KNOWLEDGE DISCOVERY FROM LAND RECORD SYSTEMS

Thaer SHUNNAR, Palestine and Michael BARRY, Canada

Key words: Land Records; Registration Statistics; Registration Problems; Data Mining; Knowledge Discovery; Outlier Mining; Fraud Detection

SUMMARY

Decision support, planning, and policy making are core processes in land management. Performing these tasks may require the examination and analysis of large numbers of land records. In current land management systems, the analysis of data is done by people, which may be slow, inaccurate and require substantial resources. This paper explores data mining as a technique for automating data processing and analysis of land records. Automated data mining may be useful in two ways. Firstly, it may assist in the discovery of important knowledge and patterns of behaviour from large digital datasets, which otherwise would be extremely difficult to do manually. Secondly, automation may reduce the level of skills required to run a system. Scenarios are explored for applying data mining techniques to detect manipulations of land records, where post conflict is the situational context. Of particular interest are fraudulent changes in the land records by the state or powerful factions to grab land from unsuspecting land holders on the ground. The results of preliminary tests on simulated data are reported. Specifically, outlier detection techniques are used to potentially identify patterns such as an unusual number of transactions in short periods of time, and periods with no or very few transactions.

KNOWLEDGE DISCOVERY FROM LAND RECORD SYSTEMS

Thaer SHUNNAR, Palestine, Michael BARRY, Canada

1. INTRODUCTION

Data mining is a potential tool for knowledge extraction and pattern identification in land record systems. We identify some of the patterns of behaviour that may occur in land records and then explore how data mining methods may identify these patterns. If successful, these methods may contribute to identifying errors in land registration processes, discovering fraudulent activities, provide new searching methods, and provide analysis tools that may assist planners and decision makers.

Data mining is being used increasingly as a method of exploratory analysis and pattern recognition in an era where data collection and accumulation is accelerating rapidly in numerous spheres. Unfortunately much of the data are being collected without careful planning as to how they may be used and little consideration of the potential information which may be gleaned from them. As Naisbitt (1986) note more than twenty years ago: "We are drowning in information, but starved for knowledge".

Data Mining has emerged as a way of finding valuable (and not valuable) patterns in data, and is now a well established methodology (Weiss *et al.* 2004). Much of this vast data collection probably contains untapped, useful knowledge. However, identifying patterns and extracting knowledge from sizable data sets using traditional search and analysis tools is often impractical (Bramer 2007).

Land management systems are too accumulating vast amounts of data every day. Integrated land information systems contain various forms of related data such as land taxation records, personal details, survey plans, deeds, titles, building plans, property management files, maps, aerial photographs and satellite images. Semi-automated analysis of large datasets may identify patterns and relationships which may provide useful land governance knowledge. In post conflict areas – as an example - existing land records require careful investigation when trying to fix or re-construct a land administration system. Part of this process might include identifying patterns in the land records which may uncover fraudulent activities (Zevenbergen and van der Molen 2004). Computer assisted analysis, in which data mining could be one of a range of tools, may contribute to improving governance.

In this paper, we use the term land records to refer to any kind of land tenure related data. In developed countries, the trend is towards automated land information systems. Legal documents, such as deeds, titles and survey plans, are now submitted in digital form in many jurisdictions (Hallett *et al*. 2003; Haanen *et al*. 2002; Nyerges 1989). In many parts of the developing world, much of the land data is still in paper format. However, these are being

converted to digital form in many jurisdictions, and it is reasonable to forecast that many of these systems will be fully electronic or at least computer assisted in the next twenty years.

The paper proceeds as follows. We provide a brief overview of data mining and text mining as exploratory data analysis and knowledge extraction methods. Following this we discuss some of the behaviours that may occur in land administration systems, especially in post conflict situations. Lastly we report on preliminary results from rudimentary initial experimental work on outlier detection to detect fraudulent activities in land records based on simulated data.

2. DATA MINING AND TEXT MINING

Data mining may be defined as the exploration and analysis, by automatic and semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules (Han and Kamber 2006). A key element is the linking together of extracted information to form new facts or hypotheses, i.e. patterns, to be explored further. It is an established data analysis methodology, which has been applied successfully in numerous applications such as hazard analysis, fraud detection (e.g. in credit card transactions), decision support in stock market trading, medical diagnostics, marketing, and weather forecasting. For example, Lai and Fu (2010) developed methods for identifying similar patterns of behaviour in stock markets at different times. DeBarr and Eyler-Walker (2006) use knowledge discovery in databases, machine learning and statistical analysis to assist the United States Internal Revenue Service (IRS) in identifying tax evaders who use offshore tax shelters. In the medical field, (Rao *et al.* 2005) developed REMIND which is a probabilistic framework for Reliable Extraction and Meaningful Inference from Non-structured Data. The idea of REMIND is to combine patients' data from multiple sources (billing data, text data, medical records data, …etc) and then remove the redundancy in the combined data which then can be used to make inferences about patients' medical health. This is a good example of how land information could be combined and used for knowledge extraction as much land related data are stored in un-structured documents. Many other case studies on the application of data mining have proven the effectiveness of data mining techniques in other fields (Giudici and Castelo 2003; Kanellopoulos *et al*. 2006; Lawrence *et al*. 2007).

