

# **Geographical and Data Mining Analysis of 1978 – 2006 Price Changes in the Residential Real Estate Market of Halifax Regional Municipality (HRM)**

**Jie Ou**  
**Centre of Geographic Sciences (COGS)**  
**Advanced Diploma Program of Geographic Information Systems**  
**May 2007**

## **Abstract**

This paper presents the analyses of sale price changes of the residential housing market in Halifax Regional Municipality (HRM), Nova Scotia, between the years 1978 and 2006. The data is divided into six sub-sets of five-year sale records with the exception of the first set. Data mining techniques are applied to individual sub-set, to discover determinants of high sale prices. Geostatistics and spatial statistics methods are used to find out changes of spatial distribution of prices.

## **Introduction**

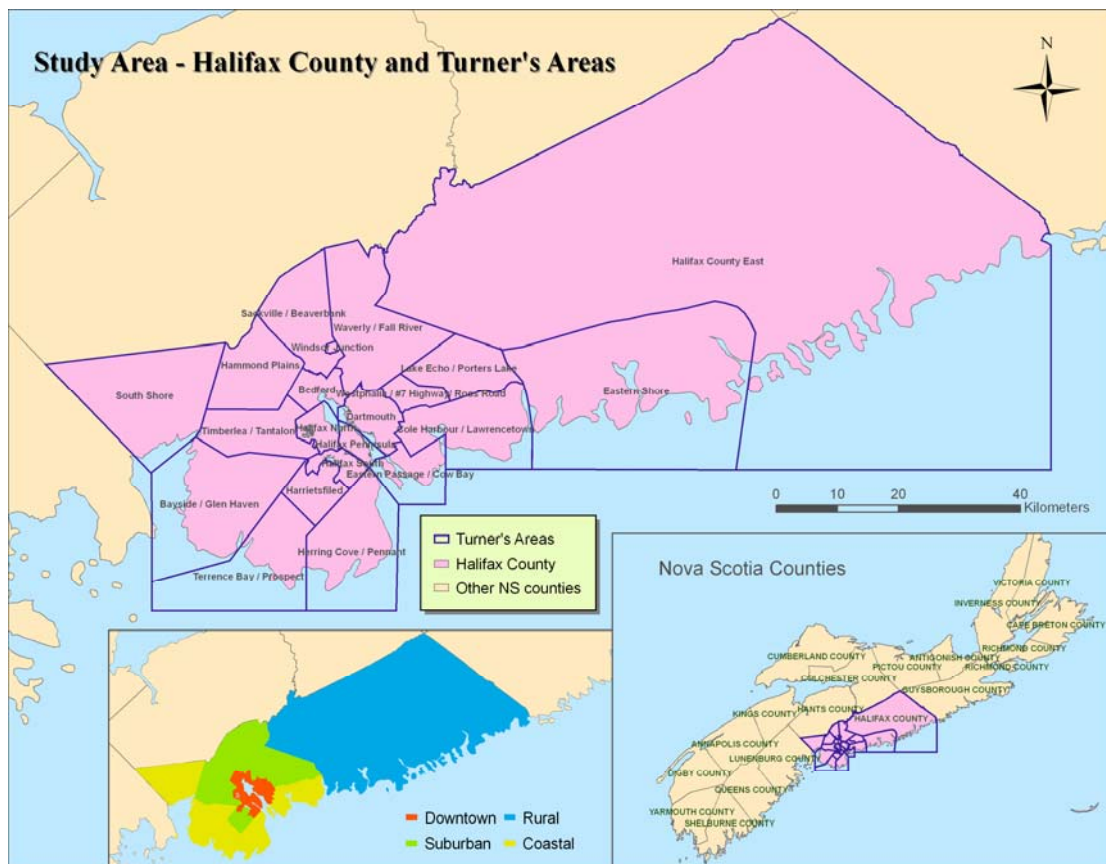
A report from Canada Mortgage and Housing Corporation (CMHC) indicates that housing prices have continuously increased since the late 1980s. Over the period from 1997 to 2002, Halifax was a hot spot in Canada that recorded the highest average annual growth rates of the real MLS (Multiple Listing Service) price. To find out the details of the real estate boom, it becomes important to examine how much and how fast the residential housing prices change during the past several decades. Price analyses help to find out housing attributes and demographic and geographic determinants for the incensement.

Turner Drake and Partners Limited is a Nova Scotia based company that offers real estate services to the business community in Atlantic Canada. It provided COGS a complete dataset containing house sale records in the HRM since 1978 including more than thirty attributes for each house. With data mining, hidden information is discovered which reveals housing attributes that are important to high price real estate in different time periods. Maps of price distribution in HRM are generated by applying spatial statistics and geostatistics analysis.

