VISUAL SENSITIVITY ANALYSIS

(APPLIED TO REAL ESTATE PREDICATION SYSTEM)

A thesis submitted in fulfilment of the requirements for the degree of

Master of Science in Computing Sciences in the

Faculty of Engineering and information technology at

University of Technology, Sydney

MASSARA DA'ANA

JANUARY 2014

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student:

Date:

Acknowledgments

Many thanks to my supervisor, A/Prof. Mao Lin Huang, who encouraged this research idea from the very beginning. It would not have been possible to accomplish this work without his constant guidance, support and invaluable advice. Thank you again for teaching me how to become a better researcher.

Thanks to Mr. Mohammed Al Alkadi and Miss Zhenya Zhang for their help while developing the prototype system.

Thanks for Dr. Esam Dana and Dr. Zenat Dana for assisting me in different ways throughout this research.

Thanks to my husband, my sisters and brothers, for their constant encouragement and support, and to my children for respecting the time I spent on this research more than I expected. To them all, I dedicate this thesis.

Table of Contents

List of Figures

List of Tables

ABSTRACT

Sensitivity analysis is the science studies the impact of independent variables on the dependant variable in the studied model, in addition to investigating relationships between those variables. Sensitivity analysis is a prevalent group of techniques and approaches has proven its feasibility in wide range of disciplines.

However, the traditional sensitivity analysis methods have the common weakness of *user interaction absence*. Furthermore; each sensitivity analysis method has its *own level of difficulty which is an obstacle for a non-expert user* to use or even to interpret the results if the analysis is conducted using a software like SPSS.

Recently, visualizations are being used to present data efficiently in terms of assisting human visual perception and reducing cognition effort. These visualization techniques when integrated with interaction will facilitate data manipulation and exploration.

This study integrates sensitivity analysis with interactive visualization into a prediction system that allows non-expert users to analyse and understand the real estate data through the visualization and direct visual interactions, which hide the complexity of the sensitivity analysis algorithms. We take advantage of the visualization that amplifies cognition in dealing with abstract data.

As shown in the outcome, the user can use the sensitivity analysis method used in this system *interactively* setting his/her preferences for the property via the visualization *without any prerequisite of sensitivity analysis knowledge.*

The use of scatter plots, one of sensitivity analysis methods, is used in studying the relationships between the predictors and the response variables to decide whether variable transformation is needed or not. Additionally, scatter plots are used to summarize all analyses conducted.

1 INTRODUCTION

1.1 MOTIVATION

Traditional sensitivity analysis methods were being used, generally, to measure the impact of input variables on a specific dependant variable and to eliminate non-influential variables from the model under study.

A researcher has many previously proposed sensitivity analysis methods to choose from to apply to his/her model under study. These methods, as presented in the related work chapter, are classified in two different ways; one based on the method capabilities and the other is based on how the method is conducted. However, each traditional sensitivity analysis method has its own strength and weaknesses, there is no super method. But all sensitivity analysis methods share the weakness of *user interaction absence*. Additionally, each sensitivity analysis method has its own complexity which is a hurdle for a sensitivity analysis domain non-expert to use or even to interpret the results if the analysis is conducted using any statistical software package.

Therefore, the research questions are; how can we overcome the user interaction absence that sensitivity analysis methods suffer from? And how can we allow non-expert users to use sensitivity analysis without any prior knowledge of the field?

Information visualization is a modern tool which provides users with advanced graphical user interface for supporting direct human-computer and human-data interactions in all kinds of data analysis processes. It plays an important role for exploring model sensitivity to facilitate decision making under uncertainty. It provides a means for graphically exploring the

relationships between the output and the inputs of a model and to determine how sensitive a model is to changes in the values of the input.

Scientists from different disciplines such as finance, biology, sociology, and many others are implementing different visualization techniques to present data efficiently to benefit from the visual perception and the human cognitive system. *These visualizations when integrated with interaction will facilitate data manipulation and exploration.*

Current property lookup systems and/or web sites offer searching capabilities to list a number of properties that satisfy a search criteria, but do not offer the ability to customize the property features that the user is interested in to match his/her preferences, taking into account that the criteria is entered by textual means.

To address this question, this research introduces a novel approach that integrates interactive visualisation with sensitivity analysis. In this approach, both interactive visualisation and sensitivity analysis work together to overcome "user interaction absence" and the hurdle for "non-expert user" to utilize sensitivity analysis in a transparent manner. While the visualisation presents the data in more perceptive manner than traditional textual formats, the interactivity feature make up for the sensitivity analysis interaction absence by allowing the user to directly interact with the data underlying the visualisation. Additionally, the visualisation equipped with interactivity will hide the complexity of the sensitivity analysis method used and produce interpreted results in a non-expert user comprehensible language. Furthermore, all produced results for each individual dimension are aggregated in a traditional scatter plot visualisation as a summary plot for the expert user.

1.2 CONTRIBUTION

This study presents a novel approach that has the following contributions:

1. Integrates interactive visualizations and sensitivity analysis; the interactive visualisation *provides interaction means* directly with the computer and the data while the visualisation assists human visual perception and reduces cognition effort.
2. IMAP: is an interactive geometrical map that employs the overview + detail and filtering concepts, providing users with an abstract and very user-friendly navigational map. For all the selected regions, IMAP shows geographical references that are essential for people daily life like; train stations, rail tracks, shopping centers, tourism icons and the current listed properties for sale. The visualisation hides the complexity of the sensitivity analysis method, so using sensitivity analysis is not limited to sensitivity analysis domain experts.
3. This approach of integration is applied to a *real estate prediction system* which presents the user with a visualisation for an abstract geographical map (IMAP) that shows important information about the properties where he/she can select his/her preferences through the interactive map (IMAP) rather than traditional textual selection methods.
4. The real estate prediction system is presenting the results of the sensitivity analysis by means of scatter plots. Each scatter plot reflects only one influential predictor and shows two series, the first series depicts the property observed selling price designated in blue coloured points and the other depicts the property predicted selling price designated in pink coloured points. The adjusted value of the influential predictor is depicted on the x-axis whereas the price is depicted on the y-axis. This scatter plot compiles the regression

analysis results and exposes the differences between the properties observed and predicted selling prices.

5. Sensitivity analysis and information visualization, are both considered interdisciplinary; hence this approach of integrating both will be interdisciplinary too and expected to have broad range of applications, where researchers will extend their communication across disciplines and their combined effort will benefit all disciplines involved.

1.3 ORGANIZATION

The reminder of this thesis is organized as follows: chapter 2 provides background information about sensitivity analysis methods and visualization approaches. Chapter 3: describes approaches of building prediction model and describes the prediction system development. Chapter 4: discusses the results. Chapter 5: presents 3 case studies to demonstrate how the system works and to provide evidence supports the discussed results. Chapter 6: compares the presented property prediction system to previous property prediction systems. Chapter 7: concludes this thesis and discusses future work.

2 Related Work

This chapter reviews two different fields, the first is sensitivity analysis, specifically graphical sensitivity analysis, and the second is visualizations techniques.

2.1 Sensitivity Analysis

"Sensitivity Analysis (SA) studies the relationships between information flowing in and out of the model" (Saltelli, Chan & Scott 2000). By definition, SA methods should have the ability to deal with a model irrespective of model linearity or additivity, expecting interaction between input variables, and to estimating the effect of individual input variable while allowing all other inputs to vary (Saltelli & Annoni 2010).

The third international symposium on sensitivity analysis of model output (SAMO 2001) elaborated clearly the developing interest in sensitivity analysis. Possibly, this interest can be clarified by (Tarantola & Saltelli 2003):

- The expansion of simulation models employment in many disciplines.
- The professional's recognition of SA methods as formulas for model use.

Sensitivity Analysis is broadly used and it has many applications including and not limited to (Pannell 1997):

- Decision making or development of recommendations for decision makers.
- Communication.
- Increased understanding or quantification of a given system.
- Model development.

Saltelli & Annoni categorized sensitivity analysis into three different categories which focus on sensitivity analysis technique capabilities (2010), these are: Local sensitivity analysis, Global sensitivity analysis and One-At-A-Time (OAT) designs (briefly highlighted in the sections 2.1.1.2). While Christopher Frey categorized sensitivity analysis into other three categories (2002), which focus on how sensitivity analysis is conducted, these are: mathematical, statistical and graphical methods (reviewed in the section 2.1.2.1 and 2.1.2.2 respectively).

2.1.1 SA Classification based on SA capabilities

In this section, local and global SA will be briefly highlighted. A special attention is paid to the one-at-a-time design, which its results simulate the adopted method results this research.

2.1.1.1 Local SA

Local sensitivity methods simply employ derivatives to examine the importance of an input variable by the derivative of the output with respect to that input variable, given that all derivatives are taken at the same point called the 'baseline' point in the hyperspace of the input variables (Saltelli & Annoni 2010), to which the local analysis relatively addresses sensitivities (Hamby 1995). It is considered suitable -to some extent- for only small changes in the input variable (Hamby 1995; Saltelli, Chan & Scott 2000).

For the analysis to be informative, local approach restricts the mathematical model to be linear or at least additive when it's far from the baseline point where the derivatives are computed. Moreover, it does not provide exploration of the rest of the space of input variables (Saltelli & Annoni 2010). It proves less useful when SA is employed to compare the effect of different factors on the output (Saltelli, Chan & Scott 2000).

This approach is favoured due to its computer time efficiency (Saltelli et al. 2008).

2.1.1.2 One-At-A-Time design

A widely used SA approach is the one-at-a-time design (OAT), which is considered the simplest SA method (Campolongo, Kleijnen & Andres 2000; Hamby 1995; Saltelli & Annoni 2010). OAT is conducted iteratively by varying one factor at a time while keeping all other factors unchanged (Campolongo, Kleijnen & Andres 2000; Hamby 1995; Saltelli & Annoni 2010). By changing one-factor-a-time, the analyst can monitor the impact on the output variable.

A recent study by Saltelli and Annoni had criticized OAT because it is non-explorative in high dimensional models and fails to activate factor interaction which is an important factor of SA by definition (Saltelli & Annoni 2010).

However, OAT has many advantages including the following:

- Low computational cost (Campolongo, Kleijnen & Andres 2000);
- All OAT sensitivities has the same starting point (Saltelli & Annoni 2010);
- Because OAT varies one factor at a time, the observed impact on the output variable will be exclusively ascribed to that varied factor even if it has no effect, no effect does not signify no influence (Saltelli & Annoni 2010);
- On the contrary, influence is implied by a non-zero effect; hence OAT will never observe non-influential factors as relevant (Saltelli & Annoni 2010);
- Practically, it is unlikely for the model to crash or even to produce unreasonable results because it varies one factor at a time, the possibility will increase when more than one factor are varied (Saltelli & Annoni 2010);

- Even if the model had crashed, the responsible factor will be immediately figured out because it is the only factor that had been changed (Saltelli & Annoni 2010).

2.1.1.3 Global Sensitivity Analysis

Global SA is "the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input" (Saltelli et al. 2008). Furthermore, Global SA methods should be defined with two attributes (Saltelli, Chan & Scott 2000):

Multidimensional averaging: implies investigating sensitivity for individual factor while varying all other factors at the same time.

The inclusion of influence of scale and shape: integrates the effect of the whole range of variation and the form of the probability distribution density function of the input.

In other words, global SA investigates sensitivity with respect to the entire parameter distribution (Hamby 1995) and takes into consideration multiple input interaction, so that it can provide more robust insights (Christopher Frey & Patil 2002).

2.1.2 SA Classification based on how the method is conducted

This classification includes statistical methods, mathematical methods and graphical methods. Because this research is concerned in visualizing sensitivity analysis, only shall the graphical sensitivity analysis methods be reviewed in the present section, in addition to the adopted method in this research, regression analysis which considered one of the statistical sensitivity analysis methods.

2.1.2.1 Graphical Sensitivity Analysis Methods

Graphical methods for sensitivity analysis or graphical sensitivity analysis (GSA) have a significant role in the analysis process. Due to the visual nature of representation, GSA provides scientists with easy approaches to describe models and get insights into variables' relationships (Cooke & Noortwijk 2000), simplify model exploration and determine model sensitivity to changes in individual input variables (McFarlane & Young 1994). GSA uses several kinds of graphs, charts or surfaces to represent the outcome of the analysis process (Daradkeh, Churcher & McKinnon 2008).

The theoretical concept underpinning GSA has not been widely researched; however, any model can employ these methods (Daradkeh, Churcher & McKinnon 2008). Graphical sensitivity analysis was applied to different areas including neural and social networks, climate studies and nuclear studies.

This thesis discusses generic graphical methods, emphasizing on the presentation and the readability of the graphs, strengths and weaknesses for each of the following:

- Tornado graphs;
- Radar graphs;
- Scatter plots and scatter plots matrix;
- Parallel coordinates plots.

Each graphical method has its own strengths and weaknesses and it will be pointed out individually. Generally, strengths could be any of the following:

- Representing complex relationships between input and output variables (Daradkeh, Churcher & McKinnon 2008);
- It can be employed as a screening method to help in choosing appropriate sensitivity analysis method (Daradkeh, Churcher & McKinnon 2008). Furthermore, they can also be used as a

complementary for statistical and mathematical methods (Daradkeh, Churcher & McKinnon 2008);

- It helps in communicating results between decision makes, users and analysts (Cooke & Noortwijk 2000).

Weaknesses may involve incapacity to handle model high dimensionality or method lack of interaction capability (Cooke & Noortwijk 2000).

2.1.2.2 Tornado Graphs

Tornado graphs are one of the graphical sensitivity analysis methods that show the most influential variables very easily on model output and rank them in order of their significance (Daradkeh, Churcher & McKinnon 2008). A tornado diagram is comprised of horizontal bars each one corresponds to the sensitivity of the predictor, these bars are stacked in descending order of their impact on the dependent variable, the length of the bar designates the total impact of the independent variable on the model output (Daradkeh, Churcher & McKinnon 2008). The left and right ends of the bar represents the upper and lower limits of the model output in regards to the upper and lower limits of the independent variable (Daradkeh, Churcher & McKinnon 2008).

Tornado graphs can represent an unlimited number of input variables (Eschenbach 1992). Tornado graphs assume that all predictors are independent, if this assumption is not valid in the real model, then a group of dependant variables may falsely look lower on the diagram than they should be (Daradkeh, Churcher & McKinnon 2008).

The information depicted on a tornado diagram shows the boundaries for each independent variable, the relative change in each variable and whether the underlying model is linear or not; non-linearity can affect the simplicity of interpreting a tornado diagram (Eschenbach 1992). Furthermore, Tornado

diagrams highlight the most influential variables, guiding the analyst through the directions to reducing uncertainties (Daradkeh, Churcher & McKinnon 2008).

A tornado diagram is shown in Figure 1. The diagram illustrates the sensitivity of the price of housing properties in Sydney regions with respect to the property size (area) and distance to Sydney CBD, distance to nearest train station, number of bed rooms, number of bath rooms, and car space. The Tornado diagram clearly shows that the property price is least sensitive to the number of bath rooms and car space, and the price is most sensitive to the property distance from Sydney CBD; followed by the distance to nearest train station.

There are three different types of common errors that analyst should be aware of while constructing a tornado diagram (Eschenbach 1992). Firstly, wrong assumption of existence of monotonic relationship between decreasing an input variable and decreasing an output variable. Secondly, assigning plus and minus to the same arbitrary value for the independent variable boundaries. Thirdly, ignoring the uncertainty in the boundaries of the input variables.

Briefly, Tornado diagrams are useful to represent a model with any number of independent variables, to specify the most significant ones, and to identify the limits for each variable. Model non linearity can complicate the interpretation process for the Tornado diagrams. Care should be taken to avoid the aforementioned three common errors while constructing Tornado diagrams.

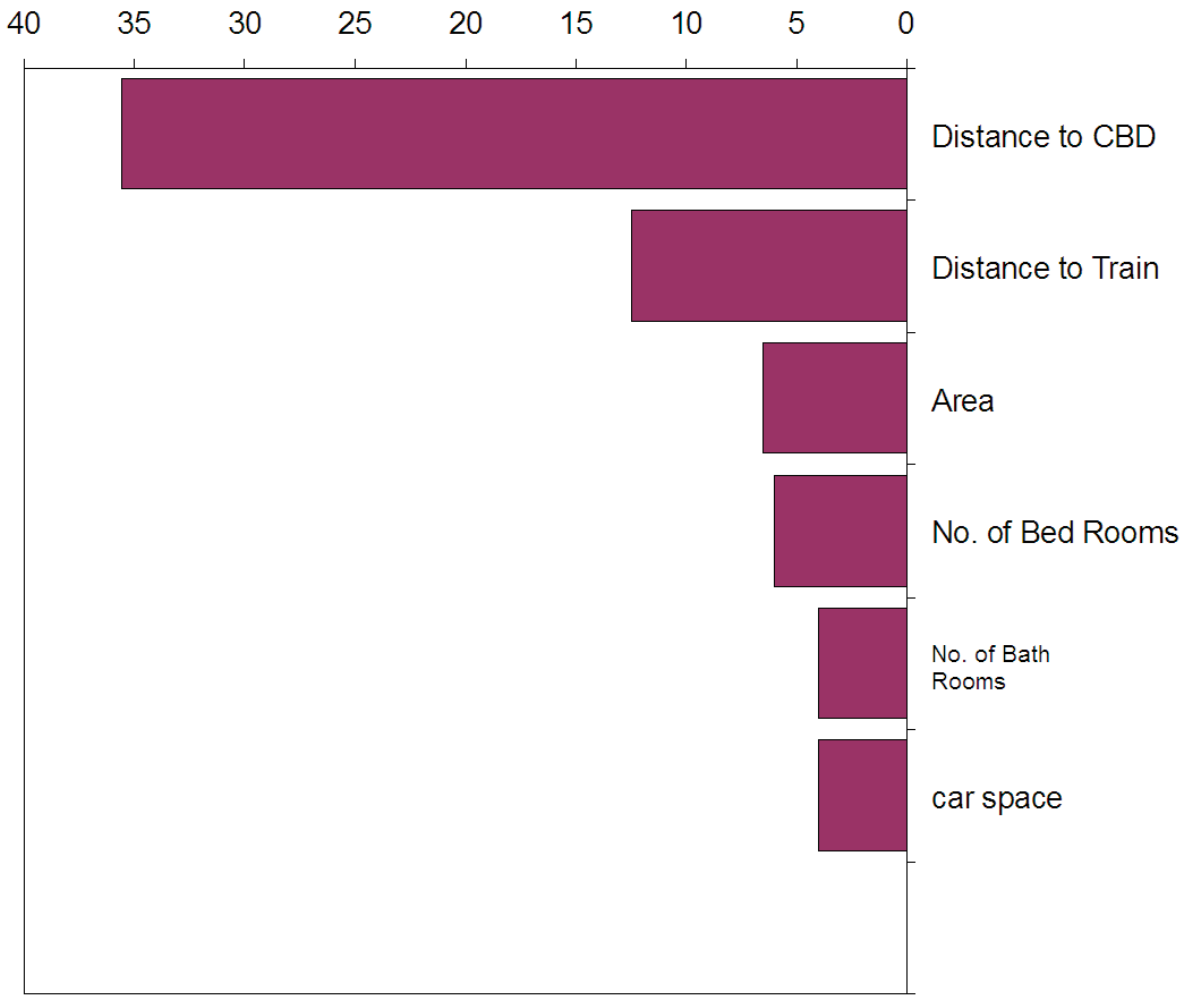


Figure 1: Tornado diagram.

