Harnessing data flow and modelling potentials for sustainable development

CODATA 22nd INTERNATIONAL CONFERENCE ON SCIENTIFIC DATA AND SUSTAINABLE DEVELOPMENT
24th -27th October-2010, Stellenbosch, Cape Town, RSA

Dr Kassim S. Mwitondi[1] and Dr Jamal B. Bugrien[2]

1. Sheffield Hallam University; Computing and Communications Research Centre; Sheffield S1 1WB United Kingdom. k.mwitondi@shu.ac.uk mwitondi@yahoo.com

2. University of Garyounis, Department of Statistics, Faculty of Science, PO Box 9480 Benghazi, Libya. E-mail: jbugrien@gmail.com

Presentation outline…

- Introduction
- Rationale and Motivation
- Aims and objectives
- Methodology
- Analyses and discussions
- Concluding remarks

Introduction

- DATA===>INFORMATION===>KNOWLEDGE is fundamental for our existence.
- We propose a fundamental approach to transforming data into knowledge.
- A generic data sharing model providing access to data utilising and generating entities.
- An unsupervised and supervised modelling demonstrated via simulated and real data
  - Accuracy and reliability
  - Multidisciplinarity
- Impact on STI-the social transformation engine

Rationale and Motivation

- Systems sustainability
  - Ecosystem - organisms, air water, soil, sunlight
  - Social infrastructure - business, environment...
- Tapping into data and information flow has always been an integral part of the human race
- Disparate approaches imply knowledge gaps…
  - Social computing (Wang et al., 2007)
  - Scientific computing (Rushing et al., 2005)
  - Ubiquitous computing as well as web and business computing discussed by many authors.
- Knowledge extraction from data remains in a finite scope of time, concept, data and location.

Aims and objectives…

- Highlighting the influence of information flow in generating knowledge from data and use it as a mean and output in social transformation via...
- A framework for implementing a coherent data flow system across disciplines and regions
- Extracting and utilising knowledge from data as a basis for effecting successful applications of STI

Some basic considerations

- Data-based decision errors are typically attributed to disparities in data sources and modelling techniques.

- Geographically diverse data, software and hardware resources can now be aggregated as a platform to create dynamic, adaptive and robust knowledge tools and products with universally acceptable attributes.

- The complex nature of socio-economic systems entails diverse knowledge domain issues which must be properly addressed for the aggregated knowledge to be recognised as a tool and product of social transformation.

Methodology – data description

- Simulated data: 500 simulations from a uniform distribution and 500 corresponding coefficients for each data point labelled -1 and 1 such that

$$\beta_i = \{-1, 1\}^K = \begin{cases} -1 \; if \; x_i \in k \\ 1 \; if \; x_i \notin k \end{cases}$$

- Real data: 199 observations on 8 variables condensed into two super-attributes.

- In both cases natural groupings are induced.

Modelling methods

Unsupervised modelling

$$P(x) = \frac{1}{NS_1S_2 \dots S_p} \sum_{i=1}^{N} \prod_{j=1}^{P} K_j\left(\frac{[x - x_i]_j}{S_j}\right)$$

$$P(x) = p(x, S)\frac{1}{N} \sum_{i=1}^{N} |S|^{-\frac{1}{2}} K\left(S^{-\frac{1}{2}}(x - x_i)\right)$$

Supervised modelling

where ***K(.)*** in SM is a ***p-***variate spherically symmetric density function and ***S*** is a symmetric positive definite matrix

Iterating to handle allocation rule errors…

$$P(Y|X_1, X_2, \dots, X_\lambda) = \frac{P(X_\lambda|Y)P(Y|X_1, X_2, \dots, X_{\lambda-1})}{\int P(X_\lambda|Y)P(Y|X_1, X_2, \dots, X_{\lambda-1})dY}$$

ALLOCATION RULE ERRORS DUE TO DATA RANDOMNESS			
POPULATION	TRAINING	CROSS VALIDATION	TEST
$\psi_{D,POP}$	$\psi_{D,TRN}$	$\psi_{D,CVD}$	$\psi_{D,TST}$

