south african
cultural observatory

South African Cultural Observatory

A Position Paper on a Forecasting Model In Regard to the Cultural and Creative Industries of South Africa

Researcher: Professor Igor Litvine

South African National Cultural Observatory, Chief Statistician

Submitted to the Department of Arts and Culture

arts & culture
Department:
Arts and Culture
REPUBLIC OF SOUTH AFRICA

Summary

The purpose of this document is to set a framework for forecasting and predictive modelling for South African Cultural and Creative industries and their impact on sustainable development of the national economy and society. This document is based on the ideas laid out in the accompanying SACO report on statistical framework for cultural and creative industries prepared by Prof. Jen Snowball (Rhodes University) and Prof Alan Collins (Portsmouth University). Ideally we would expect that the reader is already familiar with the report Towards the Development of a Framework for Cultural Statistics for South Africa [17].

Firstly we provide an overview of data (already available and likely to be available in foreseen future) and data sources that may be valuable in the task. Broadly this data may be categorised into three sets: (a) previously collected and recoded data; (b) purposely collected data (e.g. via specialised surveys) and (c) open data.

Secondly we review potentially productive methodologies and techniques that are most likely to contribute to the task of the design of scientifically motivated statistical, mathematical and computational predictive models.

Finally we elaborate on visualising techniques that support better understanding both of the data and the results of the modelling.

Acknowledgements

My sincere thanks go to:

- Prof Richard Haines, who consistently provided his guidance and support. Specifically he provide valuable recommendations on the general structure and content of this report.
- Prof Jen Snowball (Rhodes University). Her knowledge of Cultural Economics is exceptional and I do appreciate working with her in the Cultural Observatory.

Contents

List of Figures

Chapter 1

Introduction

1.1 Objectives

The UNESCO Framework for Cultural Statistics defines culture as follows: “culture is the set of distinctive spiritual, material, intellectual and emotional features of society or a social group, that encompasses, not only art and literature, but lifestyles, ways of living together, value systems, traditions and beliefs”. Due to the intangible nature of culture as beliefs and values direct statistical measurements are impossible. Therefore the measurements should be implemented “through the identification and measurement of the behaviours and practices resulting from the beliefs and values of a society [28].

Developing a South African Framework for Cultural Statistics (FCS) is an important first step in collecting information on, and building understanding of, the cultural and creative industries in South Africa [17]. The document by South African Cultural Observatory (SACO) gives six most important questions that need to be answered:

1. What is the economic and social impact and value of the CCIs?
2. How can the CCIs contribute to national and regional policy objectives?
3. What are the challenges faced by the sector and how can their potential be optimised?
4. How has the structure and contribution of the industry changed over time?
5. How effective have national, provincial and regional CCI policies been in achieving their goals?
6. Are there some CCI sectors with more potential for achieving development goals? Which sectors should be targeted and in what ways?

The forecasting and predictive models concept is a cornerstone of the Framework as it is the only tool that may provide scientifically correct and methodologically motivated answers to the questions 1, 4 and 6 above. In more detail, forecasting techniques may provide substantive answers to following questions:

- What are the trends in the development of the national cultural and creative industries (CCI)?
- What impact the CCI industries may have on the future national growth (both in economic and social context)?
- Are the cultural and creative industries are at all important and should be supported by national and local governments?

- What CCI have the key importance for South Africa and should be supported as a matter of priority.
- Will cultural and heritage activities ease social tensions and support peace and harmony in the society?
- What regulatory steps by the state are likely to be most efficient and productive?

Chapter 2

Available Data and Data Collection Strategies

2.1 Introduction

Snowball indicates that the adaptation of the UNESCO Framework on Cultural Statistics will depend, among other things on [17]:

- Policy priorities: What is it that a FCS should measure? Should the Framework be able to measure both Economic and Social effects? Are they both equally important? To what extent should the more commercial creative industries, as well as the core industries be included? How should support services and activities be treated, for example, cultural education? What are the policy and funding implications of these choices?
- Available data: What are the available data sources on the cultural and creative industries in South Africa? How recent are they? How detailed is current national accounts data? For example, does it allow the breakdown of various classification codes to the point where cultural goods, services and employment can be measured?

In addition to that we would like to see what innovative modern data collection tools may be employed in the forecasting and predictive models.
As we mentioned above, the data sources may be categorised into 3 sets:

- Previously collected and recoded data;
- Purposely collected data (e.g. via specialised surveys);
- Open data.

2.2 Previously Collected and Recorded Data

Statistics SA would play a key role in providing data of this kind. The following sources present potentially attractive opportunities.

2.2.1 Quarterly Labour Force Surveys (QLFS)

The QLFS collects data on the labour market activities of individuals aged 15 years or older who live in South Africa. The QLFS is the principal vehicle for disseminating labour market information on a quarterly basis.

2.2.2 Time-use surveys (TUS)

The first national Time Use Survey was conducted by Statistics South Africa in 2000 and again in 2010. The aim of the survey was to provide information on the way in which different individuals in South Africa spend their time. The TUS activity classification system used in the TUS has ten broad categories which are consistent with the System of National Accounts (SNA) which underlies the calculation of gross domestic product (GDP).