Most data mining techniques assume that data is well structured and organised, such as in fields in a database. Land records may contain a mix of well structured data in databases, semi-structured text such as a title deed, and unstructured text. On the whole, much of the data are seldom rigorously structured in a manner well suited to many established data mining methods For example, data in a well structured relational are well suited to many data mining techniques.

In our opinion text mining - a special form of data mining - is most suited to exploring land tenure records; general data mining and text mining use many of the same methods. Text mining evolved due to the increasing number of text documents on the web, in document management systems and other similar systems that include numerous documents (Weiss *et al*. 2004; Feldman and Sanger 2007). Applications include document classification,

information retrieval, clustering and organizing documents, information extraction from documents, and prediction and evaluation (Hearst 1999; Weiss *et al.* 2004).

Despite the large number of applications of data mining, we could not find published evidence of the technique being applied to land tenure records. We explore the potential for applying data mining methods to land records systems.

3. PATTERNS IN LAND RECORDS

The analysis of land records for decision support is a complex task drawing on vast data sets that serve as the information infrastructure for land records systems. Size is determined by the scale of a particular system, the number of different data types and classes, and the temporal aspect of land records in which historical information is stored. Complexity is best illustrated in a land registration system which models numerous relationships between people (e.g. inheritance, spousal rights in land), between people and the land, and between land parcels (e.g. easements, servitudes) that may be modelled.

We now discuss some of the problems and patterns that may occur in land records.

3.1. Patterns Indicating Fraud in Post Conflict Situations

Zevenbergen and van der Molen (2004) mention some patterns that could be found in land records following a conflict which may indicate fraudulent transactions. These include:

- An unusual number of transactions of a certain type in short time, or even on the same day.
- Transfers between members of different groups in the conflict.
- Transfers from public, common or communal properties to private persons often in the form of privatization.
- Periods without, or with few, transfers, which might indicate that certain parts of the transaction records have been removed.
- A lack of transfers especially when the overview data showing the situation just prior to the conflict is missing, and/or a new group has come to power.

3.2. Fraud in Afghanistan

In Afghanistan, the Ministry of Urban Development in Afghanistan says that every day, powerful government officials and warlords in Afghanistan are grabbing between 1,000 and 1,500 jirib (roughly equal to an acre) of land on which families reside. Most of this land grab occurred during the early years of the U.S. occupation. Government officials abused their positions during the chaos and used their privileged access to official maps and property ownership records to make counterfeit deeds during the early years of the war. Ministry of Urban Development says that in total, more than 3.5 million jirib of the lands in Afghanistan have been stolen (Commodity Market News 2009).

Thus the registration systems cannot be relied on due to the large number of counterfeit deeds. In the event that the situation stabilises and attempts are made to return lost lands to their

original owners, a number of methods may be required to clean up the titles and resolve what may turn out to be multi-layered disputes.

3.3. Errors in Land Registration and Fraud in Housing Programmes in South Africa

The Department Rural Development and Land Reform in South Africa reports that the Deeds Registration database contains 6.6 million land parcels and more than 116 million personal, property and land related records (Department of Rural Development and Land Reform 2009). In the state subsidised housing programme, Minister Tokyo Sexwale (2009) reported that the government had facilitated the delivery of a total of 2.3 million homes government since 1994 (Sexwale 2009). Fraud has occurred in the allocation of some of these houses (Mail and Guardian online 2009; The Mercury 2008)

Errors may occur in the registration, albeit South Africa applies a strict examination process. For example, the deeds registration report from Department of Rural Development and Land Reform (2009) reports that during the examination process of 1,447,180 deeds in the nine deeds offices in South Africa, an average of 26% of the examined deeds were rejected as they were found to be unregister able due to conveyancing errors, missing attachments, interdicts or legal constraints. Text mining and perhaps other data mining methods may uncover patterns which would indicate fraud or undiscovered errors in the different sets of records. There may also be interesting patterns of information about the land market (e.g. trading trends), groups or individuals dominating the market, and other knowledge.

3.4. USA Deeds Registration and Private Conveyancing

In some U.S. states, only evidence of title is recorded in the public land record office. Each time there is a change in the ownership, a Title Search is carried out by private title insurance companies. A chain of title of all previous owners is constructed which tracks the history of owners and other forms of transaction for up to 60 years. All the documents relevant to this title and its chain are uncovered and examined for fraud or errors (e.g. omissions) in the chain. (Onsrud 1989).