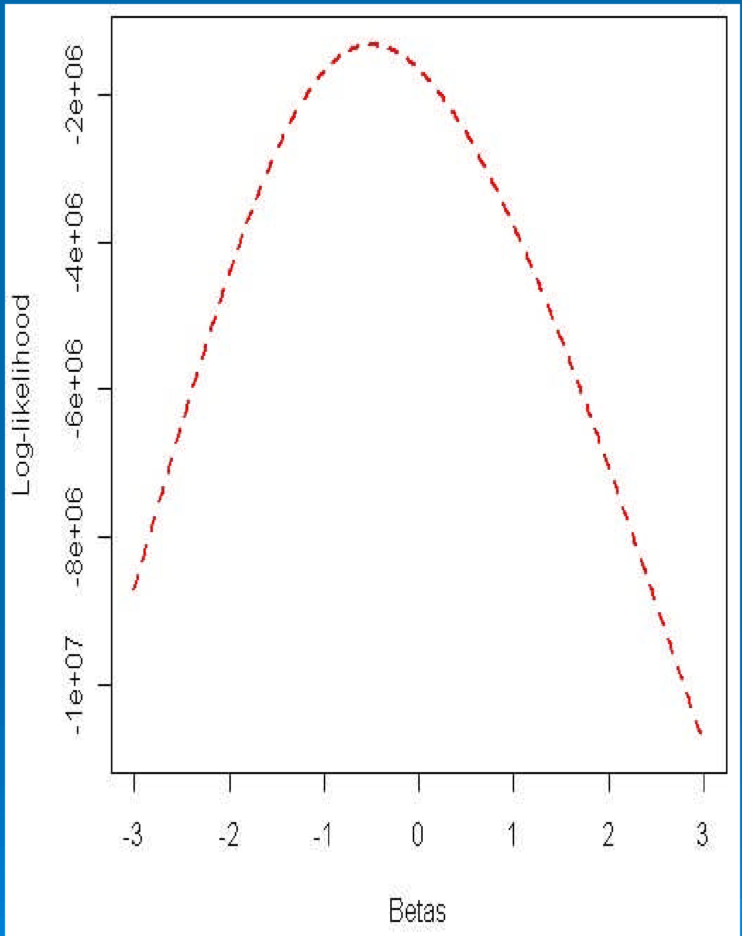
Source: Mwitondi (2003)

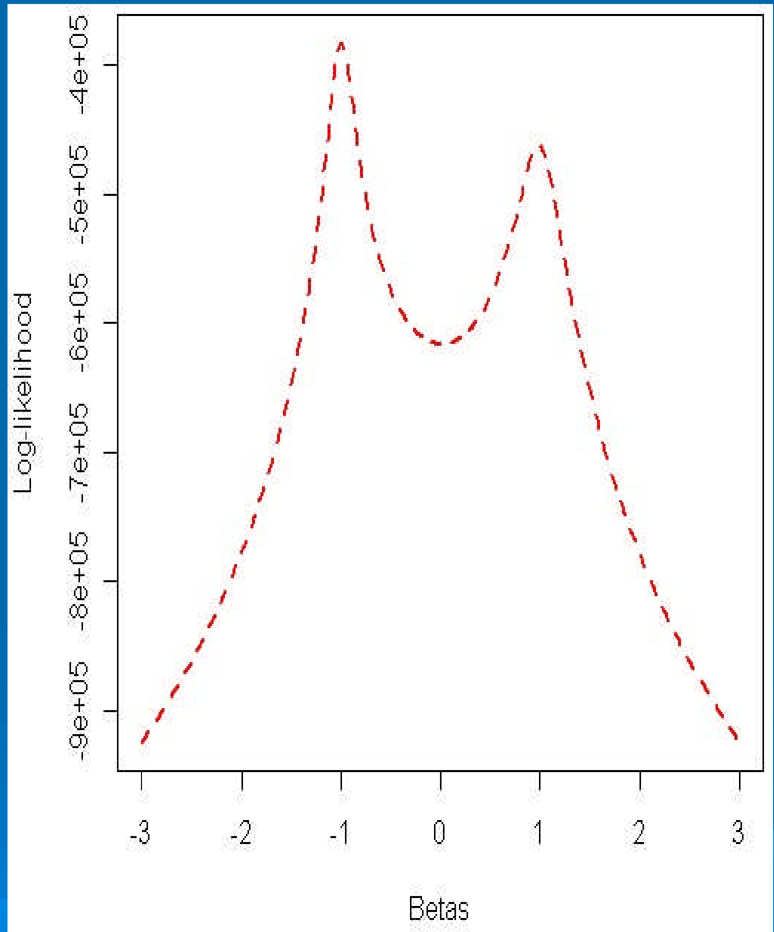
$$\Psi_{D,CVD} = \sum_{k=1}^{K}\sum_{i=1}^{N} \pi_k P(X_i \in C_k | Y \notin C_k)$$

