Knowledge Discovery from Land Record Systems

T. Shunnar & M. Barry

Technology in Land Administration :
FIG Congress 2010
Facing the Challenges – Building the Capacity
Sydney, Australia, 11-16 April 2010

Outline

- Introduction
- Problem
- Methodology
- Patterns in land records
- Fraud detection in post conflict situations
  - Simulation
  - Problem description
  - Problem formulation
  - Result
- Conclusions

Introduction

- Land Management Systems are accumulating vast amounts of data
  - land taxation records, personal details, survey plans, deeds, titles, building plans, property management files, maps, aerial photographs and satellite images
  - ICT facilitating major efficiencies in society in general (e.g. banking) but also providing opportunities for fraud e.g. identity theft, mortgage fraud
- Data Mining –potential to support land management systems?
  - Identifying errors.
  - Discovering fraud and other unusual behavior
  - Better search and analysis tools.

Problem

- Extracting knowledge from sizable land record systems using traditional search and analysis tools is often impractical.
  - Resources.
  - Error prone.
  - Hidden information

Methodology

Identify	Study	Mine
• What are the patterns? • **Fraud**, Errors, Data analysis and decision support	• Reasons • Representation in the data infrastructure	• Data • Technique • Results

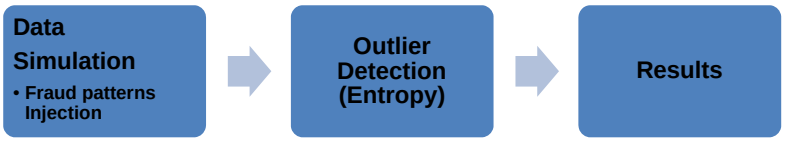
Patterns in land records

1. Fraud in Post Conflict Situations
   - Exceptional number of transactions
   - Transfers between members of different groups in the conflict
   - Privatization
2. Title & Mortgage Fraud
   - Appraisal fraud
   - Fraud by forgery
   - Oklahoma flip
   - Impersonation frauds
   - Air Loans
   - Title fraud

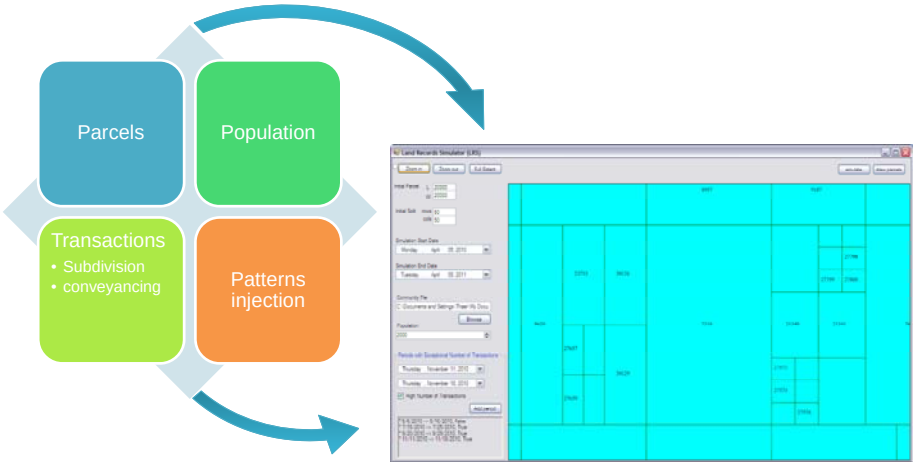
We focus on experimental work related to exceptional transactions in this presentation

Fraud detection in post conflict situations

- Detection of exceptional number of transactions that may take place during conflicts.



