Harnessing data flow and modelling potentials for sustainable development

MWITONDI, Kassim and BUGRIEN, Jamal

Published version

MWITONDI, Kassim and BUGRIEN, Jamal (2012). Harnessing data flow and modelling potentials for sustainable development. CODATA Data Science Journal, 11, 140-152.

HARNESSING DATA FLOW AND MODELLING POTENTIALS FOR SUSTAINABLE DEVELOPMENT

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

*[*1]Sheffield Hallam University, Computing and Communications Research Centre, Sheffield S1 1WB, UK*
Email: k.mwitondi@shu.ac.uk; mwitondi@yahoo.com
[2]University of Garyounis, Dept. of Statistics, Benghazi, Libya
Email: jbugrien@gmail.com

ABSTRACT

Tackling the global challenges relating to health, poverty, business, and the environment is heavily dependent on the flow and utilisation of data. However, while enhancements in data generation, storage, modelling, dissemination, and the related integration of global economies and societies are fast transforming the way we live and interact, the resulting dynamic, globalised, information society remains digitally divided. On the African continent in particular, this division has resulted in a gap between the knowledge generation and its transformation into tangible products and services. This paper proposes some fundamental approaches for a sustainable transformation of data into knowledge for the purpose of improving the people's quality of life. Its main strategy is based on a generic data sharing model providing access to data utilising and generating entities in a multi-disciplinary environment. It highlights the great potentials in using unsupervised and supervised modelling in tackling the typically predictive-in-nature challenges we face. Using both simulated and real data, the paper demonstrates how some of the key parameters may be generated and embedded in models to enhance their predictive power and reliability.

The paper's conclusions include a proposed implementation framework setting the scene for the creation of decision support systems capable of addressing the key issues in society. It is expected that a sustainable data flow will forge synergies among the private sector, academic, and research institutions within and among countries. It is also expected that the paper's findings will help in the design and development of knowledge extraction from data in the wake of cloud computing and, hence, contribute towards the improvement in the people's overall quality of life. To avoid running high implementation costs, selected open source tools are recommended for developing and sustaining the system.

Keywords: Cloud computing, Data mining, Digital divide, Globalisation, Knowledge transfer partnership, Predictive modelling and science technology and innovation (KTP)

1 INTRODUCTION

Tackling the global challenges relating to health, poverty, business, and the environment is heavily dependent on the flow and utilisation of data. However, while enhancements in data generation, storage, modelling, dissemination, and the related integration of global economies and societies are fast transforming the way we live and interact, the resulting dynamic, globalised, information society remains digitally divided. On the African continent in particular, this division has resulted in a gap between the knowledge generation and its transformation into tangible products and services which Kirsop and Chan (2005) attribute to a broken information flow. Many institutions, bodies, and individuals across Africa find themselves charged with the responsibility for improving the socio-economic prosperity of the continent. The geographical, socio-economic, legislative, technological, and intellectual diversities within and among countries and regions entail a unified development delivery team with a common goal – that is, to improve the people's overall quality of life. It is worth noting, however, that the vision, strategies, priorities, and resources among the stakeholders may vary fundamentally. Dossiers on unifying the initiatives of the disparate stakeholders committed to the development goal are well-documented (see, for instance, Howells, 2005).

The outcomes of implementing good governance, poverty alleviation, and economic integration as well as science, technology, and innovation (STI) schemes across Africa, like elsewhere else, are a function of data transformation into information and knowledge. An information flow system capable of delivering the ultimate goal of improving the people's quality of life must be one that both the stakeholders and the beneficiaries (the people) can trust. This paper seeks to answer the general question as to how data and information flow potentials can be harnessed for socio-economic prosperity. It proposes a robust, accurate, and reliable information flow and archiving system and demonstrates it via an enhanced predictive neural networks model with re-labelled classes. The paper is organised as follows. Sub-sections 1.1 and 1.2 provide the paper's motivation, its aims, objectives, and scope. The current STI and data utilisation scenario is given in Section 0and an outline of the study methodology is in Section 0. Data analyses and discussions of results are presented in Section 0and concluding remarks and recommendations are in Section 0

1.1 Study rationale and motivation

The sustainability of any system is typically dependent on the interdependence among its constituent parts as demonstrated by examples from our ecosystem and the related social infrastructure. Across Africa, cases of local people being exposed to toxic material as a result of mining activities or industrial scale fishing diverting nutritional sources from villagers are commonplace (Campbell et al., 2003). Such activities and incidents aggravate the vicious cycle of poverty, and they have been widely attributed to a range of factors including the digital divide (Warren, 2007). It is therefore imperative to establish a cohesive inter-disciplinary information flow management and utilisation system that will help explain and resolve some of the main socio-economic challenges.

Tapping into data and information flow potentials has always been an integral part of the human race, and in recent years, researchers have taken different approaches towards optimal utilisation of data and/or information. Social computing (Wang et al., 2007), scientific computing (Rushing et al., 2005; Gray et al., 2005), and web/business computing are household concepts all aimed at filling knowledge gaps in our societies. Globalisation and the associated technological developments have transformed tremendously the way we live and interact (Dreher et al., 2008). We are entering the second decade of the 21st century with a clear and full knowledge of its magnitude, direction, and influence on our societies. There is no doubt that the fast evolving computing needs, power, and growing potential require a harmonised computing environment built on solid interdisciplinary foundations.