2.1.2.2.1 RADAR GRAPHS

Rader graphs are one of the simplest generic graphical methods. Radar graphs are another graphical sensitivity analysis method which shows the impact of independent variables on the model output (Cooke & Noortwijk 2000). Radar graphs are also known as spider graphs (Christiernin 2010; Daradkeh, Churcher & McKinnon 2008; Linn Gustavsson 2010).

Radar graphs are two dimensional polar graphs, where a radial axis is drawn from the center of the graph for each input variable (Christiernin 2010; Cooke & Noortwijk 2000; Daradkeh, Churcher & McKinnon 2008; Linn Gustavsson

2010). The axes are marked with scale values increase as we move from the center towards the circumference, the variable with the highest rank correlation is plotted farthest from the center while the variable with lowest rank correlation is plotted closest to the center (Christiernin 2010; Cooke & Noortwijk 2000; Daradkeh, Churcher & McKinnon 2008; Linn Gustavsson 2010). After the data being plotted along each axis; a straight line will connect the data points forming a polygon or a star shaped pattern (Christiernin 2010).

The same information shown in figure 1 is depicted on a radar graph in figure 2. The most sensitive variable, distance to Sydney CBD, is plotted closest to the circumference and the least sensitive variable, number of bath rooms and car space, are plotted nearest to the centre. Both, the radar graph and the tornado diagram agree on the sensitivity ranking for all the variables included.

Radar graphs are useful for large set of variables while keeping the presentation compact, for example, presenting 31 variables on Excel bar chart require 7 pages of size A4, but a Radar graph can fit the presentation on one page (Cooke & Noortwijk 2000). Radar graphs are also useful in showing stakeholders consensus (Saary 2008) when all or most of the polylines meet in a point, this point is considered as the stakeholders consensus.

In summary, the two dimensional polar graph, is helpful in exhibiting the influence of large number of independent variables on the output variable in a compact display and in clarifying stakeholders consensus. The most sensitive variables will be closer to the circumference and the least sensitive will be closer to the centre. Radar graphs are not applicable for dependant variables.

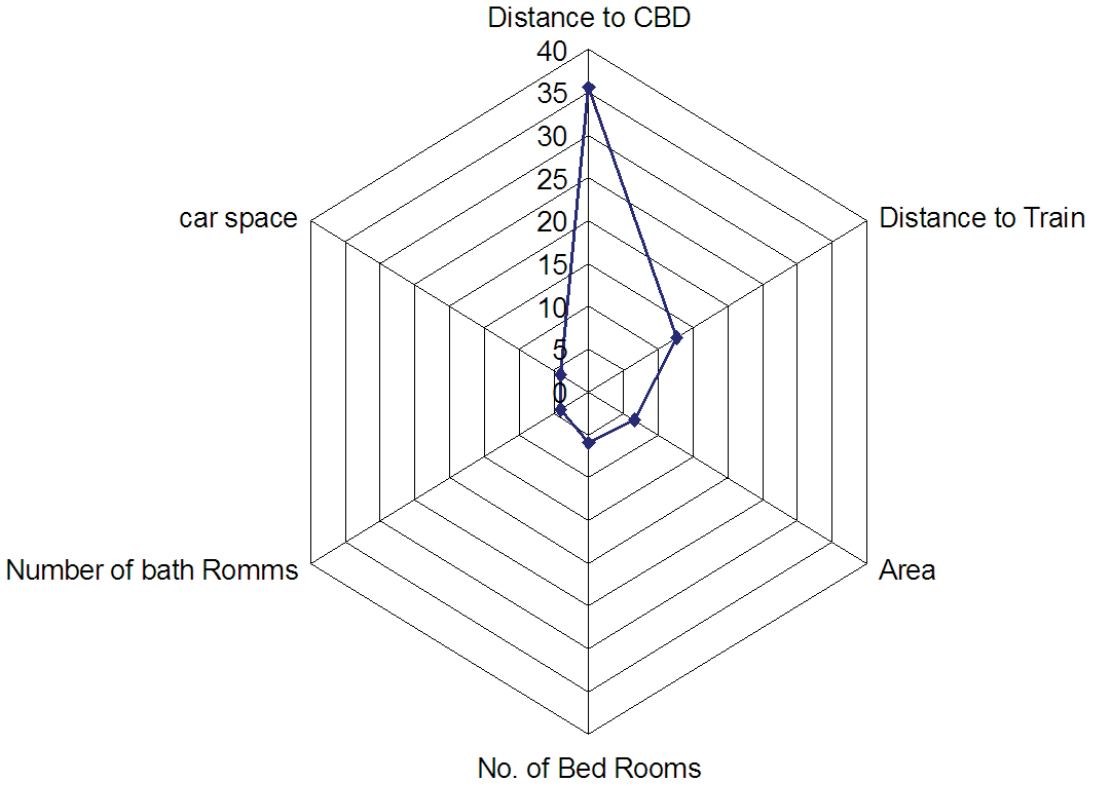


Figure 2: Radar chart.

2.1.2.2.2 SCATTER PLOTS (SCATTER DIAGRAM)

Scatter plots are widely used and are applied to various research disciplines. Scatter plots are known as the most intuitive and straightforward method in SA (Campolongo et al. 2000; Saltelli et al. 2008). Scatter plots are helpful in investigating the relationship between input and output variables (Christopher Frey & Patil 2002; Tague 1995) by plotting a single input variable on the x-axis and the output variable on the y-axis (Daradkeh, Churcher & McKinnon 2008; Tague 1995). Scatter plots also useful in identifying possible root causes of a problem (Tague 1995). Scatter plots are more informative than line graphs, they show outliers and trends are more obvious than line graphs and they show the degree of variability that each variable has at each location of the other variable (Utts 1999). Furthermore, scatter plots often show

nonlinear relationships, thresholds and input variables interactions that help in comprehending model behaviour and planning of more complex sensitivity studies (Campolongo et al. 2000).

The scatter plot in figure 3 shows a negative moderate correlation between the variables property sale price and distance to Sydney CBD and a correlation of -0.434 reasserts that. The correlation calculations agree with the scatter plot which clearly shows that the relationship between the two variables is moderate.

Investigating scatter plots is a significant part of sampling-based SA and can capture patterns that are not detected by regression-based procedures (Kleijnen & Helton 1999), and is frequently used as a guide in selecting appropriate SA methods because it has the capability to identify complex dependencies between an input and an output variables, this capability is considered as a major strength for scatter plots (Christopher Frey & Patil 2002).

Scatter plots is model independent, provide a qualitative analysis (Campolongo et al. 2000; Christopher Frey & Patil 2002) since they can not provide quantification for the variables. The computational cost depends on the number of input and output variables (Christopher Frey & Patil 2002).

Ranking input factors quickly and automatically is a challenge for scatter plots in models with many input factors (Christopher Frey & Patil 2002; Saltelli et al. 2008), due to the huge number of plots involved. Furthermore, in some models, uncertain input factors might form sets of input variables which can not be visualized simply by two dimensional scatter plots (Saltelli et al. 2008).

In summary, in virtu of scatter plots simplicity, they offer huge benefits to the analyst including identifying complex dependencies and relationships between variables, showing outliers and trends and assisting in selecting appropriate SA methods. The cost of the qualitative analysis offered by scatter plots depends on the number of model's variables. The major drawback for scatter plots is the huge number of plots produced that can slow down input variables ranking.

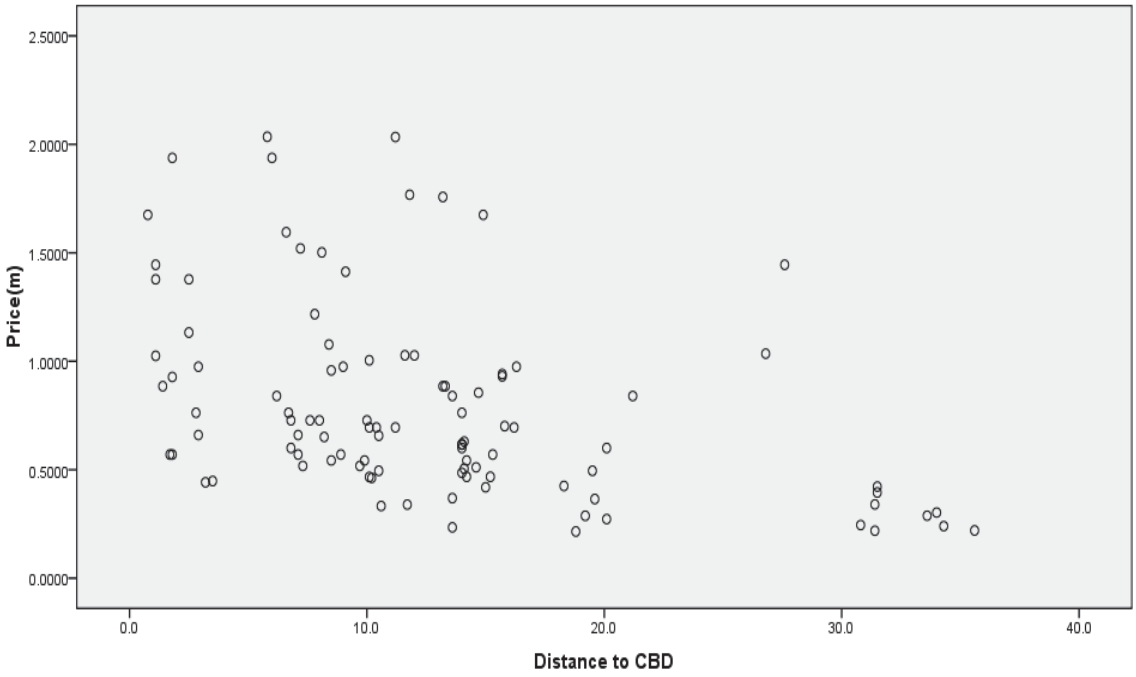


Figure 3: Scatter plot of property prices (million dollars) versus the distance to Sydney CBD (km).

2.1.2.2.3 SCATTER PLOT MATRIX

A scatter plot matrix is a combination of several scatter plots ordered in rows and columns. More precisely, a scatter plot for p variate model is the ordered display of p(p-1) scatter plots (Carr et al. 1987). Scatter plot matrix is used when there is a need to display more than two variables simultaneously,

instead of using single scatter plot which can display two variables only at a time (Daradkeh, Churcher & McKinnon 2008; Niklas 2008).

In addition to the scatter plot advantages, scatter plot matrix is considered as a high-performance interaction technique and effective approach to explore multidimensional data (Carr et al. 1987; Niklas 2008).

Figure 4 shows a scatter plot matrix produced by this research as an exploration to the problem at hand. The scatter plot matrix displays the pair wise scatter plots between the variables property price, distance to CBD, distance to train station, area, number of bed rooms, number of bath rooms and car space. Some pairs included are meaningless like number of bed rooms and distance to train station, but they are included because scatter plot matrix pair each variable with all other variables. Our interest is the pairs that have the price as one of the two variables.

Overall, scatter plot matrix offers a comprehensive display for a multivariate model with multidimensional exploration capabilities and a high quality interaction technique.

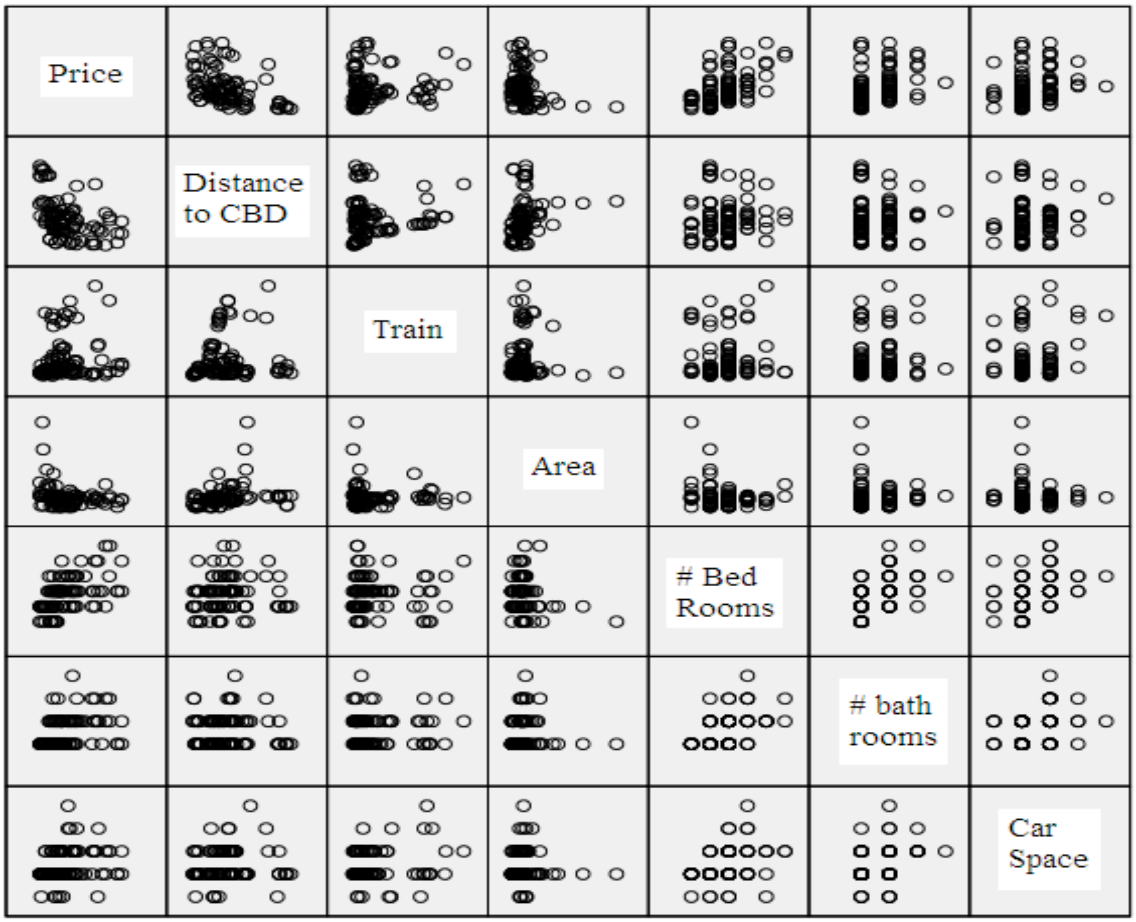


Figure 4 Scatter plot matrix.

2.1.2.2.4 PARALLEL COORDINATE PLOTS

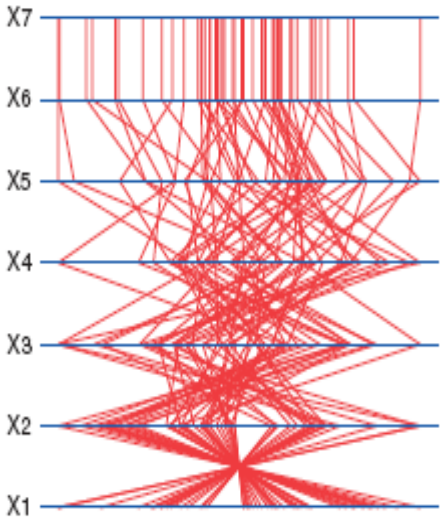
One of the most important and successful techniques to deal with multidimensional visualization is parallel coordinates plots (PCP). It displays all variables in a single plot and presents multidimensional distributions in a two-dimensional plot (Daradkeh, Churcher & McKinnon 2008). PCP are employed widely in various applications involving social networks, geographic information systems (GIS), intrusion detection and many others (Moustafa 2011).

The way PCP is constructed is by drawing a polyline for each case, the polyline goes through parallel axes, where each parallel axis represents a

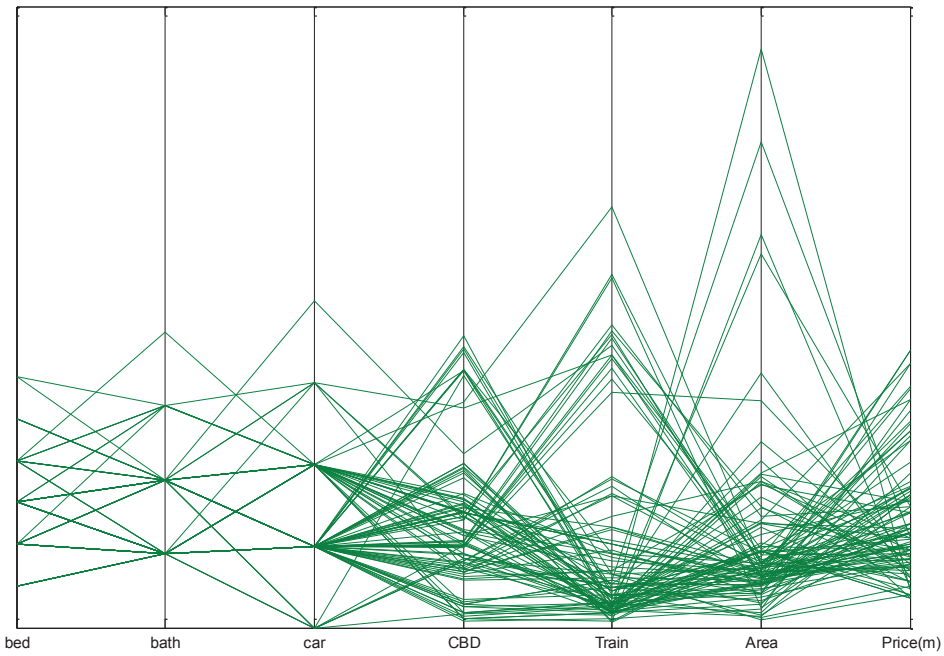
single input variable (Daradkeh, Churcher & McKinnon 2008; Moustafa 2011; Therón & De Paz 2006).

PCP has many advantages over the previously described visualization techniques, these include and not limited to its ability to providing interactive visualization (Daradkeh, Churcher & McKinnon 2008; Moustafa 2011), to exploring complex patterns between variables such as the positively correlated variables as well as the negatively correlated variables. Furthermore, PCP has become one of the most used visual data mining tools (Moustafa 2011).

Figure 5a demonstrates how PCP can clearly show the correlation between two variables; the higher line crossing between two axes means the higher negative correlation between the two variables represented by these two axes and vice versa. Accordingly, x1 and x2 are perfectly negatively correlated while x6 and x7 perfectly positively correlated (Moustafa 2011). The same applies to the PCP in figure 5b, it shows the negative correlation between the property price and the area of the property.



(A)



(B)

Figure 5: A Parallel coordinate plots

PCP has some major challenges; these include heavy overplotting, even for few thousands of data points because PCP conform to the one-to-one mapping from the original Cartesian space to the visualization space (Moustafa 2011). Other challenges are: high dimensionality which can be overcame by using a technique to reduce dimensionality, and axes permutation and rotation is another critical challenge for PCP because different axes order results in different images (Moustafa 2011).

PCP can be represented in different forms, each form can be used to overcome one of the challenges that PCP is facing.

Interactive PCP, Three dimensional PCP, Scalable PCP, Density Estimated PCP and Enveloped PCP (Moustafa 2011).

Interactive PCP: helps in identifying all possible relationships between the variables.

Three dimensional PCP: helps with revealing the hidden relationships that cannot be shown using Generalized PCP.

Scalable PCP: to increase the visual scalability of the PCP. PCP is known by its low scalability due to the high number of data points visualized.

Density Estimated PCP: is used as a better way of exposing dense regions between the axes and to improve the visual clatter in the plot.

Enveloped PCP: to subdue the overplotting issue and help in investigating the convex relationships.