2.2.3 Import-Export data

An example of import-export data is the Producer Price Index (PPI) for exported and imported commodities. The PPI is compiled from the results of a monthly survey of prices of exported and imported commodities which covers samples of exporters and importers in the South African economy. The PPI is used by the private sector for contract price adjustments and also as a deflator in the compilation of national accounts.

2.2.4 Household and Community Surveys

One such survey, is the General Household Survey (GHS) which is conducted annually by Statistics South Africa since 2002. The GHS collects information on a variety of subjects including education, health, the labour market, dwellings, access to services and facilities, transport and quality of life.

The Community Survey (CS) reflects the country's performance on education, labour markets, service delivery, poverty and access to services.

2.2.5 Tourism Satellite Account (TSA)

The Tourism Satellite Account (TSA) provides an overview of the role that tourism plays in South Africa as well as information on tourism's contribution to the South African economy in terms of expenditure and employment.

The TSA is one element of a System of Tourism Statistics (STS) that provides information for the understanding and monitoring of the impact of tourism on the South African economy over time. Other elements of the STS for South Africa include the surveys of international tourists and domestic visitors, visitor arrival statistics, tourist accommodation and food and beverages statistics.

2.2.6 Input-Output tables

The I-O table provides an overview of the economy focusing on the inter-relations between industries, represented by a symmetric matrix which contains both supply and use data.

2.2.7 Supply and use tables, etc.

The SU-tables have both statistical and analytical functions. As a statistical tool they serve as a coordinating framework for economic statistics, both conceptually for ensuring the consistency of the definitions and classifications used, and as an accounting framework for ensuring the numerical consistency of data obtained from different sources (i.e. industrial surveys, household surveys, investment surveys, foreign trade statistics).

As an analytical tool, the tables serve as a basis for calculating the economic data contained in the national accounts, and for detecting weaknesses in the economic data. Moreover, they are conveniently integrated into macroeconomic models in order to analyse the link between final demand and industrial output levels.

2.2.8 Other sources

Other potential sources of useful data include the following: government institutions and agencies, private companies, non-profit companies (NPCs) and foundations, academic research documents as well as international organisations. Examples of the above and the potential types of data they are able to produce include the following: Government Departments and Agencies:

- Department of Arts and Culture.

  Funding information: which events and organisations receive subsidy, characteristics of successful applicants, policy documents, discussion documents and legal documents.

- Department of Education and Training

  For example, The Organising Framework for Occupations (OFO), which is a coded occupational classification system. It is the key tool used by the Department of Higher Education and Training for identifying, reporting and monitoring skills demand and supply in the South African labour market.

- South African Revenue Services (SARS), including Customs. Trade statistics are published on the last day of each month.

- Department of Trade and Industry (DTI). Subsidy data (e.g. TV and film subsidies).

- South African Broadcasting Corporation (SABC). Data on media audiences (e.g. viewership and listenership statistics).

- National Treasury. Tax Statistics are published in collaboration with the South African Revenue Service (SARS) annually. The aggregated statistics presented in the Tax Statistics Bulletin provide important insights into South Africas economic trajectory. These include: how the profitability of companies is evolving, the value added by entities registered for VAT, the growth in gross income reported by individual taxpayers, and their taxable income in turn.

- Independent Communications Authority of South Africa (ICASA). The Independent Communications Authority of South Africa (ICASA) is responsible for regulating the South African communications, broadcasting and postal industries in the public interest and ensuring affordable services of a high quality for all South Africans. In order to fulfil its mandate, ICASA also aims to be the authoritative source of relevant sector statistics for consumers, government, industry and other stakeholders. An example is the Report on the state of the ICT sector in South Africa, which was published in March of 2016.

- NPCs and Foundations:

- South African Audience Research Foundation (SAARF), e.g. All Media Product Surveys

- National Film and Video Foundation (NFVF).

- Business and Arts South Africa (BASA).

- National Research Foundation (NRF).

- Private companies.

  An example of a private company is Quantec, a consultancy providing economic and financial data, country intelligence and quantitative analytical software. (http://www.quantec.co.za/). International organisations include, for instance, the United Nations (UN) and associated organisations (such as UNESCO and UNDP), The World Bank, etc. Certain international private companies also possess valuable data, e.g. Global Database (http://app.globaldatabase.co.uk/). The Global Database has a data set on the Africa Leisure and Entertainment Industry, which contains 2030 records of South African companies including the following fields:

  - Full Company Name and Address.
  - Senior Executives Personnel list.
  - E-mail addresses and phone numbers.
  - Company website.
  - Year of establishment.
  - Industry classifications.
  - Employee size.
  - Sales volumes

  Apart from records of South African companies, the same dataset stores 5437 records of other African companies. This feature makes this dataset attractive for purposes of panel regression modelling and forecasting.

- Universities.

  As far as academic research documents are concerned, South African universities in the first place and thereafter other universities around the world could potentially produce a vast array of useful data.

