Exploratory analysis of large spatial time series and network interaction data sets: house prices in England and Wales

Crispin H. V. Cooper

Final version, June 2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

NOTICE OF SUBMISSION OF THESIS FORM:
POSTGRADUATE RESEARCH

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed C. Cooper (candidate) Date 21/6/2010

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed C. Cooper (candidate) Date 21/6/2010

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed C. Cooper (candidate) Date 21/6/2010

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed C. Cooper (candidate) Date 21/6/2010

Summary

This thesis describes a combined exploratory analysis, on a fine spatial scale, of (i) England and Wales house prices, between the years 2000 and 2006; (ii) aggregate statistics taken from the UK census of 2001; and (iii) interaction statistics also taken from that census. The house price data is derived from individual transactions and analysed mainly in the form of ward level indices with a time resolution of 100 days.

The study has twin aims: firstly, to improve understanding of the data set - which is large in nature - particularly with respect to exploring the interaction statistics; secondly, to improve the methods of exploratory analysis themselves.

With respect to the aim of understanding the data, both migration and house price changes are visualised in a novel way, and regression is used to determine indicators of likely house price cross-correlations between different market areas. Ripple type effects are shown to be related both to reactive mechanisms, and to the composition of migration flows. Further visualisation shows that the market may be understood in terms of clusters with similar behaviour, or alternatively, in terms of market-driving and market-driven regions. Variables which can be used to define these clusters and regions are identified via further regression.

With respect to improving the techniques of analysis, existing methods of visualising interaction data - based on clustering and linear ordering of points in geographic space - are extended to larger, hierarchical data sets and evaluated in this context. Novel approaches are presented for (i) construction of relative house price indices with minimal hedonic data, (ii) enhancement of time series predictions using cross-correlation data, and (iii) comparison of heterogeneous data sets via unification of all relevant information in the interaction domain, making it susceptible to analysis by regression aided with principal component based dimensionality reduction.

Acknowledgements

The author would like to thank the Landmark Information Group Ltd for providing essential data for this research, and ARCCA (Advanced Research Computing @ Cardiff) for providing the resources to process it.

To my parents

> who must have wondered why - shortly after employment levels peaked in 2005 - both of their offspring reverted to being students;

to my sister,

> who promptly finished her PGCE and is now in her third year of teaching, thus experimentally validating her course of action;

to my supervisors Chris and Chris,

> without whom I would not have been able to hop between academic fields with anything like the ease that I did;

to everyone who has enriched my life in South Wales,

> particularly through their substantial contributions to the fields of rock climbing, mountain biking and electronic music;

to Ami,

> to whom I am eternally grateful for helping me through the darker moments of the dreaded fourth year of study;

and to the many inspirations provided by the Welsh hills.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 On the adventure of mountaineering - or, the aims of the study

This body of work is presented in the spirit of a first ascent of a vast mountain of data, a mountain formed by the collision of two pre-existing data ranges. One of these is the ancient massif of the UK Census, well trodden and traversed throughout history - a significant exploration of which was completed as early as the year 1085, as recorded by William the Conqueror in his Domesday book, though this was almost certainly not the first. Meanwhile in modern times, the region's tourist board - the Office for National Statistics - has been keen to encourage visits to the area for all who display an interest in doing so.

The second mountain range is a newer batholithic eruption of house price data, rising from the hot magma of widespread individual home ownership, and in very recent years collected by the computers of the England and Wales Land Registry. In contrast to the former, this range has been fiercely guarded by its inhabitants, who would seldom allow outsiders to tread upon its slopes, let alone undertake a detailed expedition into its midst. Even now, hefty summit fees are charged for the privilege of doing so.

Due to the limited accessibility of the entire second range, it is believed that few researchers, if any, have yet climbed the interesting peak where the two ranges meet. This thesis, then, is an account of the exploration of that summit. The ascent has yielded evidence of some interesting features of the data set - plants,

beasts, rock and ice formations which had not yet been recorded by any explorer. Some of these were already suspected to exist (and their discovery therefore helps to validate the usefulness of the exploration) - most of the specimens collected, however, remain mysterious. The explorer is a specialist in computation, not housing markets; akin to a mountaineer who possesses only a basic knowledge of botany, zoology, geology and glaciology. Therefore, the principal contribution of this trip is a variety of techniques for climbing, traversing and descending the mountain, which in future might be used by specialists in other fields to explore it (or indeed other mountains) more fully.

George Mallory, who died in 1924 when descending from what might have been the first ascent of Everest, famously gave as his reason for climbing the mountain: "Because it's there". This thesis is likewise driven by the exploratory analysis of a largely unknown socio-economic data set, simply because it is there. However it is not, primarily, a thesis on the housing market; the real motivation is better expressed by Edmund Hillary, who along with Tenzing Norgay in 1953 returned from the summit alive. "It is not the mountain we conquer", he said, "but ourselves".[1] Climbing the mountain of data effects improvement of the self; the creation of better tools and techniques which can be applied to other problems in future - and therein lies the primary purpose of the undertaking.

The aim, then, is not to build and interpret models for the sake of a comprehensive understanding of phenomena, but simply to use such models to explore some very large data sets which would otherwise remain untractable.

1.2 Background to the study: the state of computational social sciences and the case for exploratory data analysis

The author hopes it is clear from the previous section that the main focus of this thesis is on the development of new exploratory data analysis techniques. The secondary focus is on the UK housing market. Instead of writing about these topics at length in the introduction, their background literature will be reviewed

[1] Actually, this also echoes the words of Mallory - who wrote in an Alpine Club journal: "Have we vanquished an enemy? [...] None but ourselves".

where it is most relevant to the technical discussion: exploratory data analysis in chapter 3, and the housing market in chapters 4 and 5 - though some literature on the census and housing market, where it relates to the data sets and their aggregation, is also reviewed in chapter 2.