## **Study Area**

The study area is the Halifax County, Nova Scotia, as shown in Figure 1.

Figure 1. Study area



The Turner's Areas are defined by Turner Drake and Partners Limited, containing twenty-one polygons representing sub-area of the Halifax County. The dataset has one attribute "AREA" which is used to record the location of houses within one of those twenty-one polygons. To simplify the data mining process and improve its running efficiency, the Turner's Areas are aggregated into four regions: downtown, coastal, suburban and rural. The relations between Turner's Areas and the four regions are shown in Table 1. A new field is added to the dataset for storing the region information on houses.

Table 1. Aggregation of Turner's Areas

Region	Turner's Areas
Downtown	Bedford, Dartmouth, Halifax North, Halifax Peninsula, Halifax South
Coastal	Bayside / Glen Haven, Cole Harbour / Lawrencetown, Eastern Passage / Cow Bay, Herring Cove / Pennant, South Shore, Timberlea / Tantalon
Suburban	Hammond Plains, Harrietsfield, Timberlea / Tantalon, Lake Echo / Porters Lake, Sackville / Beaverbank, Waverly / Fall River, Westphalia / #7 Highway / Ross Road, Windsor Junction
Rural	Eastern Shore, Halifax County East

## Data preparation

Before any analysis, the raw data is examined. Missing value and duplicated records are removed. Some of the fields in the original dataset are not useful and therefore are eliminated. The reserved fields include: record ID that uniquely identifies each house being sold, housing attributes that describe the characteristics of houses, sale price of houses, and sale date of houses.

The dataset contains both continuous and categorical attributes. All the categorical attributes are defined with value domains. A frequency statistics analysis is performed to test if all the records fall into the domains. Records with errors are deleted from the prepared dataset. Some records contain unrealistic values for continuous attributes. For example, in some records, the number of bathroom is 1.80. Those records are also removed from the data file.

The final step of data preparation is to divide the dataset into six sub-sets of five-year sale records with the exception of the first set. Table 2 summarizes information of each sub-set.

**Table 2. Information of six data sub-sets**

Time Period	Year	Number of Records	Number of Geocoded Records
1	1978 - 1981	5096	4824 (94.66%)
2	1982 - 1986	15463	14646 (94.72%)
3	1987 - 1991	21973	20731 (94.35%)
4	1992 - 1996	22573	21267 (94.21%)
5	1997 - 2001	26117	24640 (94.34%)
6	2002 - 2006	29340	27142 (92.51%)

## Statistics of Price

Figure 2 on the next page shows the line chart of mean price and Table 3 shows some descriptive statistics grouped by time periods. From the first to the third time period, the rate of price incensement is high with the slope about 1. From the third to the fifth time period, the speed of inflation slows down. From the fifth to the sixth time period, the rising rate is higher than in any previous time period.

Figure 2. Line chart of mean price grouped by time period, generated from SPSS

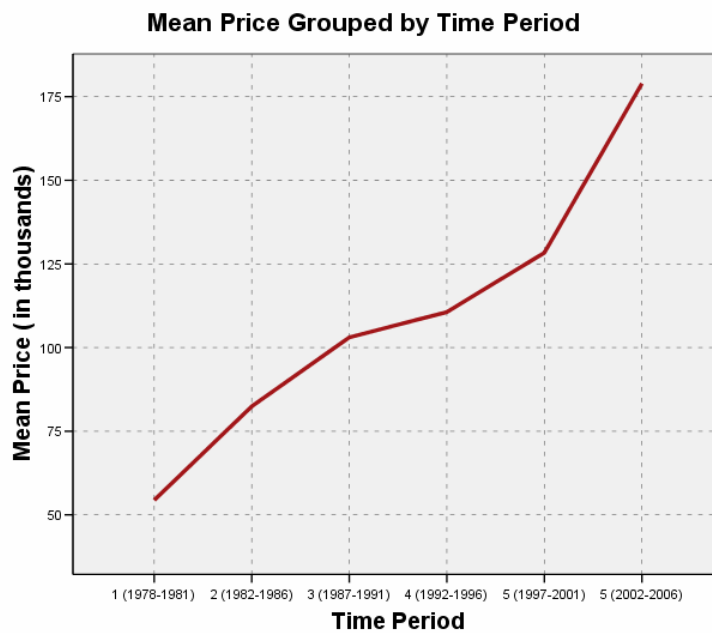


Table 3. Descriptive statistics for house prices grouped by time period

Price per thousand

	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
1 (1978-1981)	5096	54.39	22.616	.317	3	228
2 (1982-1986)	15463	82.43	38.587	.310	0	1120
3 (1987-1991)	21973	103.06	51.001	.344	0	1319
4 (1992-1996)	22573	110.61	48.216	.321	0	870
5 (1997-2001)	26117	128.35	63.966	.396	6	1037
6 (2001-2006)	29340	178.92	95.857	.560	4	3100
Total	120562	123.71	74.313	.214	0	3100