The challenge is to minimise the error while maintaining model reliability

Simulated data results

Initial iteration

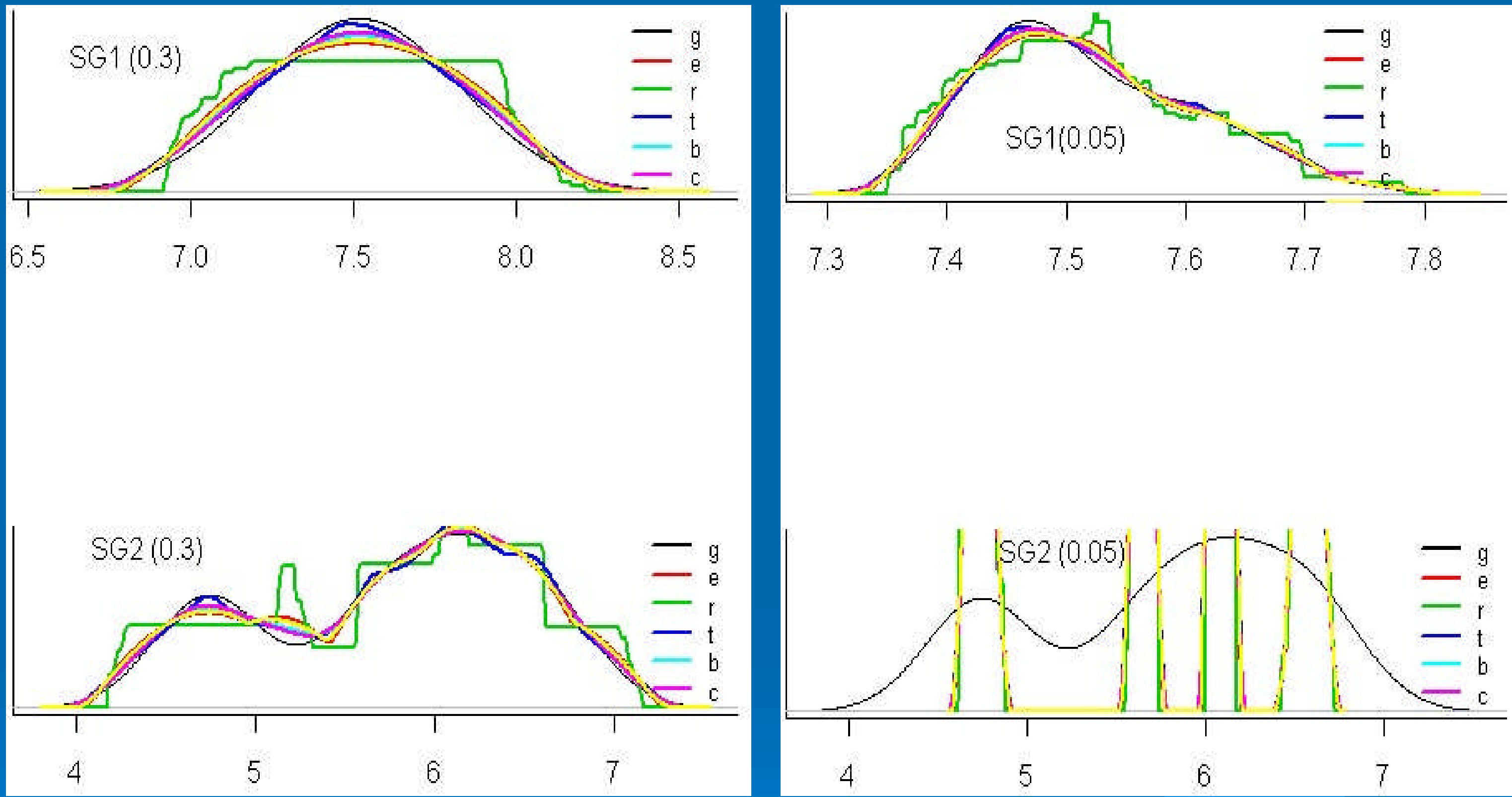




Final iteration

Real data results

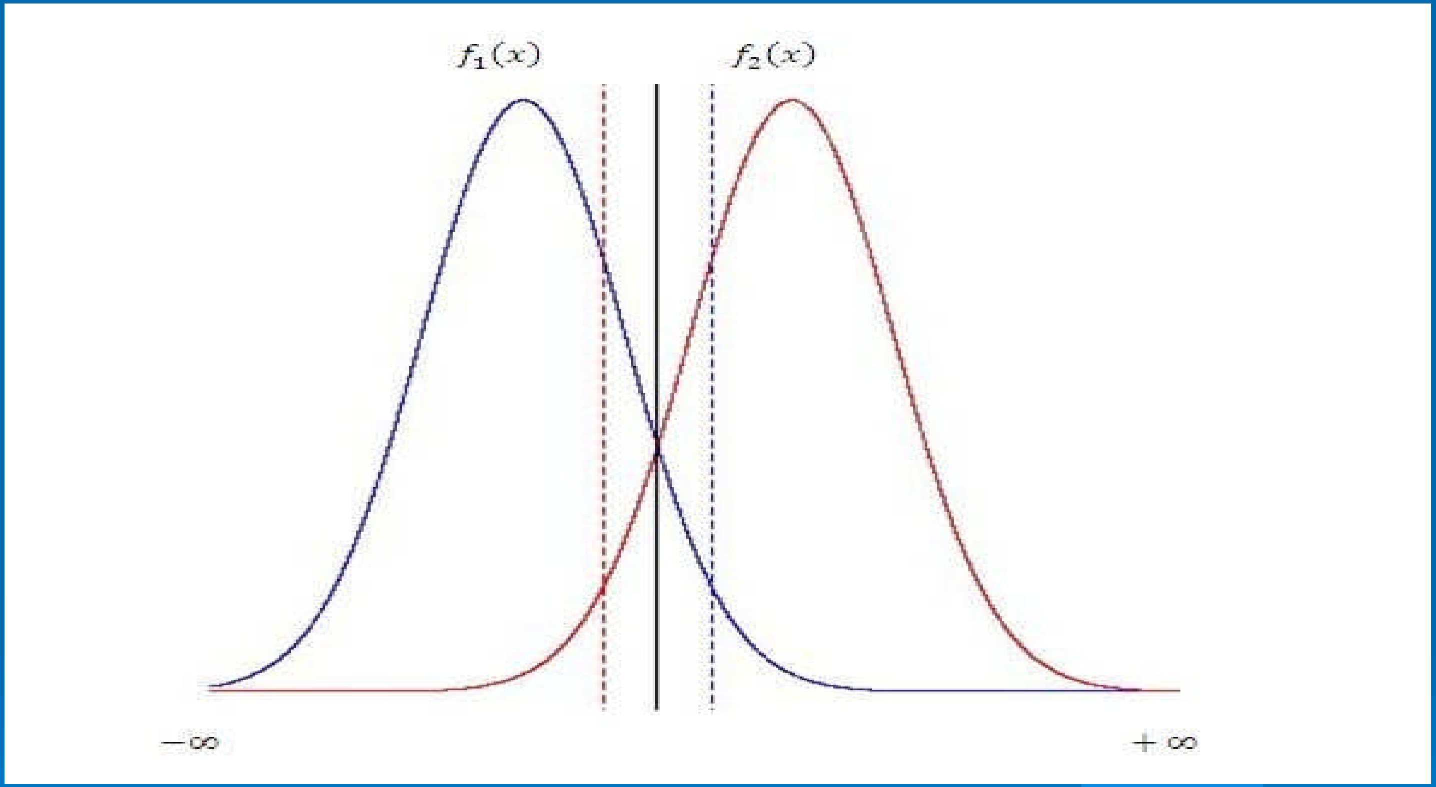
Group densities at 0.3



Group densities at 0.05

Seven kernels - Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine and the Optcosine. Focus is on the choice of the key parameters (eg bandwidth) method-data relationship