Such treatment of the existing literature, however, misses the question of why the study is necessary in the first place. This section therefore seeks to answer that question, through analysis of the current state of computational social sciences. This analysis discusses three areas of literature in turn: classical spatial economics (section 1.2.1), agent-based computational modelling (section 1.2.2) and network formalisms (section 1.2.3). Section 1.2.4 explains how the present study fits in with this literature.

1.2.1 Classical spatial economics

Modern approaches to analysis of the spatial distribution of human activity are generally considered to have started with the bid rent theory of Von Thunen, in the early part of the 19th century, which predicted patterns of land use based on who could best afford to pay for each unit of land. Much of mainstream spatial economics (or economic geography) owes its heritage both to this, and to the General Equilibrium framework developed in the wider field of Economics during the second half of that century - though also, to some extent, to the related field of Game Theory that emerged during the 1940s. McCann (2001) gives a wide overview of spatial economics, while Kreps (1990) covers game theory. The underlying philosophy of the field could be summarised by saying that prices, spatial locations and human behaviour are best understood by the study of theoretical states of equilibrium, either defined as a balance between quantities and prices (in models of supply and demand) or structured agreements which no individual would have the incentive to defy (as is the case with game theory). Furthermore, it is generally accepted that analytical solutions to the equations that describe these states are a worthwhile contribution to knowledge. A good case study is the spatial location game of Hotelling (1929), which has been applied to the explanation of a huge variety of phenomena, from the locations of shops on a street to the positioning of political parties on a left-to-right spectrum. The theory has been progressively refined over the years (Gabszewics & Thisse 1992, gives a good overview), including solutions for arbitrary numbers of competitors (Eaton

& Lipsey 1975) and in multidimensional space (Irmen & Thisse 1998).

The mainstream is not without its problems, however. Kreps (ibid) notes the difficulties of game theory, for example: on what basis is an equilibrium chosen if there are multiple equilibria? And what if players make moves which run counter to theory? Day (1993) points out that the founders of classical economics, including Adam Smith himself, were well aware that not all of human behaviour was rooted in balance and rationality. Atkinson (1969, in Ormerod 2005, p. 21), states that it may take over 100 years for economic growth equilibria to stabilise - meaning that the systems we observe are largely in disequilibrium in any case. And in the latter half of the 20th century, the advent of chaos theory undermined the idea that even the simplest behavioural foundations would necessarily result in an analytically tractable outcome. The sentiment is succinctly expressed by Strogatz (1994): "If you listen to your two favourite songs at the same time, you won't get double the pleasure!"[2]

1.2.2 The advent of agent-based models and evolutionary economics

The fields of social, economic and geographic agent-based modelling were founded on the principle that real economic systems "change slowly and irreversibly over time, which means that they do not lend themselves well to equilibrium analysis" (Andersson et al. 2006). Instead of taking the mathematically tractable, analytical approach, agent modelling seeks to simulate the actions of large numbers of independent entities by explicit stepwise computation. Axelrod (2006) clarifies the goals of agent modelling as (i) assisting theoretical understanding of the fundamental causes in social systems, (ii) assisting empirical understanding of why certain features have come to exist, and (iii) assisting normative understanding by helping us to design better systems.

The seminal work in this field is the model of ethnic segregation presented in Schelling (1971), which demonstrated that segregated societies can easily form in spite of a reasonable quantity of inter-racial tolerance on an individual level. This, combined with the advent of more powerful, yet cheaper computers in the past few decades, inspired a substantial body of similar work - some explicitly spatial,

[2]Notwithstanding Drummond & Cauty (1988), which describes the production of a number one hit single based on the application of just such a procedure.

some not. Some examples are Epstein & Axtell (1996), in which a simulated 'sugarscape' provides a theoretical model for culture and trade; Tesfatsion (1997) which models trade via game theory; Page (1999) which deals with the location of cities; Axtell et al. (2000) which uses game theory to explain discrimination in a hierarchy of social classes; Lai (2006) which models urban planning through chance combinations of problems, solutions and decision makers; Webster (2001) which simulates competitive and cooperative behaviour over public goods; and Lake (2001) which investigates the effects of cultural learning on hunter-gatherer survival. Overviews of further literature are provided by Lebaron (2006), Kollman & Page (2006) and Batty (2001) in the fields of finance, politics and pedestrian modelling respectively. Li (2005) contains some examples of cellular automata integrated with geographic information systems, and applied to agricultural zoning. Finally, echoing a return to the origins of spatial economics, Heikkila & Wang (2006) describes the application of an agent model to examine the implications of unifying the theory of bid rent with Hotelling's spatial location game.

A field that often overlaps with agent-based modelling is that of evolutionary economics. While Darwin's *Origin of Species* was published in 1859 - and had a widespread impact on subsequent understanding of not only the biological, but also the social world - it wasn't until the 1990s that serious attempts were made to explain society through actual simulations of Darwinian behaviour. As was the case with agent modelling, the reason for this probably relates to the advent of cheap computing power. Witt (1993) gives a good introduction to evolutionary economics, and also criticises general equilibrium from the point of view of evolutionary fitness and optima, which are in the real world are ever-changing and adaptive (Witt 1992). Examples of the application of the field extend into financial markets (Xu 2006), innovation (Cowan et al. 2006) and the adoption of new technology (Schwoon 2006, Sandberg 2007).