There are many similar forms of deeds systems around the world, where the chain of title is often not examined with the same rigour, or not examined at all. Thus there are possibilities for fraud and other errors. First, forged documents may be recorded in the land record offices as valid documents. An owner might fraudulently sell the same property to several different people which results in many documents, but only one valid transaction. From a technical perspective, two problems can be addressed regarding the process. First, the examination of the chain of title can be a long and complex process that requires experts to do the work. Also, this process is done every time there is a change in the land ownership. Assuming that a digital form of all the documents is accessible, an automated knowledge discovery system might be able to trace all the recent titles, create title chains, and identify forged documents or errors in the system.

4. FRAUD DETECTION IN LAND RECORDS

In this section, we provide an example of data mining to discover fraud. In our design and testing process, we have simulated land record data using a random number generation process, then biased parts of the data set with a pattern, and used data mining concepts to identify these patterns.

As mentioned in section 3.1, there are patterns in the land records that may indicate the presence of fraud. Two of these patterns are an unusual number of transactions in short periods and periods with no or few transfers. In data mining, unexpected behaviour patterns may be identified using Outlier Mining techniques.

4.1. Outlier Detection

We used an optimization model for outlier detection developed by He *et al* (2005). Many data mining techniques aim to find general patterns inside the data, while outlier detection aims to find small groups of data that indicate unusual or exceptional behaviour (He *et al.* 2005). In this way, we can identify abnormalities in a system. In other words, outlier detection is used to differentiate between incoherent observations and general patterns, or to highlight exceptional cases in the data (Fan *et al.* 2008). Many outlier detection algorithms have been proposed. In general, these algorithms follow two approaches; the supervised approach and the unsupervised approach (Ye *et al.* 2009). In our experiments, we are using the unsupervised approach as we do not rely on previous training samples of data whereas the supervised approach needs to learn from a sample set of data which has labels identifying each record as outlier or not to be able to detect outliers in new samples of data.

An outlier inside a data set is a data object that does not comply with the general behaviour of the data (unusual pattern). Sometimes, it can be considered as noise or an exception or indicates an error. However, it could be quite useful in fraud detection and rare event analysis. Outlier mining has been used for fraud detection in credit card transactions and credit approval, weather prediction, data cleaning, and the identification of suspicious activities in e-commerce.

4.2. Problem Description

Our problem can be summarized as the need to discover groups of records in the data set where each record in the group might be suspicious and perhaps indicates fraud. In our first trial, we simulated the two types of deviations described by Zevenbergen and van der Molen (2004); 1) an unusual number of transactions of certain type in short time, or even on the same day 2) periods without, or with few, transfers, which might indicate that parts of the transaction records have been removed or deleted.

4.3. Land Records Simulator (LRS)

As a first stage, we developed a Land Records Simulator system (LRS). Access to real data is very difficult if not impossible. The second reason is as a first stage we are able to control the records that we generate. As mentioned above we simulate synthetic land records and introduce certain patterns into them. We can then calibrate and test the data mining algorithms and methods to see if they extract the patterns which we know exist and perhaps patterns in the supposedly random data; i.e. patterns that that we did not manufacture.

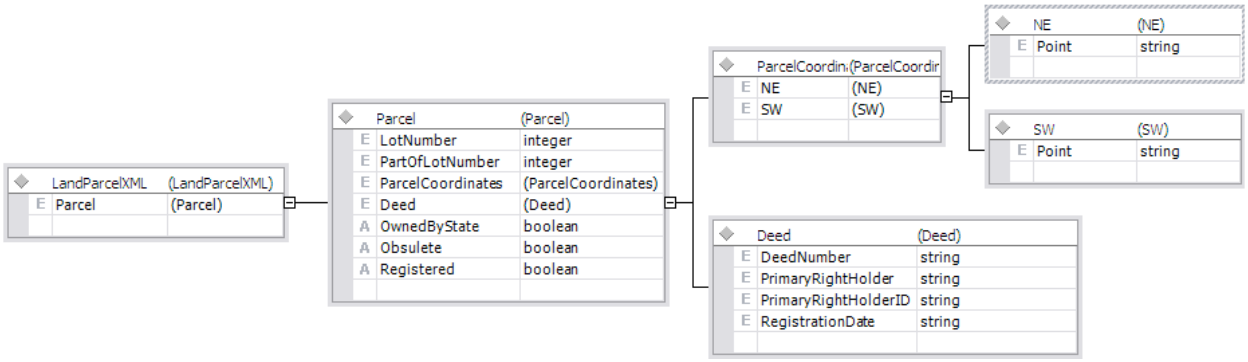

Figure 1: XML schema diagram for the output of LRS system.

The LRS is a Windows based desktop application. It is being developed using the C# programming language, and Microsoft Visual Studio 2005 as the Integrated Development Environment. The user interface consists of two parts: The controls that are used to set the different parameters for the simulation and a mapping area that can be used to visualize a map of the land parcels for which the records are generated. For the mapping, the SharpMap library is used. The primary reason of choosing SharpMap was it is an open source mapping library written in C# and based on Microsoft.NET 2.0. The Library also provides access to many types of GIS data and enables spatial querying of that data (SharpMap 2009).

