchers since the actual data is independent from the application using it. The data can be shared; it controls for error, integrates security issues and ensures that the data rules are maintained.

3. What are the problems faced today by researchers?

i) data redundancy (many copies)
ii) unknown versions being recorded
iii) issues of data standards
iv) issues of data consistency
v) issues of data security

4. What are Database Management Systems (DBMS)?

General purpose computer programmes aimed at making a database work. A database holds only ONE structure and presents the user with a query system that can be based on any one of the drawers' resident index. The trick is in the linkages and in the electronic indexing.

5. Mention one major way in which a database differs from a spreadsheet.

The spreadsheet cannot link to any number of external datasets; a database can.

6. List the main functions of a database.

i) Create, maintain and delete data structures inclusive of data definitions and file structures.
ii) Data importation.
iii) Edit data structures (ex: adding and deleting records).
iv) Allow searching for and extracting information from data.
v) Establish security protocols in terms of data security and maintenance and access management.
vi) Would include a programming language.

7. What is the main issue in database creation?

The establishment of a sturdy Design Process that reflects the organisation's requirements and implements them within a digital structure.

8. Briefly describe how one could establish a sturdy design process.

i) List organisation's requirements (the relationship between the different entities and attributes drafted into a conceptual data model).
ii) Move from the conceptual design into a physical structure.
iii) Test it.
iv) Implement it.

9. Mention three different types of DBMS.

i) Relational Database Management Systems (RDBMS).

ii) Object-Orient Database Management Systems (OODBMS).
iii) Object-Relational Database Management Systems (ORDBMS).

10. What is an EAR Diagram?

An EAR symbolises an Entity Attribute Relationship Diagram. This is very similar to a mind map but includes the relationships and entities as structured within a digital system.

11. List the main steps one needs to take when using an EAR Diagram.

i) Draw the different elements.
ii) Identify the relationship between the elements.
iii) Identify those attributes that are common in different elements.
iv) Rename the elements into words that are readable to a database system.
v) Queries can then be run once the database has been populated by the raw data.

12. Mention one major advantage of using databases as well as the two major issues that it gives rise to.

Advantage: the same database can have its components distributed over a number of computers. In turn, these components could be distributed anywhere in the globe.
Issues: Access & Security.

13. Why is it not a good idea to integrate many datasets together?

i) A dataset can become obsolete if not updated regularly.
ii) Organisations may not be able to deliver the complete datasets.

14. How can you avoid integrating many datasets together?

What could constitute an acceptable solution?

Link the databases through a distributed database approach where the datasets are linked through one or more attributes.

15. List the steps one needs to take to set up the dataset linkages.

i) Structure the links into the main theme, the database topic, the variable one is going to use and the source.
ii) Acquire access through a series of protocols and agreements between organisations.
iii) Set up the new database and ensure that the linkages work.
iv) Create a query tool based on the mind map created as part of the process.
v) Run the relevant queries.

16. What is SQL?

SQL is Structured Query Language. SQL is pronounced as 'Sequel'. SQL allows researchers to carry out most queries based on the W6H as it filters the attributes for data that falls within the respective structures as outlined in the commands it was given.

17. What is Microsoft Office Access?

- Forms part of a suite of applications targeted for office use.
- Based on a particular data storage system employing the Access Jet Database Engine.
- Can import or access data in other database and applications.

18. What is PostGreSQL?

- Developed on the ORDBMS model.
- A multi platform solution.
- The system is supported by many third party GUI tools.

Chapter 11

1. Why is statistical testing important?

Statistical testing helps researchers to control and validate the analysis carried out in their studies. These tests ensure that errors are not committed during the course of an analytical process. Also, one should be able to identify the quantity of errors generated.

2. Briefly explain what descriptive statistics are used for, providing examples.

Descriptive statistics are used to describe a data set quantitatively through summarisation rather than through the usage of probability analysis ex: mean, median and mode; standard deviation and variance.

3. Briefly explain what inferential statistics are used for, providing examples.

Inferential statistics employ probability tests; comparative tests that allow one to infer on a population ex: the Z-score; the T-test, the ANOVA and the Chi squared.

4. What do you understand by independent variables?

Independent variables serve as predictor variables, cause changes in other variables when a value is changed and are manipulated and the resultant changes on the other variables (dependents are established).

5. Why are dependent variables also known as criterion variables?

Dependent variables are also known as criterion variables in that they are tested for changes that occur when a value in the Independent variable has changed. Therefore dependent variables are dependent on the independent variables.