1.2.3 Network formalisms

Throughout the majority of spatial computational models viewed above, the representation of space is limited mainly to physical topology. In the late 1990s, however, a new concept of space started to emerge: that of *network space*. A network, whether deduced from transport links, trade relations or human contact, can be defined as a structure representing the degree to which each pair of

entities in a system is connected.

Watts & Strogatz (1998) demonstrated the creation of a *small world* network, in which the majority of nodes are not directly connected, but can nevertheless be reached from one another in a limited number of steps. Such networks are thought to be widely prevalent in the real world. Several network formalisations followed, extending the theory to allow network links of varying strength, and enabling the abstraction of any such network to a limited number of parameters (Barrat et al. 2004, Vragovic et al. 2005, Latora & Marchiori 2003). Also, alternative processes were proposed that could explain the formation of networks (Masuda et al. 2005). Such networks soon became incorporated into agent-based models, e.g. of migration (Silveira et al. 2006) and the labour market (Tassier & Menczer 2001), while the properties of real-world networks were investigated (e.g. Jiang 2004, Faust et al. 2000). Notably, it is sometimes the case that abstract parameters of nodes derived from their positions in the network can be shown to correlate with real-world characteristics, for example in the De Montis et al. (2007) study of inter-urban traffic.

Such network models have now gained widespread acceptance as a relevant depiction of social reality - one that is now mirrored in the significant commercial success of collaborative filtering systems such as *stumbleupon.com*, the recommendations feature of *Amazon*, and social networking websites such as *Facebook*.

1.2.4 Concerns of the present day

To date, then, computational models, dealing with both physical and network space, have been developed to address the limitations of classical spatial and economic theory. In the 21st century however, it has become apparent that - despite hundreds of citations of Schelling's original paper on segregation[3] - agent based modelling is not without its own problems. While Andersson et al. (2006) was quoted at the start of this section as a proponent of theoretical agent modelling, the same paper notes that

> "the lessons that have been learned have not easily been carried over to applied models".

[3]The ISI Web of Knowledge records 375 as of October 2008

Indeed, while theoretical models have often been enlightening as to the emergent nature of society, they have not often been followed by accurate predictions of future changes. Pontius Jr et al. (2007), comparing a few different predictive models, concludes that many contain more error than truth. Wu (2005) states that validation of such models is still a problem. Indeed, it is often the case that several different plausible models can produce behaviour similar to that observed in the real world, and we are left wondering which one is correct. Models can lead to qualitative, as well as quantitative errors in understanding. Batty (2006) provides an example by visualising changes in the city hierarchy over time. When viewed at a single stage in time, existing models of city growth are *quantitatively* consistent the data, but the change in rank orderings of cities revealed by an extended historical study is *qualitatively* irreconcilable with those models.

The question, then, is where to go from here. One approach to the *quantitative* problem is calibration with real data, to ensure that the model successfully reproduces known test cases before using it to extrapolate to future behaviour. An example of this is Whalley & Zhang (2007), which studies labour mobility restrictions in China. Another approach is to demonstrate that the model quantitatively outperforms humans, as in the case of the auction trading agents of Cliff (2003): such behaviour is hard to argue with. However, none of these approaches can protect against the problem of *qualitative* error, whether in the form of failure to consider future effects of variables not considered in the model, or in the form of a misunderstanding of the true causes of existing scenarios.

One method for improving *qualitative* accuracy of the rule-set is to build models on better-developed discourse, such as the description of spatial co-operative behaviour in Webster & Lai (2003). Alternatively, empirical measurements can be taken of the motivation of individual agents, as in the interviewing techniques of Berger et al. (2006), which were applied before creating a model of the adoption of innovations in developing world agriculture. Ultimately however, these approaches still cannot guarantee protection against such error.

What then is to be done? Perhaps qualitative error is best seen as an inevitable part of the rise and fall of scientific theories, which are progressively established and refined through making real-world predictions and checking them against actual data. It is this second component - the real-world data - which although not entirely absent, is perhaps underrepresented in the field of agent

modelling. While data plays a part in many of the studies cited above, the fraction used in these models arguably lags behind the vast information resources available to us in the information age of the 21st century. Additionally, it would seem that the use of data to calibrate existing models, rather than to inform the choice of model in the first place, is potentially a waste of a valuable resource. The approach taken here, then, is to temporarily abandon modelling in favour of exploring the data already collected.

Rather than using computers to execute a model - which is usually geared towards testing a particular hypothesis - processing power can instead be used to extract likely bases for future models from empirical measurement. This is achieved through exploratory analysis. Cliff & Miller (2006) provides an example in evolutionary economics, which, despite studying the output of yet another model rather than collected data, serves to illustrate exploratory analysis fairly well. The work uses a novel visualisation to verify the assumption implicit in many genetic algorithms, that the most recent evolved population is in fact the most optimal when compared to the entire simulation history: however, this visualisation also enables the discovery of a diverse array of patterns which may not have been previously suspected in such simulations.

On the other hand, an excellent example of extensive real-world data analysis used to inform future simulations is the migration work of Stillwell (2008) (and the technical report *Development of a migration model* 2002), in which a vast quantity of existing hypotheses and known data are reviewed with the ultimate aim of model creation. These studies also inherently make use of network data, mirroring the focal interest of the formal network literature discussed in section 1.2.3. However, this data is not analysed through explicit abstraction and formalism but instead, through a direct attempt to model a network phenomenon.