As the movement of financial and human resources, stores, and market stalls continues to defy geographical boundaries, stiff competition remains part and parcel of global businesses. As classrooms, lecture theatres, scientific laboratories, consultancies, and call centres span across the continents of our increasingly dwindling world, new challenges and opportunities arise. Thus, in the current bumpy and murky global terrain, our socio-economic development and prosperity depends much on how we face the emerging challenges and grab the arising opportunities. In the course of doing so, we generate and depend on knowledge - the generation, dissemination, and utilisation of which require an accurate and reliable access to ubiquitous computing methods and resources. These are the main focal points of this paper.

1.2 Study aims, objectives, and scope

The paper's key objective is to highlight the influence of information flow on the development and prosperity of the African continent and identify the framework for sustainable implementation of sharable data repositories and sources. Indeed, a unified development delivery team must clearly identify the needs, means, and ends of the development subject. It can only achieve this by having in place a coherent information flow system capable of being interrogated to a level of detail appropriate for addressing or predicting any of the issues therein. To explore the influence of information flow on the development and prosperity of the African continent through various channels, we focus on knowledge as a complex function of data input, a mean, and an output in a social transformation system. Building on the views of innovation in Mwitondi (2010) and Mwitondi and Ezepue (2008), the paper seeks to utilise old and current ideas to promote new ones. The ultimate goal is to create and

utilise knowledge from data as a basis for effecting successful applications of science, technology, and innovation for socio-economic prosperity.

2 AN OVERVIEW OF THE CURRENT SCENARIO AND DATA UTILIZATION POTENTIALS

Across the African continent, the direction of computing, data, and information manipulation and flow still largely depends on what governments do or don't do. Thus, many researchers have recently focused on ICT policies and how they impinge on data acquisition and utilisation within the continent and beyond. There are many examples of what African governments have done or could do to promote science, technology, and innovation (STI) and realise their development objectives through ICT (Juma, 2005; Ludvall & Borras, 2005). Yet, with all the hype about the way the information age has transformed our lives, despite the numerous unilateral and multilateral pacts, declarations, and policy adoptions we have seen over the years, the digital divide, poverty, and social imbalances are still the key discriminating parameters between the societies in the two hemispheres (Mwitondi, 2009).

Extracting knowledge from data has never been an easy task, and many decisions that go wrong can typically be attributed to disparities in data sources and modelling techniques. Thus, the key ideas in this paper derive from globalisation and its associated attributes, such as great enhancements in computing power. For instance, developments in cloud computing - described by Grossman (2009) as a computing environment based on clusters of distributed computers providing on-demand resources and services over the Internet - provide us with new challenges and opportunities in all matters relating to information processing and utilisation. In particular, Buyya et al. (2009) cast a light on how geographically diverse data, software, and hardware resources can be aggregated as a platform to create dynamic, adaptive, and robust knowledge tools and products with universally acceptable attributes. However, the complex nature of socio-economic systems entails the presence of issues relating to diverse knowledge domains that must be properly addressed for knowledge to be recognised as both a tool and product of social transformation. In the following exposition, we highlight how the foregoing ideas can be transformed into practice.

3 METHODOLOGY

Despite emerging patterns from data analytics forming fundamental bases for decision making processes in almost all disciplines, most remain defined in a finite scope with respect to time, concept, data, and location. This paper's methodology is an extension of earlier ideas on data harmonisation developed in Mwitondi and Ezepue (2008) and Mwitondi (2009). Most of the problems we face in real life are predictive in nature, and so they are dependent on existing knowledge to generate new knowledge. The two most common approaches to knowledge extraction from data are supervised and unsupervised modelling. The former describes knowledge extraction from labelled data while the latter refers to identifying natural groupings in data without much information relating to the exact nature of the groupings. A typical illustration of the latter approach is via the kernel density estimation (Webb, 2005)

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

where each of the $K_j(.)$ is the kernel function and S_p is the corresponding smoothing parameter on the adjustment of which depends the modality of the univariate data. In the supervised context, the foregoing formula transforms to

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

where the kernel ***K(.)*** is a ***p***-variate spherically symmetric density function and ***S*** is a symmetric positive definite matrix typically defined as $S_k^2\widehat{\Sigma}_{\mathrm{k}}$ for the $\boldsymbol{k^{th}}$ class with the two components representing class scaling and the sample covariance matrix. In both unsupervised and supervised modelling, the main issue is choosing the smoothing parameter for which various methods have been suggested including cross-validation.

We focus on Duin (1976) in which the cross-validation-based smoothing parameter is chosen to maximise the product of probabilities $\prod_{\mathrm{j}=1}^{\mathrm{P}}\mathrm{P_i}(\mathrm{x_i})$. Indeed, to carry out both tasks, we need access to data - some of which may have to be estimated from available data sources. To do so we employ estimating methods, and we often rely on prior expert knowledge that inevitably impinges on modelling accuracy and reliability (Mwitondi & Ezepue, 2008). Thus, in both unsupervised and supervised modelling, how we proceed in the case of *"the unknown"* depends much on what we would do in the case of *"the known"* (Mwitondi et al., 2002). Without loss of generality, let us consider the predictive model in (3)

$$P(Y|X_1,X_2,\ldots,X_\lambda) = \frac{P(X_\lambda|Y)P(Y|X_1,X_2,\ldots,X_{\lambda-1})}{\int P(X_\lambda|Y)P(Y|X_1,X_2,\ldots,X_{\lambda-1})dY} \quad (3)$$

where $\boldsymbol{Y}$ represents some target phenomenon and $X_{\lambda=1,2,3,\ldots}$ are the attributes on which it depends. If we assume that $\boldsymbol{Y}$ and $X_{\lambda=1,2,3\ldots}$ are continuous and that all predictors are independent, we can obtain the expected posterior probability by iterating λ times, starting with the prior probability of observing the fundamental parameter $P(Y)$.