PCP is an effective interactive visualization technique to present moderate multidimensional models in two dimensional space using parallel axes. It has the capacity to expose complex patterns; however, it has some challenges which can be overcome by using the different forms of PCP mentioned above.

2.1.2.3 Statistical Sensitivity Analysis methods

Statistical sensitivity methods are the methods which engage running simulations in which input variables are assigned probability distributions and evaluating the effect of variance in input variables on the output distribution(Christopher Frey & Patil 2002). It is method dependent whether one input variable is varied or more at a time in statistical methods which allow the designation of the effect among multiple input variables(Christopher Frey & Patil 2002). The following section reviews in detail regression analysis, specifically multiple regression analysis which will be adopted in this research.

2.1.2.3.1 MULTIPLE REGRESSION

Multiple regression is used to describe the relationships between variables, to control the input variable for a specific value of the target variable and to predict the target variable (Christopher Frey & Patil 2002).

To guarantee successful regression analysis, a relationship between input and target variable need to be determined before the analysis depending on the understanding of the functional form of the simulated model or by using scatter plots. Regression analysis may not be helpful when extrapolating outside the range of values used for each input variable when simulating the model (Christopher Frey & Patil 2002).

The general multiple regression model uses the mean function below to fit a relationship between input variables and the target variable.

$$E(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad 3.1$$

Where:

$E(Y \mid X)$ is the estimated regression of Y on X;

X represents the terms included in the model;

B is the regression coefficient; and

ε is the statistical error.

For the above linear model, the regression coefficient βj, can be interpreted as the change in target variable Yi when the input variable Xj,i for a specific value of j increases by one unit and the values of all other input variables remain constants. Therefore, regression coefficients can be used as a form of One-At-A-Time sensitivity (Christopher Frey & Patil 2002).

A major advantage for regression analysis is allowing the assessment of individual sensitivity for the input variables while taking into consideration

the simultaneous effect of other input variables on the target variable (Christopher Frey & Patil 2002).

The potential disadvantages for regression analysis are: a prior assumption of the functional form for the relationship between the input and target variables should be available, possible ambiguity in interpreting the results and if potential key assumptions are not met, the results are likely to lack vigorousness (Christopher Frey & Patil 2002).
Regression analysis may generate statistically insignificant results and it may be attributed to the narrow range of variation of the input variable. Therefore, regression analysis results can be sensitive to the range of input variable variation in the data used to simulate the model and may not constantly unleash clearly a relationship that really exits (Christopher Frey & Patil 2002).

The significance of the regression can be evaluated by the F-test value provided by the analysis of variance (ANOVA). This value should be sufficiently larger than F(p,n-p), where p is the number of predictors and n is the number of observations. (Weisberg 1947).

The significance of individual terms in the regression model is evaluated by the t-test, which is equivalent to the F test in multiple linear regression, if the significance (p-value) of t-test for a specific term or predictor is less than 0.05 within 0.95 confidence interval, the it asserts that the coefficient estimation is reliable and statistically significant (Weisberg 1947).

2.2 Visualization Approaches

Visualization is utilized to display interactive information with spatial or graphical representations simultaneously on a single screen, thereby facilitating comparison, measurement, simulation or achievement.

The fundamental goal of visualization is to represent data using visual features such as colour and texture or a combination of both, using the fast recognition capability of human vision, thereby supporting efficient visual analysis (Hsiao 2010). Visualization is usually applied to analyse large datasets, thus it is always difficult to display the total amount of data in limited screen space.

Many approaches have been proposed to resolve this issue. Cockburn, Karlson and Bederson (2008) proposed three different methods to do so, these are:

- Overview + detail
- Zooming
- Focus + context

They categorized all overview + detail methods with spatial separation while categorized zooming with temporal separation. However, based on their work, Hsiao (2010) reorganized them as two categories with two subcategories:

- Overview + detail
    - Space multiplexing
    - Time multiplexing
- Focus + context

This report will review both approaches.

2.2.1 Overview + Detail Approach

Overview provide users with the overall outline or structure of the dataset, enabling users to obtain a clear idea of the whole size of the dataset and the relationships between items and easily recognize objects of interest (Cushing et al. 2006).

Only through overview displays, however, users cannot acquire particular information about individual items of interest, which detail views can provide. However, when navigating through detail views of large datasets, users may easily become lost (Cushing et al. 2006). Therefore, a combination of both types called overview + detail has been gaining in popularity in recent years.

According to Cushing et al. (2006), the overview + detail approach can be divided into two large categories. One is space-multiplexing; the other one is time-multiplexing. This report will review both of them later.

2.2.1.1 Space Multiplexing

The first subcategory of overview + detail approach is space multiplexing. Space multiplexing is used to show both views simultaneously in different parts of the screen (Cushing et al. 2006). This is known as the traditional and standard overview + detail displays.

2.2.1.1.1 STANDARD OVERVIEW + DETAIL VIEWS

This technique is one of the most popular visualization techniques used so far. It uses two images to visualize a dataset. The big one is a high-resolution detail views and the small one is a low-resolution image to display the overview (Hsiao 2010). There are many examples of these types of overview + detail displays, including image browsers, maps. A Google map, shown in figure 6, illustrates this technique.

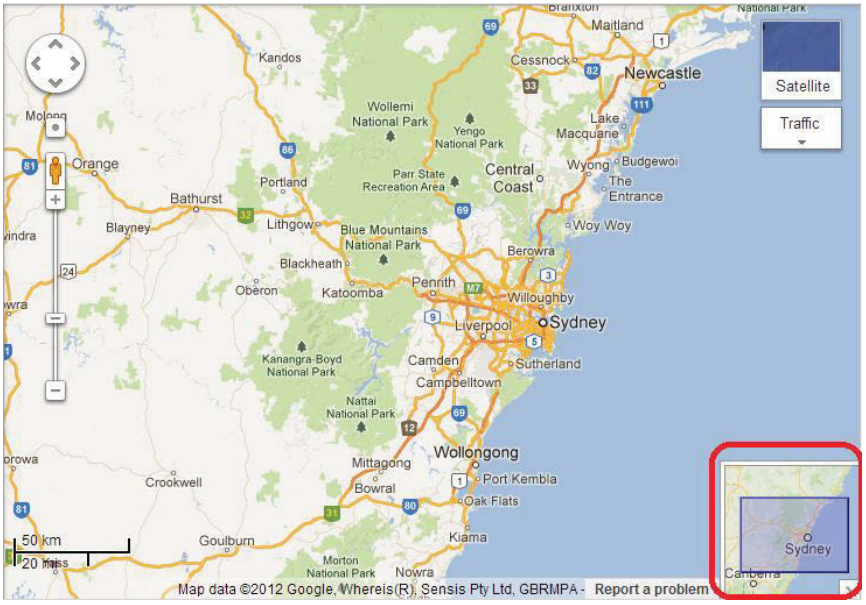


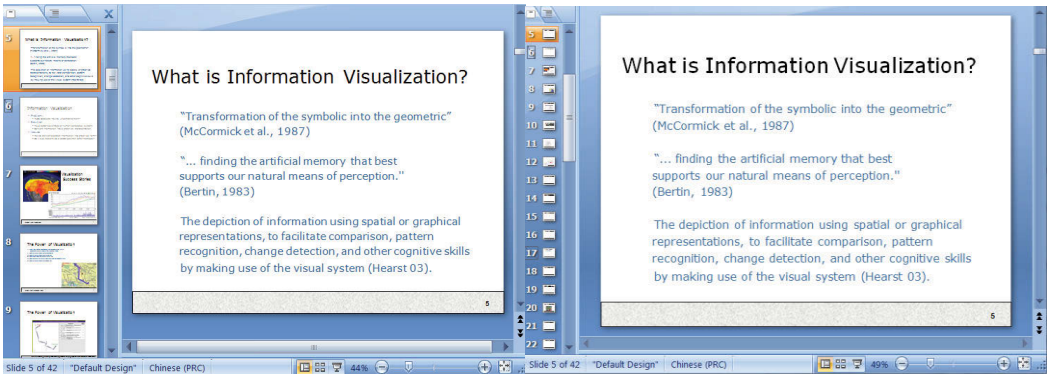
Figure 6: Google Maps

Google Maps (Figure 6) has a small inset overview window (red rectangle on the bottom right hand). It includes an interactive blue rectangular subregion that corresponds to the area displayed in the detailed view (Cockburn, Karlson & Bederson 2008).

The downside of providing both views on the same screen is more requirement of space; however, it reduces the frequency of navigational switching (Cushing et al. 2006).

2.2.1.1.2 SCROLLBARS AND THUMBNAIL OVERVIEWS

Scrollbars are a common component of user interfaces and the proportion of the visible document can be displayed by the length of the scrollbar (Cockburn, Karlson & Bederson 2008). The typical examples are Microsoft PowerPoint and Adobe Acrobat.



(a) Five thumbnails in the overview (b) 18 thumbnails in the overview

Figure 7: Microsoft PowerPoint's overview and detail interface

According to Figure 7, it is easy to see that PowerPoint address a trade-off in scale ratios: users can identify the content of thumbnails easily from low scale ratios (a) which, however, increased overview scrolling to access distant data. On the other hand, high scale ratios (b) provide users to access more distant data while at the cost of visual clarity (Cockburn, Karlson & Bederson 2008). As a result, many applications allow users to resize the overview region directly when needed.

Cockburn, Karlson & Bederson also mention another problem related to synchronization. Some applications only implement one-way synchronization. Take Microsoft PowerPoint as an example, when users scroll the detail view, the overview will synchronize correspondingly, while the detail is unsynchronized with the overview, which may cause disorientation.

2.2.1.1.3 THE TREE-MAP

The tree-map is a well-known space multiplexing system (Shneiderman 1992). Tree-maps display hierarchical data as a collection of nested rectangles. Each branch of the tree is set as a rectangle, and then recursively divided into smaller rectangles representing subbranches on a sublevel of the hierarchy, in

both the horizontal and vertical directions (Hsiao 2010). See the example below:

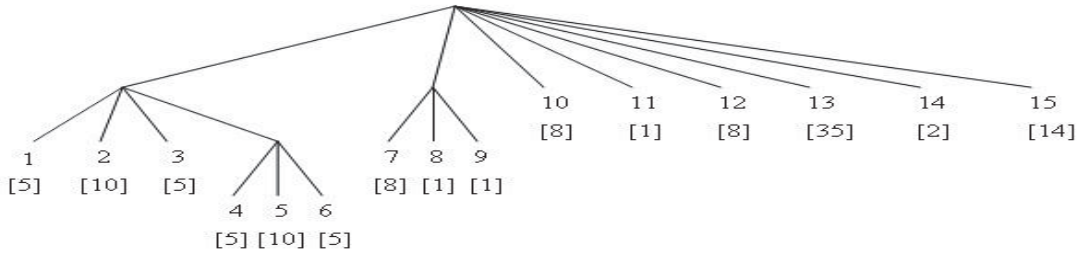


Figure 8: A Tree Diagram (Hsiao 2010)

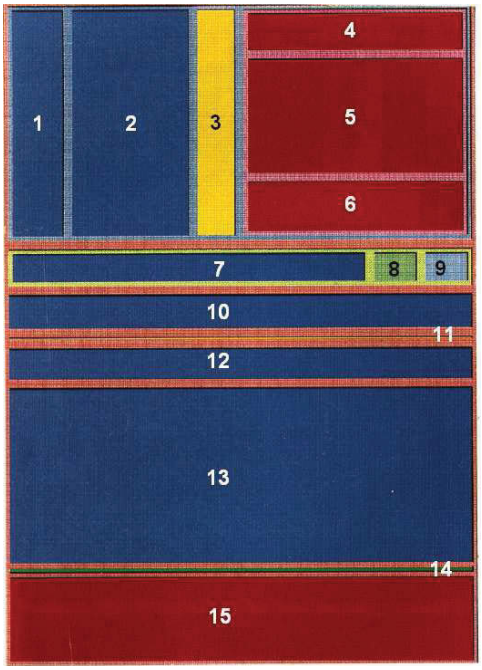


Figure 9: A Tree-map for the tree in Figure 8 (Hsiao 2010)

In Figure 8, numbers in brackets under nodes represent the nodes' weights. In Figure 9, the number in each area is the node number in tree in Figure 8.

A leaf node's rectangle has an area, which has the same proportion with the node's weight. Often the leaf nodes are coloured to show a separate dimension of the data.

2.2.1.2 Time Multiplexing

The other subcategory of overview + detail is time-multiplexing. Time multiplexing is showing either overview or detail displays at a time. The typical technique applied in this category is zooming.

2.2.1.2.1 ZOOMING

Time multiplexing techniques allow users to zoom in or out of visualization. Additionally, provides mechanisms to switch easily between various views therefore minimizing loss of context (Hsiao 2010). A typical case is Google Earth:

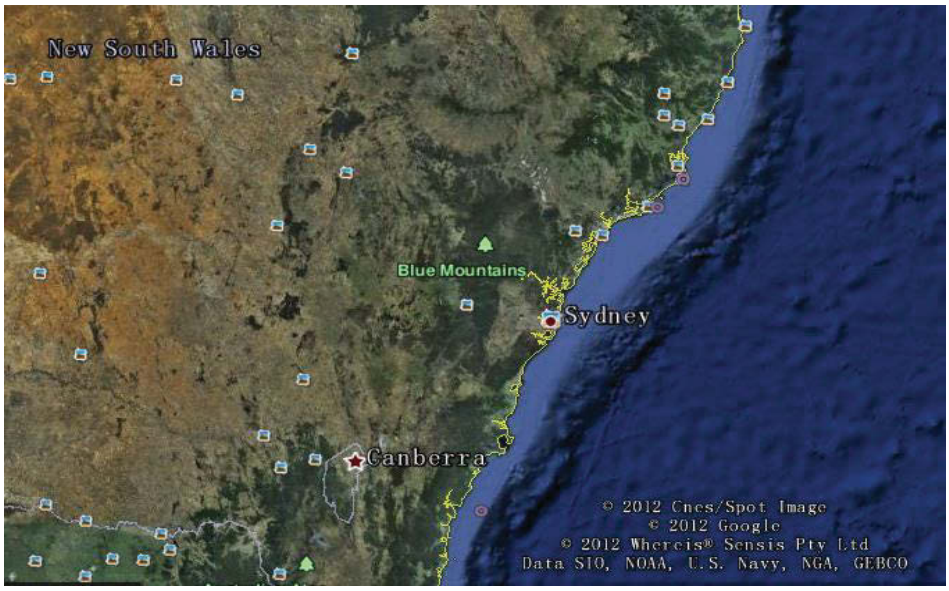


Figure 10: Map of New South Wales

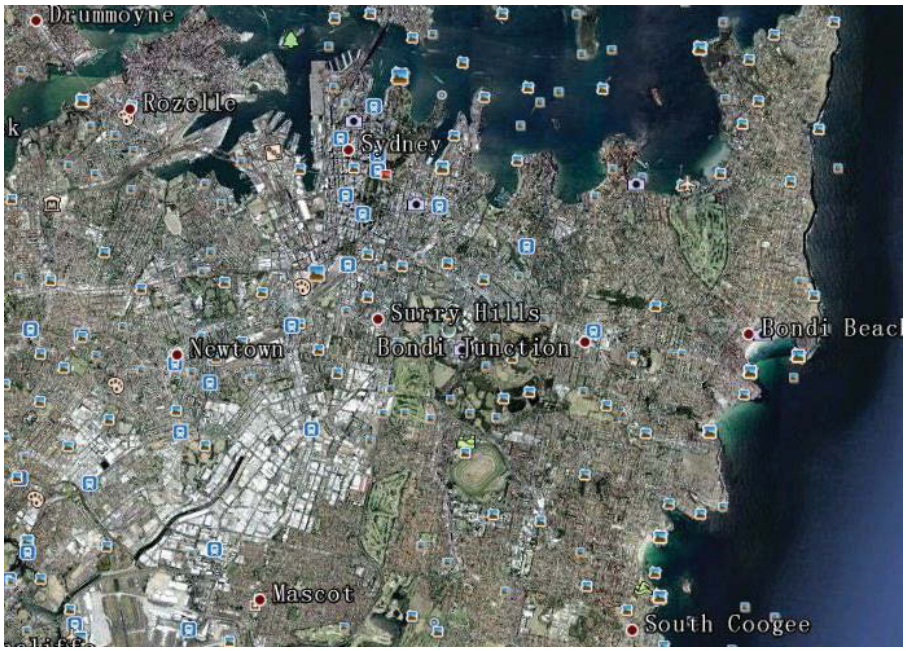


Figure 11: Map of Sydney with remote suburbs

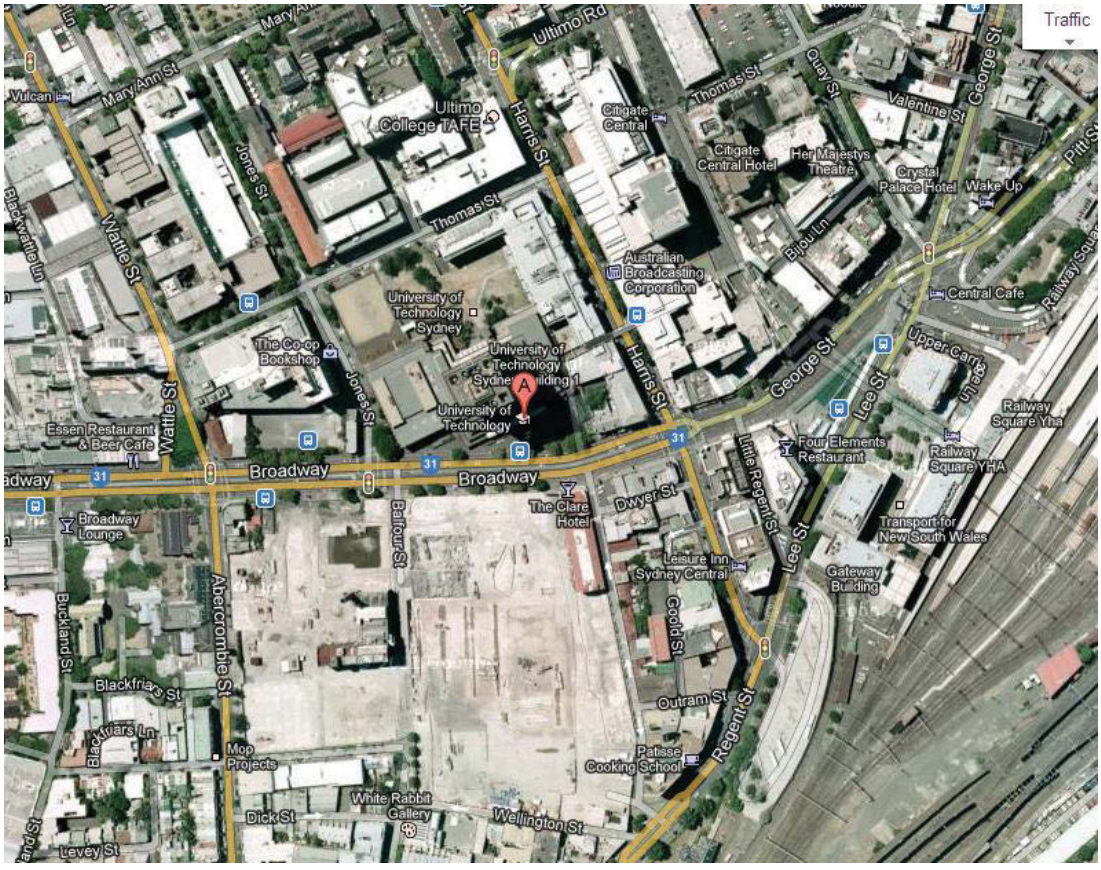


Figure 12: Map of the city of Sydney

Figures 10 to 12 show Sydney with several views in Google Earth. In Figure 10, the viewer is far from the target. At this distance, the relative locations of neighbouring cities such as Canberra are displayed. In Figure 11, the viewer is closer to the target. Detailed information such as the names of some suburbs and some railway stations are displayed in this view. Figure 12 is the most detailed view, which only displays part of the area of Sydney city, providing street and railway station names, building and park images and all other facilities' names.