The current study, therefore, does not aim to directly produce models of human interaction, but simply to comprehend it better through the exploration of existing data sets in ways which have not been achieved before. Additionally, in order to address the rising recognition of the importance of network models, analysis of network interaction data will play a part. The techniques chosen will be general, and hopefully not limited to use in the domains in which they are initially applied. It is hoped that in future, more accurate models may be informed by any ensuing deductions, thus furthering the field of computational

social science; however, it may well be that the realities unveiled by exploratory analysis alone constitute useful results.

Finally however, it should still be remembered that the focus of this thesis is on the methods, rather than outcomes, of analysis.

1.3 Limitation in scope: the housing market of England and Wales 2000-2006

For the most part, this study concerns the housing market of England and Wales[4] during the years 2000-2006. Figure 1.1 shows the overall behaviour of the market before, during and after this period. Historically, the UK property market has tended to rise in the long term, but to cycle through phases of boom and bust in the short term. The period of this study was characterised entirely by a small segment of a rising 'boom' phase. This particular boom started during the mid nineties with house price increases in London and the South East, continued with rapid gains nationwide (which slowed in 2004 after the interest rate started climbing from its level of 3.75% for the first time since 2000), and ended in 2008 with what the media now calls a global recession caused by the sub-prime lending fiasco.

It can be assumed that any specific findings of this study, unless otherwise stated, apply only to the time period of the study.

1.4 Structure of the thesis

A mountaineer would be foolish to leave base camp without a map, therefore in the spirit of aiding the reader of this thesis, figure 1.2 provides a visual representation of its structure. The perspective view chosen, however, is not topographical: instead, a data-flow diagram is provided. A key is given in figure 1.3, which introduces four types of data: unprocessed lists of transactions, geo-spatial data (typically modelled as a two-dimensional entity), geo-spatial-temporal data (shown here as three-dimensional) and finally interaction data (which, being defined between a 2-d space of origins and a 2-d space of destinations, is four

[4]Scotland and Northern Ireland have separate land registries.

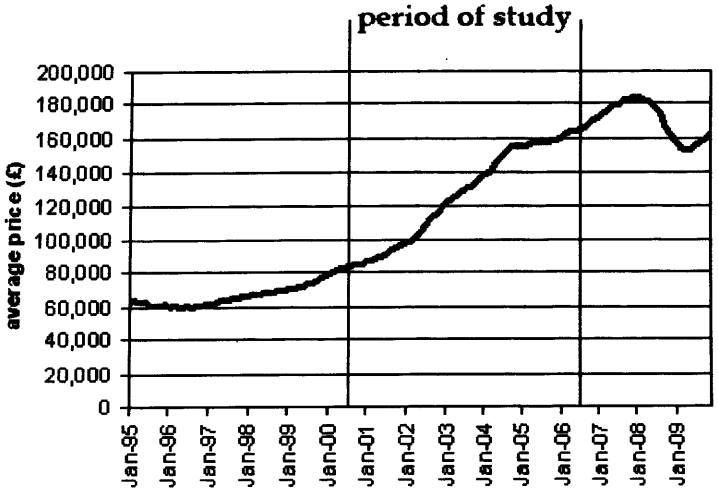


Figure 1.1: Average England and Wales house prices 1995-2010. Source: HMLR website

dimensional).

The remainder of the study is structured as follows.

Chapter 2 , to the left of figure 1.2, covers the first stage of analysis - the consideration of data issues and its spatial and temporal aggregation into the desired units of analysis. A novel technique is presented which enables increased accuracy in time aggregation of housing transaction data; this is useful for the detailed analysis conducted later on. The outputs from this stage are the housing, census aggregate and census interaction data used in the subsequent chapters.

Chapter 3 presents techniques used to visualise the data. Visualisation is a key part of this research, as it allows presentation to the human mind of a vast quantity of data while requiring minimal assumptions. Existing techniques for displaying interaction data via linearisation are incrementally improved through use of colour, interactive software, a better ordering algorithm and pre-processing of time series into cross-correlations. The techniques thus developed are also found to be useful in chapter 5.

Chapter 4 deals with the development of a general purpose, multivariate, multi-level regression engine to complement the visualisation techniques. It is tested on some of the simpler types of input data: the 2-d geospatial sets of