Table 4, 5 and 6 are the output tables from the one-way ANOVA test. ANOVA stands for Analysis of Variance. It is a method of testing the equality of three or more attribute means based on the sample information. The significance level in the Test of Homogeneity of Variances table is 0.00, which is less than 0.05. It means that the variances of price for each time period are significantly different. The significance level in the ANOVA table is also less than 0.05. It indicates that the means of prices for each time period are not equal. The Multiple Comparisons table lists the difference between each pair of means of price. The asterisks printed after the numbers indicate that the mean difference is significant.

Table 4. Test of homogeneity of variances

Price per thousand

Levene Statistic	df1	df2	Sig.
2110.155	5	120556	.000

Table 5. ANOVA table

Price per thousand

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.5E+008	5	30817936.22	7260.708	.000
Within Groups	5.1E+008	120556	4244.481		
Total	6.7E+008	120561			

Table 6. Multiple comparison table

Dependent Variable: Price per thousand

Tukey HSD

(I) period	(J) period	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-28.043*	1.052	.000	-31.04	-25.04
	3	-48.671*	1.013	.000	-51.56	-45.78
	4	-56.225*	1.010	.000	-59.10	-53.35
	5	-73.965*	.998	.000	-76.81	-71.12
	6	-124.536*	.989	.000	-127.35	-121.72
2	1	28.043*	1.052	.000	25.04	31.04
	3	-20.627*	.684	.000	-22.58	-18.68
	4	-28.181*	.680	.000	-30.12	-26.24
	5	-45.922*	.661	.000	-47.81	-44.04
	6	-96.493*	.647	.000	-98.34	-94.65
3	1	48.671*	1.013	.000	45.78	51.56
	2	20.627*	.684	.000	18.68	22.58
	4	-7.554*	.617	.000	-9.31	-5.79
	5	-25.294*	.596	.000	-26.99	-23.59
	6	-75.866*	.581	.000	-77.52	-74.21
4	1	56.225*	1.010	.000	53.35	59.10
	2	28.181*	.680	.000	26.24	30.12
	3	7.554*	.617	.000	5.79	9.31
	5	-17.740*	.592	.000	-19.43	-16.05
	6	-68.312*	.577	.000	-69.96	-66.67
5	1	73.965*	.998	.000	71.12	76.81
	2	45.922*	.661	.000	44.04	47.81
	3	25.294*	.596	.000	23.59	26.99
	4	17.740*	.592	.000	16.05	19.43
	6	-50.571*	.554	.000	-52.15	-48.99
6	1	124.536*	.989	.000	121.72	127.35
	2	96.493*	.647	.000	94.65	98.34
	3	75.866*	.581	.000	74.21	77.52
	4	68.312*	.577	.000	66.67	69.96
	5	50.571*	.554	.000	48.99	52.15

\*. The mean difference is significant at the .05 level.

Table 7 presents means of prices grouped by region and time period. For all six periods, means of price in the downtown area are the highest. The difference of means between the downtown and the rural area becomes larger as time goes by.

**Table 7. Means of prices grouped by region and time period**

Time period	Price per thousand							
	Region							
	Downtown		Coastal		Suburban		Rural	
	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
1 (1978-1981)	58.69	26.509	51.16	16.797	50.87	16.359	32.11	16.145
2 (1982-1986)	90.65	43.302	76.14	34.815	73.92	24.733	41.65	20.205
3 (1987-1991)	117.35	58.475	91.21	37.465	89.59	36.138	55.86	30.183
4 (1992-1996)	123.77	56.364	99.22	35.595	99.02	30.996	67.82	36.315
5 (1997-2001)	140.22	72.826	112.25	51.131	119.73	45.294	77.07	49.399
5 (2002-2006)	199.57	106.521	152.37	71.603	168.01	83.368	106.00	67.762
Total	137.05	83.062	106.06	55.765	116.40	64.704	73.45	52.342

## Decision Tree Analysis

A decision tree is a commonly-used predictive model in data mining and machine learning. It partitions data based on the relationships between predictor variables and a target variable. A successful resulting tree reveals predictor variables that are most strongly related to the target variable. SPSS AnswerTree is a software package that generates decision trees. It allows users to choose different algorithms and input parameters. For the analysis of price changes, six decision tree models are produced by AnswerTree for six time periods respectively using the CHAID algorithm. The purpose is to discover important predictors and corresponding values for high price houses in each time period.