6. List the three measures of central tendency.

   i) mean or average value – interval/ratio data
   ii) median or middle value – ordinal level data
   iii) mode or most frequent value – nominal data

7. There are two types of Mode. Name them.

   i) the Unimodal: has one peak
   ii) the Bimodal: has two peaks

8. How would you define “the range” in statistics?

The range is defined as the difference between the two extremes in the data range: the minimum (smallest number in the data set) and the maximum (largest number in the data set).

9. Briefly describe standard deviation, explaining its function in statistics.

   - A widely-used measure to calculate the deviation (dispersion) of the data around the mean.
   - Helps researchers to understand the structure of their data in terms of how the individual observations deviate from or vary around the mean of that variable.
   - The larger the spread of the data, the larger the standard deviation.

10. What is the variance (in statistics)?

The variance is defined as the sum of the squared deviations from the mean, divided by n-1. It is the square of the standard deviation.

11. What does the Z-score do?

The Z-score defines the distance the sample value is from the mean, always in terms of standard deviation.

12. Mention five statistical tests and very briefly describe each of them.

    i) F-test: compares the ratio of the two variances which, if equal should result in a value of 1.
    ii) T-tests: employed for testing standard deviations when the population is normally distributed.
    iii) Regression analysis (described as the Line of Best Fit): used to establish the existence of a linear relationship.
    iv) ANOVA (the analysis of variance): determines the existence of differences in datasets that contain two or more sample means. A two-way ANOVA is tested for when two independent variables are chosen.
    v) Chi-Squared (x^2): a critical test that investigated, looking for the frequencies of category (nominal) presence in a sample and analysis whether they represent the predicted frequencies in the total population.

Chapter 12

1. How frequently is the Census of Population and Housing of the Maltese Islands carried out?

The Census of Population and Housing of the Maltese Islands is carried out every 10 years although in some special circumstances it is held earlier.

2. Who conducts this laborious survey (the Census of Population and Housing of the Maltese Islands)?

This survey is carried out by the National Statistics Office.

3. List the six main steps done by the Malta National Statistics Office before the actual census starts.

    i) Drafting a list of enumerators.
    ii) Identification and mapping of routed to be followed and which household to interview (over 160,000).
    iii) Interviewing enumerators.
    iv) Monitoring their progress.
    v) Inputting and double-checking the questionnaire replies.
    vi) Chasing persons not found at home, etc.

4. Who are the people (mention just their roles/official nomenclature) that comprise the Malta Census Management Team?

    i) Census officer
    ii) Chief Coordinator
    iii) Census Officers
    iv) District managers
    v) Supervisors
    vi) Enumerators

5. List the four main steps a researcher would take when using the Census for research.

<u>Step 1</u>: getting to grips with the terminology.
NUTS: a common classification of territorial units for statistics. Ensuring the classification of territorial units into comparable levels.
LAUS: Local Area Units. Cater for the smaller administration units ex: local councils.
<u>Step 2</u>: Sourcing the Data.
<u>Step 3</u>: Analysing the Data.
<u>Step 4</u>: Visualising the Results.

6. Briefly describe the main problems, an archival researcher might encounter.

   i) Access: the researcher needs the authorities' permission; sensitive information might be in archives; protecting the privacy of those mentioned.
   ii) State of records: organised? Legible? Deteriorated?
   iii) Validation of authenticity: Real? Genuine?
   iv) Different categorisations over the years.

Chapter 13

1. John is compiling a study on the British period in Malta. Which are the best archives for him to visit?

Santo Spirito Hospital, Rabat, Malta.

2. Melinda is researching particular events that occurred in Malta and Gozo between 1530 and 1899. Which are the best archives for her to visit?

Banca Giuratale, Mdina, Malta.
Gozo archives, Victoria, Gozo.

3. Francesca is researching particular events that occurred in Malta between 1107 and 1800. Which are the best archives for her to visit?

Bibjoteka, Valletta, Malta.

4. What would one find at the Department of Information (Malta)?

   - Important documents.
   - Media releases from 1957.
   - Government gazette from 1813.
   - Archive of films from 1959.
   - Photo archive from 1970.

5. List at least five other public archives in Malta.

   - Archives of the Notaries from 15th Century to date.
   - Public Registry Archives from 1863 to date.
   - Archives of the Courts from 1900 to date.
   - Archives of the Lands Dept.
   - The Records and Archives of the Department of Works from 1800.
   - The Parliament Records from 1849.
   - The UOM archives from the 11th century.
   - The PBS Ltd Archives from 1970s.
   - The Central Bank Archives from 1964.
   - The Medical Records from 1978.
   - The National Museum of Arts from 1798.
   - The centre of documentation for Teachers from 1851.
   - The Archives of Administration of Burials.
   - The EneMalta Archive from 1853.
   - The Maltacom Archive from 1943.