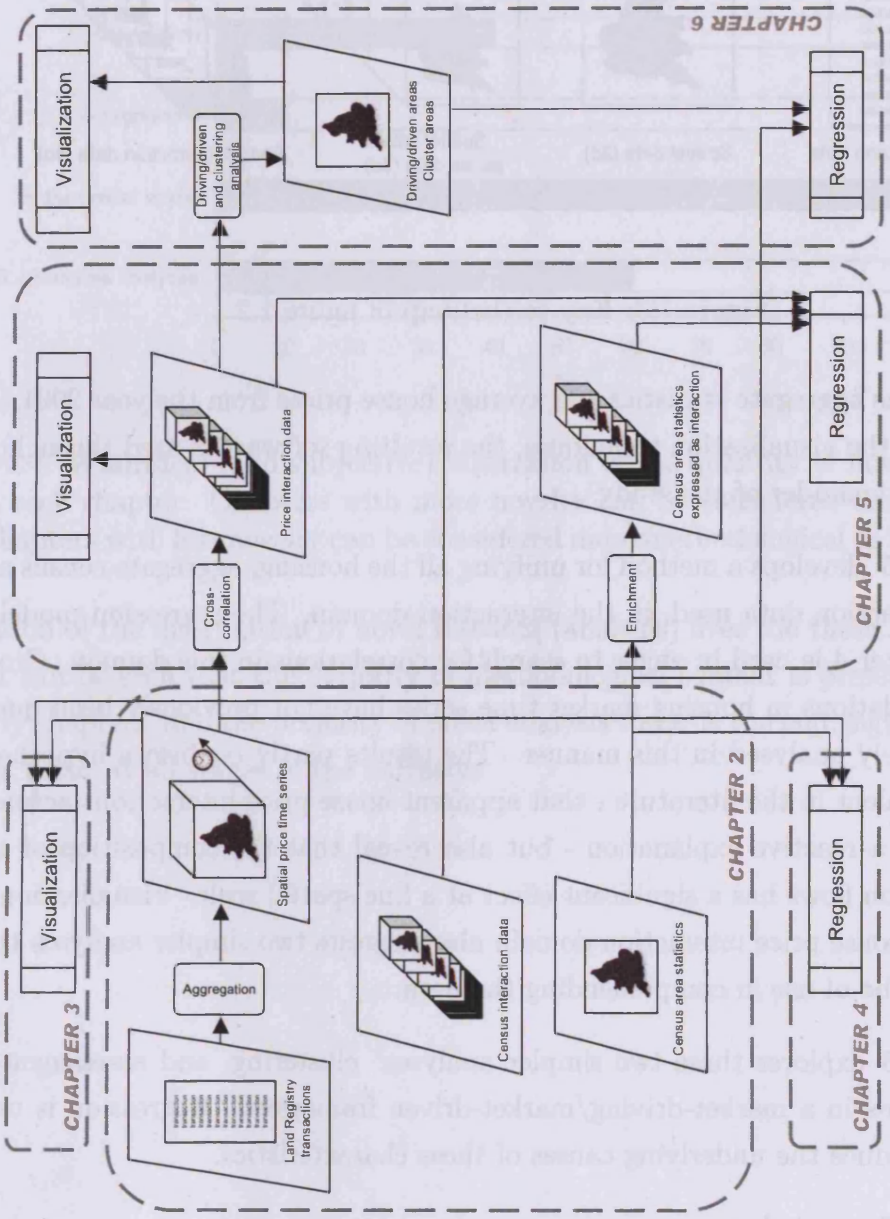


Figure 1.2: A map to assist in navigation of the thesis, based on data flow. See key in figure 1.3.

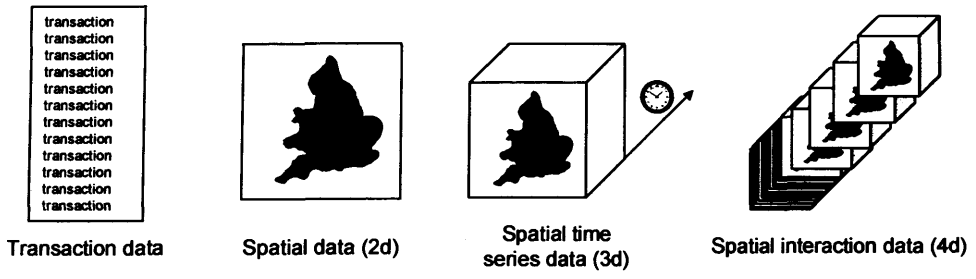


Figure 1.3: Key to the map of figure 1.2.

census aggregate statistics and average house prices from the year 2001. As with the visualisation techniques, the resulting software is used throughout the remainder of the study.

Chapter 5 develops a method for unifying all the housing, aggregate census and interaction data used, in the interaction domain. The regression model of chapter 4 is used in anger to search for correlations in this domain. Cross-correlations in housing market time series have not previously been qualitatively analysed in this manner. The results partly confirm a hypothesis prevalent in the literature - that apparent house price interactions actually have a reactive explanation - but also reveal that the composition of migration flows has a significant effect at a fine spatial scale. Visualisation of the house price interaction domain also suggests two simpler analyses that may be of use in comprehending the data.

Chapter 6 explores these two simpler analyses: clustering, and assessment of regions in a market-driving/market-driven framework. Regression is used to deduce the underlying causes of these characteristics.

Chapter 7 concludes.

It is common practice in research, to separate methodology from analysis, however for the purpose of this thesis, the distinction between the two is not rigidly maintained: this enables instead, the presentation of relevant topics together and in appropriate sequence. Figure 1.4 shows a somewhat tongue-in-cheek