Now, if we further assume that ***Y*** represents the class variable reflecting socio-economic prosperity and that the arbitrary predictors are defined across disciplines and regions, then we can set a goal of capturing and measuring the impact of any interactive effects among the predictors on the target variable $\boldsymbol{Y}$. However, as a rule, in both unsupervised and supervised modelling we do expect to make prediction errors simply because in both cases we can distinguish two types of allocation rules - theoretical and empirical. The formulations above are based on the former, which assumes known priors and densities and can therefore be tested on a notionally infinite test dataset. The overall empirical error arises from randomness due to the allocation region and randomness due to the allocation rule assessment by random training and validation data (Mwitondi, 2003), as shown in Table 1.

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

Table 1. Error types associated with unsupervised and supervised modelling

Thus, the overall misclassification error for each one of the errors in Table 1 is the sum of the weighted probabilities of observing data belonging to a particular class given that we are not in that class. For instance, the overall cross-validation error can be defined as

$$\Psi_{D,CVD} = \sum_{k=1}^{K} \sum_{i=1}^{N} \pi_k P(X_i \in C_k | Y \notin C_k) \qquad (4)$$

where C_k and π_k represent the partition region and the class priors respectively. The challenge is to minimise the error - that is, attain accuracy while maintaining model reliability. Rather than relying on sequential data interrogating algorithms as in Gray et al., (2005), we recommend a combination of parallel iterative models capable of generating posterior information in each problem domain λ while monitoring the behaviour of the two parameters. In the next section we describe the proposed model prototype.

4 DATA DESCRIPTION, ANALYSES, AND RESULTS

To illustrate the practical implementation of the foregoing methods, we use both simulated and real data. The former is based on $x_{i=1,2,\ldots,500}$ simulations from a uniform distribution and 500 corresponding coefficients for each data point labelled -1 and 1 such that

$$\beta_i = \{-1, 1\}^K = \begin{cases} -1 \; if \; x_i \in k \\ 1 \; if \; x_i \notin k \end{cases} \qquad (5)$$

We then create a dependent variable $\mathbf{Y}$ and add to it a random noise. In the latter case, 199 observations on 8 variables were obtained from a leading African stock exchange. The 8 attributes – selected as potentially strong socio-economic drivers - were condensed into two super-attributes, SG1 and SG2. In both cases the aim was to induce natural groupings.

4.1 Simulated data example

Thus, in this example we illustrate the maximisation of the target class $\mathbf{Y}$ by maximising its likelihood $\mathcal{L}(Y|x_i) = \prod_i^N P(x_i|Y)$ where the data are held fixed at observed values. Taking logarithms to both sides we obtain $log\mathcal{L}(Y|x_i) = \sum_{i=1}^{N}[P(x_i|Y)]$ and so we can maximise the sum (log-likelihood) rather than the product (likelihood). The two maximisations can reach their maximum values for the same value of Y. Note that the formulations in Eq. (1) through Eq. (3) describe likelihoods justifying the use of the foregoing simulations because it is known that the likelihood estimator coincides with the most probable Bayesian estimator given a uniform prior distribution on the parameters (Tan et al., 2009).

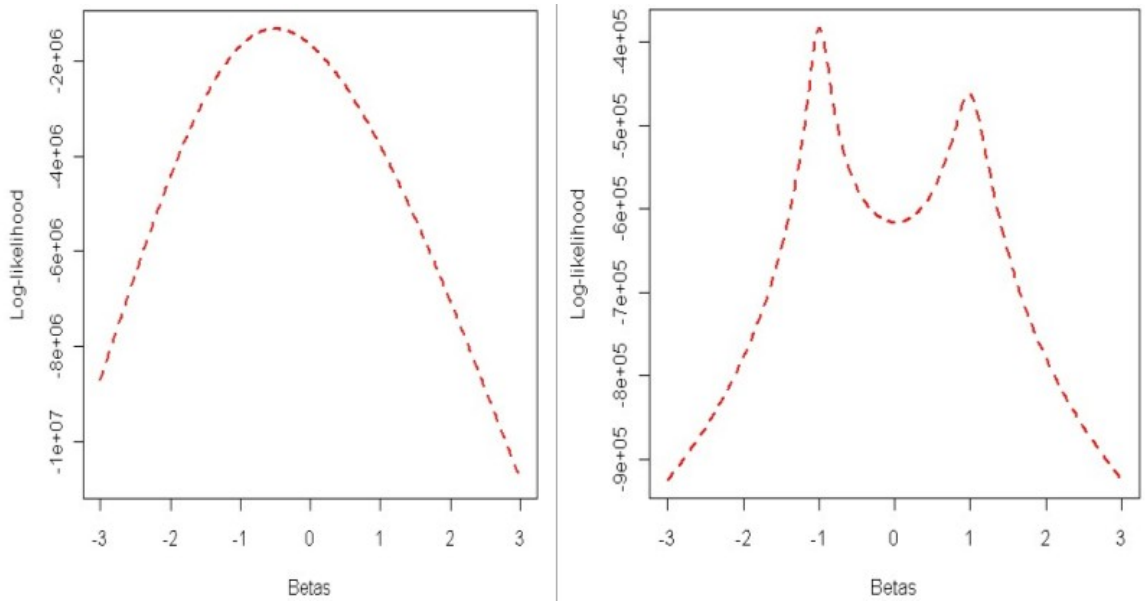

Figure 1. Log-likelihood plots reached after 1 and 500 iterations (LHS and RHS respectively)

On the basis of the above mentioned simulated data, the plots in Figure 1 were generated by a simple iterative algorithm written in the statistical language **R.** The uni-modal left hand side panel was obtained at the initial iteration while the bi-modal right hand side was reached after 500 iterations. As expected, data variability was fundamental in determining the shapes of the plots with lower and higher variability exhibiting masking and swamping effects.