In the basic operation of the LRS system, the user creates an initial long, wide parcel. This initial parcel is then subdivided into a number of child parcels that will be defined as owned by the state. The parcels are then transferred to individuals. The user enters a start date and an end date for the simulation process which will define the period during which the land transactions take place, the population who will occupy the land, and the required patterns to be introduced to the generated transactions. At the time of writing, only the two patterns mentioned in section 4.2 can be superimposed on randomly generated data. We will be extending this once we have completed a survey of different unusual patterns of behaviour drawing on the knowledge and experience of a number of experts from around the world.

The simulation randomly generates land transactions between community members. Two types of transactions have been simulated: conveyancing and subdivision. During each transaction, information regarding the transaction is generated and saved in an object oriented class structure. At the end of the simulation process, all transaction data is written into an XML structure as portrayed in figure 1 which is generated in a single XML file.

4.4. Methods

4.4.1. Data set simulation

For the first experiment two data sets were generated, Land Records Dataset 1 (LRDS1) and Land Records Dataset 2 (LRDS2). The attributes are used to generate the datasets shown in table 1. Table 2 shows periods of fraudulent behaviour that were introduced into both datasets. The simulator generates a very high or very low number of transactions and superimposes these on the randomly generated transaction pattern.

Dataset	Initial subdivision	Simulation start date	Simulation end date	Population	Records
LRDS1	2500 parcels	Jan 1, 2010	Dec 31, 2012	50,000	59398
LRDS2	2500 parcels	Jan 1, 2010	Dec 31, 2012	20,000	52727

Table 1: Attributes used to generate the datasets (LRDS1 and LRDS2)

Dataset	Periods of fraudulent behaviour		
	Start date	End date	Number of injected outliers
LRDS1	March 3, 2010	March 22, 2010	32
	April 3, 2010	April 9, 2010	
	May 11, 2011	May 27, 2011	
LRDS2	April 14, 2010	April 14, 2010	9
	September 15, 2010	September 15, 2010	
	July 15, 2011	July 15, 2011	
	November 23, 2011	November 28, 2011	
	September 19, 2012	September 20, 2012	

Table 2: Periods of fraudulent behaviour injected into the datasets (LRDS1 and LRDS2)

4.4.2. Problem formulation

A dataset that contains many outliers will have a high degree of disorder. So, our target is to identify a subset of the data set such that removing this subset from the data set will result in less degree of disorder in it. To solve this problem, we adopt the concept of entropy as discussed by (He *et al.* 2005). Entropy in information theory represents the amount of uncertainty attached to a random variable (Shannon 1948) and so entropy can be used as a measure of information disorder. Hence, we use entropy as the objective function in our algorithm to measure the amount of disorder in the data sets. The entropy H(X) is defined as shown in equation (1) where X is a random variable, $S(X)$ is the set of values that X can take and $p(x)$ is the probability function of X.

$$H(X) = -\sum_{x \in S(X)} p(x) \log p(x) \qquad (1)$$

Based on the above, our problem can be formulated as follows. Assuming we have a data set *DS* of *n* records, and given an integer *k* representing an expected number of outliers. Our solution would be a subset of outliers *O* contained in *DS* such that the entropy *H(DS – O)* is minimized. It is important here to mention that *k* has to be input to the algorithm to represent the desired number of outliers that the algorithm should look for. Thus we should have some idea of the number of outliers or we can run a series of experiments from which we can extract the value of *k*.

4.4.3. Algorithm implementation

There are two inputs into the algorithm; *k* and a data set of land records. Once we have identified the outliers, we need to identify the dates when the unexpected records were created. A new data set is created which contains two attributes; the registration date, and the number of transactions that took place on that date. The new data set is *DS* which contains the information for which the entropy will be calculated.

```
Input:      D     //land records dataset
            k     //number of desired outliers

Output:     O     //a subset of D with k outliers

Begin:
/* initialization of O */
        For counter equal 0 to k
                record = select a random record from D
                O[counter] = record
                remove record from D

/* evaluating the initial entropy*/
        MinEntropy = Entropy(D)

/*iteration: minimizing the entropy for D*/
        foreach record x in O
                foreach record y in D
                        if Entropy(D-y+x) less than MinEntropy
                                MinEntropy = Entropy(D-y+x)
                                Swap x and y
End
```

Figure 2: Outliers detection algorithm (in pseudocode)

In the algorithm, (see figure 2), initially, a new data set is created from *DS*. This data set is the outliers' data set *O* and contains *k* randomly selected records from *DS*. Any record that is randomly selected from *DS* will be removed from it and stored in *O*. for processing the data

sets; we use Arrays of a simple structure that contains two variables; RegistrationDate and NumberOfTransactions.