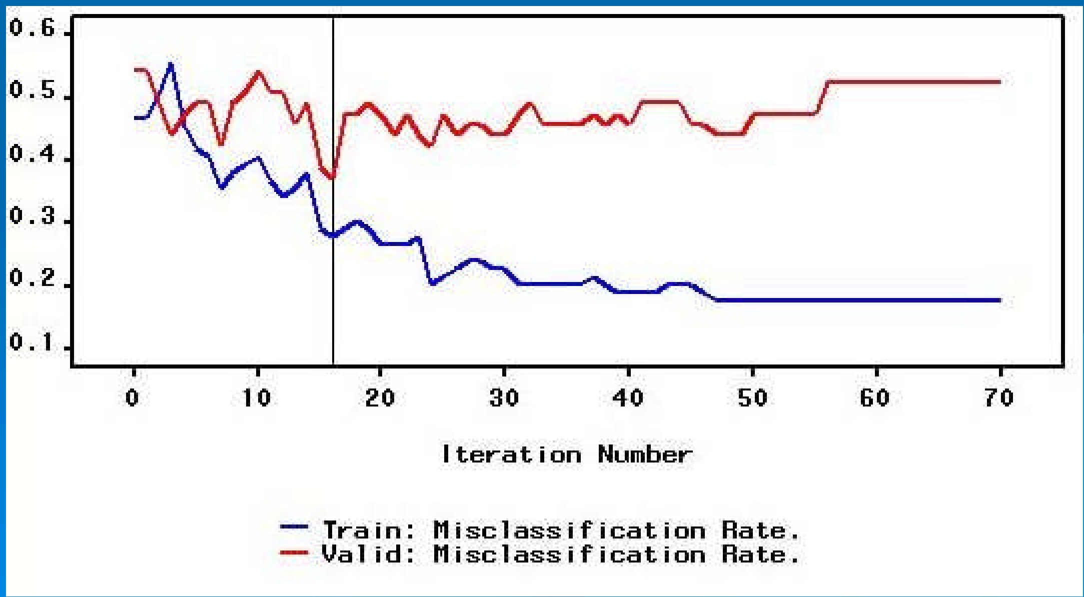
Predictive modelling…



The challenge is to minimise the error while maintaining model reliability

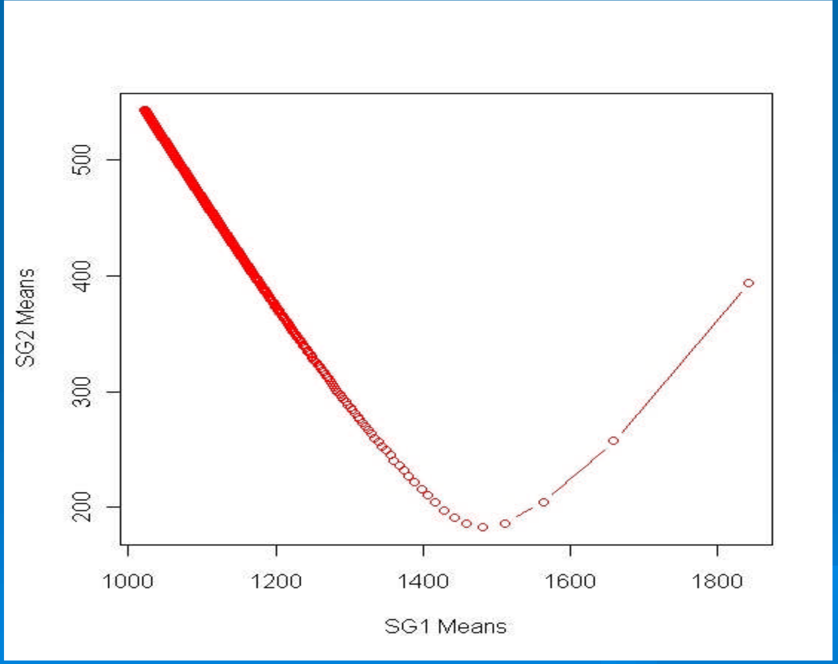
Data-generated parameters for model updating: Neural networks model 1