4.2 Real data example

This example highlights some of the challenges in identifying naturally arising structures in data. The plots in Figure 2 represent the log transforms of the data variables based on seven kernel densities – the Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine - corresponding to the two groups at bandwidths 0.3 and 0.005. In both this and the previous example, the main challenges come not only from the choice of the key parameters such as the bandwidth but also from the methods used. Note how the kernels are influenced by the data in both cases - particularly how the 0.05 bandwidth works much better with SG1 for all kernels than with SG2 while at 0.3 the rectangular kernel does better with SG2 than with SG1.

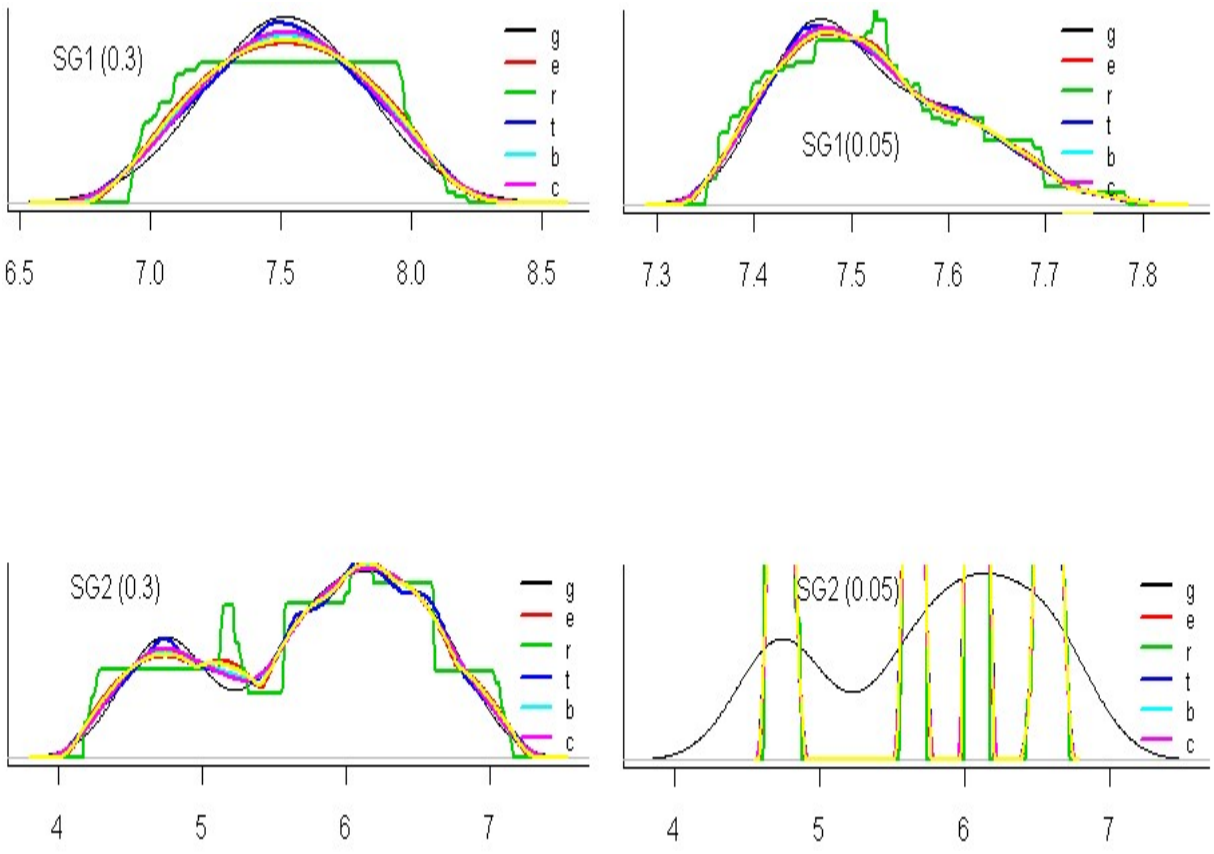

Figure 2. Kernel density estimation for the two groups of data attributes

Both the simulated and real data examples can be used as inputs in determining class labels, which we illustrate in the following example using neural networks.

4.3 predictive modelling example

In the predictive scenario, the challenge is to allocate new cases to previously known groups in which, in the case of two over-lapping normal densities, the typical problem would be to accurately estimate and separate the densities and their underlying parameters. Even in the unlikely scenario of known densities and class priors, the optimal classifier yields a natural class overlap between the upper tails of one density and the lower tails of the other with data randomness typically swaying the classifier into either direction of the over-lapping region.