6. List the five church archives (in Malta) mentioned in the text.

   - The archive of the Cathedral from 11th century.
   - The Archbishop Curia from the 16th century.
   - The Bishop's Curia (Gozo) from 1554.
   - The Gozo Cathedral from 1623.

- Various archives of different religious orders: Domenican Archive from 15th century; Archives of Sisters of Charity from 1868.
- Others: The Archive in the Wignacourt Museum.

7. List the three private archives (in Malta) mentioned in the text.

- Dr Albert Ganado's – from 1296.
- The Times start in 1930.
- The Lanfranco family start in 1540.

8. Give at least two examples of English archives that could be of interest to a researcher compiling a study on Malta and Gozo.

- The National Archives (near Kew Gardens) in London, England (The British Period).
- The Family Records Centre.
- The Historical Manuscripts Comm.
- The British Library.

9. Give at the two examples (mentioned in the text) of Italian archives that could be of interest to a researcher compiling a study on Malta and Gozo.

- Magistral Library
- Archives of the Order of Malta

10. Give at least one example of American archives that could be of interest to a researcher compiling a study on Malta and Gozo.

National Archives and Records Administration.

11. Give at least four examples of important libraries in Malta that could be of interest to a researcher compiling a study on Malta and Gozo.

- NSO
- MEPA
- UN International Institute on Ageing
- The National Archives
- UOM

12. Briefly describe how a researcher's request for data should be.

Researchers are advised to be specific in their request for data (to agencies). Agencies have databases which hold a lot of data parameters (most would be irrelevant to the researcher). So, the researcher's request should not be generic and unrealistic! The request should be planned and specific.

Chapter 14

1. What does ethics (or moral philosophy) provide us with?

Ethics provides us with recommended guidelines of what is right and what is wrong. These guidelines secure the nature of human well-being.

2. Why are ethical considerations important for researchers?

These are important because they go along creating new knowledge, researchers cannot leave a trail that leads to the subjects of of their own study! Researchers cannot base their actions on the principle that the end justifies the means.

3. Mention two cases of research studies that attest breach of ethics.

Case 1: The Stanley Milgram Experiment.
Case 2: The Stanford Prison Experiment.

4. List the four main criterions for ethical research.

   i) informed consent
   ii) confidentiality and anonymity
   iii) objectivity
   iv) deception

5. What is plagiarism?

Plagiarism refers to the unauthorised use or close imitation of the language and thoughts of another author and the representation of them as one's own original work (http://dictionary.reference.com/browse/plagiarism)

6. In Criminology, referencing should be compiled on the guidelines provided by a particular association. Which association is this?

Referencing should be compiled on the guidelines provided by the American Psychology Association (2002) – A.P.A.

7. With regards to referencing, state what needs to be done if the same author has different publications?

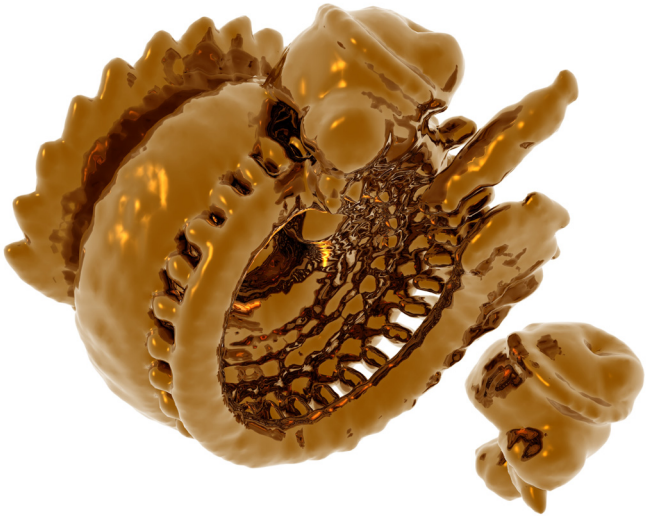
The sources of evidence need to be arranged according to the year of publication with the earliest one first.

8. With regards to referencing: what does one need to do if the same author has two or more publications in the same year?

One needs to use the lower case letters just after the year of publication so as to distinguish between the two different sources.

Ending

BEBX GLASSY OR
3D Fractal - S. Formosa (August 2010)



Anybody who has been seriously engaged in scientific work of any kind realizes that over the entrance to the gates of the temple of science are written the words: *Ye must have faith*. It is a quality which the scientist cannot dispense with.

Max Planck

Where is Science Going?, translated by James Vincent Murphy (1932), 214.