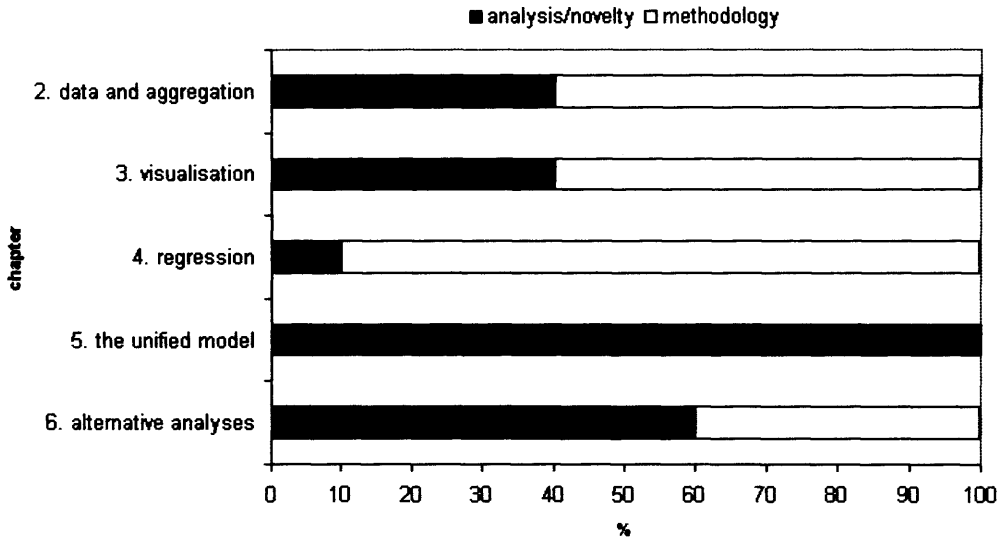


Figure 1.4: Whimsical (and subjective) illustration of the quantity of novel content in each chapter. Chapters with more novelty can be considered analytical while chapters with less novelty can be considered more methodological in nature.

illustration of the distribution of novel research (analysis) over the thesis. Thus, while it can be seen that the majority of methodological content is presented in the early chapters, and the majority of novel analysis towards the end, significant findings occur at all stages of the narrative.

Chapter 2

Data issues and aggregation

2.1 Introduction

This chapter discusses the data sets used in the thesis, and their aggregation into the units used for analysis. Such aggregation forms a significant part of the research methodology, and therefore merits in-depth discussion. Defined as the combining of a large number of data items into a smaller number of items, it is conducted here for three reasons: to increase data comprehensibility, to make computation more tractable, and to increase the accuracy of predictions. One might argue that all of scientific method is aggregation: the reduction of all our sensory inputs (a vast number of data items) to simple scientific models explaining those inputs (a much smaller number of data items). Certainly a thesis such as this could not exist without such a process - perhaps instead, a student would simply print out the entire Land Registry and Census datasets in their raw form, and submit them. If only it were that easy!

However, much though as researchers we may be thankful for any short-cuts to the completion of a project, if our long term aim is to better understand the world around us then aggregation is definitely one of those short cuts. Therefore any processes which reduce the size of a dataset, but maintain its comprehensibility and (so far as possible) its information content, are most welcome. The scope of this chapter is restricted to such processes.

This chapter therefore introduces the Land Registry and Census data sets used (section 2.2), and discusses their pre-processing - especially their aggregation both in space and time. Spatial aggregation is dealt with in section 2.3. Section 2.4

discusses time aggregation for the Land Registry data only (a more complex task meriting its own section). In the case of Census data, no time aggregation is in fact conducted; section 2.5 explains why this is the case. Section 2.6 concludes.

2.2 Introduction to the data sets

This section introduces each of the two data sets in turn.

2.2.1 The UK Census of 2001

A full census of the UK is conducted every 10 years, and this study makes use of two types of data from that carried out in 2001: aggregate statistics (which record a very wide range of statistics for each area), and interaction data (which records a more limited range of interaction statistics, such as the number of moving households or commuters, for each distinct *pair* of areas). These are both employed at three levels of aggregation: Output Area (OA), Ward and Local Authority.

2.2.1.1 Census Geographies

The smallest aggregation area is the OA, which is

> ...built from clusters of adjacent unit postcodes, but as they [reflect] the characteristics of the actual Census data ... [they are not] generated until after data processing. [OAs are] designed to have similar population sizes and be as socially homogeneous as possible (based on tenure of household and dwelling type) ... Urban/rural mixes [are] avoided where possible ...
>
> The minimum OA size is 40 resident households and 100 resident persons but the recommended size [is] rather larger at 125 households ... In total there are 175,434 OAs in England (165,665) and Wales (9,769). (*UK Census Geography description web page* (n.d.))

The next largest area of aggregation is the census ward, of which there are 8850 in England and Wales. The largest area of aggregation used in this study is the Local or Unitary authority (a political division) - of which England and

Wales contain 376. In chapter 3 some visualisations of the entire UK, including Scotland and Northern Ireland will be shown; in this case there are 426 local authorities present.

2.2.1.2 Issues of privacy

In order to preserve the privacy of individuals, the census office ensures that it is difficult to identify specific people in their statistics. Stillwell & Duke-Williams (2007) gives a good overview of the techniques used to achieve this. Besides pre-tabulation thresholding (only publishing values above a certain threshold) and record swapping (random exchanges of the data on individuals from similar areas), a Small Cell Adjustment Method (SCAM) is applied to interaction data.

SCAM operates by randomly adjusting table cells containing values of 1 or 2 to either 0 or 3 with probabilities defined as follows:

$$\begin{aligned} P(1 \rightarrow 0) &= 2/3 \\ P(1 \rightarrow 3) &= 1/3 \\ P(2 \rightarrow 0) &= 1/3 \\ P(2 \rightarrow 3) &= 2/3 \end{aligned} \tag{2.1}$$

The intention of this Monte-Carlo-style publication is that such small differences should cancel when aggregating over large numbers of areas, as the expected value of each cell remains unchanged by the random adjustments. In other words, if x is a value pre-adjustment, and x' its corresponding value which is published post-adjustment, then

$$\forall x.E(x') = x \tag{2.2}$$

Fortunately, the property of differences cancelling during aggregation also applies to the regression analyses conducted in chapters 4-6. However, it is possible that they will cause occasional errors to appear in the high-resolution visualisations of chapter 3. This cannot be avoided; however, in an era when personal privacy is increasingly eroded both by private corporations and in the name of national security, the census office should be applauded for at least making an effort to preserve it in their published statistics.