In the following example, we focus on how we can make use of data-generated parameters as guidelines in updating class labels. The two attributes SG1 and SG2 were labelled by values in the performance index variable with a variance of 1.85 and range 6.8. To use it as a target, we discretised the variable into three classes following the two-class logic of the normal before fitting a neural networks model with 5 hidden neurons, a logistic activation additive combination function for a maximum likelihood function. The prediction results in Figure 3 show that the optimal model is reached after just 17 iterations with high training and validation errors at 27.85% and 37.29% respectively. As a rule, variations are likely to be observed across data and methods. For instance, re-labelling the classes would yield different patterns as new allocation rules are generated.

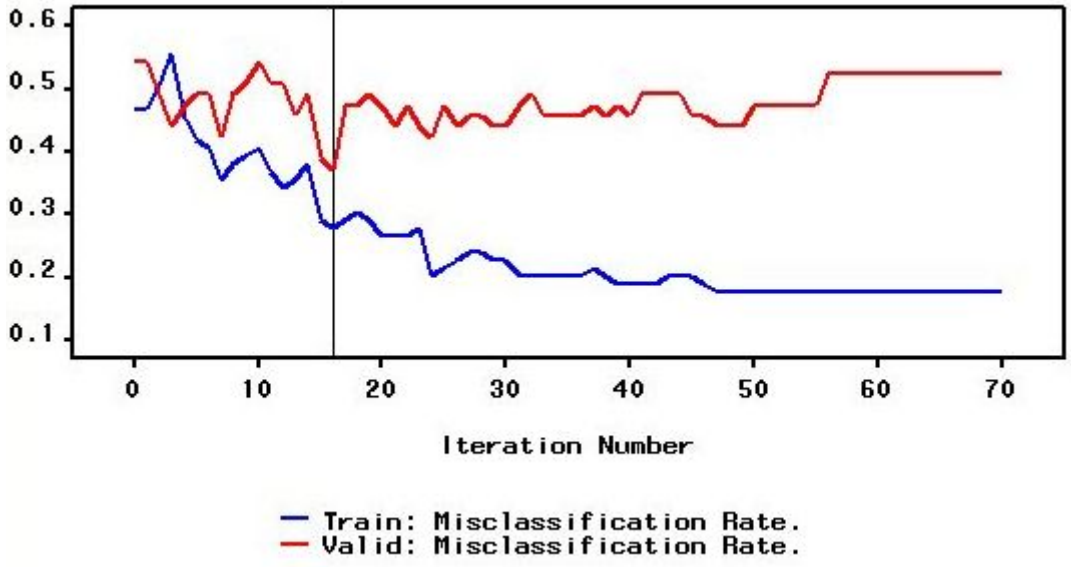

Figure 3. Neural networks results in class prediction of SG1 and SG2

Data labelling or re-labelling is typically guided by the approaches discussed in the methodology. Further, increasing or decreasing training and validation sets would also impinge on the model accuracy and reliability. Conventional solutions in managing the error in Eq. (4) include choosing from multiple cross-validation models and/or implementing the model in Eq. (3). To obtain the mean convergences, the two attributes were treated as a random sample of observations from a parametric finite mixture density with corresponding group means and a common variance. Group proportions were then computed and used to iteratively generate and update new group means as shown in Figure 4.