The aforementioned data hubs are freely available. Furthermore, they represent a variety of data categories. The above sources would be a logical starting point in the study of data with the view to establish a South African cultural framework.

2.3 Purposely Collected Data

In this category we situate data which was (and will be) collected either by the Department of Arts and Culture (DAC), the Cultural Observatory (SACO) or businesses and individuals contracted by the former. At the moment we have two datasets of this kind, both collected as part of so-called Mapping study. The mapping study was conducted by a private company Plus 94 as a contract job for the DAC [21]. The first dataset includes 2478 data records and provides quite detailed information on a sample of the cultural sector constituents. The second data set has less fields (variables) however has much more records (24757). The first data set is quite useful for forecasting purposes as it contains time variable (data of establishment of the business/organisation).

2.4 Open/Free Data

This method of data collection is becoming increasingly popular. For example we have evidence that Internet users may be diagnosed for cancer by analysing their bing-searches (http://www.ibtimes.co.uk/

can-microsofts-bing-predict-cancer-by-analysing-your-search-history-1564803).

Needless to say that this approach requires advanced IT skills. More about it may be found in [2].

We have used this approach to collect data on South African municipalities (municipal data). This dataset has records for two years (2001 and 2011), 234 municipalities, 21 demographic, social and economic variables.

2.5 Crowdsourcing

Google defines crowdsource as: obtain (information or input into a particular task or project) by enlisting the services of a number of people, either paid or unpaid, typically via the Internet.

The term crowdsourcing was introduced in 2006 by Merriam-Webster (Merriam-Webster, Incorporated, is an American company that publishes reference books, especially dictionaries). The company defines the term as the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, especially an online community, rather than from employees or suppliers.

Apart from advanced IT skills this approach requires lots of creativity from the data collecting team.

2.6 Statistical Data: General Concepts

There are several ways in which statistical data may be categorised. Firstly the data may be raw and organised. It will be wrong to say that raw data is data which is not clean or messy or of poor standards! Perfectly clean and best quality data is still raw if it was not processed [10]. Let us be more specific. One of the datasets collected during the mapping study [21] contains data of 2478 cultural entities (businesses, organisations informal groups and individuals in South Africa (see a screen shot of the fragment in figure 2.1).

	A	B	C	D	
1	Respondent Unique ID	Q1a.Street Address <ADDRESS> (VERY IM	STREET_ADRESS	Latitude	Longitude
2	66000	83 6th Street Parkhurst	83 6th Street Parkhurst	-26.143303	
3	66008	31 JELLICOE AVE ROSEBANK JOHANNESBURG	31 JELLICOE AVE ROSEBANK JOHANNESBURG	-26.141358	
4	66028	40 Von brandis street	40 Von brandis street	-26.160369	
5	66047	2 Price rosser Street, Prolecon Cidy Deep	2 Price rosser Street, Prolecon Cidy Deep	-26.204103	
6	66066	no 100 President ,5th floor Belmoral Building	no 100 President ,5th floor Belmoral Building	26°12'37.4"S	28°09'33.3"E
7	66120	19 Sebakwe road selcourt	19 Sebakwe road selcourt	-26.302007	
8	66135	COEN SCHOLTZ CENTRE MOOIFONTEIN ROAD BIRCHL	COEN SCHOLTZ CENTRE MOOIFONTEIN ROAD BIRCHL	-26.250656	
9	66168	15/Monumentstr/Glenmarais/Kempton park/gauter	15/Monumentstr/Glenmarais/Kempton park/gauter	-26.077426	
10	66180	johannesburg antletics stadium 124 van beek street	johannesburg antletics stadium 124 van beek street	26°11'54.8"S	28°03'30.1"E
11	66184	22 Bezuidenhout street	22 Bezuidenhout street	-25.698463	
12	66204	87 De korte Str Braamfontein	87 De korte Str Braamfontein	26°11'35.8"S	28°02'10.4"E
13	66212	1st floor ellis house no 23 voorhout street ..2049	1st floor ellis house no 23 voorhout street ..2049	26°12'05.0"S	28°03'13.7"E
14	66256	41 Brodigan	41 Brodigan	26°08'59.8"S	28°19'39.5"E
15	66269	northmead square,fourteenth ave,benoni	northmead square,fourteenth ave,benoni	33°47'22.9"S	150°59'46.1"E
16	66282	11 Ribork road 22118 Randburg	11 Ribork road 22118 Randburg	26°06'56.4"S	27°58'52.1"E
17	66332	Pioneers Park 193 Rossettenville Road La rochelle	Pioneers Park 193 Rossettenville Road La rochelle	26°13'56.7"S	28°03'11.8"E
18	66333	16 Southern Klipriviersberg RD The hill Johannesbu	16 Southern Klipriviersberg RD The hill Johannesbu	26°14'35.8"S	28°05'12.9"E
19	66389	St Andrew's Street Parktown	St Andrew's Street Parktown	26°10'55.9	28°02'17.3
20	66489	54 7th Avenue, Parktown North	54 7th Avenue, Parktown North	26°08'42.2"S	28°01'52.9"E
21	66508	42 Dauley Park Wood 2193 Johannesburg	42 Dauley Park Wood 2193 Johannesburg	26°08'55.7"S	28°02'31.4"E

Figure 2.1: A fragment of the raw data from the Mapping Study.