The initial entropy of *DS* is calculated and then an iterative process starts. First, a record is selected from *O*. This record then is swapped with a record from *DS* and the entropy of *DS* is measured again. If the new entropy is less than the previous measured one, the swap is considered final and the algorithm proceeds to the next record of *DS*. When all the records of *DS* are processed, a new iteration is started by selecting the next record of *O* and this continues until the algorithm reaches the end of *O*.

4.4.4. Experimental Results

The algorithm was applied to the two simulated datasets; LRDS1 and LRDS2. For each dataset, we did four runs of the algorithm using four different values for *k*. For the tests on the LRDS1 we use 20, 40, 60, and 80 as values for *k*. LRDS2 contains fewer days of abnormal numbers of transactions in one day and so we used 10, 20, 40, and 60 as values for *k*.

Figures 3, 4, 5 and 6 show the results of the tests. In general, we were able to identify the outliers in the data sets in most of the tests. However, the success rate of the algorithm in finding the outliers injected into the data sets varies based on the value of *k* that the user selects.

In LRDS1, we inserted 32 outliers. In figure 3, we can see that when $k = 20$, the algorithm fails to find any of the outliers between March 3 and April 9 2010, but it does identify those in 2011. If we set $k = 20$, the algorithm finds 13 outliers of the 32 injected outliers. But, using $k = 40$, the algorithm is able to find 30 or the total of 32 expected outliers. This is a good percentage but we are investigating why 100% success rate was not achieved. Finally, at $k = 60$ and $k = 80$ (figure 4), the results show 100% detection rate. A 100% detection rate also achieved for LRDS2 at k = 40 and k = 60 as we can see in figure 6.

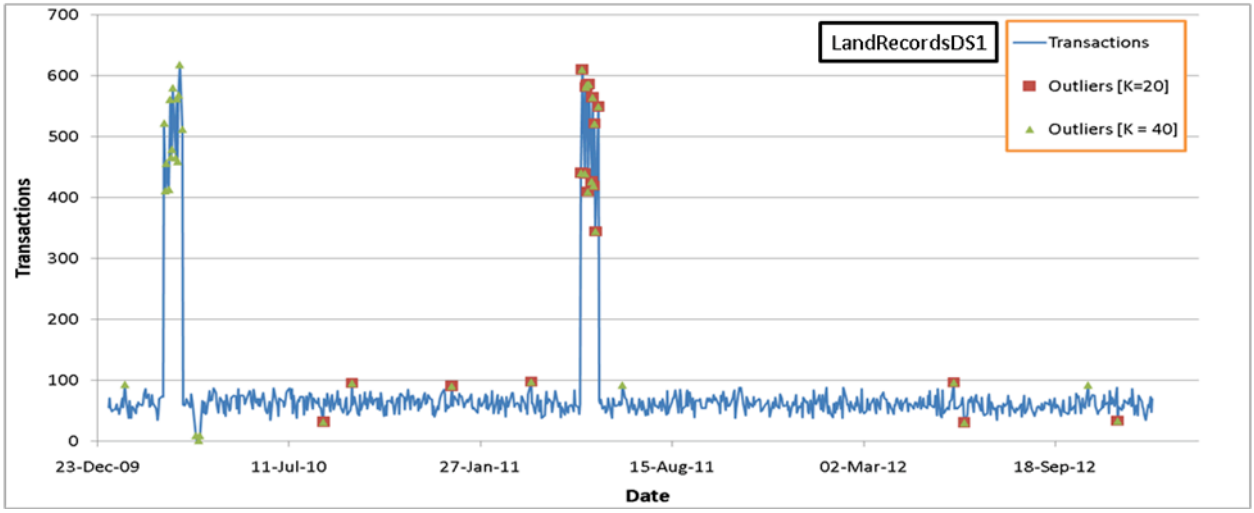

Figure 3: Detected outliers from LRDS1 data set for k=20 and k=40

The algorithm also identifies some points as outliers that were not inserted as outliers in the simulation. However, these were generated by the random transaction generator. This does reflect reality as there will be times when unusual numbers of transactions which are all above board. The datasets may have other outliers. For example; in figure 4 and figure 5, only the long spikes are the outliers that were inserted into the simulation, the other outliers were generated by the random transaction simulator.