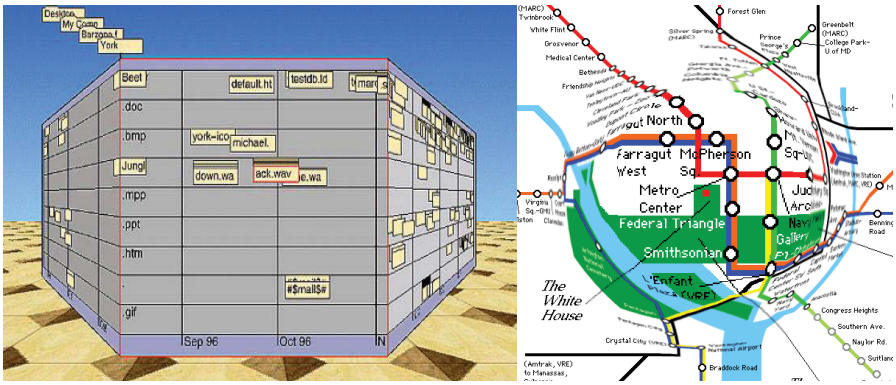
Although a closer view produces a more detailed view and it saves screen space, while global information is lost. Furthermore, it requires additional time for users to reorient themselves through switching focus between views with different degrees of detail (Hsiao 2010).

All the previous methods have separated two views to manage focused and contextual information in either space or time.

2.2.2 Focus + Context Approach

Focus + Context, which makes the whole dataset visible to users (Cockburn, Karlson & Bederson 2008): seamlessly combines a detail view with the distorting part of the overview (Hsiao 2010).

Differential scale functions across the information surface are used in most of focus + context interfaces, resulting in intentional distortion (Cockburn, Karlson & Bederson 2008) such as that shown in Figure 13.



(a) The Perspective Wall (b) Fisheye View

Figure 13: Two typical techniques of focus + context approach

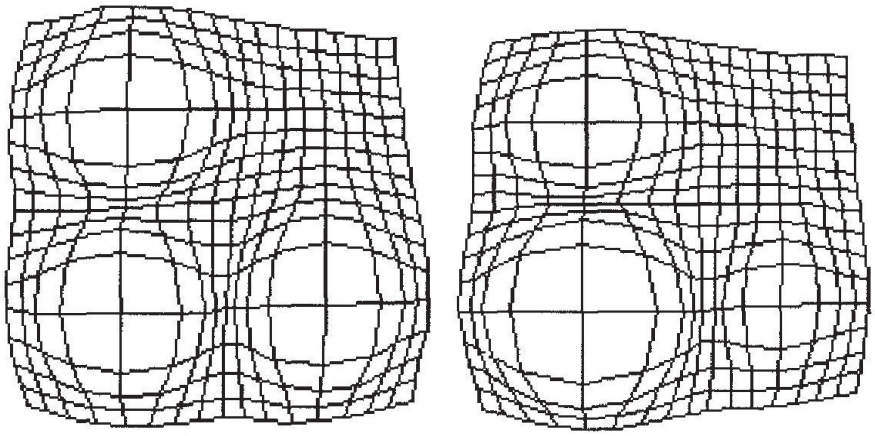
Consequently, the term 'distortion-oriented view' (Leung & Apperley 1994) is also commonly applied in this category.

2.2.2.1 Fisheye View

A well-known focus + context technique is the fisheye view. Furnas (1986) established the theoretical foundations for focus + context interfaces by describing a generalized fisheye view formula to calculate the user's degree of Interest (DOI), which will not be discussed in detail in this report.

As Figure 13 (b) shows, the overview (context) and the detail view (focus) are displayed on the screen simultaneously. The overview presents a view of the entire dataset, while the detail view simulates the effect of magnifier lens to display the subregion chosen by a user (Hsiao 2010).

The advantage of using this technique is that, through maintaining the spatial orientation of the overview, smooth transitions are provided between overview and detail. Furthermore, multiple lenses are possible (Figure 14), and the viewer is able to manipulate all of them easily (Hsiao 2010).



(a) (b)

Figure 14: A multiple-focus view of the projection with the same parameters for each focus point (a) and with various parameters (b) (Leung & Apperley 1994)

A summary of the advantages and disadvantages of the above mentioned visualization approaches are listed in table 1

. Approach	Advantages	Disadvantages
Standard Overview + Detail Views	▪ Provides overview of the whole dataset and the detailed information ▪ Reduces the frequency of navigational switching.	▪ Requires more screen space ▪ Not suitable for a huge dataset
Scrollbars and Thumbnail Overviews	▪ Users can choose a different part of the whole dataset in the overview and the size of the data they want to view. ▪ Reduces the frequency of navigational switching.	▪ Requires more screen space ▪ Difficult to manage the trade-off between the clarity and the size of overview data ▪ Synchronization might be a problem ▪ Not suitable for map data
Zooming	▪ Produces a more detailed view ▪ Users can decide the degrees of detail information they want ▪ Saves screen space	▪ Requires additional time to switch focus between views with various degrees of detail. ▪ Easily become lost when in the detail view
Fisheye View	▪ Provides smooth transitions between overview and detail	▪ Images are distorted (not suitable for map data)

	▪ MORE FLEXIBLE TO SWITCH FOCUS ▪ MULTIPLE LENSES ARE POSSIBLE	

Table 1: Comparison of visualization techniques

After the comparison shown above, I decided to abandon the fisheye view approach. The system at hand is a geographical map, which requires accurate distance scale ratios. Users should be able to distinguish directly the distance between the property and the railway station. Fisheye view will distort the image, causing viewer's visual error and influencing the data accuracy, so it is not suitable for visualizing map data.

The second eliminated technique is zooming. Zooming provides users with the function of zooming in and out freely. However, when zooming into the detail view, users may easily get lost in the whole map. Therefore, they may only focus on a small part, instead of the whole map.

The reason why we did not choose scrollbars and thumbnail overviews is that it is not suitable for visualizing map data either. A geographic map is not like slideshows, which are separated, and has thumbnails. A geographic map has its own orientations and all regions/areas interconnect.

Therefore, in my property prediction system, standard overview + detail technique is applied to the interface. When users select a particular region from the overview map, the selected region will be zoomed in as detail view while the overview map will be zoomed out with highlighting the selected region. This technique provides users with easy access to particular properties

through the detailed region, while the overview map guarantees users' orientation.

2.3 SUMMARY

As section 2.1 shows, different sensitivity analysis methods have diverse usages and advantages and they all share a common draw back which is "user interaction absence", in other words, sensitivity analysis methods prevent the user from interacting with it and expressing his/her preferences. Additionally, applying sensitivity analysis and interpreting the results are restricted to an expert user, in other words, if a non-expert user used any software to apply sensitivity analysis, interpreting the results will still be an obstruction.

Also section 2.2 shows different visualisation techniques and states the advantages and clearly shows that one of those is user interaction ability and displaying complicated data in more simplified format. Thus, in this research, a visualisation technique is combined with a sensitivity analysis method in order for the visualisation to make up for sensitivity analysis user interaction absence and to hide the complexity of the underlying data and the sensitivity analysis method used, so a non-expert user will have the ability to employ sensitivity analysis and comprehend the results without prior knowledge.

3 Research Methodology and System Description

An empirical methodology is applied to achieve this study.
Conducting this study entails two parts; part 1: developing a prototype for a system that combines interactive visualization and sensitivity analysis, and part 2: utilizing multiple regression to develop a prediction model.

3.1 Scope

The interactive visualisation will include an abstract map for some regions of Sydney. The abstract map will involve the following geographical regions: city of Sydney, Inner west, eastern suburbs, northern beaches, Liverpool and Fairfield, Bankstown and district, northern metropolitan, Parramatta and district and St George/Sutherlands.

All housing data is collected mainly from Australian Property Monitors (APM) web site, In particular, saved appraisal section, and some is collected from Domain web site, sold section.

The investigated predictors are: number of bedrooms, number of bathrooms, car space, internal area, land size, distance to Sydney CBD, distance to nearest train station or bus stop and property price. Other property features are displayed including property type (house, unit, studio, … , etc), suburb and some extras like in-house pool, spa, and gym. This research will not tackle features attribute such as bedroom area or kitchen installed appliances. Predictors' interaction is beyond the scope of this study.

3.2 PART 1: PREDICTION SYSTEM DEVELOPMENT.

This study is applied to a real estate prediction system, the prototype development has the following phases; phase 1: Creating the first visualization, phase 2: Adding user interaction feature to the overall visualization, phase 3: Data Collection, phase 4: Embedding the outcome of part 2 "the regression model" into the prototype, and phase 5: Visual outcome presentation, second visualisation.

The system framework is illustrated in Figure 15.

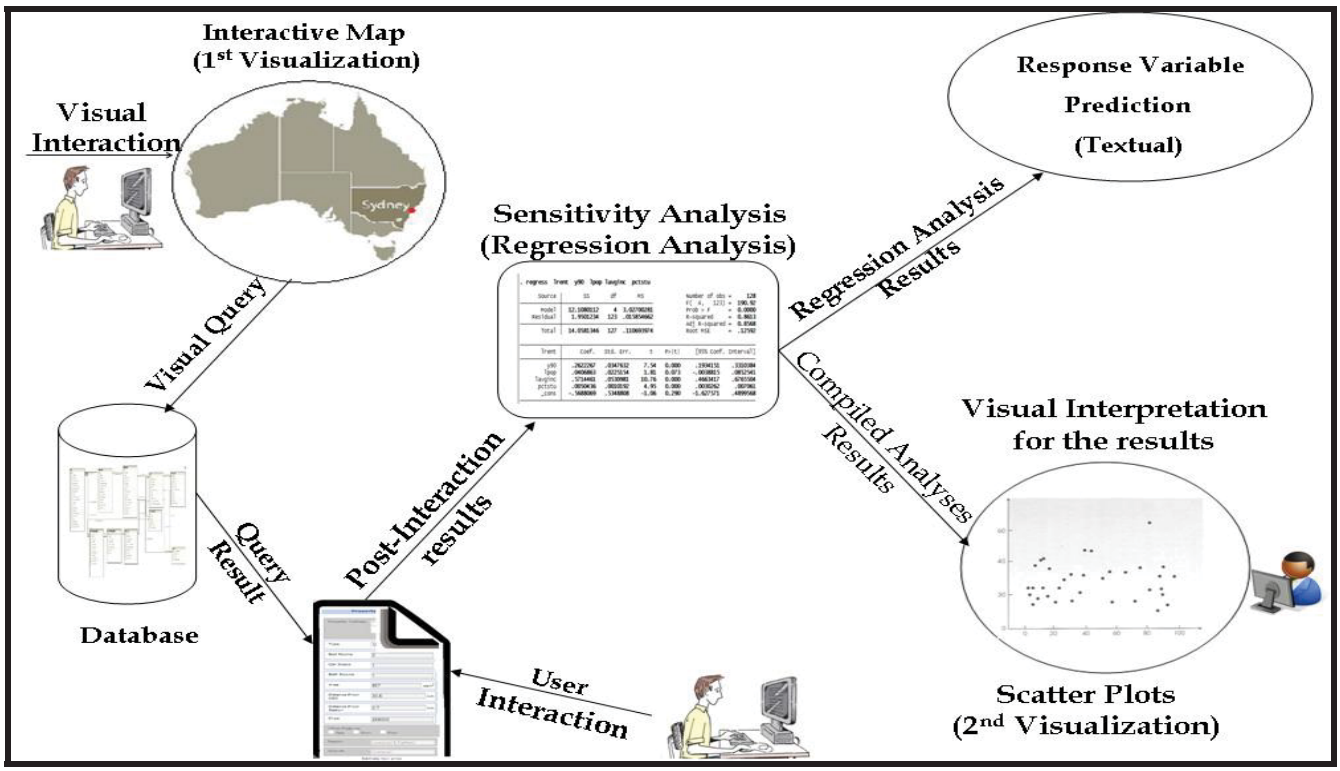


Figure 15: Proposed property predication system.

Phase 1- Graphical Map Design (IMAP) - First Visualization: this study follows Ben Shneiderman's information visualization mantra: "Overview, zoom and filter, details on demand" in the process of creating the visualization.

We create the first visualization to present the user with an overview for an abstract geographical map of some Sydney regions, mentioned in section 3.1 above. Figure 16 shows the abstract map for the selected regions.

Phase 2- Interaction Design - Any map usually has a huge amount of data, therefore, to clearly present this data without losing focus of both the overall view of the map and the detailed part that has the user interest, we apply Overview + Detail approach. This approach which is called spatial separation too, is characterized by presenting both the overview area (the structure of the dataset) and the detailed area with a separate independent space for each one, this will allow the user to interact with each area separately, i.e. the overview area presents the whole picture, when detailed information is required, navigation to the detail area should reveal this information.

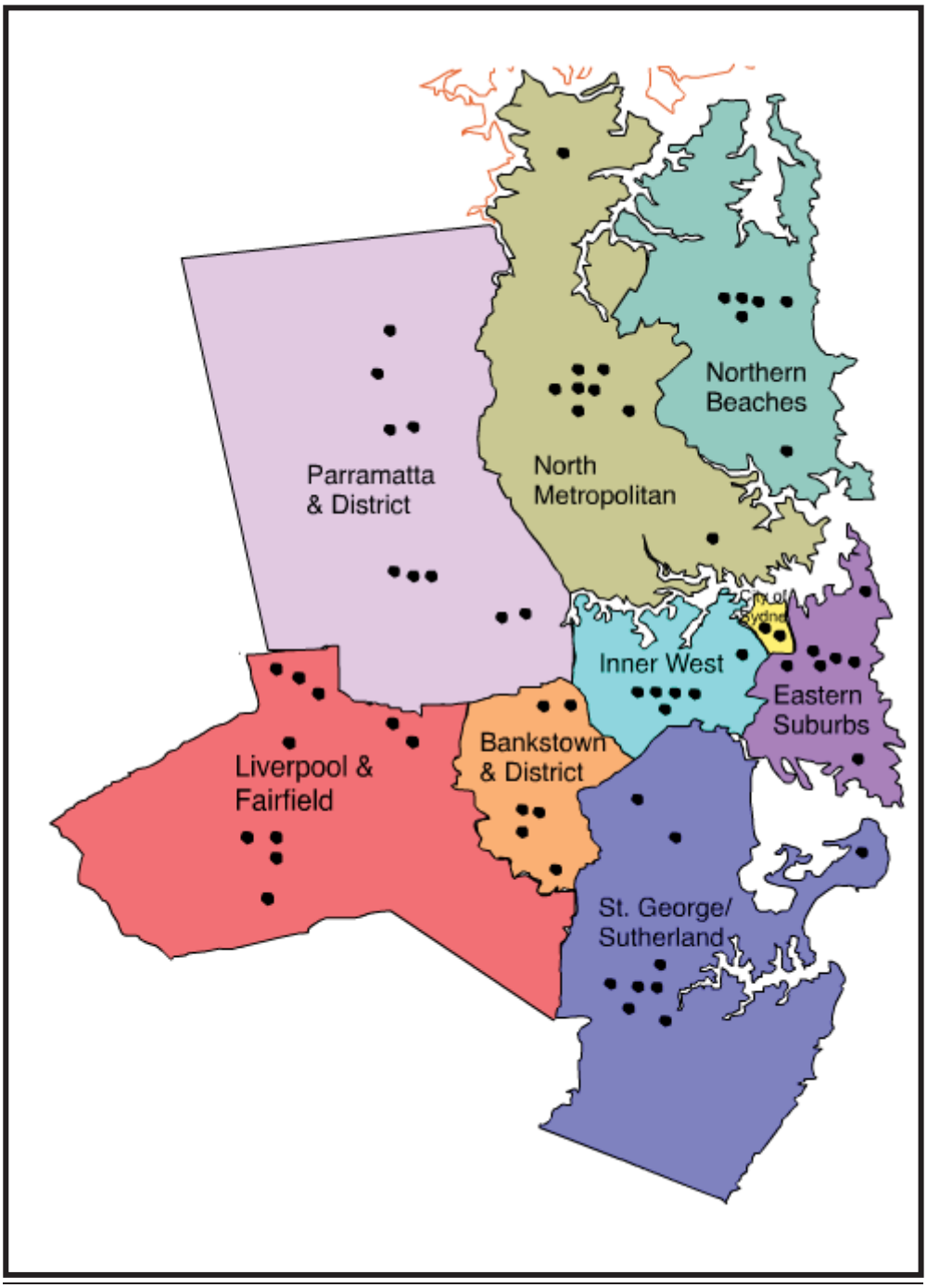


Figure 16: The overview of IMAP

We apply overview + detail to the prototype, so the user can obtain the required information easily and can setup his/her query visually in no time. The user needs to zoom in his/her region of interest by clicking

the region itself, and then he/she needs to locate a property on the selected region so he/she can obtain the details for that property. As a result to the visual query, the user is presented with the property detailed features, as shown in figure 17, such as number of bed rooms, distance to Sydney CBD, distance to transport station and price among other features.

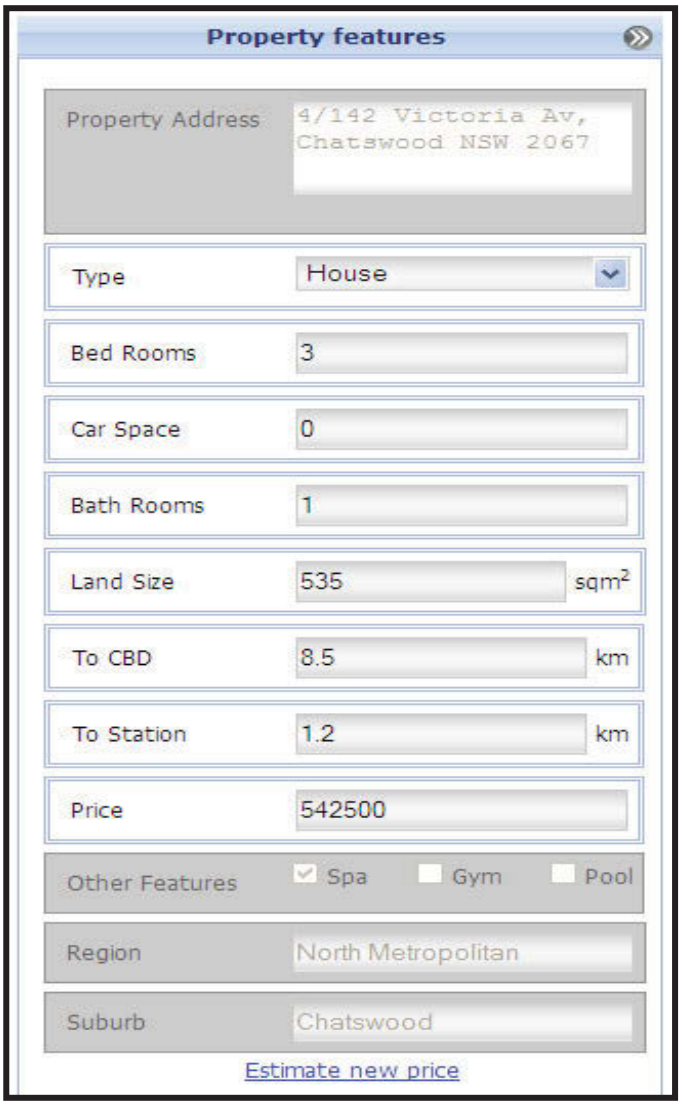


Figure 17: Visual query outcome, detailed property features.