The raw data is typically organised in a table where the columns represent statistical variables, that is observed characteristics of the elements included in the study (note in database context the columns are referred to as fields. In this case the elements are cultural entities (or respondents,

since the data was collected via telephonic interviews). Note, that the first column is headed as "Respondent's Unique ID" - this is usual standard requirement of the data collection practice.

Each row of the raw data represents all data for a single element. In Statistics rows in the raw data are referred to as observations - one raw per observation, in database context one may find that raws of a data-table are called records.

The main feature (and advantage) of the raw data is that it has all information collected. However this advantage is also a drawback:

- huge tables are difficult to publish;
- data is not summarised and therefore difficult to comprehend;

Raw data are also referred to as primary data.

Processed data is called summarised data. The idea is to present important features of the data in a short but easily understandable way. The data may be summarised in numerous ways and at different depth. The two most important summaries of a single numeric variable are measure of central location and measure of variability [10]. The most common practice is to use expected value (mean) and standard deviation for this purpose. However these are not enough if more than one variables are studied. In this case we have to consider as well measures of associations between the variables, and the Pearson's correlation coefficient [10] has widest use for this purpose.

Another way to summarise data is to use certain tables. Most common tables in this case are frequency tables and cross-tabulation (or contingency tables). The advantage of the tabular method is that it is applicable to categorical (non-numeric) data.

Another method for data summarising is ranking. Kendall's tau may be used as a measure of association between two rankings [19]. We see huge opportunities for the ranking techniques in cultural statistics, as these methods are highly robust - they are very insensitive to minor errors and are valid under very general assumptions.

In any case one should remember that summarising data inevitably results in information loss. So the summarised data is a great tool for presenting your findings and results, however any reliable and trustworthy analysis should use raw data as the input.

As was already mentioned above the variables in the dataset may be of different nature. In brief the three most important types of variables are:

1. Numerical variables.
2. Categorical variables.
3. Ordinal variables.

Generally speaking numerical variables are variables which allow meaningful mathematical operations, including calculation of mean, variance, standard deviation, etc. Categorical variables (or qualitative variables) define certain characteristics of the elements which cannot be measured numerically, for example race, gender, home language, etc.

Ordinal variables are (in a sense) a compromise between the numerical variables and categorical ones. Values of the ordinal variables may be ordered (in other words ranked) according to some scale. Good example of ordinal variable is level of education where we can assign highest rank to the highest level (e.g. PhD) and the lowest rank to the lowest level (no formal education). Summarising, ordinal variables allow ordering of the values, however any mathematical operations with ranks are meaningless. For example if one object has rank 1 and another is ranked 12, it does

not mean that the first one is twelve time better than the other one.

It will be wrong to ask a question which variables are better. All types of variables are important and equally used in Statistical analysis as long a researcher uses relevant and appropriate techniques for each kind. For example, classical regression is impossible for categorical variables and categorical regression (with dummy variables) should be used instead [12]

2.7 Errors in Data

Unfortunately data often may contain errors. Most common errors are:

1. Misprints (or typos).
2. Variations in spelling (e.g. "color" and "colour", "modelling" and "modeling")
3. Variations in geographic names (e.g. "Free State" and "Freestate", "Cape Town" and "Kaapstad").
4. Variations in formats (e.g. degree-minute-second format and degree-decimal formats for GCS coordinates).
5. Corrupt data.
6. Missing data.

More detailed classification may be seen on figure 2.2 [5].

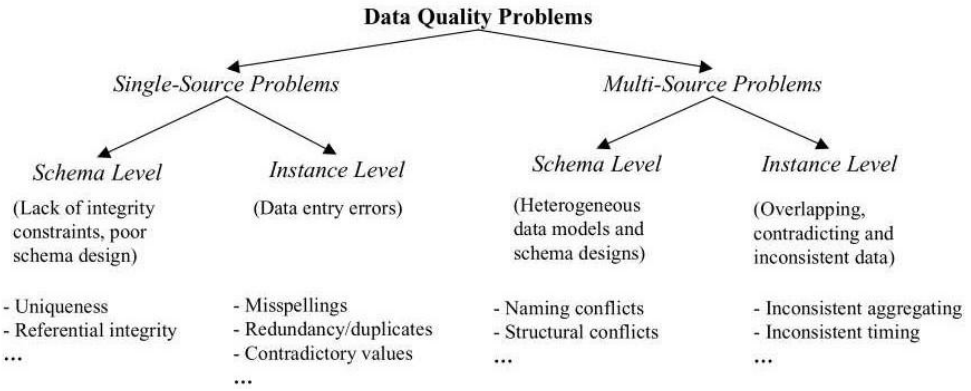

Figure 2.2: A classification of data errors (source [5])

There are many techniques which allow fixing the errors in data automatically (or semi-automatically) in cases 1-5 above [5]. However it is not guaranteed that all errors may be fixed (especially in case of corrupt data). In this case, and also in case of missing data one may use use one of the imputation techniques - that is, replacing corrupt or missing data with a reasonable guess (see [4]). In case of very poor data quality manual cleaning/correcting may be required. In exceptional cases a respondent may be contacted again (if the survey was not anonymous).