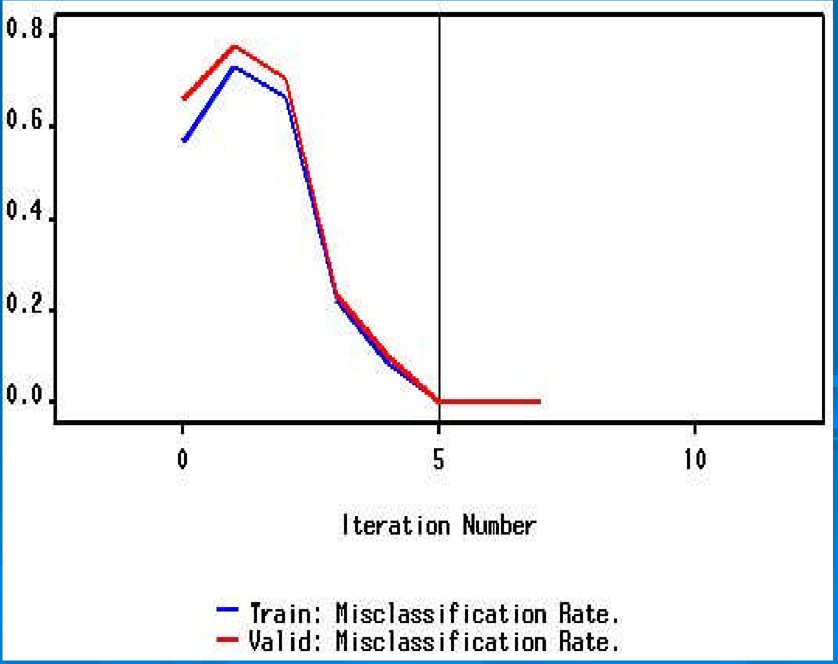
- An NN model with 5 hidden neurons, a logistic activation and an additive combination functions was fitted for a maximum likelihood function. Optimal model reached after 17 iterations - low accuracy (27.85% and 37.29%)



Re-labelled date: Neural networks model 2

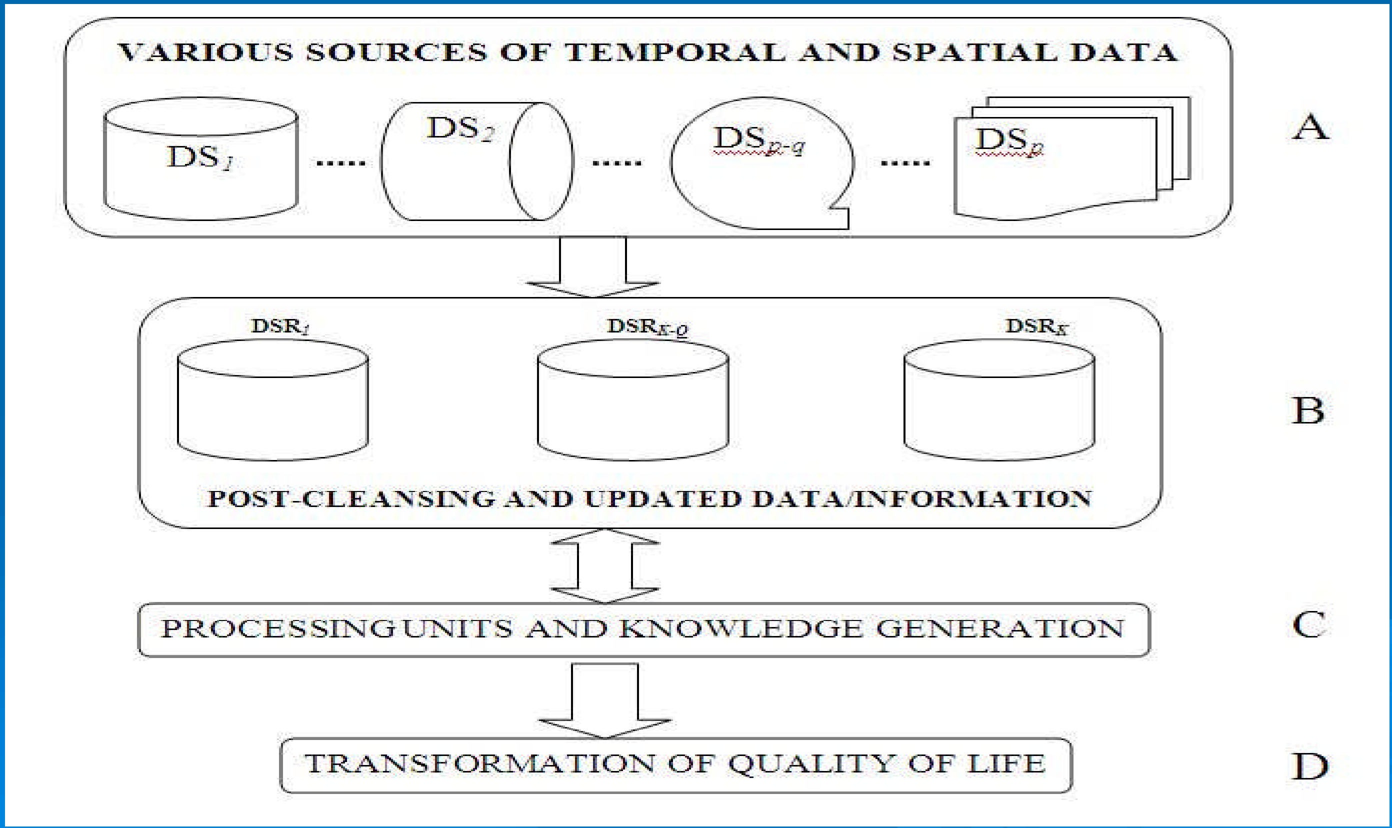
- Re-labelling data using means patterns yielded a very high accuracy of approximately 0.5% and an important feature – resisting over-fitting