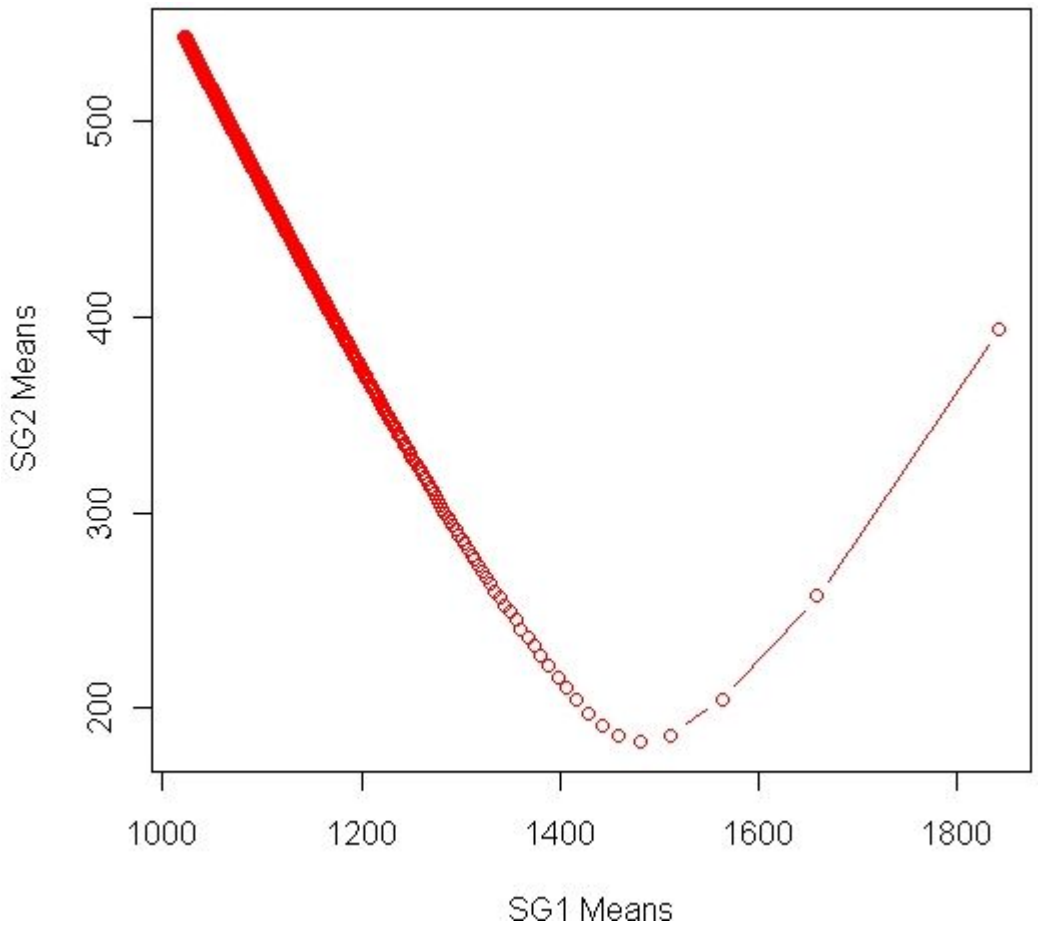

Figure 4. Convergence of the means for the two attributes

From the initial values of SG1=1841.96 and SG2=394.075, both means monotonically decreased up to SG1=1450 and SG2=180 when the latter started ascending. We used the two mean values as inputs to generate a new standardised variable in which we discretised three classes based on the same rule used to generate Figure 3 and implemented a neural networks model with exactly the same settings as the one used in the previous example.

The neural networks model output based on the newly generated class labels is shown in Figure 5. Note the great improvement in prediction accuracy for both training and validation error rates - from 27.85% and 37.29% respectively to under 0.5%. The most interesting feature observed is that the model is resistant to over-fitting despite the high accuracy rate.

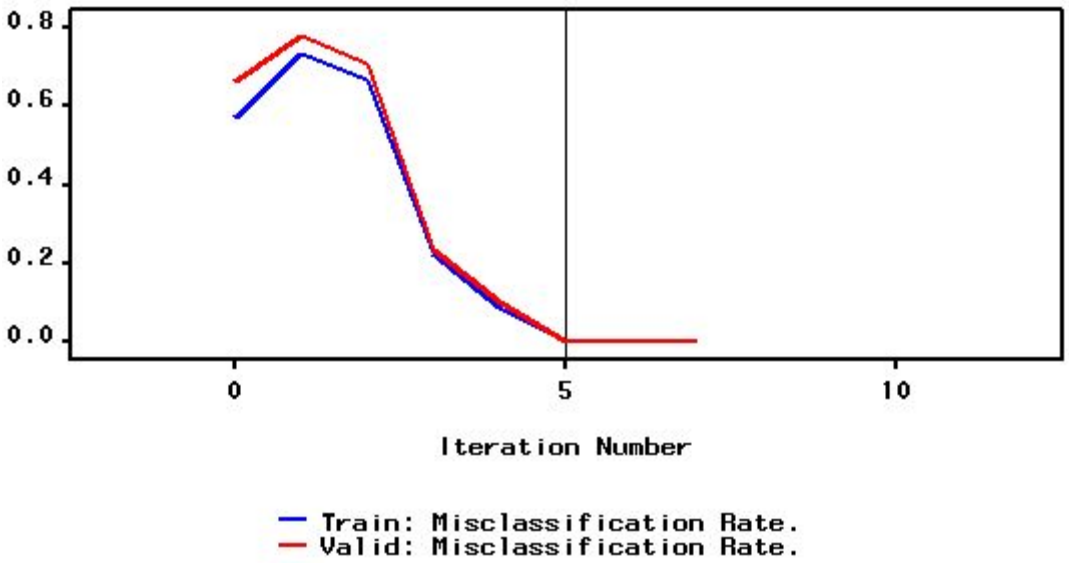

Figure 5. Neural networks prediction patterns using re-labelled data

The performance enhancement in Figure 5 as a derivative of the original model in Figure 3 and parameter inputs from Figure 4 summarise what we referred to in earlier developing models with the capability of generating posterior information while observing the model behaviour. In this case model complexity may be addressed through tracking and monitoring the neural networks weights and architecture and the outcomes from applications of a similar nature.

5 CONCLUDING REMARKS AND DISCUSSIONS

This paper followed Mwitondi (2010) in which science, technology, and innovation were identified as key drivers in socio-economic transformation and knowledge was presented as both a tool and output. It is built on the premise that most of the problems we face are predictive in nature. Today, the volume of data generated by web applications as cloud computing gathers momentum far outstrips analytical capacities - hence we simultaneously face challenges and opportunities to extract knowledge from data. The main idea is to demystify the large volumes of data into comprehensible concepts - typically by developing integrated systems capable of consensually formulating these concepts from various data sources and technologies that lead to innovations. Given the infinitely large number of data-driven knowledge generating activities across the continent and beyond, the paper proposes a cohesive data modelling approach towards harnessing the potentials of data flows. The ideas highlighted in the methodology section of this paper and the data analysis results could be embedded into a cohesive knowledge generating system as graphically illustrated in Figure 6.