Figure 18 demonstrates clearly the application of overview+detail approach. Once the user selects a specific region by clicking it from the

overview area, the overview area will zoom out, and the detail area will zoom in, the clicked and/or the selected region, in this example the detail area is Bankstown & district.

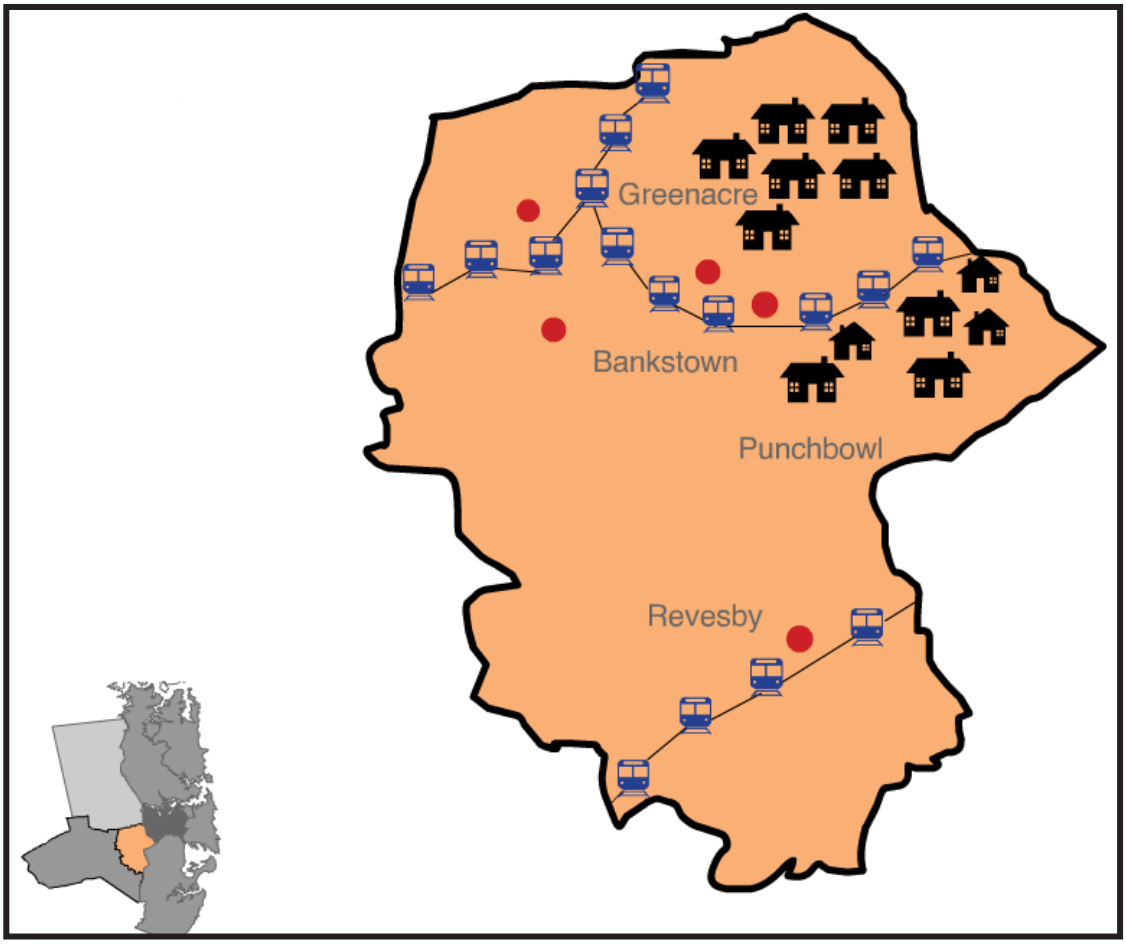


Figure 18: IMAP revealing filtered information in the detailed area. The overview area is zoomed out to the bottom left.

Figure 19 shows another example of the application of overview+detail approach, the zoomed in area is North Metropolitan.

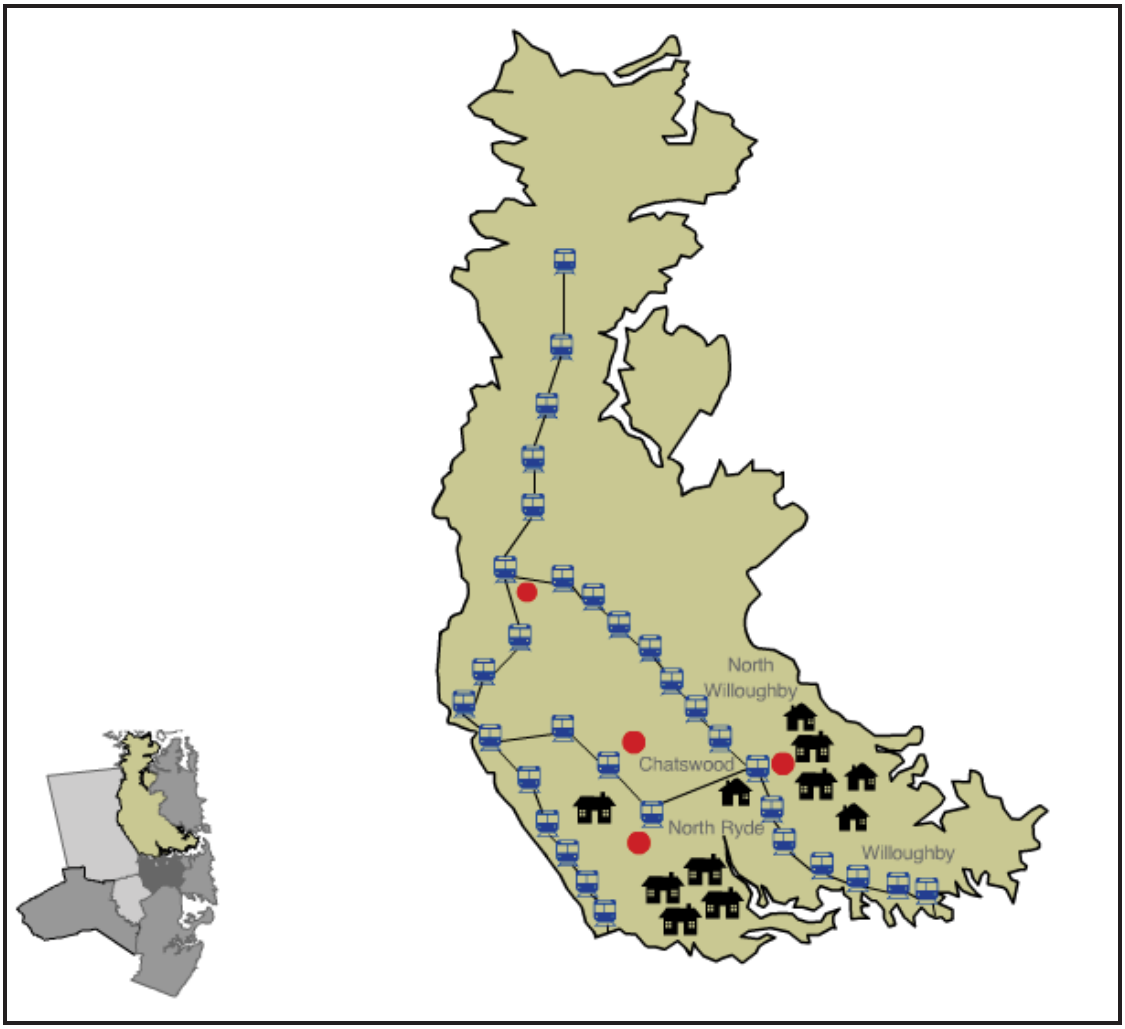


Figure 19 : IMAP revealing filtered information in the detailed area. The overview area is zoomed out to the bottom left.

Additionally, the detail area is filtered so it highlights only the important information which directly impacts the housing price and is most interested by house buyers. It does not distract the user with any unneeded information. Both figures 18 and 19 show the filtration clearly as it displays only the market-listed properties. Furthermore, both figures show important data such as the train stations and some landmarks such as major shopping centres and other tourism attractions.

Phase 3: Data collection. The housing information is collected from Australian property Monitors (APM) and Domain web site. In the process of collecting housing data, the property features that are tackled by this research are property type, number of bedrooms, car space, internal area, land size, distance to Sydney CBD, distance to train station, suburb, and the property price. The property prices presented in this study are the average values for the prediction interval presented by domain. The internal area is approximate as the floor plan shows for some properties. For the properties that does not have floor plan, it was assumed to be the same as other properties of the same features.

Phase 4: Embedding the regression model.

This study adopts multiple regression analysis to run sensitivity analysis. Regression analysis is useful in describing the relationships between input variables and the target variable, will control input variables for specific values of a target variable, and is useful in predicting a target variable based on input variables. The regression model is the outcome of part 2 of this study described in the section 3.2.

Phase 5: Visual Outcome presentation.

Apart from displaying the predicted price instantly in textual format for the unobserved (customized) property. In this the study, all individual analyses done are logged into a database table to compile all the results together to produce scatter plots visualization for each predictor along with the response separately. These scatter plots will visually present the results so that the user can easily compare between the response observed values and the predicted values, this visualization will expose

the differences between the response observed values and the predicted values.

To briefly describe the graph; a single predictor will be depicted on the x-axis while the response will be depicted on the y-axis. A single property will have two points representing it, the first point represents the response observed value and the other point represents the response predicted value, the two points are distinguished by two different colours, which makes it easier to compare and investigate large number of analyses' results in a two dimensional visual space. Figure 20 demonstrates this graph.

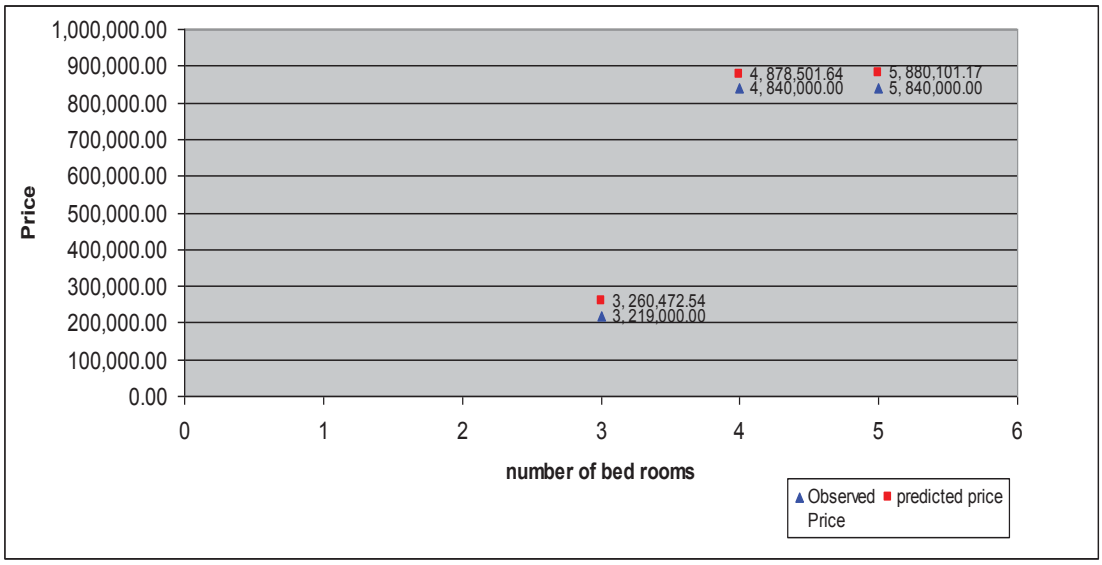


Figure 20: Compiled results visualization. The predicted price is for the property with one unit increase in the predictor "number of bed rooms".

These traditional non-interactive scatter plots are accessible for all users and are easy to interpret by a non-expert user if he/she has a basic general knowledge of scatter plots; however they are directed to expert users for further analysis.

3.3 PART 2: MODEL DEVELOPMENT

This part employs multiple regression analysis to achieve the objective of predicting the response when changing a value for one of the predictors at a time while other predictors held fixed. The general form for the mean function for multiple linear regression, shown in (1), is used to guide the construction of our prediction model.

$$E(Y \mid X) = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad (1)$$

Initially, we chose the following predictors to construct the price prediction model for the real estate prototype: *Number of bed rooms (BedR), car space (CarS), number of bath rooms (BathR), property area (InternalArea), land size (LandSize), distance to Sydney CBD (CBD) and distance to nearest train station and/or bus stop (Dtrain).*

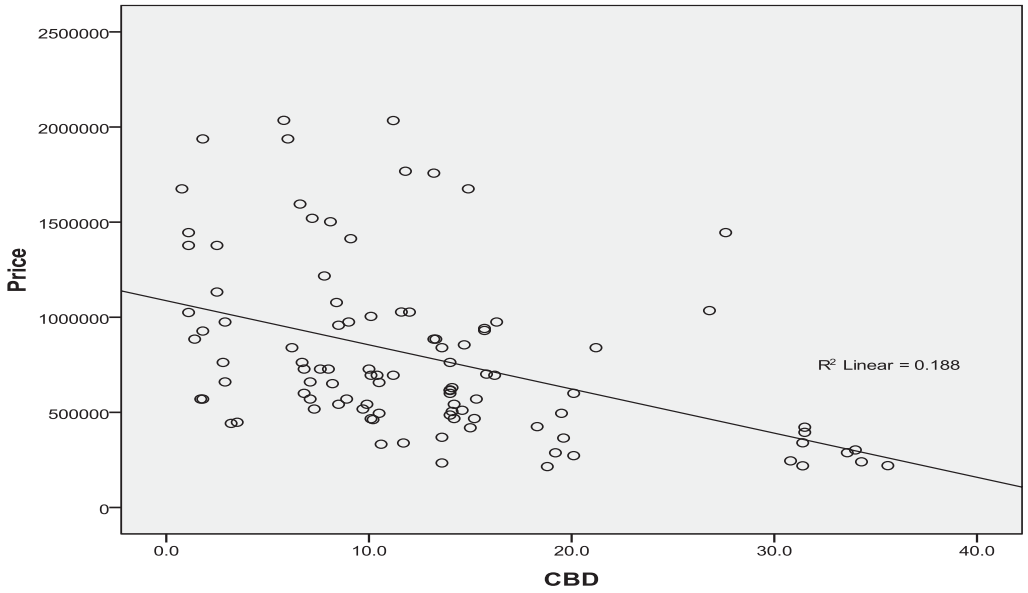
Model construction underwent the following steps:

Step 1: Pre-processing, Variable Transformation. The linear regression model expects a linear relation to achieve best results. Sometimes the predictors involved in the model do not satisfy the linearity feature implied. To overcome this issue, variable transformation is frequently used to obtain a mean function that is linear in the transformed scale.

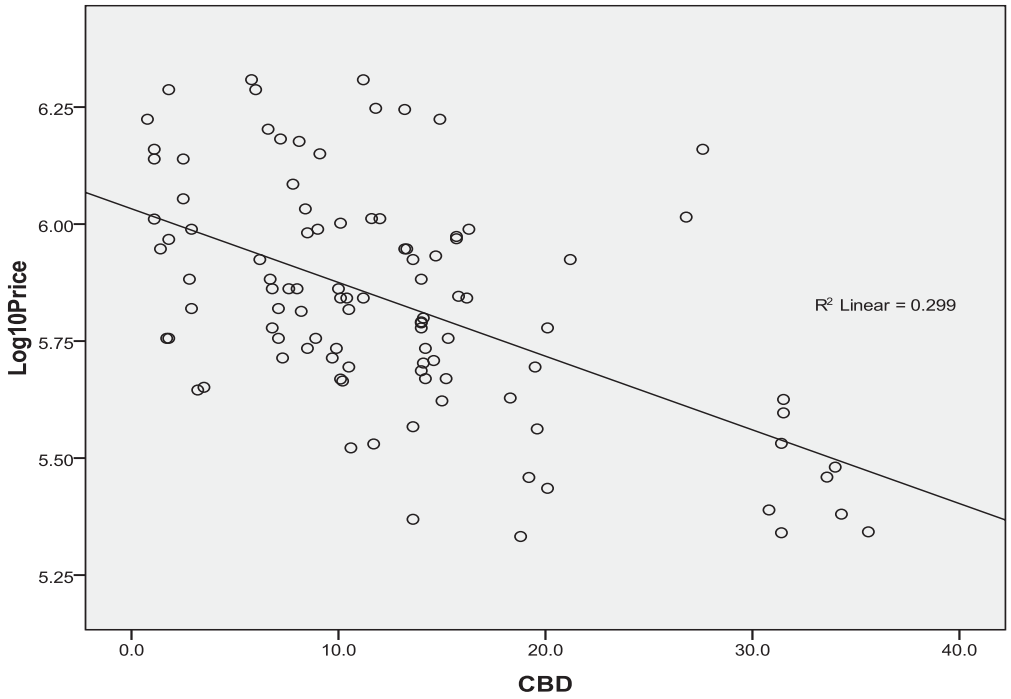
This step examines model linearity prior to conducting the analysis. An appropriate tool to examine linearity is scatter plots. For example, the scatter plot in figure 21-a shows the property price in dollars (plotted on the vertical axis) and distance to Sydney CBD (CBD) in Km (plotted on the horizontal axis) using the data collected in part 1 phase 3. This graph shows how the points are scattered from the line fit which indicate weak linearity. Linear $R^2 = 0.108$ confirms the relatively weak linearity that

needs to be improved. This graph also shows the moderate negative correlation between the two plotted variables and the correlation calculations of -0.434 reasserts that, which means there is a moderate association between them and the relation between is worth further investigation.

Common logarithm transformation for the response showed it is the best to enhance linearity between the two variables (price and CBD) among other transformations including power transformation and reciprocal transformation. The scatter plot in figure 21-b demonstrates how the points are closer to the line fit which indicates linearity boost with linear $R^2 = 0.299$. Additionally, correlation calculations of -0.547 confirms a better association between $\log_{10}$(price) and CBD than the one between price and CBD.