Table 8 summarizes the most important predictors of each of the six decision trees. For the first two time periods, "AESTHETIC APPEAL" and "NUMBER OF BATHROOMS" are the most important determinants for high price houses. In later time periods, "LIVING AREA" takes over the role.

**Table 8. Important predictors for high price houses for each time period**

Time Period	1st Important Variable	2nd Important Variable(s)
1 (1978-1981)	AESTHETIC APPEAL	NUMBER OF BATHROOMS
2 (1982-1986)	NUMBER OF BATHROOMS	AESTHETIC APPEAL,
3 (1987-1991)	LIVING AREA (sq ft)	AESTHETIC APPEAL, CONSTRUCTION QUALITY, NUMBER OF BATHROOMS
4 (1992-1996)	LIVING AREA (sq ft)	AESTHETIC APPEAL, NUMBER OF BATHROOMS
5 (1997-2001)	LIVING AREA (sq ft)	STYLE, BUILDING AGE
6 (2002-2006)	LIVING AREA (sq ft)	BUILDING AGE, SOLD YEAR

Table 9 lists a summary of important predictors and their corresponding values from AnswerTree analysis result.

**Table 9. Summary of AnswerTree result.**

Predictor	1 (1978-1981)	2 (1982-1986)	3 (1987-1991)	4 (1992-1996)	5 (1997-2001)	6 (2002-2006)
AESTHETIC APPEAL	Good   excellent	Unknown   good   excellent	Unknown   average  good   excellent	Unknown   average  good   excellent		
NUMBER OF BATHROOMS	>1	>1.5	>2	>2	>2	
NUMBER OF LAUNDRY ROOMS	>=0					
NUMBER OF ROOMS	<=6					>7
NUMBER OF FIREPLACES			>0		>0	
BASE SIZE (sq ft)	>1100					
LOT AREA (sq ft)		>3600		>4845		
LIVING AREA (sq ft)			>1345	>1579	>1643	>1500
REGION	Downtown   rural			Downtown	Downtown   rural   suburban	
STYLE					Unknown   apartment   cottage   mobile home   split   over 2 storey	
CONSTRUCTION QUALITY		Unknown   low   fair   good   excellent	Good   excellent	Unknown   low   fair   good   excellent		
DRIVEWAY		Paved   unpaved double	None   paved   Unpaved double			
GARAGE	Built-in or carport   single   triple   over triple			Attached/detached double or above	None   built-in or carport   attached/detached single, double and above	
SUNDECK SIZE						None   unknow   small   medium   large
SOLD YEAR		1986				2005   2006
BUILDING AGE				<=5	<=9	<=10

There are a few observations resulting from Table 4:

1) In the first two time periods, expensive houses have good or excellent aesthetic appeal. As time passes by appeal becomes a less important factor determining the costs of expensive houses. A house may be sold at a high price even if its looks are adequate.

2) Desired number of bathrooms, in expensive houses, keeps increasing during the first three time periods, and maintains to a fixed number in the recent time periods.

3) Desired living area of expensive houses increases in the third, fourth and fifth time periods, but drops slightly in the last time period. It indicates that the larger the living space is, the higher price a house may be sold.

4) House age becomes an important predictor of high price houses in the recent fifteen years.

## **Probability Analysis**

Probability analysis produces a surface that gives the probability that the variable of interest is above or below some specified threshold value. Figure 3 – Figure 8 are the probability maps generated from ArcMap for the six time periods. Each map reveals the geographical distribution of probability that a house sale price is higher than 150,000 dollars. The threshold of \$150,000 is the third quartile of all the house prices in the dataset.

In the map of the first time period, the probability is low with a range from 0.01 to 0.04. The area of such probability is small. The spots are the south of Bedford and along the coast of the Northwest Arm. For the second time period, the probability goes up slightly. The spots found in the first time period have higher probability values and larger area on the second map. There are also some high probability areas in the Timberlea/Tantalon area. The maps of the third and the fourth time periods look more complicated than the maps of the first two time periods. High probability areas are mostly located in the southwest of the Halifax County. At a glance, the color shown in Figure 5 is darker than that in Figure 6. This indicates that the general probability in the third period's map is even higher than the one in the fourth time period's map. It probably is one of the reflections of the line section with a low slope in Figure 2. The distribution of probability in Figure 7 and Figure 8 look more irregular. The contrast of light color and dark color in Figure 7 is obvious. This may indicate that in the fifth time period, the gap between high prices and low prices is significant. In Figure 8, the dark color is more intensive especially around the Halifax-Bedford-Dartmouth region. The area of light color is much smaller than in any of the maps of previous time periods. Comparing with the line chart in Figure 2, it is not difficult to understand that house prices increase very fast.