Chapter 3

Towards Cultural Forecasting Models

3.1 Introduction

Non-professionals (and sometimes professionals) often confuse forecasting and prediction. We may note here that in many languages there is only one word that has both meanings. However in English these two concepts stand for two separate (still related) tasks. Both concepts refer to making scientific guess about the future. However in different senses:

- Forecasting is a statistical technique that aims at estimating future value of a certain parameter or variable. There are some examples of forecasts:
  - Maximum temperature tomorrow will be 22 degrees Celsius.
  - Next quarter GDP will increase by 1% (year to year).
  - The state subsidy towards cultural growth will increase by 5% next year.
- Prediction aims at estimating the likeliness of future events and suggesting the most likely one. Some examples of predictions are:
  - The Kaiser Chiefs will win the national championship next year.
  - Chances of rain tomorrow are 30%
  - Most likely the share of cultural industries in the national GDP will grow in the next 5 years.

Neither forecasting nor prediction is possible without a model. In both cases we refer to predictive modelling.

3.2 Basics of Predictive Modelling

All models are wrong, but some of them are extremely useful. This famous saying of George Box [8] emphasises two important features of a scientific model:

- A model represents a simplification (in simulacra) of a real object or process. From this prospective any model is "wrong", as it disregards certain properties and features of the reality.

- Models are useful because they are supposed to disregards only irrelevant, unimportant features of an object or process making it much easier to understand, visualise, analyse and simulate the reality. Not to forget that studying a model is normally less dangerous, much cheaper and risk-less than studying an object/process itself.

A perfect model is the result of a balance between the principles of parsimony and adequacy. A parsimonious model is as small and simple as possible in given circumstances (some scientists also appreciate aesthetic appeal). A model is adequate if it retains all properties of the real object which are essential for a given study. So the art of modelling is in making a perfect separation between irrelevant, which may be discarded and important, which should be retained and analysed. Especially important is the preservation of existing causal mechanisms.

There is a variety of approaches in scientific modelling: deterministic and probabilistic, scaling, iconic (pictorial), operational, structural, in vitro, etc. However in this study we deal exclusively with mathematical models, that is models which may be presented by numbers, variables, parameters, functions, inequalities and equations (of all kings) and other mathematical objects.

3.2.1 Linear Models

Among the mathematical models the most popular are linear models, that is, models which contain only linear functions and relationships. The reasons for that are obvious:

- These models are simplest (remember about parsimony!).
- Many non-linear models may be reduced or transformed to linear ones.
- Often linear models represent acceptable approximation of nonlinear ones (e.g. Generalised Linear Regression models [22]).
- Linear models make a great departure point. Subsequently they may be upgraded and generalised to more complex models.

Despite of their seeming simplicity, linear models are not yet fully discovered. Quite simple new linear model (LL-model or double linear model), which was introduced and analysed only recently [11].

3.3 Static, Stationary and Non-stationary Models

With respect to time-dependence the mathematical models may be categorised as Static, Stationary and Non-Stationary.

- Static models do not include time as a variable. This does not mean however that static models are simple. For example Structural Equation Models (SEM) represent quite complex construct and are considered as a tool for revealing causality. These models are widely used in Economics (e.g. equilibrium models).
- The other two categories (Stationary and Non-stationary Models) represent the class of dynamic models and they include time as a variable.
- Stationary system moves in time (changes the system state), however the parameters of the model (e.g. drift, volatility, autocorrelation, etc.) remain constant.
- A non-stationary system not only moves in time but also evolves itself.

3.3.1 Static Models

Static models are described by systems of algebraic equations. In economics and mechanics such models represent certain state of equilibrium. The most known static model is a SEM model, which has its roots in the multivariate factor analysis [27]. While the SEM models are believed to describe and reveal causal relationships, they have serious limitations and nowadays their use is almost entirely limited to psychometrics [18].

3.3.2 Dynamic Stationary Models

These models are usually represented by systems of differential equations (continuous time) or systems of difference equations (discrete time). The coefficients of the equations are constants. The continuous-time models are mainly used in Technometrics, where the readings may be taken very frequently, so that one may assume continuous time.

In Econometric models time is usually discrete (mainly due to inability to increase the frequency of observations). Commonly the discrete-time dynamic systems are also known as Auto-regressive systems. The most popular class of stationary discrete time models is so-called ARIMA model (ARIMA stands for Auto-Regressive Integrated Moving Average), introduced by George Box [7]. ARIMA model assumes that after the process is differenced d times it becomes ARMA process, that is:

$$y(i) = c + \sum_{j=1} a_j y(i-j) + \sum_{k=1} b_k e(i-k) + e(i)$$

where c, a_j, b_k are constants and e(i) are random errors.