(a)

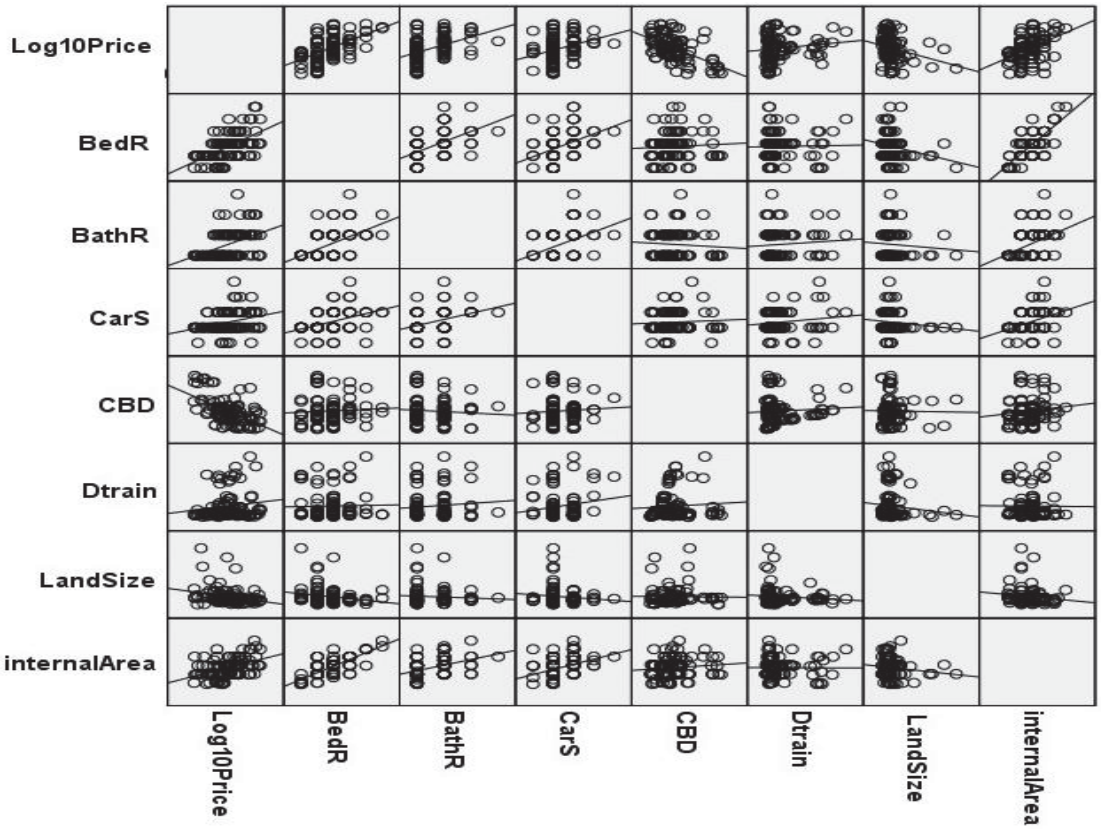


(b)

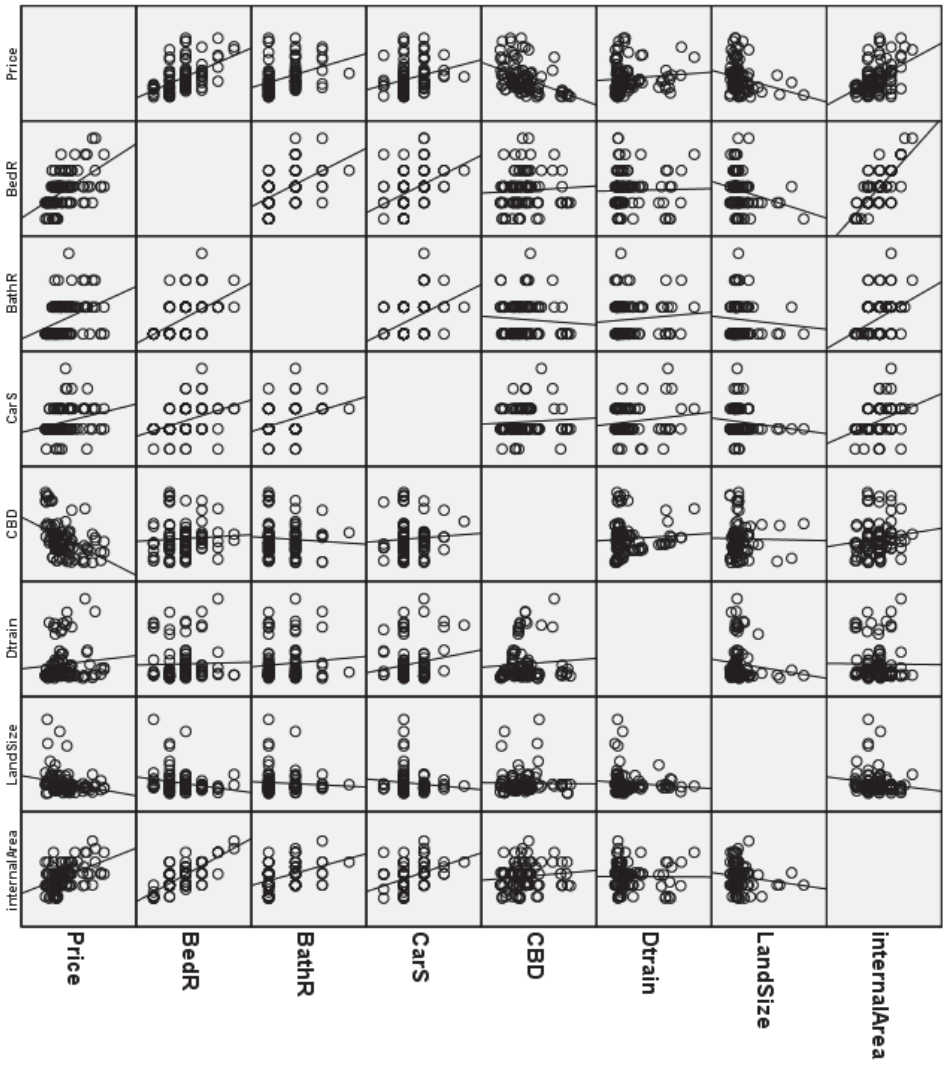
Figure 21 : a) Scatter plot for the response variable and CBD

b) Scatter plot for the transformed response variable log10(price) and CBD.

I repeated the same comparisons for all other predictors against the price in its untransformed form and against the transformed $\log_{10}$(price), and it showed that the model will produce better results with the transformed form of the response because, in general, the transformation had enhanced the linearity . The scatter plot matrix in figure 21-a shows the scatter plot for the response $\log_{10}$(price) and all independent predictors individually, while figure 21-b shows the scatter plot for the response price and all independent predictors individually, when comparing the first top rows of both matrices, we find out that the points in the scatter plots in figure 21-a are closer to the fit line than the points in figure 21-b which indicates that the response transformation has enhanced the linearity between the response and the independent predictor.


Log10Price
BedR
BathR
CarS
CBD
Dtrain
LandSize
internalArea

(a)



(b)

Figure 22 : a) Scatter Plot matrix for the transformed response and all other predictors. B) Scatter plot matrix for the response in its original form and all other predictors.

Step2: Regression Coefficient Estimation

To fit the values of the target variable, I need to estimate regression coefficients (β). To obtain those estimates, I use the ordinary least squares method (OLS) in which the parameter estimates are chosen to minimize the residual sum of squares function.

I use the statistical package SPSS to perform the regression analysis since it uses the OLS method. I run the analysis assuming

$$NH: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

$$AH: \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5 \neq \beta_6 \neq \beta_7 \neq 0$$

After running regression analysis, I obtain the following results

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.843[a]	.710	.688	.13545

a. Predictors: (Constant), internalArea, Dtrain, CBD, LandSize, CarS, BathR, BedR

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.140	7	.591	32.238	.000[a]
	Residual	1.688	92	.018		
	Total	5.828	99			

a. Predictors: (Constant), internalArea, Dtrain, CBD, LandSize, CarS, BathR, BedR

b. Dependent Variable: Log10Price

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.659	.052		108.705	.000
	BedR	.086	.022	.389	3.976	.000
	BathR	.028	.025	.077	1.133	.260
	CarS	.023	.022	.065	1.019	.311
	CBD	-.017	.002	-.597	-10.429	.000
	Dtrain	.013	.005	.146	2.546	.013
	LandSize	-3.347E-5	.000	-.139	-2.403	.018
	internalArea	.001	.000	.132	1.447	.151

a. Dependent Variable: Log10Price

To construct the price prediction model, I use the coefficients table, and the model becomes as follows:

$$\text{Log}_{10}(\text{Price}) = 5.659 + (0.086 * \text{BedR}) + (0.028 * \text{BathR}) + (0.023 * \text{CarS}) + (-.017 * \text{CBD}) + (0.013 * \text{Dtrain}) + (-3.347\text{E-}5 * \text{LandSize}) + (0.001 * \text{Internalarea}) \quad (2)$$

Step3- Predication

To predict a property price, I employ the outcome model (2) from previous step to predict a property price, so after substituting the property details, I obtain the value of $\text{Log}_{10}(\text{Price})$, and by exponentiating I obtain a prediction point.

To obtain a %95 ($\alpha = 0.05$) confidence interval for the predicted point, I apply the following before exponentiating to find the confidence interval limits.

$$\text{Confidence Interval limits} = \text{Log}_{10}(\text{Price}) \pm 0.05$$

4 VISUAL SENSITIVITY DATA ANALYSIS.

The presented model in (2) in chapter 3 to obtain a predicted point is analysed to demonstrate the capability of the interactive visual system for sensitivity analysis.

4.1 MODEL EXPLAINED VARIABILITY.

The coefficient of determination, R^2, provides evidence about the goodness of fit of a model. A value of 1 indicates that the regression line perfectly fits the data. In other words, R^2 indicates the overall explained variability of the model. The model summary table shows that R^2 equals 0.710 which means that the model explains 0.710 of the overall variability of the model. And just 0.29 of the variability is not explained by the model. A measure of 0.710 for the coefficient of determination indicates a good fit of the model.

4.2 OVERALL MODEL SIGNIFICANCE.

The value for the F distribution associated with the regression assesses the overall performance of the regression analysis. From the ANOVA table, we obtain the F value equals to 32.238. This value should be greater than the critical value for the F distribution F(df1,df2), where df1 is the degrees of freedom in the numerator and df2 is the degrees of freedom in the denominator. For the data underlying this regression analysis, the critical value is obtained for F(6,93) with ($\alpha = 0.05$) from the F distribution tables equals to 2.19.

The F value (32.238) is sufficiently greater than the critical value obtained (2.19) therefore we reject the null hypothesis which assumes that all means are equal and do not contribute to model variability.

The F test assures that the overall model is statistically significant, but this does not imply the significance of each predictor and it should be examined separately.

4.3 Individual Coefficient Significance

The value of the t-test associated with each regression coefficient determines the significance of the coefficient and the p-value. For a coefficient to be statistically significant, the absolute value of the t-test should greater than 2 or the p-value should less than the significance level ($\alpha = 0.05$) .

The coefficients table shows the t-test value and p-value (Sig. column) for each coefficient individually. It shows that the predictor BedR is statistically significant since the associated t-test absolute value is 3.976 (greater than 2) and p-value is 0 (less than 0.05). The same applies to the predictor CBD with t-test absolute value 10.429 and p-value 0. These two predictors along with constant (t- test = 108.705 and p-value = 0) are the most significance predictors of the model since the p-value is almost 0.

Other statistically significant predictors are Dtrain, LandSize but they are not influential as BedR and CBD because of the p-value, 0.013 and 0.018 for Dtrain and Landszie respectively.

The analysis shows that Dtrain predictor, which represents the distance to nearest transport, is positively correlated to the value of the response.

But this generally contradicts with the results when we apply the regression model to predict a property price. This is because of one of the regions included in the study, Northern beaches, has an expensive property market, and all properties in that region is very far from transport means when compared to the other properties in other regions. Therefore, the properties in this region which have high observed prices with far distances from transport means has an impact on the results and showed a positive correlation between Dtrain and the property price.

The analysis shows that Landsize is a significant predictor negatively correlated with the response value, but applying the prediction model while examining Landsize produced about half of the cases agrees with the result while the other half disagrees with this result. The regression model has been affected by the fact that the observed land size is related to the property type, house and unit mainly, house type properties has smaller land sizes and higher observed prices and units has larger land sizes a lower observed prices.

BathR, CarS, and Internalarea are the model's insignificant predictors therefore they do not contribute to the response variability, since the t-test absolute value for these three predictors is less than 2 and the p-value is greater than 0.05.

To summarise, the overall model explained variability is determined by $R^2 = 0.710$ indicates a good fit of the model.

The F distribution value assures that the overall model is statistically significant.

The t-test and p-values indicate the significance of individual predictors; the most influential predictors are BedR, CBD followed by Dtrain and Landsize. While BathR, CarS and Internalarea are statistically insignificant.

4.4 Implementation with Real Estate Data.

This study utilises scatter plots present the overall results because it have the ability, in this study, to present multidimensional data in two dimensional space.

4.4.1 Adjusting variable One with resulting visualization.

Figure 23 presents the scatter plot that illustrates the difference between the predicted values of the response and the observed values of the response when adjusting the predictor BedR while holding all other predictors fixed. The adjustment involves increasing BedR by 1 unit for all the observed properties underlying this study.

The scatter plot in figure 23 depicts the number of bed rooms on the x-axis while the y-axis depicts the numeral value for the response. A blue coloured point presents the predicted response value and a pink coloured point presents the observed response value. In general, we observe that most of the blue points have greater values than the pink points which means that predicted prices has risen and confirms what we had discussed before, that the predictor BedR is an influential predictor positively correlated with the response.

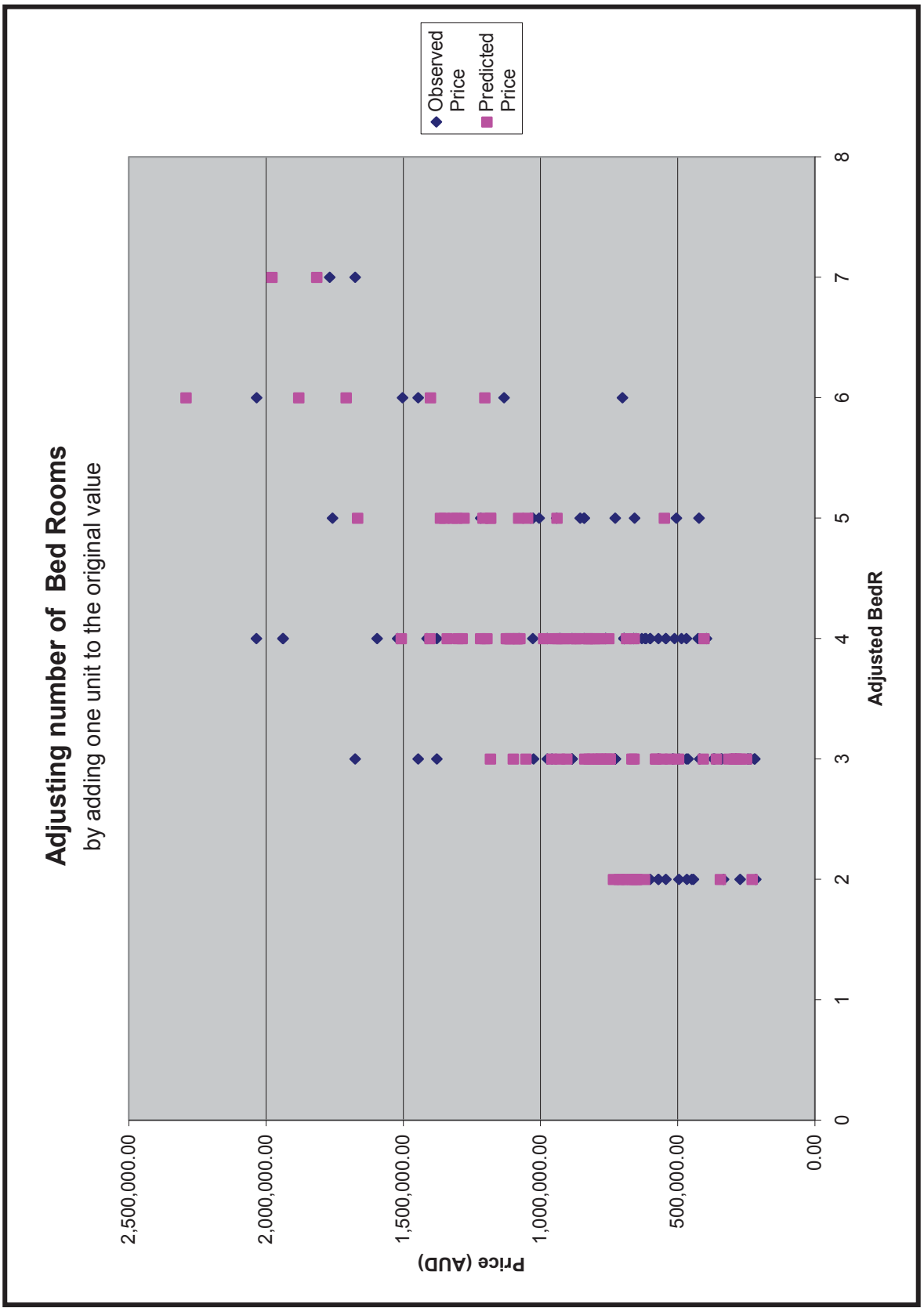


Figure 23: Compiled Results, number of bed rooms vs. predicted and observed price

4.4.2 Adjusting variable Two with resulting visualization.

Figure 24 presents the scatter plot that illustrates the difference between the predicted values of the response and the observed values of the response when adjusting the predictor CBD while holding all other predictors fixed. The adjustment involves increasing CBD by 2 km for all the observed properties underlying this study.

The scatter plot in figure 24 depicts the distance to CBD on the x-axis while the y-axis depicts the numeral value for the response. A blue coloured point presents the observed response value and a pink coloured point presents the predicted response value. In general, we observe that most of the blue points have greater values than the pink points which mean that the predicted prices has declined and confirms what we had discussed before, that the predictor CBD is an influential predictor negatively correlated with the response.

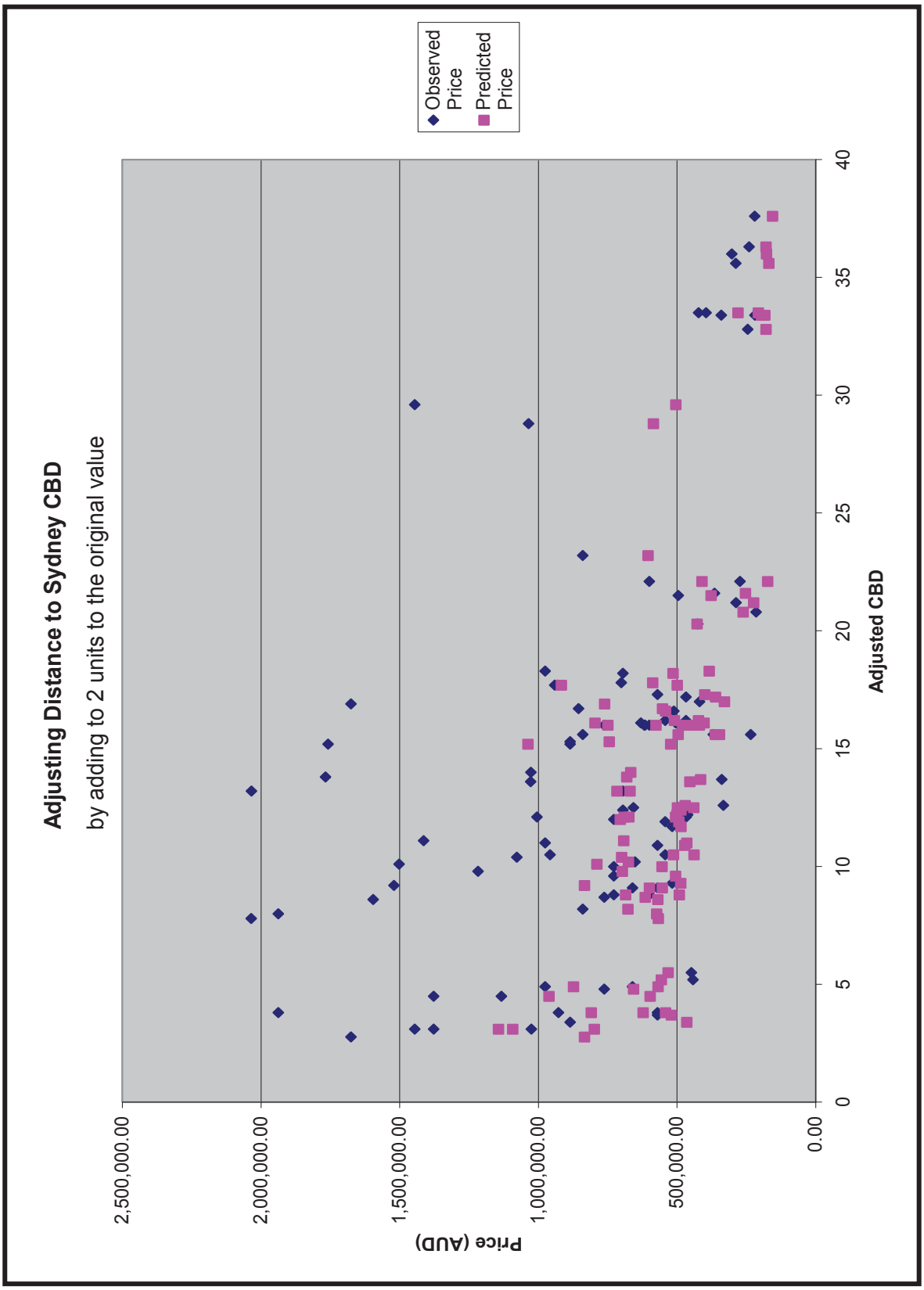


Figure 24: Compiled Results, distance to Sydney CBD vs. predicted and observed price

4.4.3 ADJUSTING VARIABLE THREE WITH RESULTING VISUALIZATION.

Figure 25 presents the scatter plot that illustrates the difference between the predicted values of the response and the observed values of the response when adjusting the predictor Dtrain while holding all other predictors fixed. The adjustment involves increasing Dtrain by 2 km for all the observed properties underlying this study.