Figure 3. Probability map of time period 1

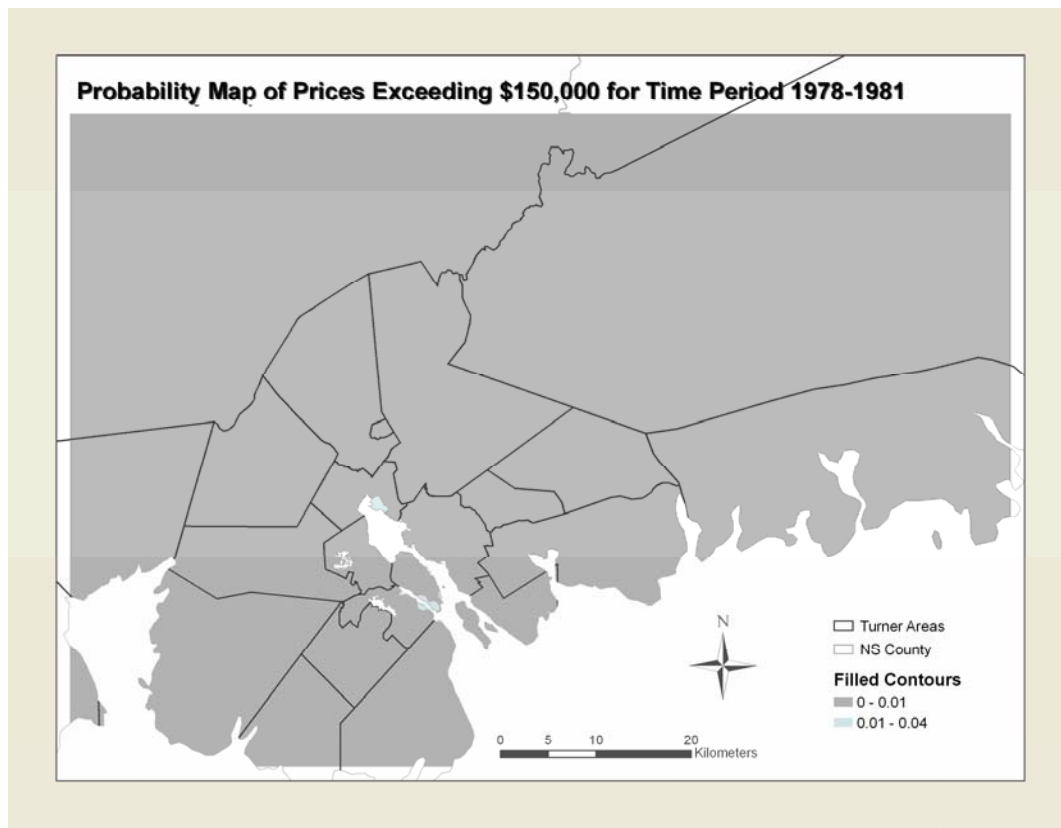


Figure 4. Probability map of time period 2

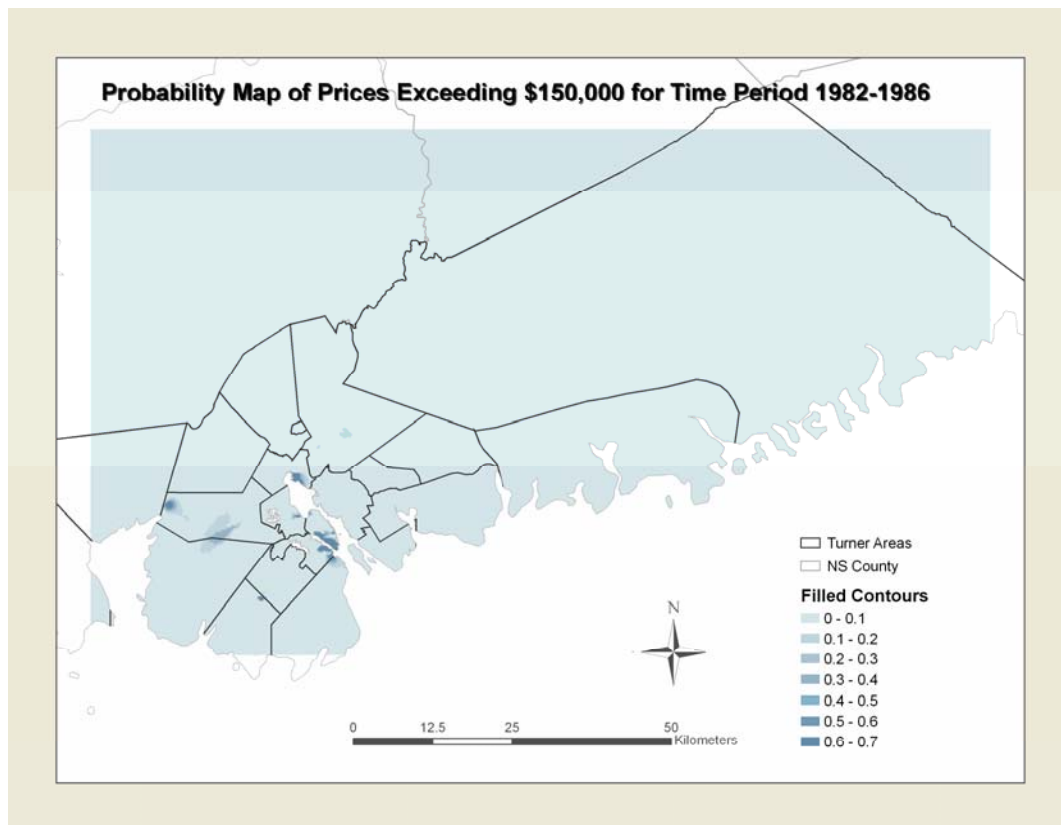


Figure 5. Probability map of time period 3

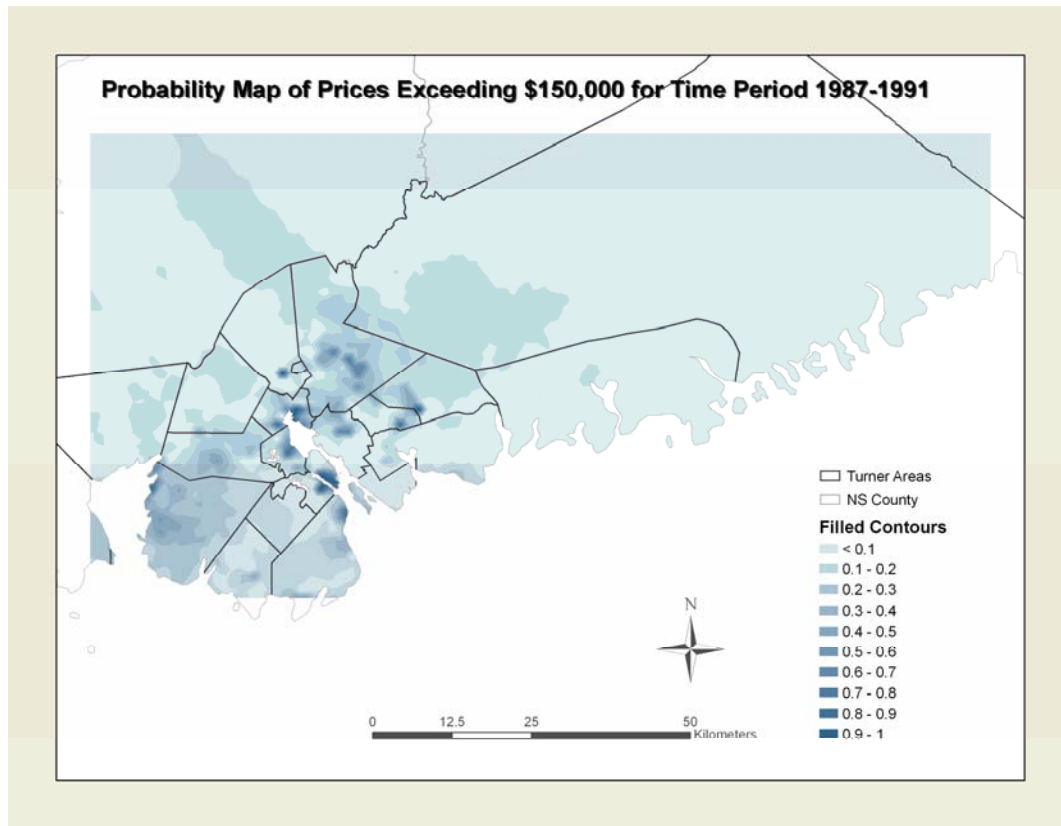


Figure 6. Probability map of time period 4

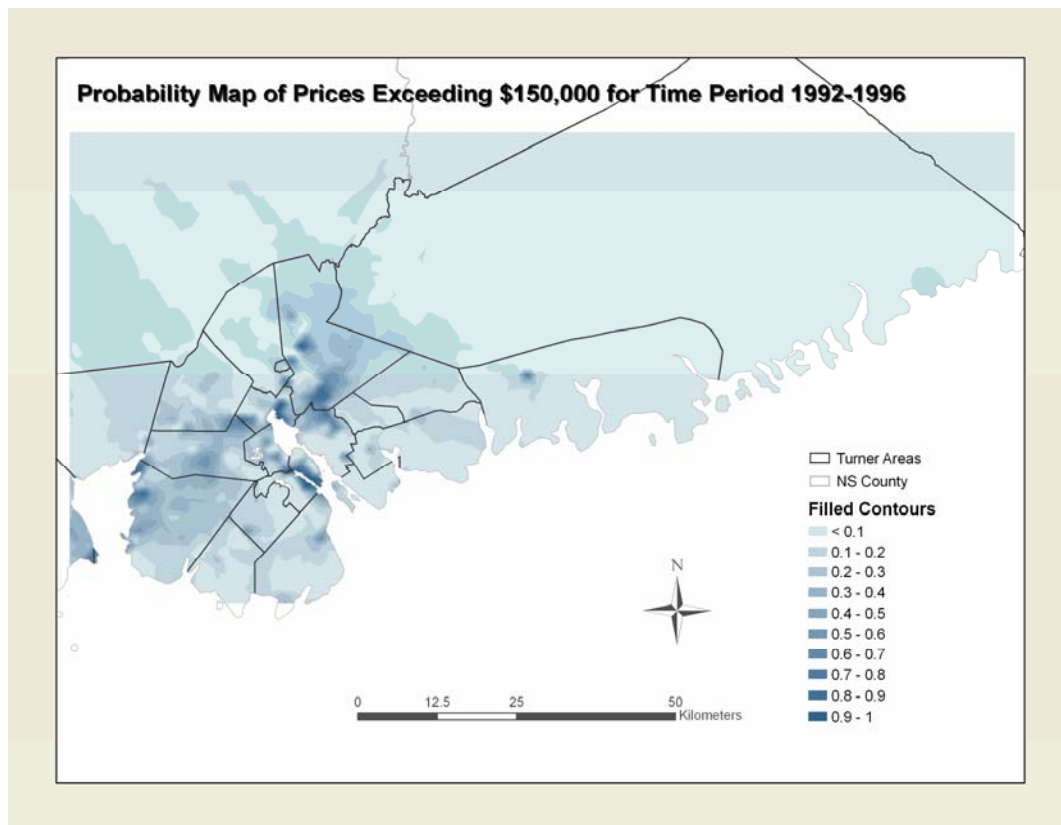


Figure 7. Probability map of time period 5