As we mentioned this model is very popular in Econometric applications, however has serious limitations. Box was a chemist and developed his models mainly for chemical technology experiments where (a) experiments may be repeated several times under strict control of required conditions; (b) if an experiment goes wrong, it may be repeated (c) measurements may be taken practically continuously. A these are not true in Economics and Econometrics. The danger of over-fit (due to insufficient data pints) is more than real. Over-fit models are highly inadequate and present serious dangers in used for forecasting and prediction.

3.4 Dynamic Non-Stationary Models

Another issue with stationary models is the assumption that the system does not change with time. This is seldom the case in an economics context. This is why the revolutionary model by Robert Fry Engle III (known as GARCH process) brought him Nobel Prize in 2003. The GARCH process (Generalised Auto-Regressive Conditional Heteroskedasticity) is not stationary, because the variance of the error term (and, therefore the volatility) of the process is not constant. In fact the error of the variance of GARCH model follows ARMA model [23].

The area of non-stationary models is vast. We can mention that the GARCH model has numerous off-springs: NGARCH, EGARCH, IRARCH, GARCH-M, to name but few. Another popular example of non-stationary series is the FARIMA process, that is Fractional ARIMA and Fractional Brownian motion processes. These processes involve a new parameter called Hurst exponent [20]. The Hurst exponent is a parameter which describes persistence in the series. The persistence is regarded as a synonym to trendiness and opposite to jaggedness.

A detailed review and classification of other models used in Econometric research may be found in [24] and [25].

3.5 Computational Intelligence

Computational Intelligence may be defined as follows: [26] "a system is called computationally intelligent if it deals with low-level data such as numerical data, if it has a pattern-recognition component and if it does not use knowledge as exact and complete as the Artificially Intelligent one".

The computational intelligence methods are usually nature-inspired methodologies and approaches that address complex real-world problems to which mathematical or traditional modelling can be ill-suited for a a number reasons: the processes might be too complex for mathematical reasoning, they might contain some uncertainties, or simply be stochastic in nature [1].

Currently popular approaches in Computational Intelligence include biologically inspired algorithms such as swarm intelligence and artificial immune systems; evolutionary computation, Neural Networks models, image processing, data mining, natural language processing, fuzzy logic, learning theory. Probabilistic thinking is essential in Intelligent Computing as most of the methods are randomised.

3.6 Spatial Models

This innovative type of modelling is also considered as potentially exceptionally promising. The cultural data (as discussed in the previous chapter) is organised by units of time (e.g year of establishment) as well as by geographical constructs such as provinces, towns/cities or GCS coordinates. This is usually the case for governmentally collected economic medical and social data. Today these methods are successfully used in all fields of forecasting, e.g. real estate pricing, epidemics propagation, wild-fires, etc.

3.7 Panel Regression

This is also a reasonably new technique. There are 3 types of longitudinal data: (a) time-series (b) pooled cross-sections and (c) panel data. In all three cases the data contains time variate. In panel data (contrary to time series) we do not have many points in time, however number of units observed is large. This approach is particularly useful in cases when we have data on organisations and firms at different time points or we have aggregated regional data over time. This is exactly what we have both in case of mapping study and municipal data.

The literature on the subject specifically emphasises the importance of the panel regression for:

- describing change over time;
- revealing trends in social phenomena;
- discover casual relationships;
- policy evaluation;

It is important to note that the spatial models may be naturally combined with spatial models [9].

3.8 Causality

Causal analysis is a significant role-playing field in the applied sciences such as statistics, econometrics and financial analysis. Probability-raising models have warranted significant research interest, most of which is philosophical in nature. Contemporarily, the econometric causality theory, developed by C.J.W. Granger ([15] [16]), is popular in practical time series causal applications. While this type of causality technique has many strong features, it has serious limitations. The processes studied, in particular, should be stationary and causal relationships are restricted to be linear.

The probability-raising causality ideology, as postulated by Good [13], [14], suggests that causal analysis methodology is applicable in a stochastic, non-stationary domain. Litvine and Mlambo [6] proposed a Good's causality test based on transforming the originally specified probabilistic causality theory from random events to a stochastic time-series. This new test can be used to detect whether one stochastic process is causal to the observed behaviour of another, probabilistically.

Let us consider some technicalities of the above approaches. The Granger's approach is based on the concept of "all information in the Universe". Suppose this information is contained in the variables $x_1(t), ..., x_N(t)$, where t is time. Then the following model is considered:

$$y(t) = a_1 x_1(t) + ... + a_M x_M(t) + e(t)$$

where the dependent variable y(t) represents the effect (in the cause - effect settings). $x_1(t), ..., x_M(t)$ represent the variables $x_1(t), ..., x_N(t)$ as well as their lags in time, however this set excludes the variable y(t) but includes the lags $y(t-k)$ (as a process may be caused by itself in the past); e(t) is a random error.