The proposed data housing shell

- All the foregoing ideas could be embedded into a cohesive knowledge generating system



Data housing shell levels explained

LEVEL	ELEMENTS	PREREQUISITES
A	Primary and secondary data sources generating cleansed data and data repositories across regions and disciplines.	Supportive national policies, appropriate knowledge and skills, data acquisition tools, R&D and KTP initiatives.
B	Cleansed data, updatable data repositories, model and parameter-related information.	Tools for data capturing, storage and dissemination.
C	Research centres, public, private firms, academia, R&D, KTP, individuals for enforcing collaboration.	Computing and data mining tools, methodologies and techniques preferably within the cloud computing environment.
D	Transforming knowledge into tangible outputs - patents, publications, products and services	Financial, human, and technical resources. Supportive policies, legislative, social and technological infrastructure.

Summary

- Natural and social dynamics cause changes in socio-techno attributes - government policies, consumer behaviour, gene mutation, carbon emission and related technologies.
- Result - concept drift (see Karnick, 2008) - key properties of predictive model outputs change
- Apparently, these dynamics impinge on the overall accuracy and reliability of the models which is what the focus of the proposed model

And before you ask...

- Only a handful issues have been addressed in this paper. Challenges remain - model complexity and inter-regional policies/issues.
- The money? We try free lunch, when we can

TOOL/S	USABILITY/AVAILABILITY
MySQL, PHP, PERL, APACHE (From XAMPP) http://www.apachefriends.org/en/xampp.html http://www.php.net	Connectivity/Open
R: http://www.r-project.org	Analytical/Open
LaTex: http://www.latex-project.org Open Office: http://www.openoffice.org	Documentation(Reporting)/Open
BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi	Heuristic search/Open Access

Bibliography...

- Karnick, M., Ahiskali, M., Muhlbaier, M. and Polikar, R. (2008). Learning Concept Drift in Nonstationary Environments Using an Ensemble of Classifiers Based Approach; World Congress on Computational Intelligence/IEEE International Joint Conference on Neural Networks, Hong Kong, 1-6 June 2008, pp 3455-3462, ISNN 978-1-4244- 1821-3.
- Mwitondi, K. (2009). Tracking the Potential, Development, and Impact of ICT in Sub-Saharan Africa; In: Science, Technology, and Innovation for Socio-economic Development: Success Stories from Africa; Published by (ICSU-ROA), ISBN 978-0-620-45741-5.
- Mwitondi, K. and Ezepue, P. (2008). How to appropriately manage mathematical model parameters for accuracy and reliability: A case of monitoring levels of particulate emissions in ecological systems; International Conference on Mathematical Modelling of Some Global Challenging Problems in the 21st Century; Proceedings of NMC-COMSATS Conference on Mathematical Modelling of Global Challenging Problems - 26th-30th, Nov. 2008; pp 24-36, ISBN 978-8141-11-0.
- Mwitondi, K., Taylor, C. and Kent, J. T. (2002). Using Boosting in Classification; Proceedings of the Leeds Annual Statistical Research (LASR) Conference; July 2002; pp. 125 – 128, Leeds University Press.
- Rushing, J., Ramachandran, R., Nair, U., Graves, S., Welch, R. and Lin, H. (2005). ADaM: A data mining toolkit for scientists and engineers; Computers & Geosciences, Vol. 31, Issue 5, pp 607-618, ISSN 0098-3004.
- Wang, F-Y, Carley, K., Zeng, D. and Mao, W. (2007). Social Computing: From Social Informatics to Social Intelligence; Intelligent Systems, IEEE, Vol. 22, Issue 2, pp 79-83, ISSN: 1541-1672.

ANY QUESTIONS?