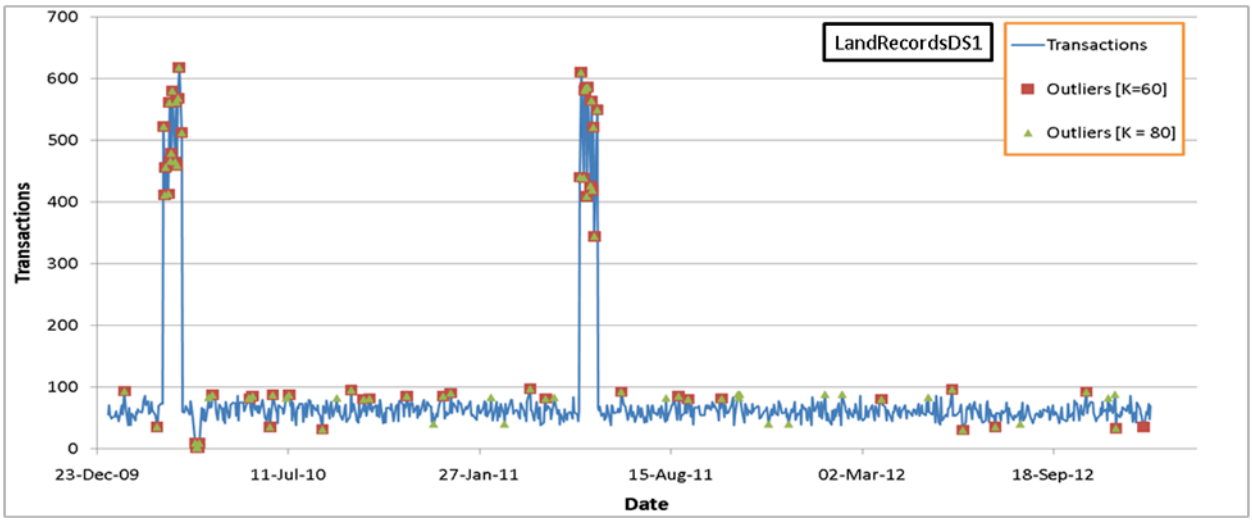

Figure 4: Detected outliers from LRDS1 data set for k=60 and k=80

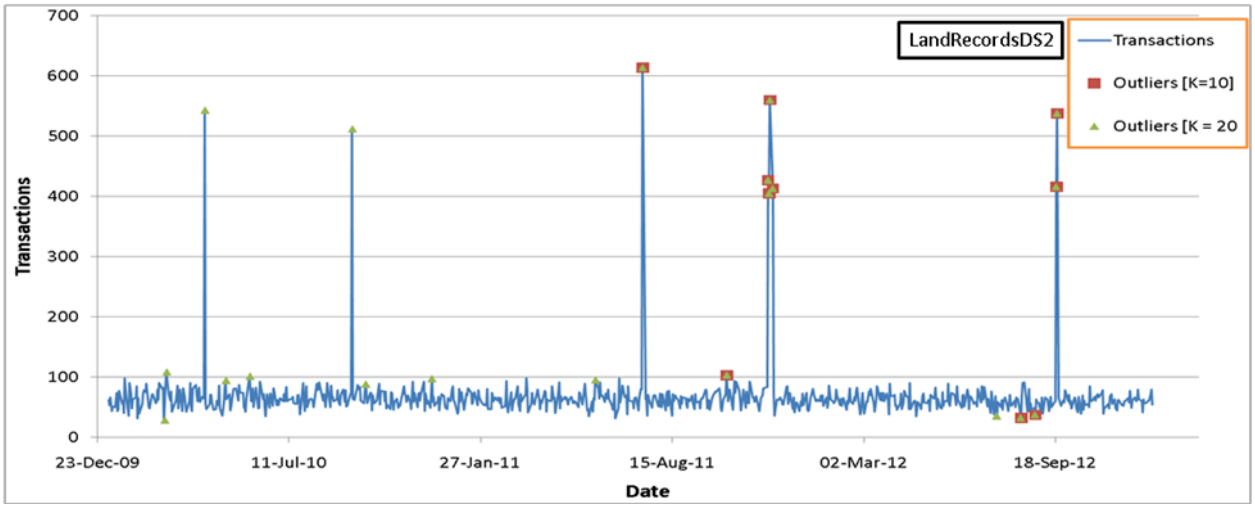

Figure 5: Detected outliers from LRDS2 data set for k=10 and k=20

An important point here is that the algorithm will not guarantee the detection of the top *k* outliers as we can see in figure 3 for *k* equal 10. The implementation of the algorithm should estimate a local minimum value of the entropy and not the global minimum. The local minimum depends on the initial selection of the random *k* records from the dataset for the initial outliers set. This is an area that requires further work.