Further Granger states that (by definition) x_i has causal impact on y if excluding this variable from the above equation reduces the fit of the model. Specifically Granger suggests that the fit should be measured by the sum of squared residuals (or equivalently, by the standard deviation of then residuals). The reduction in the standard deviation may be easily confirmed by a standard chi-square statistical test.

The shortcomings of this approach were discussed by Granger himself:

- It is impossible to include "all information in the Universe" a practically applicable model. Instead only variable which are reasonably expected to have relevance are included.
- The model assumes linear relationship in causality. In reality the casual mechanism may be non-linear (e.g. Markov-switching processes).
- We need quite long series of observations if we want to include in the model all variable which are "suspected" to be relevant.

From this prospective Good's causality may look more attractive in many practical instances. Let's assume that E and F denote random events; the same could be thought of as propositions asserting the occurrence of the events, respectively. In a way of defining when an event is said to be a cause of another, Good [13] takes a propositional approach. Let's assume further that:

- H_1 represents all the true laws of nature either known or unknown.
- H_2 represents all the essential physical circumstances usually taken for granted.
- Then $H = H_1 \cdot H_2$ represents all true laws of nature and essential physical background information.

Then the definition of the causal event is: the causal event (F) raises the probability of the effect (E). The effect is more likely to happen when the cause occurred than otherwise. Hence

$$P(E|F.H) > P(E|\overline{F}.H)$$

While this approach has own limitations, it still looks exceptionally attractive for revealing indirect impact of culture and cultural industries on economic and social growth, because:

- The Goods definition does not refer to time. This means that we may use it when we have substantial data, but for relatively short period.
- It does not introduce any specific model (linear, non-linear or others). Therefore the methodology is robust.
- Standard statistical tests for probability may be utilised.

As was shown [6] the approach may be generalised for time series (if more data becomes available in future).

Chapter 4

Data Visualisation

We see very important role of data visualisation in cultural industries forecasting. Therefore the data of the mapping study is so critically important.

The following two figures represent visualisations of: (a) wild-fires in USA and (b) restaurants in Port Elizabeth.

Figure 4.1: Wild-fires in the USA - Public Information Map (by ESRI, USA).

Figure 4.2: Restaurants in Port Elizabeth (capture from tripadvisor.co.za).

Even non-scientist may easily see trends and patterns and can make certain obvious judge-

ments about the underlying data and even make reasonable predictions.

Finally we mention such new (and very popular) method of visualisation as ”word-cloud”. The following figure represents the word-cloud of this report.

Figure 4.3: Word-cloud of the present report

In modern Statistics and Computerised Data Analysis, data visualisation plays a critical role (espe- cially in terms of spatial and big data). Existing and emerging methodologies allow to demonstrate the analytical findings in a format understandable to management, decision-maker and even the general public. SACO will benefit by implementing these within various types of presentations, including on the website of the South African Cultural Observatory.

Chapter 5

Conclusions

The predictive modelling process has the following major steps (http://blog.aureusanalytics.com/steps-to-building-a-statistical-model/):

1. Abstraction of a real problem to a data science problem.
2. Data collection from reliable sources.
3. Exploration of fields in the data set and understanding what they mean (exploratory data analysis)
4. Checking the data for any abnormalities like missing data.
5. Making sense of the data using descriptive statistics.
6. Fixing outliers.
7. Imputation of missing values.
8. Deciding on the type of models to be used, based on the nature of problem and data..
9. Calculating derived variables where required and making sure that the model and variables adequately capture real problem.
10. Dividing data into training, test and validation sets.
11. Estimating the parameters of the model ("training") using the "training" set.
12. Checking accuracy and adequacy of the model using both "training" (in-sample)and "test" (out-of-sample) sets.
13. If the results are not satisfactory, go back to point 8, tune the model and and remove variables (try-and-error process) until good results are obtained on both sets.
14. Validate the model on the validation set.

Similar recommendations may be found in [3]. Graphically the process is illustrated on the flow-chart 5.1.

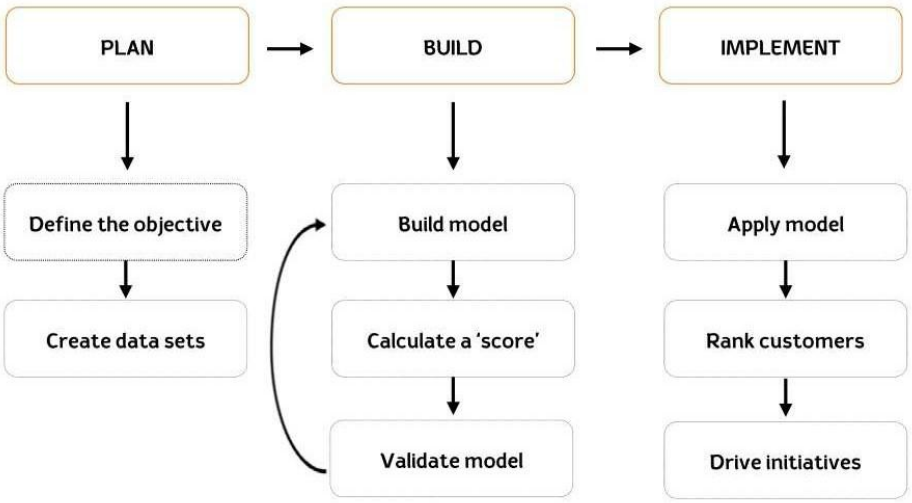

Figure 5.1: Predictive Model Design Process

Concluding, the results of this investigation are summarised the with following conclusions and recommendations seeming the most feasible:

- The process of modelling is currently at stages 7 - 8. Taking into account the considerable size of the mapping study dataset and the amount of inconsistencies and missing data, we would recommend engaging additional labour resources for fixing these.
- As regarding the models which look most promising we would recommend a hybrid spatial-panel regression model for measuring the direct impact; and
- Good's causality model for measuring both direct and indirect impact.

Possibly other models discussed earlier may come to fore, depending on the new data that may arrive from various sources.

Bibliography

[1] Engelbrecht A.P. Computational Intelligence: An Introduction. Wiley, 2 edition, 2007.

[2] Davis C.B. Making Sence of Open Data: From Raw Data to Actionable Insight. FSC, 2012.

[3] Datamine. Predictive modelling process. http://www.datamine.com/site/datamine/files /White%20Papers/predictive_modelling_process_whitepaper.pdf.

[4] Rubin D.B. Multiple Imputation for Nonresponse in Surveys. New York: Wiley & Sons, 1987.

[5] Rahm E. and Do H.H. Data cleaning: Problems and current approaches. http://dbs.uni-leipzig.de or http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.1530.

[6] Mlambo F. F. and Litvine I.N. Causality Test for Non-Stationary Time Series. In European Simulation and Modelling Conference, Porto, Portugal, pages 29–31. EUROSIS-ETI, 2014.

[7] Box G. and Jenkins G. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day, 1970.

[8] Box G. and Draper N.R. Empirical Model-Building and Response Surfaces. Wiley, 1987.

[9] Badi H.B., Bresson G., and Pirot A. Forecasting with spatial panel data. Technical report, Forschungsinstitut zur Zukunft der Arbeit Institute for the Study of Labor, 2009. IZA DP No. 4242.

[10] Litvine Igor. Moneymetrics: Numerical and Statistical Methods for Financial Analyst. Nautilus, 2007.

[11] Litvine I.N. LL model - theory and applications. In ITISE 2014, 2014. Granada, Spain.

[12] Cohen J., Cohen P., West S. G., and Aiken L. S. Applied multiple regression/correlation analysis for the behavioural sciences. New York, NY: Routledge, 3 edition, 2003.

[13] Good I. J. A Theory of Causality. The British Journal for the Philosophy of Science, 9(36):307–310, 1959.

[14] Good I. J. A Causal Calculus (I). The British Journal for the Philosophy of Science, 11(44):305–318, 1961.

[15] Granger C. W. J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. Econometrica, 37(3):424–438, 1969.

[16] Granger C. W. J. Some Aspects of Causal Relationships. Journal of Econometrics, 112(2003):69–71, 2003.

[17] Snowball J. Towards the development of a framework for cultural statistics for South Africa, 2016. Project report, Cultural Observatory of South Africa.

[18] Bollen K.A. and Pearl J. Eight myths about causality and structural equation models. In Morgan S.L., editor, Handbook of Causal Analysis for Social Research, pages 301–328. Dordrecht: Springer, 2013.

[19] Kendall M. A new measure of rank correlation. Biometrika, 30 (12):8189, 1938.

[20] Benoit Mandelbrot and Richard L. Hudson. The Misbehavior of Markets: A Fractal View of Financial Turbulence. Basic Books, 2004.

[21] Department of Arts and Culture. National mapping study of the cultural and creative industries, 2014. Prepared by Plus 94 Research.

[22] McCullagh P. and Nelder J.A. Generalized Linear Models. Chapman and Hall/CRC, 2 edition, 1989.

[23] Engle R.F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica, 50, (4):987–1007, 1982.

[24] Micali V.A. Ryabchenko O.O., Litvine I.N. Review of publications in energy forecasting modeling. part i: Energy models classification. INTERNATIONAL JOURNAL of ACADEMIC RESEARCH, 5:131–138, 2013.

[25] Micali V.A. Ryabchenko O.O., Litvine I.N. Review of publications in energy forecasting modeling. part ii: Energy models comperative analysis. INTERNATIONAL JOURNAL of ACADEMIC RESEARCH, 5:161–168, 2013.

[26] Adeli Siddique and Hojjat Nazmul. Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing. John Wiley and Sons, 2013.

[27] Anderson T.W. An Introduction to Multivariate Statistical Analysis. Wiley, 3 edition, 2003.

[28] UNESCO. UNESCO framework for cultural statistics, 2009. http://www.uis.unesco.org/culture/Pages/framework-cultural-statistics.aspx.