2.2.2 England and Wales Land Registry transactions 2000-2006

This is the set of housing property transactions collected by the England and Wales Land Registry (officially Her Majesty's Land Registry or HMLR). In the case of each transaction, the following information is available:

- Property sale value in pounds
- Sale Date
- Type of property: Detached, Terrace, Semi-Detached, Flat or Null
- Type of ownership: Leasehold, Freehold
- New build information: New, Not New, Null
- Address of the property
- Postcode of the property

The author has been provided only with the subset of transactions to which it was possible to assign a grid reference (geo-code), which constituted 89% of the original data set. Similarly, leaseholds are discarded from the data as information on the ground rent on all such properties is not available, and it is therefore hard to calculate a basis for comparison between leasehold and freehold prices. In both of these categories of transactions, flats are likely to be over-represented, as they tend to be sold on a leasehold basis and are not well handled in the OS ADDRESS-POINT data used for geocoding. The omission of such properties can be expected to bias the results of analyses conducted in chapters 4-6, in which the size of any effects relating to the 'flat' submarket will be underestimated.

2.3 Spatial aggregation

2.3.1 The need for spatial aggregation

Pre-aggregation of data is a prerequisite of some, but not all, computations carried out during this thesis. In particular, it is essential for the cross-correlation analysis to be carried out in chapter 5. Aggregation is also needed at some stage of the visualisation process of chapter 3 (though not necessarily beforehand). Aggregation is *not* needed for the regression models of chapter 4 per se. However,

in order to produce directly comparable results from all stages of the study, a common aggregation step is carried out prior to the analysis.

2.3.2 Choice of the appropriate scale for spatial aggregation

Part of the purpose of this study is to make use of extensive datasets which have not been combined before, so to fully exploit the large quantity of data, it is proposed that analysis is carried out at the maximum sensible resolution. But what does *sensible* mean? The following points are proposed:

1. we must consider the resolution of data available, both from the Census office and the Land Registry.

   The Land Registry data set is the more spatially accurate of the two, providing individual addresses and postcodes, and is therefore not a limiting factor in this respect.

   The census data is available on three levels: either at local authority, ward or output area (OA) level. The latter equates approximately to postcode areas, therefore this represents the highest resolution available from the data. However, not much information of interest is provided at OA level: for example, when looking at interaction data, we are told the number of migrants and commuters between each pair of OAs but not much else. Conversely if the resolution of interest is decreased to ward level, we are told the age and sex of migrants, and their occupational classifications, among other information.

2. we must consider at what level useful patterns will appear in the census data. Existing research shows that structure is visible at all levels: output area (Propper et al. 2005), wards (Titheridge & Hall 2006), London boroughs (Congdon 2006), and counties Boyle (1993). In the case of Titheridge & Hall (2006) and Boyle (1993) these patterns emerge in interaction data as well as census area statistics. Multi level modelling approaches have also been used to find meaningful patterns, e.g. Manley et al. (2006). Overall therefore, it seems reasonable to assume that patterns of interest can emerge at all levels.

3. we must consider at what level useful patterns will appear in the Land Registry data. This may well be the limiting factor in the choice of spatial resolution. For example, a typical Cardiff output area has approximately 30 property transactions recorded since the year 2000 - an average of less than 4 per year. This is unlikely to provide enough information to produce even a smooth annual price index (let alone a monthly index) for the output area. Therefore, some spatial aggregation of Land Registry data will be required.

4. we must consider the needs of the cross-correlation algorithm will be run on the output of aggregation. Computing cross-correlations for every pair of the 8850 census wards in England and Wales takes several hours. As this time is proportional to the square of the number of areas studied, practical considerations dictate that working at a finer granularity than the ward level would be inconvenient at best and impossible in the worst case.

It should be noted that it is not essential to work with only the areal units defined in the data sets, and that custom units could be defined at whatever level is deigned to be most appropriate. This could be performed by some kind of automated clustering process. However, this option has been rejected in favour of using established boundary data, as this data itself provides some socioeconomic information about the represented space. Section 2.3.3 will discuss this point in greater detail.

Balancing the considerations listed above, areal units approximately the size of (i) wards and (ii) local authorities, are deigned to be appropriate for study. As the census office provide data pre-aggregated to units of census wards and local/unitary authorities from the year 2001, these are the units of analysis chosen for the remainder of this work.[1]

2.3.3 Choice of method for spatial aggregation

A variety of methods are available for spatial aggregation, principally simple averaging, kriging and clustering.

[1]Smaller areas will in fact be used to increase accuracy when constructing then index (to be discussed in section 2.4) however the analysis will thereafter be conducted at ward level or above.

1. Simple averaging is the most obvious method for calculating an aggregate area statistic from given point data. Likewise, simple assignment is the most obvious method for calculating point data from a given aggregate statistic. This is used in e.g. Orford (1999) where Housing Condition Survey data, collected at a sub-street level, is assigned to individual properties.

2. Kriging covers a wide array of more sophisticated techniques used to estimate the value of unknown point data by interpolation from known neighbouring points. This can be extended to calculate statistics for areas based on the points within them, perhaps reflecting the underlying unmeasured data better than a simple average (Isaaks & Srivastava 1989).

3. Clustering processes group data in space based on similarity to neighbours, as in Slater (1976) or Bourassa et al. (1999, in Meen 2001). In the context of this study, for example, three neighbouring OAs containing a similar type of housing and resident could be clustered together. The similarity criterion - a measure of how similar two areas have to be, in order to be combined - would be varied at will, which would have the effect of changing the spatial resolution. Meen (2001) notes that areas constructed by these methods seldom match Local Authority areas. To call this a problem would be missing the point of the clustering process, the very aim of which is to define areas derived from the data itself. However, such aggregation is likely to complicate comparisons between different data sets, and comprehension of the results.