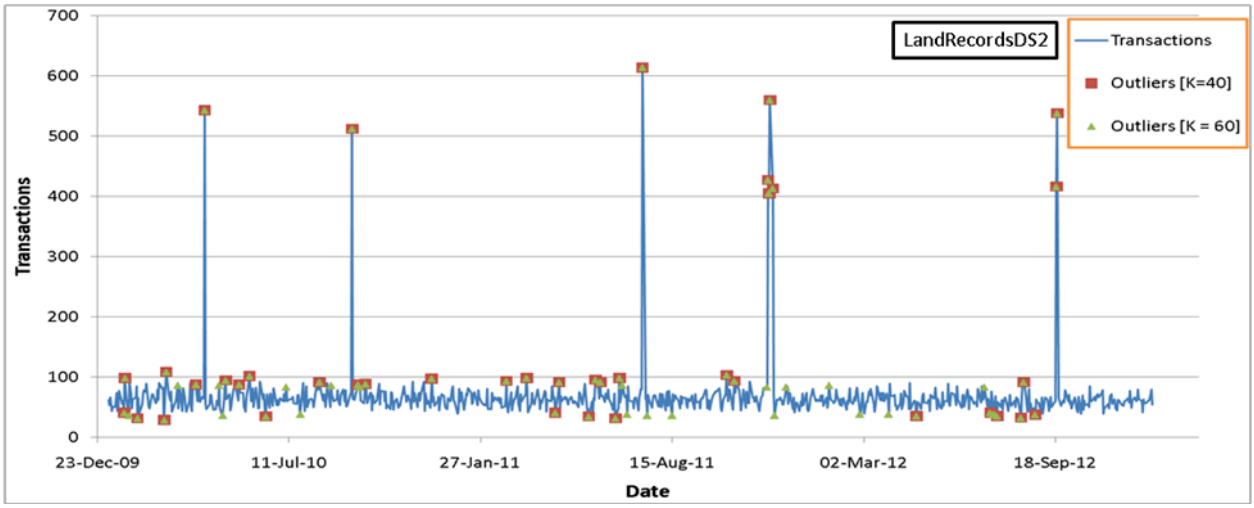

Figure 6: Detected outliers from LRDS2 data set for k=40 and k=60

5. CONCLUDING REMARKS

Land records may incorporate patterns and relationships that, if identified, may contribute to the governance of the land. A good example is patterns which indicate fraudulent activity.

We applied outlier detection using the concept of entropy to discover some of the fraud patterns that might take place in land records, using post conflict as the situational context. Our initial experimental results support the hypothesis that this technique of fraud detection may useful as one of a variety of fraud detection tools.

The entropy algorithm we tested in this paper was developed for simple one dimensional datasets. As we can see in the results, the algorithm evaluates the entropy for the transactions distribution over time and no other attributes were used.

A major limitation in using this algorithm is the need for a pre-defined number of outliers (*k)* to be input to the algorithm. In the experiments we have done, we used values for *k* that are close to the number of outliers since we have previous knowledge of the number of outliers in the datasets. However, when using real datasets where no previous knowledge about the number of outliers in them, this approach is heuristic and the results are not guaranteed. However, as the results show, using a value of *k* that is greater than twice the number of outliers, 100% of the outliers in the data sets are detected.

In our initial tests, we used this approach and applied it to one dimensional data as support for the concept. Our goal is to build the model up and extend it to be applied on multi-dimensional datasets, perhaps in which both continuous and categorical attributes are present.

Applying this technique in a real world situation may help in identifying the probable days in which illegal land transactions have taken place. As a result less effort is needed to examine the suspicious transactions because much fewer documents need to be examined and analyzed manually.

6. FUTURE WORK

We have tested one algorithm in a very simple uni-variant model to date. Future work will look at unusual patterns of behaviour in transactions involving partial rights in land. Early indications from our survey of experts suggest that fraud often happens in the mortgage market, where the true owner of the land may not even be aware that a mortgage has been taken out on his or her property. There are also cases of identity theft, where company A is registered in the company register with a name very similar to a land owning company, company B. Company A then engages in a land transaction in company B's land. Errors of omission may also occur where a partial right in land may be left of a title. For example, the institution (e.g. government registrar or a private conveyance) that creates a new title may omit mineral rights from a new title or deed.

Thus we have done initial tests involving single variable data. The next stage is to examine data which are related to each other and which involve ownership rights and partial rights in the land, such as mortgages.

ACKNOWLEDGMENTS

This study has been conducted with the support of the John Holmlund Chair in Land Tenure and Cadastral Systems and the Canadian National Science and Engineering Research Council Talking Titler project

REFERENCES

Bramer, M, 2007, *Principles of Data Mining, Undergraduate Topics in Computer Science*, Springer, London.

DeBarr, D & Eyler-Walker, Z, 2006, Closing the Gap: Automated Screening of Tax Returns to Identify Egregious Tax Shelters. *SIGKDD Explorations*, vol. 8, no. 2, pp.11 – 16.

Department of Rural Development and Land Reform, Deeds Registration, Republic of South Africa, viewed 26 August 2009, <http://www.dla.gov.za/land_planning_info/deeds_registration1_copy(1).htm>.

Fan, H, Kim, H, AbouRizka, S & Han, S, 2008, Decision support in construction equipment management using a nonparametric outlier mining algorithm, *Expert Systems with Applications*, vol. 34, no. 3, pp. 1974–1982.

Feldman, R & Sanger, J, 2007, *The Text Mining Handbook: advanced approaches to analyzing unstructured data*, Cambridge University Press.

Future and Commodity Market News, 2009, Afghanistan: Counterfeit deeds enable sales of public lands, viewed 29 August 2009, http://news.tradingcharts.com/futures/8/5/128108558.html

Giudici, P & Castelo, R, 2003, Improving Markov Chain Monte Carlo Model Search for Data Mining. *Machine Learning*, vol. 50, no. 1-2.