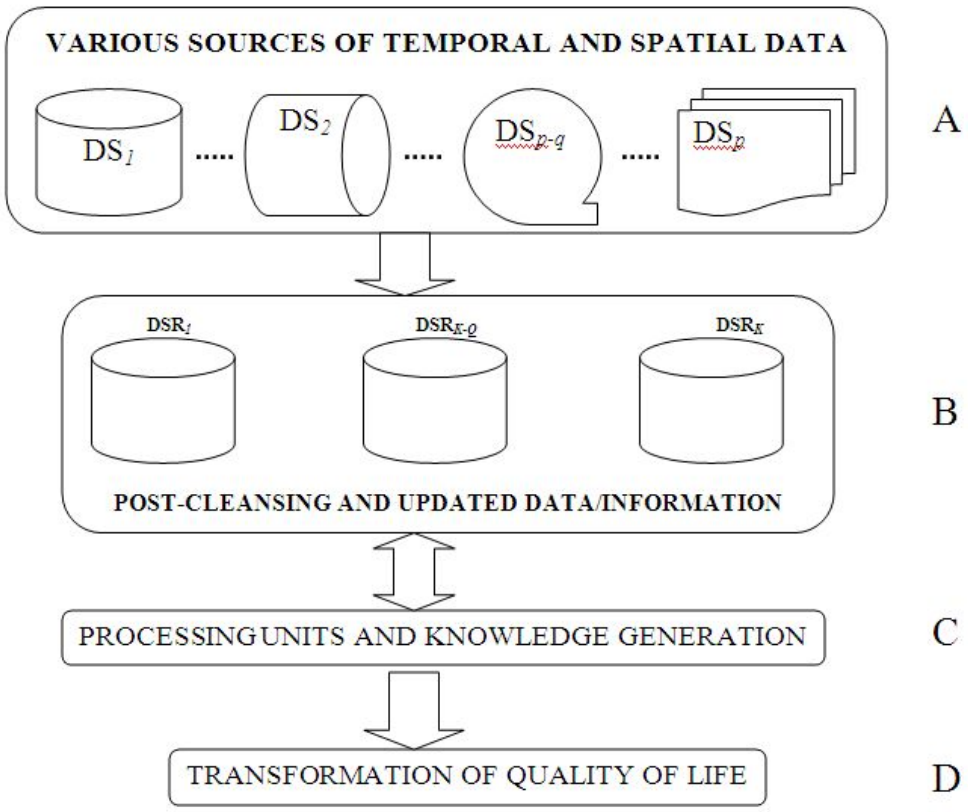

Figure 6. Recursively utilising and building data sources for knowledge generation

Each of the four elements is an indispensable component of the cohesive system as the change in one enhances a change in the others with the overall change having a more valuable contribution to the overall output. Thus, the final outcomes in D may enhance availability of data in A and lead to improvements in C as summarised in Table 1.

LEVEL	ELEMENTS	PREREQUISITES
A	Primary and secondary data sources processed to generate cleansed data and data repositories with cross-disciplinary data sharing potentials.	Resources - supportive national policies, appropriate knowledge and skills, data acquisition tools, research and development (R & D) initiatives.
B	Cleansed data, updatable data repositories, model and parameter-related information.	Tools for data capturing, storage, and dissemination.
C	Research centres, public, private, and academic institutions, R&D and knowledge transfer partnerships (KTP), individual researchers, consulting firms, NGOs, etc. Making possible the sharing of data and/or information across disciplines/regions.	Computing tools, methodologies, and techniques preferably within the cloud computing environment. Data mining tools, methodologies, and techniques.
D	Transforming knowledge into tangible outputs - patents, products, and services	Financial, human, and technical resources. Supportive policies, legislative, social, and technological infrastructure.

Table 2. Summary of the components in Figure 6

Natural and social dynamics potentially lead to changes in socio-techno attributes, such as government policies, consumer behaviour, gene mutation, carbon emission, and related technologies. The result is what is commonly referred to as concept drift (see Karnick, 2008) - whereby the key properties of the predictive model outputs change over time. Apparently, these dynamics impinge on the overall accuracy and reliability of the models, which is what the proposed recursive model seeks to minimise. The nature of the processing units at D potentially minimises the error in Eq. (4) while ensuring that the models in Eq. (1) through Eq. (3) do not over-fit the data, which is what this paper showed in Figure 5.

To develop the unified approach proposed in Figure 6, more work needs to be done on a wide range of models of much more comprehensive datasets than used in this study. The creation of a continuous flow of knowledge through academic, research, and private channels entails facilitating the development of knowledge transfer among these institutions via a range of activities such as workshops, study groups, visits, etc. Thus, supporting Knowledge Transfer Partnerships (KTP) and Research and Development (R&D) initiatives would greatly enhance the transformation of socio-economic prosperity. By combining the power of predictive tools and the fast growing cloud computing, we can enhance the mechanics of the multi-layer environment in Figure 6. Although the paper does not delve further into the mechanics of cloud computing (see Grossman, 2009 for details), it is important that we highlight its key ideas of providing on-demand computing instances and providing on-demand computing capacity. Apparently, the demand for resources will be high, but we can take comfort in the abundantly available open source tools as summarised in Table 3.

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

Table 3. A list of selected open source tools available to researchers

The proposed framework comes with a number of challenges, only a handful of which have been addressed in this paper. They include challenges relating to model complexity, inter-regional policies, infrastructural, skills, and resources. Its successful implementation will require not only standardization in data capturing mechanisms but also universal implementation of such standards. In most parts of the continent, for instance, data flow for many applications still is not available in real-time. Thus, there is still lack of timely data capture, inconsistent information updates, and lack of input-output quality assurance. The foregoing issues are attributable to the continent's poor ICT and general infrastructure. The solutions can be expected to come from co-ordinated direct investment in the sector from both public and private sectors. The suggested open access resources above can only marginally alleviate the problem as there are no direct and co-ordinated interventions in the ICT infrastructure at the continental, regional, and national levels. At national levels additional initiatives, such as retention of the key skills required to support the promoted model, will be necessary.