The scatter plot in figure 25 depicts the distance to nearest transport Dtrain on the x-axis while the y-axis depicts the numeral value for the response. A pink coloured point presents the predicted response value and a blue coloured point presents the observed response value. In general, we observe that most of the blue points have greater values than the pink points which mean that the predicted prices has declined which disagree with the results. The analysis shows a positive correlation between the predicted price and the Dtrain predictor, but most of the analyses conducted contradict this result. As discussed earlier, the expensive properties in the region of Northern beaches that is far from transport means affected the regression results to show a positive correlation between the predicted price and Dtrain.

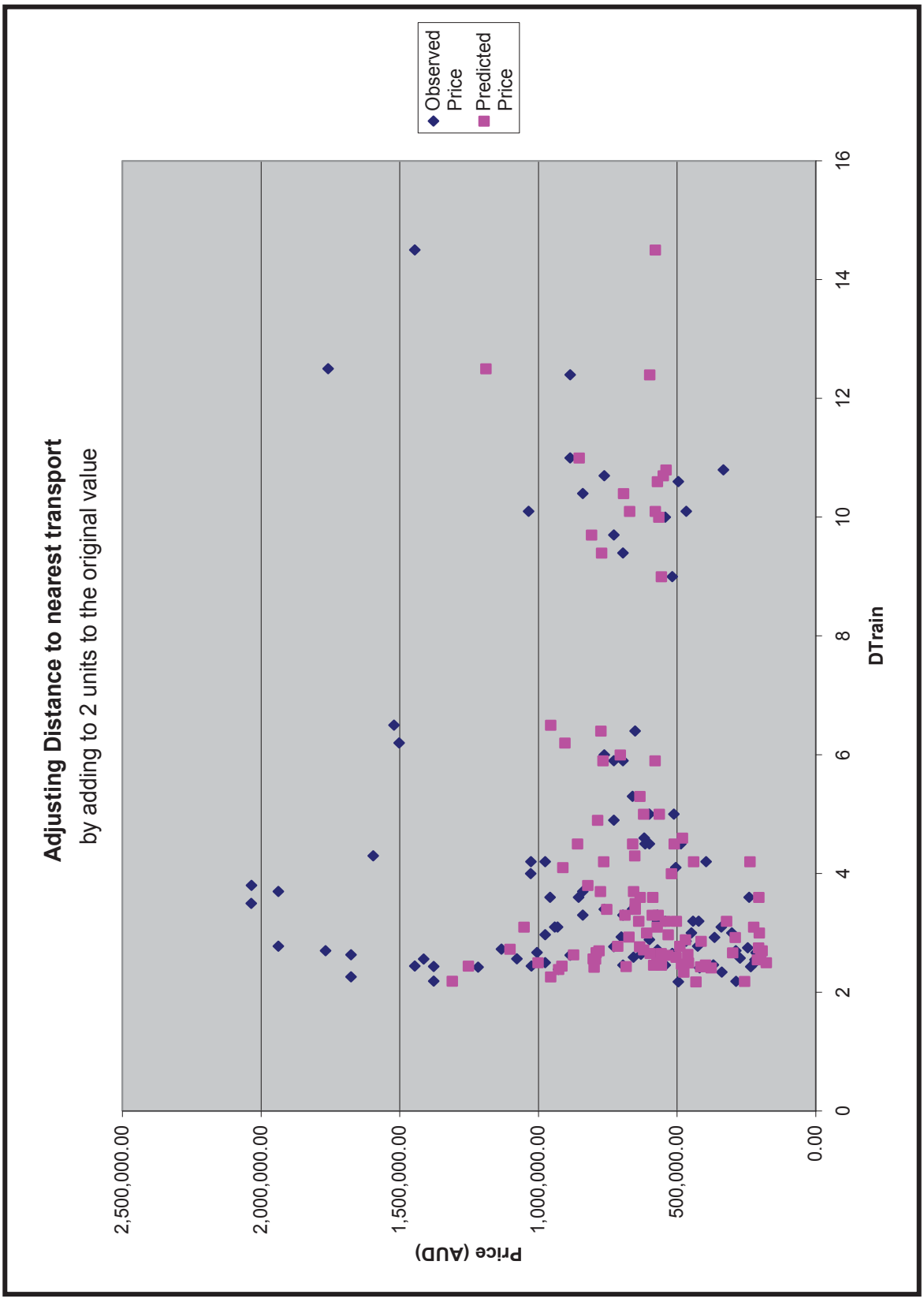


Figure 25: Compiled Results, distance to nearest Transport (DTrain) vs. predicted and observed price

5 CASE STUDIES

We conduct three case studies for interactive visualization for sensitivity analysis. All images and analyses presented can be reproduced using the following link

http://www.imap2013.dx.am/index_new.html

5.1 CASE STUDY 1: PREDICTION WHEN ADJUSTING NUMBER OF BED ROOMS.

The Canaans, a young family with two children of school age, are looking to buy their first home, so they can avoid paying expensive weekly rent, and save the property for themselves whilst paying reasonable mortgage.

They use the interactive visualization that empowers our system; they browse all regions in a rapid manner utilizing the overview + detail approach by switching back and forth from a single region to the overall map. And by hovering over the property icon to glimpse the property features in a form of a tooltip as in figure 26, and by clicking the property icon, all property features are displayed in the property features form as in figure 27.

They locate a property with features closest to what they are looking for in the region of Liverpool & Fairfield; a two bed room unit, with one bath room and one car space. Figure 26 shows the relative location of the property highlighting some features in the tooltip.

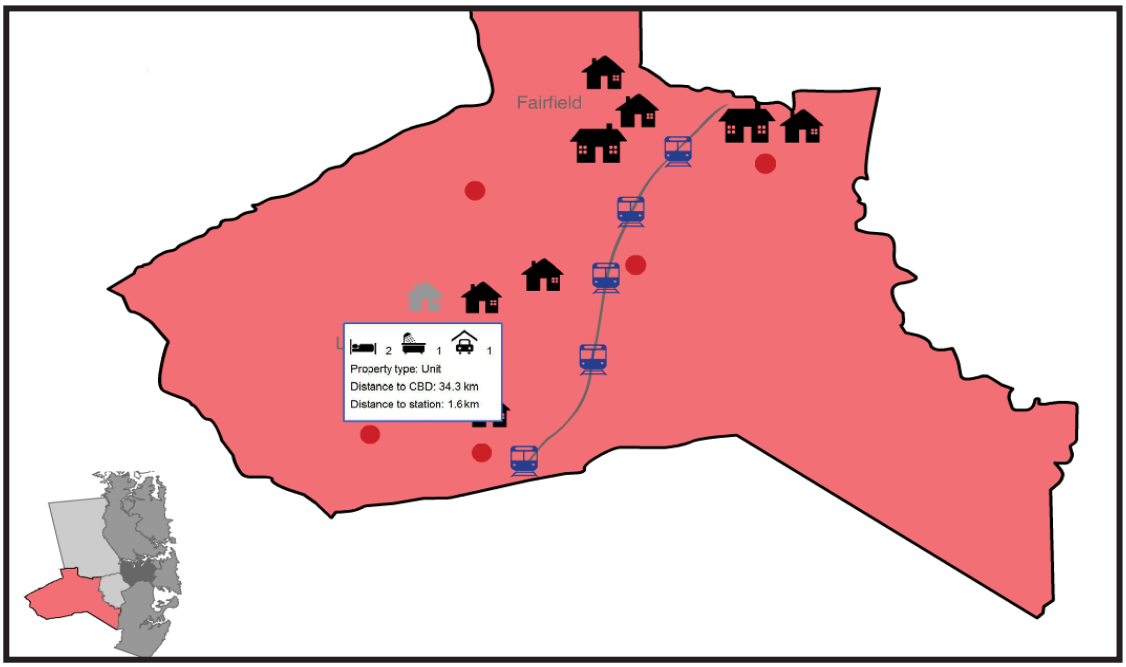


Figure 26: The selected region "Fairfield and Liverpool", with the property of interest highlighted. The tooltip shows some of the property features.

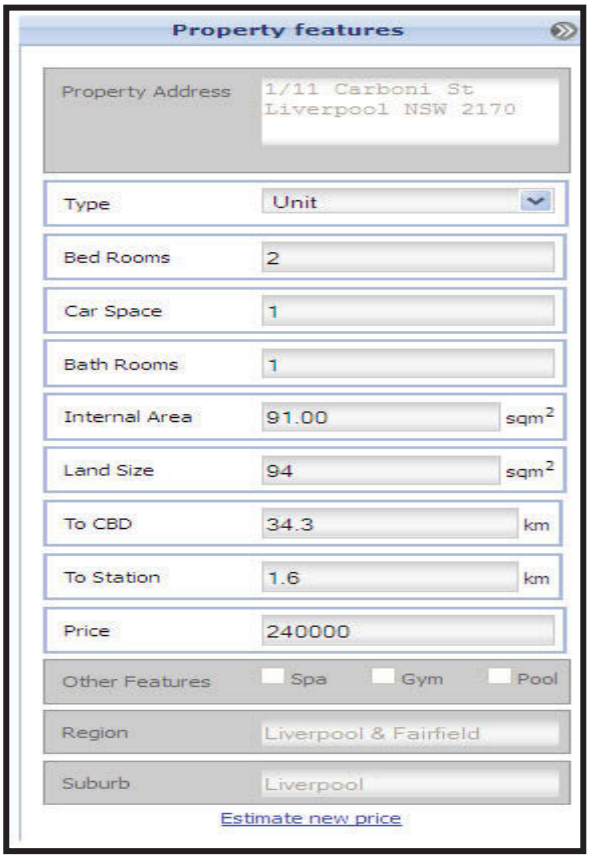


Figure 27: The property features form shows all available information about the property of interest for the Canaans.

The Cannans liked all the features including the price, AUD 240,000, but they want an extra room so that the children use it as a study room. To predict a price for a property with the same features but with three bed rooms instead of two bed rooms, they used the prediction ability of our system to obtain a price estimate for such property. They click on "Estimate new price" link; they are presented with "Customize Property" form, to adjust the property feature of interest. The features they have the ability to adjust are listed below and shown in figure 28.

- Number of bed rooms;
- Land Size;
- Distance from Sydney CBD; and
- Distance to nearest transport.

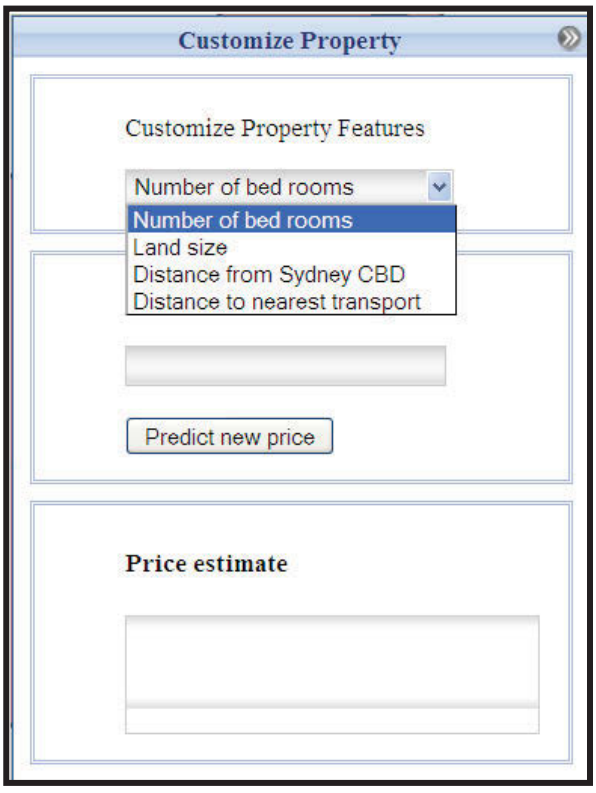


Figure 28: "Customize property" form. The drop down list shows the features that the user can adjust prior to the estimation of the price.

After entering the desired value for the chosen property feature to be adjusted, the predicted price is displayed in price estimate text box. Figure 29 shows a price estimate of AUD 286,467 within the confidence interval [255,314 and 321,421] for the same property features but with three bed rooms instead of two bed rooms.

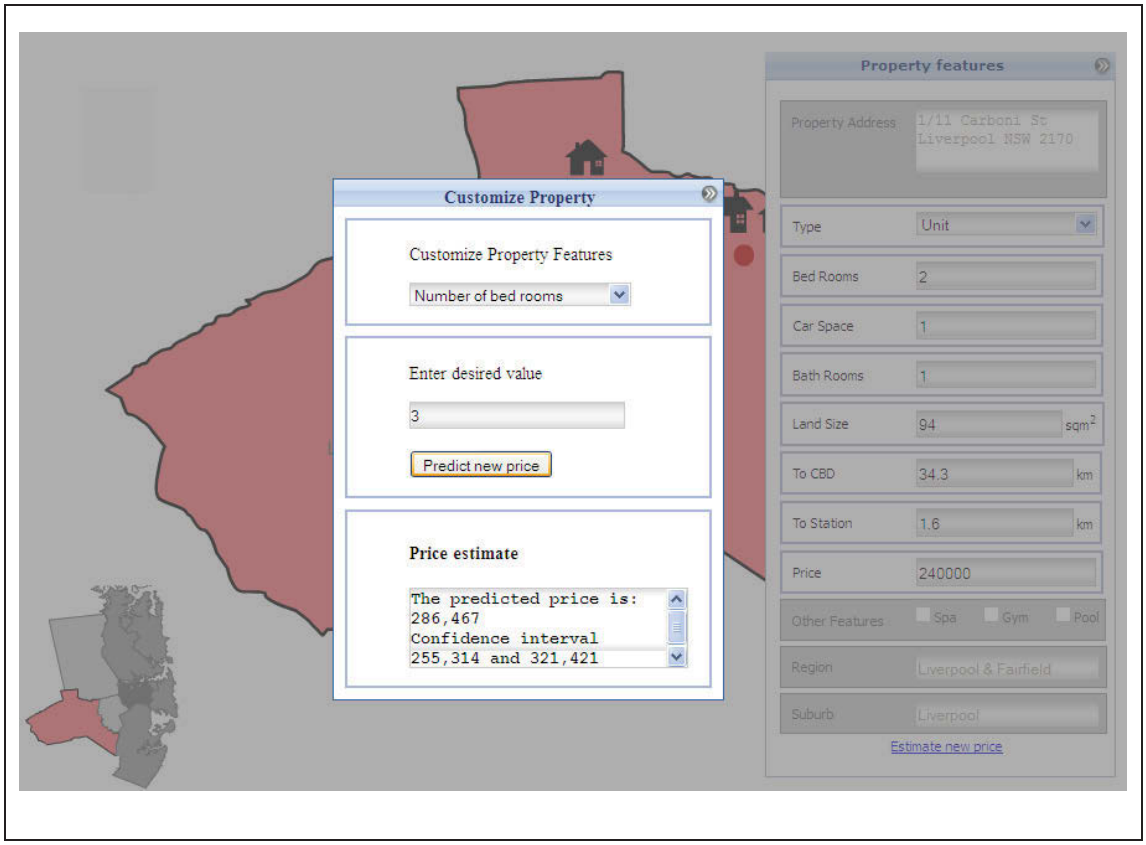


Figure 29: The predicted price for the property with adjusted BedR.

5.1.1 Discussion

Number of bed rooms (BedR predictor) is an influential predictor of the prediction model with positive correlation with the response ($\log_{10}$(price)), therefore an increase in number of bed rooms while

keeping other predictors fixed has the impact of increasing the predicted price, even the lower limit of the confidence is greater than the observed price for the selected property. The predicted price increase when adjusting number of bed rooms by one unit is AUD 46,467

5.2 Case Study 2: Prediction when Adjusting Distance to Sydney CBD.

A husband and wife who both work in the city of Sydney, are looking to buy a property that is close to their work so that they can save in commuting time and expenses, at the same time they are looking for a property with extra bed rooms so they can lease it for university students which will help them in paying their mortgage.

Using the regional abstract map of Sydney (IMAP) that the system offers; they directly choose the region of "City of Sydney", and using the tooltips that has the summary features for each property individually, they come across a property that attracts them, the property features are: 3 bed rooms, 2 bath rooms, 2 car spaces and it is just 1.8 km far from Sydney CBD and all are shown in the tooltip as figure 30 demonstrates. The selling price for that property is AUD 927,500. The couple wants a cheaper price so they decide to slightly compromise the distance to Sydney CBD.

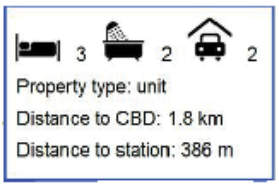


Figure 30: A tooltip shows the summary of property features

Using the prediction capability of the system, they adjust the distance to Sydney CBD (CBD) to 7 km and by estimating the price; they obtain a

price of AUD 867,518 with a confidence interval of AUD 773,176 and 973,371 as figure 31 illustrates. The couple likes the price, they record the details of the property along with adjusted distance to Sydney CBD and they book an appointment with their real estate agent to find them a property that matches the details of the adjusted property feature they obtain using the prediction system.

5.2.1 Discussion

Distance to Sydney CBD (CBD predictor) is an influential predictor of the prediction model with negative correlation with the response ($\log_{10}$(price)), therefore an increase in the number of kilometers between the property and Sydney CBD while keeping other predictors fixed has the impact of decreasing the predicted price, even the lower limit of the confidence is lower than the observed price for the selected property. The predicted price dropped by AUD 59,982 when adjusting the distance to Sydney CBD to be further than the observed value for the original property.