Land Records Simulator (LRS)

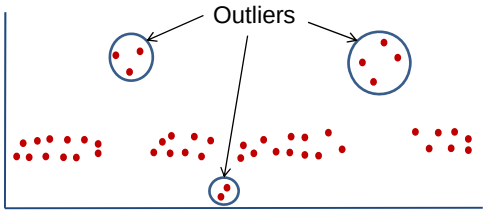


Problem description

- Detecting outliers in land transactions dataset
  - Simulate two types of outliers described by Zevenbergen and van der Molen (2004)
    - Unusually high number of transaction
    - Periods with no or few transactions
  - Entropy based outlier detection
    - Measures the level of organization in the data
    - Objective is to minimize entropy of the dataset

Problem Formulation

- Goal: find a subset of records such that removing this subset will minimize the entropy of the remaining set.
  - The size of the subset is predefined *k*
  - The subset is to be called the outliers which may reflect fraudulent activities

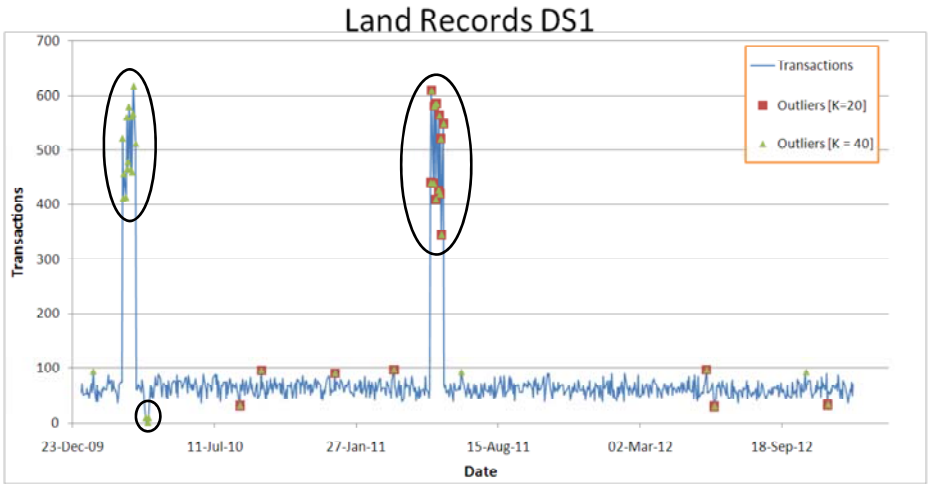


Data with high level of disorder (High Entropy)

Results 1/4

Department of Geomatics Engineering
Schulich School of Engineering
University of Calgary

Position Yourself Ahead of the Crowd

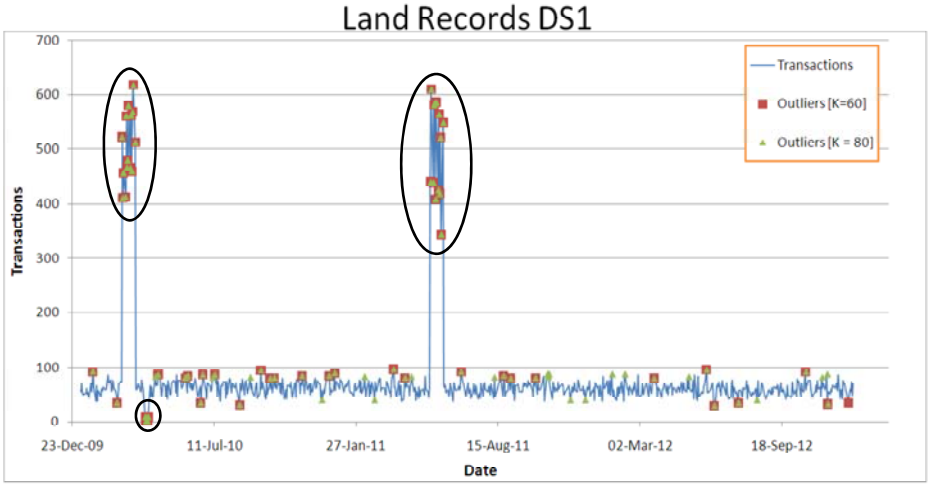


Detected Outliers from LRDS1 for k=20 and k=40

Results 2/4

Department of Geomatics Engineering
Schulich School of Engineering
University of Calgary