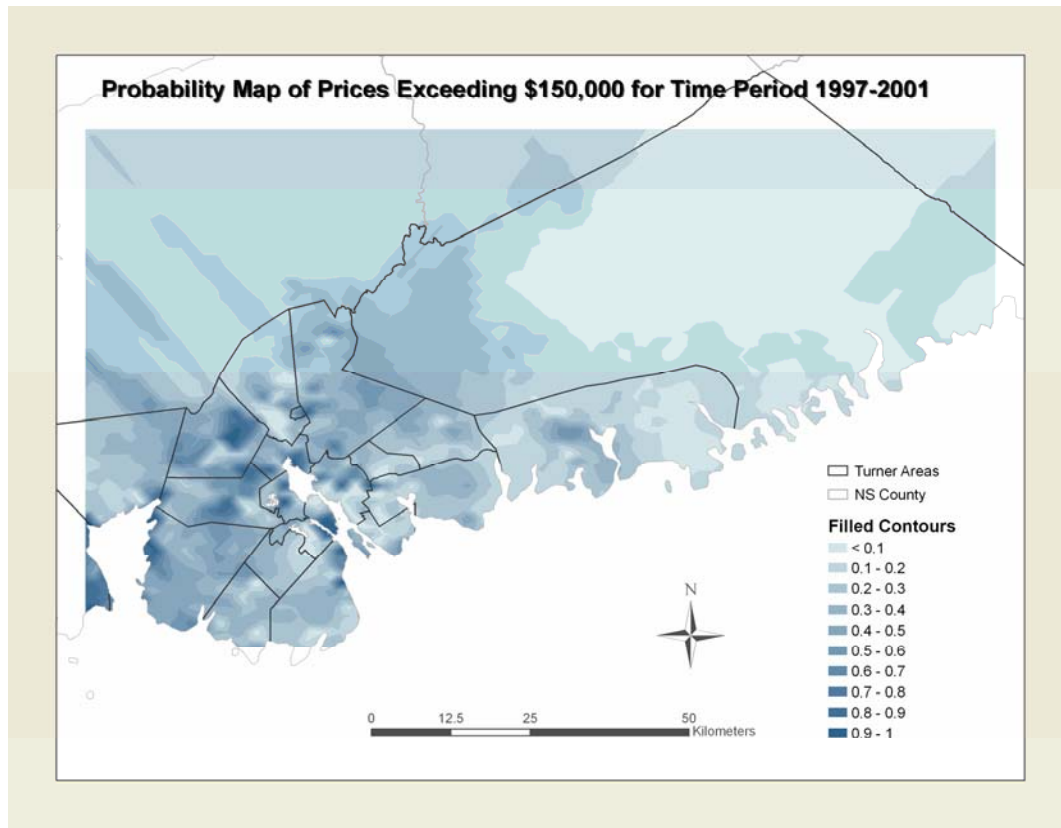
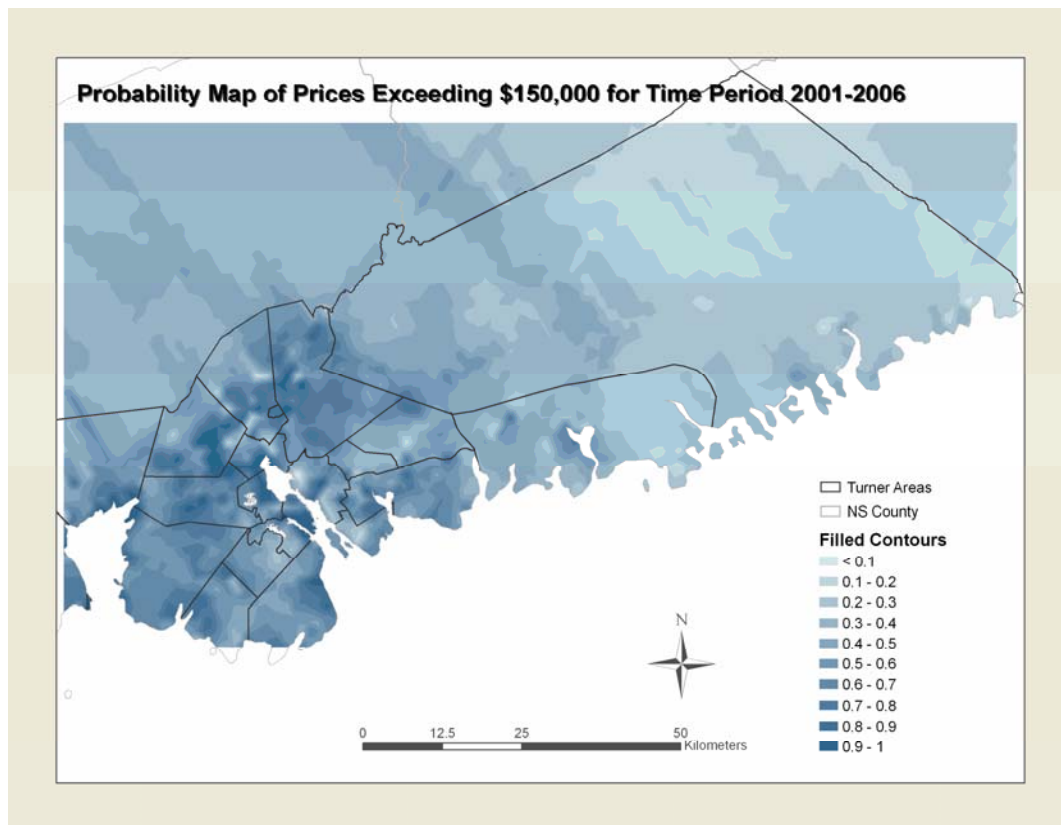


Figure 8. Probability map of time period 6



## **Conclusion**

This article discusses some analyses of sale prices of houses in the Halifax County. A dataset of sale records with columns and rows does not always present sufficient information to observers. Techniques of data mining, general statistics and geostatistics help to discover the change of house prices in different time periods, and reveal some hidden information that can not be seen readily from the numbers. Such information is useful for decision makers giving a deeper understanding their markets. In the future, more work can be done to combine census and demographic data which will yield interesting research scenario.

## **Acknowledgement**

The work shown in this article is part of the fulfillment of a continuous project of GIS for Business. The house price dataset is geocoded and provided by Turner Drake and Partners Limited. A confidential agreement has been signed to protect individual house sale price being revealed. I wish to thank Alexandra Allen for her help with data dictionary and geocoding issues.