Haanen, A, Bevin, T & Sutherland, N, 2002, e-Cadastre - Automation of the New Zealand Survey System. *Presented at the Joint AURISA and Institution of Surveyors Conference, Adelaide*, South Africa.

Hallett, S H, Ozden, D M, Keay, C A, Koral, A, Keskin, S & Bradley, R I, 2003. A land information system for Turkey - a key to the country's sustainable development. Journal of Arid Environments, vol. 54, no. 3, pp. 513-525

Han, J & Kamber, M, 2006, *Data mining: concepts and techniques*, Morgan Kaufmann, San Francisco, USA.

He, Z, Xu, X & Deng, S, 2005, An Optimization Model for Outlier Detection in Categorical Data, Department of Computer Science and Engineering Harbin Institute of Technology.

Hearst, A M, 1999, Untangling Text Data Mining, *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics,* June 20-26.

Kanellopoulos, Y, Dimopulos, T, Tjortjis, C & Makris, C, 2006, Mining source code elements for comprehending object-oriented systems and evaluating their maintainability, *ACM SIGKDD Explorations*, vol. 8 , no. 1, pp. 33 – 40.

Lai, P & Fu, H, 2010, A polygon description based similarity measurement of stock market behavior. *Expert Systems with Applications*, vol. 37, no. 2, pp. 1113-1123.

Lawrence, R, Perlich , C, Rosset, S, Khabibrakhmanov, I, Mahatma, S & Weiss, S, 2007, Analytics-driven solutions for customer targeting and sales-force allocation, *IBM Systems Journal*, vol. 46 , no. 4 , pp. 797-816.

Mail and Guardian online, 2009, Gauteng ex-councillor arrested for fraud, viewed 15 January 2010, < http://www.mg.co.za/article/2009-07-27-gauteng-excouncillor-arrested-for-fraud>
Naisbitt, J, 1986. *Megatrends : ten new directions transforming our lives*, Warner Books, New York.

Nyerges, T L, 1989, Information Integration for Multipurpose Land Information Systems, *URISA Journal*, vol. 1, no. 1, pp. 27-38.
Onsrud, H J, 1989, The Land Tenure System of the United States. Forum: Zeitschrift des Bundes der Offentlich Bestellten Vermessungsingenieure, Jan 1989, Viewed 24 August 2009, http://www.spatial.maine.edu/~onsrud/pubs/landtenureabstract07.htm

Rao, R B, Krishnan, S & Niculescu, R S, 2006, Data Mining for Improved Cardiac Care, *SIGKDD Explorations*, vol. 8, no. 1, pp. 3-8.

Sexwale, T, 2009, Housing budget vote speech, Department of Human Settlement, Republic of South Africa, viewed 25 August 2009, http://www.housing.gov.za/Content/Media%20Desk/Speeches/2009/29%20June%2009.htm

Shannon, C E, 1948, A mathematical theory of communication, *Bell System Technical Journal*, vol. 27, pp. 379-423.

SharpMap, 2009, SharpMap - Geospatial application framework for the CLR, viewed 10 January 2009, < http://www.codeplex.com/SharpMap>

The Mercury, 2008, KZN public servants linked to RDP scam, Advertizer 07 April, p. 3.
Weiss, S M, Indurkhya, N, Zhang, T & Damerau, F, 2004, *Text mining: predictive methods for analyzing unstructured information*, Springer, New York.

Ye, M, Li, X & Orlowska, E M, 2009, Projected outlier detection in high-dimensional mixed-attributes data set, *Expert Systems with Applications*, vol. 36, no. 3, pp. 7104-7113.

Zevenbergen, J. & van der Molen, P., 2004. Legal aspects of land administration in post conflict areas. *Symposium on Land Administration in Post Conflict Areas,* Geneva, April 29 – 30.

BIOGRAPHICAL NOTES

Thaer Shunnar is an MSc student in the Department of Geomatics Engineering at the University of Calgary. He has a B.Sc. in Computer Engineering from An-Najah National University (Palestine). Subsequently he worked as a software developer and software engineer. His research interests are in data mining and knowledge discovery from land registration systems.

Michael Barry holds the John Holmlund Chair in Land Tenure and Cadastral Systems in the Geomatics Engineering Department at the University of Calgary, where he has been working since 2002. Prior to that he was at the University of Cape Town, South Africa. His research interests are analysing and developing systems to support land tenure security and land administration in general.

CONTACTS

Mr. Thaer Shunnar
Department of Geomatics Engineering
University of Calgary
2500 University Drive N.W., Calgary, Alberta, T2N 1N4
Tel: +1 (403) 890 9133
Email: taishunn(at)ucalgary.ca

Dr. M. Barry
Department of Geomatics Engineering
University of Calgary
2500 University Drive N.W., Calgary, Alberta, T2N 1N4
CANADA
Tel. +1 403 220 5826
Fax. +1 403 284 1980
mbarry(at)ucalgary.ca
Web site: www.ucalgary.ca/mikebarry