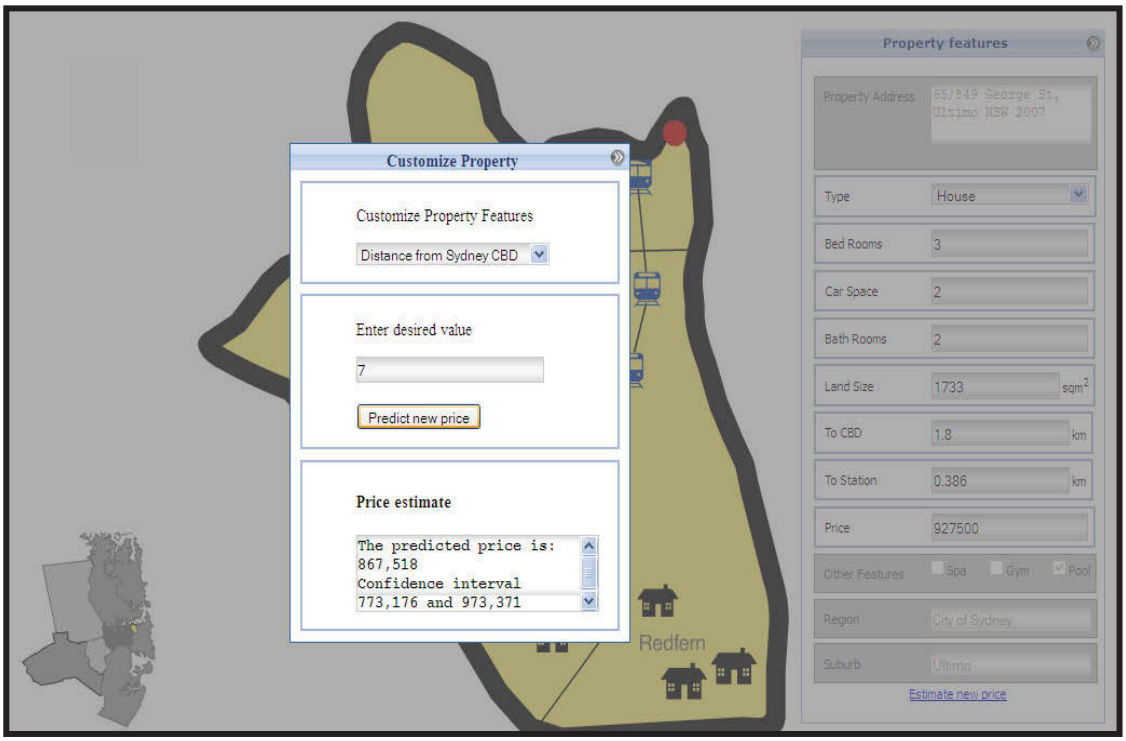


Figure 31: The predicted price after adjusting CBD predictor.

5.3 Case Study 3: Prediction when Adjusting Distance to nearest transport.

Two doctors who plan to start their own medical center are looking for a property, specifically a house that has 4 bed rooms and 2 bath rooms. They plan to use two rooms as clinics for patient visits, the third as a treatment room, and the last room for x-ray machine.

They utilize the simplicity of the interactive visualization for the geometrical map that the prototype offers and locate a house with the features in demand in the region of "Inner west". The only thing that was undesirable in that house is that it is far from transport means; because it will be hard for the patients to reach it especially if they do not have their own transport. Therefore they use the prediction ability of the system to adjust the distance to nearest transport (Dtrain) to be 0.5 km instead of 2.2 km.

Figure 32 shows what the price they obtain, a price of AUD 1,018,725 within the confidence interval 907,939 and 1,143,028 for a property that is far just 0.5 km from nearest transport but all other features kept the same.

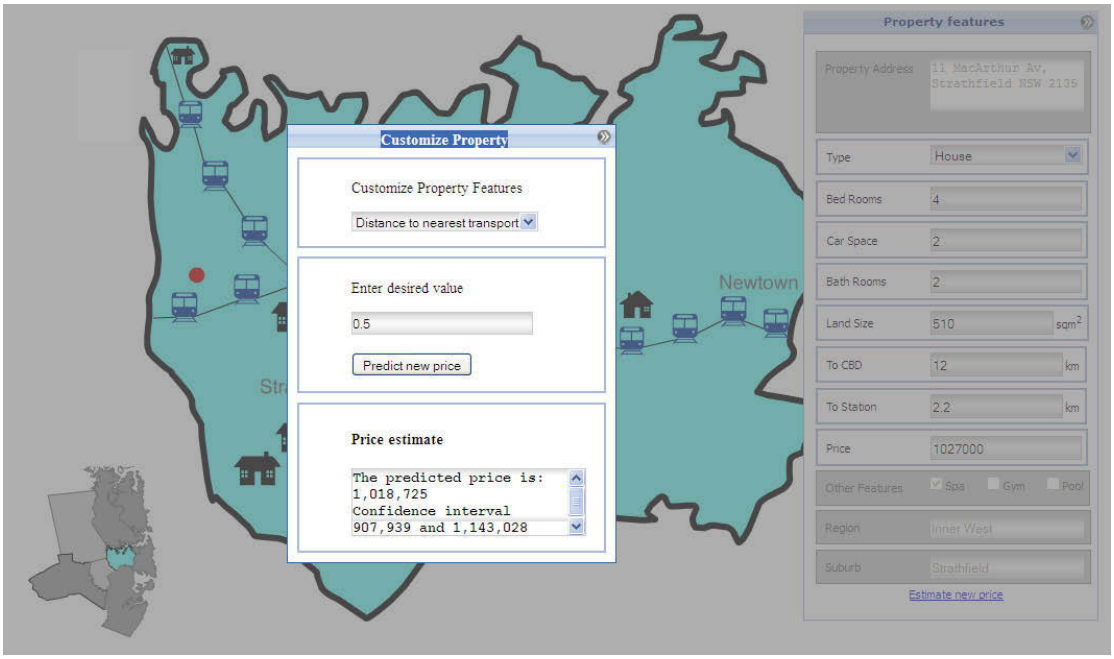


Figure 32: The predicted price after adjusting Dtrain predictor.

5.3.1 Discussion

Another significant predictor of the model is the distance to nearest transport (Dtrain), it is has less significance than CBD and BedR predictors but has more significance than LandSize. It is positively correlated with the response ($\log_{10}$(price)).

Therefore a decrease in the number of kilometers between the property and the nearest transport yields a drop in the predicted price. The predicted price is dropped from AUD 1,027,000 to AUD 1,018,725; a total drop of AUD 8,275.

This case agrees with regression model and we obtain a drop in the predicted price, but as discussed in the section 4.4.3, this result is unrealistic.

6 COMPARATIVE STUDY

The problem of property price prediction was tackled in many previous researches. In this chapter, I present three different real estate prediction system.

6.1 SYSTEM 1: USING LME ALGORITHM

A research proposed the latent manifold estimation algorithm (LME) which combined two models to predict the property price (Christiernin 2010). The first model was a smooth, non-parametric model of the latent desirability manifold. The second is a parametric model that only considers the features of the property and estimates the property intrinsic price. The two models are trained simultaneously producing an intrinsic price and desirability value. The product of the desirability and the property intrinsic price produces the property predicted price. The training set had around 750,000 transactions obtained from First America. Property attributes included geographical information, financial information like taxable land value and the property features itself, like number of bed rooms, bathrooms, and car space and others.

LME was compared to other standard techniques to predict the property price, it shown that LME outperformed all techniques compared to it.

The comprehensive discussion of the results included a color-coded sensitivity maps proved that LME is able to capture the hidden non-linearities

in the predication function with respects to a specific attribute, i.e. number of bed rooms.

It requires an expert user to run LME algorithm and to interpret the results. LME suffers from the absence of user interaction; hence user preference cannot be encompassed in the process.

6.2 System 2: Using Gamma test.

Another property prediction approach utilized the Gamma test as their approach to attribute selection and dependence modelling to solve the problem of predicting property prices (Wilson et al.). The Gamma test was used in a two-phase procedure. The first phase used the Gamma test to guide a Genetic algorithm for selecting a helpful subset of attributes from a large dataset. The second phase involved generating a predictive model using Artificial Neural Network that forecasts the changes in the House Price Index. The developed dataset, which focused on the prediction of values at a national level, where data has been recorded systematically for 30 years, included economic statistical series of historical measures that thought to have impact on the House Price Index. The attribute selection process, conducted by a simple form of the Genetic algorithm generated eight input measures; these measures included retail price index, bank rate, average earnings, durable goods and other four measures. Attribute selection phase concluded a histogram for each input measure and one pie graph summarizes the relative importance of these measures. Forecasting using artificial neural networks phase also summarized the results graphically using a multi-line graph presenting the annual percentage change forecast in the House Price Index.

An expert is required to conduct the Gamma test, even for the data pre-processing and preparation. The historical nature of the dataset does not

suggest any kind of user interaction; therefore user preferences are not accounted for in this approach.

6.3 SYSTEM 3: DOMAIN.COM

Domain.com.au offers property reporting services. Property reporting at Domian.com has three levels; postcode or suburb report, street report and property report.

Both the street report and the postcode report highlight the same information, these include:

- Market snapshot: Auction clearance rate, number sold by auction versus private treaty and total sales value.
- Median forecast : 12 month forecast for houses and units (price and percentage change)
- Last property sold in the street or the suburb along with an image for it.
- Most expensive property the street or the suburb along with an image for it.
- Sales history for 12-24 months
- Market trends
- Price summary which includes the median price, the average, the highest and lowest prices.
- Property features like number of bed rooms, number of bath rooms, size and others.

The property report highlight almost the same information but it is more specific to the property itself introducing the current market price estimate, probable value range and future price estimate.

The report mentions that the current price estimate is driven by APM's proprietary DMV modeling (Dynamic Multi-Variant Modeling) technology.

Domain reporting provides useful information for the three levels of the report, given that the report is comprehensible for a non-expert user. Ordering the report requires a form to be filled with the property full address, type

(house, unit, duplex, …) and features. This form is the only interaction that the user has through the prediction process by which the user cannot express his/her preferences.

Table 1 summarizes the features of interest for each of the reviewed algorithms and systems. The Interactive visualisation for sensitivity analysis approach is the only to involve user preference in the price prediction process and is the only to offer user visual interaction.

	Using Gamma test	Domain.com	Using LME algorithm	Using Interactive visualization for SA
User Interaction	Not available	Filling an online form	Not available	♦ Visual interaction via the map ♦ Interaction while updating a form
presumed Knowledge level	Expert level	Non-expert level	Expert level	♦ Non-Expert level for the price prediction ♦ Expert level for interpreting the scatter plots
Visualization	Histograms, pie graph and multi-line graph	Histogram	Color-coded maps	♦ Sydney regional map ♦ Scatter plots
Results interpretation	Discussion of results and graphs interpretation	Textual interpretation.	Discussion of results and maps interpretation	♦ Instant predicted price ♦ Scatter plots compiling all analyses conducted.
Prediction encompasses user preference	Not available	Not available	Not available	♦ User sets his preference to affect the price prediction.

Table 2: Reviewed algorithms/systems for property price prediction features summary.

7 Conclusion and Future Work.

7.1 Conclusion

In this thesis, I present a system that proposes a novel approach to integrate interactive visualizations with sensitivity analysis. I apply this approach to a real estate prediction system.

7.1.1 Interactive Visualization

The visualization is employed to allow user interaction with the sensitivity analysis coupled with simplicity to hide the complexity of the sensitivity analysis. The visualizations enriched the system with newly introduced features such as the distance to nearest transport as well as the distance to the center of the city and a network like graph that shows the properties that are less than a pre-specified distance from the selected train station.

7.1.2 Sensitivity Analysis

The sensitivity analysis method used, multiple linear regression, produced a statistically significant prediction model that suggests four significant predictors and three insignificant predictors. The most two influential predictors produce realistic results while the least significant predictors fail to do the same.

The presented approach is expected to have a wide range of applications since the two fields involved are interdisciplinary.

7.2 Future Work.

This thesis presents a real estate prediction system that is equipped by the integration of interactive visualization and sensitivity analysis. Upgrading this system to a recommendation system which utilizes machine learning algorithms is of a significant importance.

The real estate prediction system results will be an input to the property recommendation system along with following:

- All saved user information including age group, profession, income and others.
- Property data

Hence, the recommendation system will have the ability to learn from this information to present the user with the proper recommendations.

To conduct a user study to evaluate the usability of the real estate prediction system, this can be achieved by some activities including:

1. Observing and interviewing users and getting their feedback, this helps in identifying needed functionality or design flaws that are not anticipated.
2. Studying the user behaviour on the web site to ensure a low tolerance for any difficult designs that prevents the users from grasping the functionality of the system immediately.
3. Investigating the learnability indicators such as the easiness of accomplishing basic tasks, system efficiency, user memorability of the system, user error recovery and user satisfaction.

Another direction is to construct the regression model for the real estate prediction system taking in consideration the following:

- To study each region separately so that expensive property regions do not affect cheaper regions and vice versa.
- To study the property type effect on the model or to study each property type separately.

Appendix

Real Estate Prediction System - User Guide

To Access the system supporting this research, please visit:
www.imap2013.dx.am/index_new.html

Once lunched, the home page will show the IMAP.

1. **To use the prediction ability, please follow the following steps:**

- Click on the desired region. This will zoom in the selected region.
- Using the mouse, hover over the properties in the region represented by the house icon. Placing the mouse over the house icon will show important information about the property.
- Click the house icon for the property of interest. You will be presented with the "Property feature" form showing all available information about the property.
- To predict a property price, please click "Estimate new price" link, you will be presented with the "Customize property" form.
- Select one of the property features that you need to change from those available in the list.
- Enter the new value in the text box below "Enter desired value"
- Click on "Predict new price", then the predicted price will be shown in the text area below "Price Estimate"

2. **To view the scatter plots that compiles all the analysis, please follow the following steps:**

- Click on "Visual Results" Link.
- You will be asked to download an Ms Excel file, please open it or save to your computer.
- The file contains three different scatter plots, you can navigate as required.

Hint: To close any form at anytime, please use the icon at the right top of the form itself.

Bibliography

Campolongo, F., Kleijnen, J. & Andres, T. 2000, 'Screening Methods', in A. Saltelli, K. Chan & E.M. Scott (eds), *Sensitivity Analysis*, John Wiley & Sons Ltd, West Sussex.

Campolongo, F., Saltelli, A., Sorensen, T. & & Tarantola, S. 2000, 'Hitchehker's Guide to Sensitivity Analysis', in A. Saltelli, K. Chan & E.M. Scott (eds), *Sensitivity Analysis*, John Wiley & Sons Ltd., West Sussex.

Carr, D.B., Littlefield, R.J., Nicholson, W.L. & Littlefield, J.S. 1987, 'Scatterplot Matrix Techniques for Large N', *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 424-36.

Christiernin, L.G. 2010, 'Guiding the designer: A radar diagram process for applications with multiple layers', *Interacting with Computers*, vol. 22, no. 2, pp. 107-22.

Christopher Frey, H. & Patil, S.R. 2002, 'Identification and review of sensitivity analysis', *Risk Analysis*, vol. 22, no. 3, pp. 553-78.

Cockburn, A., Karlson, A. & Bederson, B.B. 2008, 'A review of overview+detail, zooming, and focus+context interfaces', ACM Computing Surveys, vol. 41, no. 1, p. 13.

Cooke, R.M. & Noortwijk, J.M. 2000, 'Graphical Methods for Uncertainty and Sensitivity Analysis', in A. Saltelli, K. Chan & E.M. Scott (eds), *Sensitivity Analysis*, John Wiley & Sons Ltd.

Cushing, J., Janssen, L.D., Allen, S. & Guerlain, S. 2006, 'Overview+detail in a Tomahawk Mission-to-Platform Assignment Tool: applying information visualization in support of an asset allocation planning task', Information Visualization, vol. 5, no. 1, pp. 1-14.

Daradkeh, M., Churcher, C. & McKinnon, A. 2008, 'Interactive Visualization Techniques for Exploring Model Sensitivity', paper presented to the *Proceedings of New Zealand Computer Science Research Conference*, Christchurch, New Zealand.

Furnas, G.W. 1986, 'Generalized fisheye views', ACM Conference on Human Factors and Computing Systems, ACM, New York, pp. 16-23.

Hamby, D.M. 1995, 'A comparison of sensitivity analysis techniques', *Health Physics*, vol. 68, no. 2, pp. 195-204.

Hsiao, P. 2010, 'Visualization of Large Document Collections', 3442642 thesis, North Carolina State University, North Carolina, United States

Kleijnen, J.P.C. & Helton, J.C. 1999, 'Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: Review and comparison of techniques', *Reliability Engineering & System Safety*, vol. 65, no. 2, pp. 147-85.

Leung, Y.K. & Apperley, M.D. 1994, 'A review and taxonomy of distortion-oriented presentation techniques', ACM Transactions on Computer-Human Interaction, vol. 1, no. 2, pp. 126-160.

McFarlane, M. & Young, F.W. 1994, 'Graphical Sensitivity Analysis for Multidimensional Scaling', *Journal of Computational and Graphical Statistics*, vol. 3, no. 1, pp. 23-33.

Moustafa, R.E. 2011, 'Parallel coordinate and parallel coordinate density plots', *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 2, pp. 134-48.

Niklas, E. 2008, 'Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation', *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1141-8.

Pannell, D.J. 1997, 'Sensitivity analysis of normative economic models: theoretical framework and practical strategies', *Agricultural Economics*, vol. 16, no. 2, pp. 139-52.

Saary, M.J. 2008, 'Radar plots: a useful way for presenting multivariate health care data', *Journal of Clinical Epidemiology*, vol. 61, no. 4, pp. 311-7.

Saltelli, A. & Annoni, P. 2010, 'How to avoid a perfunctory sensitivity analysis', *Environmental Modelling & Software*, vol. 25, no. 12, pp. 1508-17.

Saltelli, A., Chan, K. & Scott, E.M. 2000, *Sensitivity Analysis*, John Wiley & Sons Ltd.

Saltelli, A., Marco, R., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. (eds) 2008, *Global Sensitivity Analysis. The Primer*, Book, John Wiley & Sons.

Shneiderman, B. 1992, 'Tree Visualization with Tree-Maps: 2-d Space-Filling Approach', ACM Transactions on Graphics, vol. 11, no. 1, pp. 92-99.

Tague, N.R. 1995, *The Quality Toolbox*, ASQC, Milwaukee.

Tarantola, S. & Saltelli, A. 2003, 'SAMO 2001: methodological advances and innovative applications of sensitivity analysis', *Reliability Engineering & System Safety*, vol. 79, no. 2, pp. 121-2.

Therón, R. & De Paz, J. 2006, 'Visual Sensitivity Analysis for Artificial Neural Networks, Intelligent Data Engineering and Automated Learning – IDEAL 2006', in E. Corchado, H. Yin, V. Botti & C. Fyfe (eds)vol. 4224, Springer Berlin / Heidelberg, pp. 191-8.

Utts, J. 1999, *Seeing Through Statistics*, Second edn, Brook/Cole.

Weisberg, S. 1947, *Applied Linear Regression*, Third Edition edn, Hoboken, N.J. : Wiley-Interscience, c2005.

Wilson, I.D., Jones, A.J., Jenkins, D.H. & Ware, J.A., 'Predicting housing value: Attribute Selection and Dependence Modelling utilising the Gamma Test'.

Publications

Visual Sensitivity Analysis in Real Estate Prediction System.

Massara Dana and Mao Lin Huang,
In Proc. of IEEE Int'l Conference on Computer Graphics, Imaging and Vision (CGIV'13), pages 100-105, held in Macau, China, August 6-8, 2013. DOI: 10.1109/CGIV.2013.25