Recent information related developments on the continent highlight new paths towards information generation and utilisation. The unprecedented growth of mobile network coverage across the continent and the novel applications it has brought provide promising signs for harnessing information flow across the continent (Williams et al., 2011). The fact that these simple technologies have been emulated in financial applications in

banks across the developed world indicate that implementation of our framework has the potential to yield desirable outcomes. The starting point remains influencing policy makers. It is expected that our study will contribute to the existing knowledge base, which in turn will be employed in interpreting the primary and secondary data. On the other hand, data analyses will enable this study to provide a critical evaluation of the theory upon which the analyses have been made.

6 REFERENCES

Babcock, C. (2010) *Management Strategies for the Cloud Revolution: How Cloud Computing is Transforming Business and Why You Can't Afford to Be Left Behind*. McGraw-Hill, ISBN-13: 978-0071740753.

Bullinger, H-J, Auernhammer, K., & Gomeringer, A. (2004) Managing innovation networks in the knowledge-driven economy. *International Journal of Production Research, Vol. 42*, Issue 17, pp 3337 – 3353, ISSN: 1366-588X.

Buyya, R., Yeo, C., Venugopal, S., James, B., & Brandic, I. (2009) Cloud computing and emerging IT platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. *Future Generation Computer Systems Vol. 25*, Issue 6, pp 599-616, ISSN: 0167-739X.

Campbell, P. G. C., Hontela, A., Rasmussen, J. B., Giguère, A. Gravel, A. Kraemer, L., Kovesces, J., Lacroix, A., Levesque, H., & Sherwood, G. (2003) Differentiating Between Direct (Physiological) and Food-Chain Mediated (Bioenergetic) Effects on Fish in Metal-Impacted Lakes. *Human and Ecological Risk Assessment, Vol. 9*(4), pp. 847-866, Taylor & Francis, ISSN 1080-7039.

Duin, R. (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions; *IEEE Transactions on Computers, Vol. 25*, Issue 11, pp 1175-1179.

Gray, J., Liu, D., Nieto-Santisteban, M., Szalay, A., DeWitt, D. and Heber, G. (2005) Scientific Data Management in the Coming Decade. *Association for Computing Machinery (ACM), Vol. 34*, Issue 4, pp 34 - 41, ISSN:0163-5808.

Grossman, R. (2009) The case of cloud computing. *IT Professional, Vol. 11*, Issue 2, pp 23-27. IEEE Computer Society, ISSN:1520-9202.

Howells, J. (2005) Innovation and regional economic development: A matter of perspective? *Research Policy, Vol. 34*, Issue 8, pp 1220-1234, Regionalization of Innovation Policy.

Juma, C. (2005) Going for Growth: Science, Technology and Innovation in Africa; Report, The Smith Institute.
Karnick, M., Ahiskali, M., Muhlbaier, M., & 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.

Kirsop, B. & Chan, L. (2005) Transforming Access to Research Literature for Developing Countries. *Serials Review, Vol. 31*, Issue 4, December 2005, pp. 246-255.

Lundvall, B-A. & Borrás, S. (2005) Science, technology and innovation policy (Chapter 22 in Fagerberg, et al., (editors): *Innovation Handbook*. Oxford University Press, pp 599-631.

Mwitondi, K. S., Taylor, C. C., & 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.

Mwitondi, K. S. (2010) Science, Technology and Innovation for Development (STID): A proposed framework for implementing integrated knowledge transfer and research and development partnerships; Accepted to the *Science with African Conference; United Nations Economic Commission for Africa*, 23-25 June 2010, Addis Ababa.

Mwitondi, K. S. (2009) Tracking the Potential, Development, and Impact of Information and Communication Technologies in Sub-Saharan Africa; A book chapter in: *Science, Technology, and Innovation for Socio-*

economic Development: Success Stories from Africa. International Council for Science (ICSU-ROA), ISBN 978-0-620-45741-5.

Mwitondi, K. (2003) *Robust Methods in Data Mining.* PhD Thesis, School of Mathematics, University of Leeds, Leeds: University Press.

Mwitondi, K. S. & Ezepue, P. O. (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.

Rushing, J., Ramachandran, R., Nair, U., Graves, S., Welch, R., & Lin, H. (2005) ADaM: A data mining toolkit for scientists and engineers. *Computers & Geosciences, Vol. 31,* Issue 5, pp 607-618. ISSN 0098-3004.

Tan, M. T., Tian, G-L., & Wang, K. (2009) *Bayesian Missing Data Problems: EM, Data Augmentation and Non-iterative Computation*. Chapman & Hall. ISBN-13: 978-1420077490.

Wang, F-Y, Carley, K. M., Zeng, D., & Mao, W. (2007) Social Computing: From Social Informatics to Social Intelligence. *Intelligent Systems, IEEE, Vol. 22*, Issue 2, pp 79 - 83, ISSN: 1541-1672.

Warren, M. (2007) The digital vicious cycle: Links between social disadvantage and digital exclusion in rural areas; *Telecommunications Policy; Vol. 31*, Issues 6-7, pp 374-388.

Webb, A. (2005) *Statistical Pattern Recognition*. Wiley, ISBN 0-470-84514-7.

Williams, M., Mayer, R., & Minges, M. (2011) *Africa's ICT Infrastructure: Building on the Mobile Revolution (Directions in Development)*. World Bank Publications. ISBN: 9780821384541.

(Article history: Received 17 November 2010, Accepted 10 December 2012, Available online 19 December 2012)