Position Yourself Ahead of the Crowd

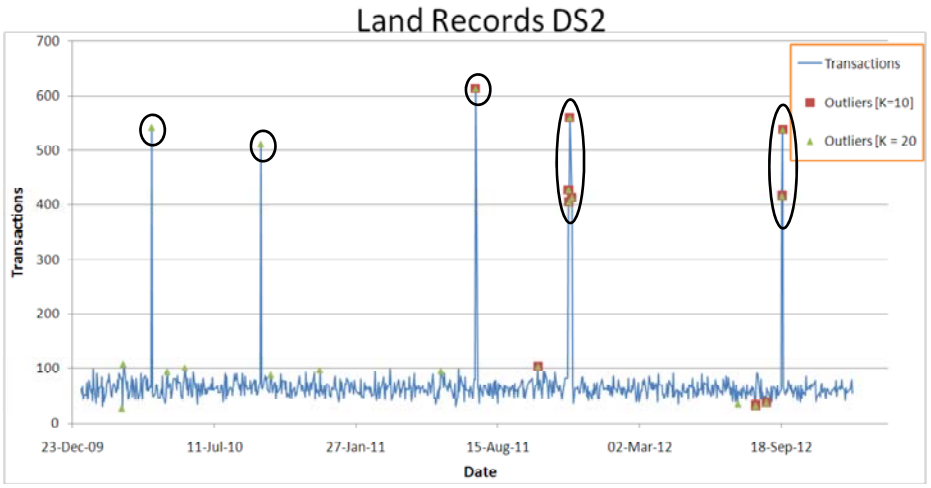


Detected Outliers from LRDS1 for k=60 and k=80

Results 3/4

Department of Geomatics Engineering
Schulich School of Engineering
University of Calgary

Position Yourself Ahead of the Crowd

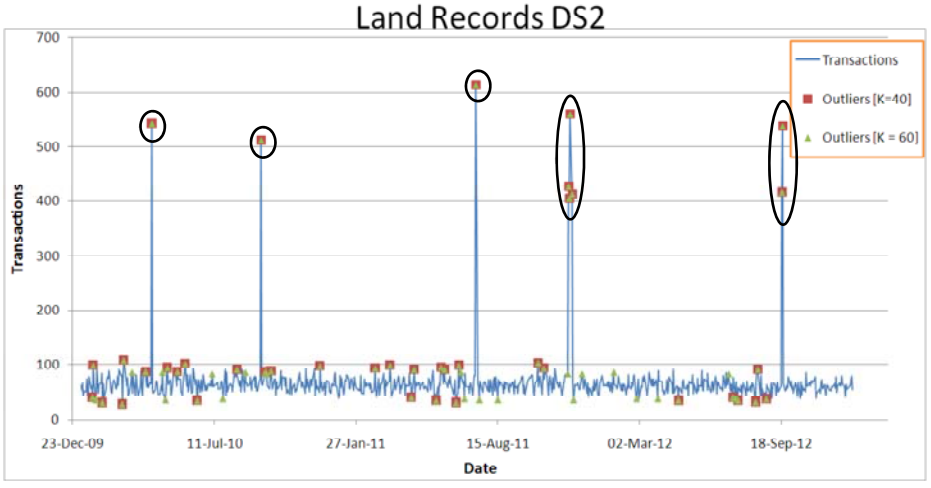


Detected Outliers from LRDS2 for k=10 and k=20

Results 4/4

Department of Geomatics Engineering
Schulich School of Engineering
University of Calgary

Position Yourself Ahead of the Crowd



Detected Outliers from LRDS2 for k=40 and k=60

Conclusions

- We identified a wide range of patterns in land and real estate transactions.
- Outlier detection could help in detecting fraudulent activities in land transactions that may take place during a conflict.
- A limitation of the entropy algorithm is the need for the pre-defined number of the wanted outliers *k*
- The entropy algorithm was developed for simple one dimensional datasets.
- The goal is to build the model up and extend it to be applied on multi-dimensional datasets

Department of Geomatics Engineering
Schulich School of Engineering
University of Calgary
Position Yourself Ahead of the Crowd

THANK YOU

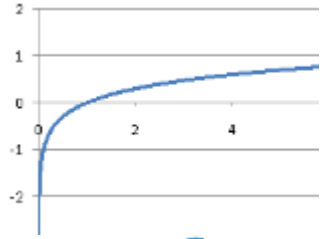
Algorithm


```
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
```


Entropy

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

- *X:* a random variable.
- *S(x):* the set of variables X can take
- *p(x):* probability function of X
- *H(X):* Entropy



Simulated data

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)