For the purposes of this study, the option of simple averaging was chosen. This is because, as noted in Isaaks & Srivastava (1989),

> "estimation requires a model of how the phenomenon behaves at locations where it has not been sampled; without a model, one has only the sample data and no inferences can be made about the unknown values at locations that were not sampled." (Chapter 9)

For geological data, it may be reasonable to assume some kind of spatial continuity and therefore apply kriging; however in the case of sociological data, the underlying unmeasured locations may change rapidly, e.g. in the case of

a railway line dividing rich and poor areas of a city. Fortunately, the results of published work on the spatial distribution of social divisions is already available in the form of the Census output area geographies, which at least in urban areas are deliberately chosen so as to each contain, so far as possible, a consistent type of housing (Orford & Radcliffe 2007). Therefore, if a model is to be chosen, it is not unreasonable to assume homogeneity among urban output areas (Rural areas are by their nature, far more varied, and in this study the problem of rural variability remains unaddressed; however it should be noted that approximately 80% of the UK population resides in an urban area). If each area, then, is assumed to be homogeneous then the appropriate technique for estimating an area statistic is a simple average of all known points within that area.

Simple averaging carries an additional advantage, over clustering, of making results easier to interpret and communicate: it is easier, for example, for the reader to relate to results concerning "output areas where the population exceeds N", compared to results concerning "areas where the special density function used for this study exceeds N".

Finally it should be noted that broadly speaking, kriging is a form of averaging with additional data smoothing. Smoothing of spatial data will be carried out in subsequent computations - the processes that were listed in section 2.3.1 all contain a spatial smoothing function already. Regression is itself a form of smoothing, fitting a straight or curved line to a number of discrete points. Visualisation leaves the process of smoothing data to the human eye and brain, and while computing cross-correlations requires no spatial smoothing, all results from the process will be either visualised or used for regression. Additional smoothing stages may therefore be unnecessary.

2.4 Temporal aggregation of Land Registry Data

With spatial data, the main question to answer was at which level to conduct aggregation, and a secondary question was how to produce data at this spatial level. In the dimension of time, the importance of these questions is reversed. Section 2.4.1 discusses the need for temporal aggregation of data, and section 2.4.2 discusses the choice of time resolution - the easy question - while the remaining

sections deal with how to aggregate data appropriately: section 2.4.3 reviews existing techniques, section 2.4.4 proposes a new solution to the problem, section 2.4.5 deals with testing the validity of this solution and section 2.4.6 concludes.

2.4.1 The need for temporal aggregation

Aggregation of data in time is needed for two of the three computations mentioned in section 2.3.1:

1. Data visualisation. As there are many more transactions in the data set than pixels on the average computer display, a single pixel of computer display must represent a number of individual transactions, and aggregation is therefore needed at some point in the visualisation process.

2. Computation of cross-correlations between housing market areas. It is the nature of the cross-correlation algorithm that it takes a time series, rather than individual point data, as input; therefore aggregation of individual data points into a time series is needed.

2.4.2 Choice of time resolution

The question of what time resolution to use is refreshingly simple: as studying fine grained house price data in this manner is unprecedented, it is not known at which level patterns will emerge, so all possible levels will be studied - within reason.

Common sense would suggest that for movements in the property market, the lower bound of 'within reason' cannot be much less than a month as it typically takes this long to complete a property transaction. Section 2.4.6.1 will present empirical evidence which points towards sensible lower limits of time resolution being in the region of 20 days; however, accuracy is decreased at this level, therefore section 2.4.6.2 considers the selection of an optimal unit of time resolution.

The upper bound of time resolution that can be studied, meanwhile, is fixed by the six year timespan of the available data.

2.4.3 Existing methods for producing house price indices

2.4.3.1 Simple Averaging

The most simple and obvious method for collating Land Registry transaction data into a time-price index for a given area, is to divide the time period of interest into a number of *slices*, and then for each slice to compute the average sale price of transactions that occurred during that time slice and in that area.

$$index^t_{AREA} = \frac{1}{|AREA^t|} \sum_{trans \in AREA^t} \text{transaction price} \tag{2.3}$$

where $AREA^t$ is the set of transactions that took place in the area of concern during time slice t. Alternatively, as in Meese & Wallace (1997), the median can be used instead of the mean, thus removing noise caused by outliers - particularly expensive or cheap houses which are unlikely to be representative of the market overall:

$$index^t_{AREA} = \text{median}(\text{transaction price for } trans \in AREA^t) \tag{2.4}$$

However, both of these methods suffer from the *property basket* problem, where the index price in each time slice is subject to variation based on the 'shopping basket' of properties which happen to sell during that time slice. One approach to this problem is to create a *mix-adjusted index*, where average prices for each type of housing are computed separately and later combined into an index. Another approach is to employ hedonic modelling.

2.4.3.2 Hedonic modelling

Hedonic modelling (Rosen 1974) makes use of regression to estimate a wide variety of parameters which may affect the price of a property. Predictor variables may include for example, property-specific factors such as the floor area, number of bedrooms or access to off street parking; or neighbourhood-related factors such as the proximity of schools, parks or transport links. Thus, rather than trying to explain the value of each particular sale, it is possible to remove property-specific factors and analyse instead the changing values over time of the premiums associated with particular spatial regions.

Unfortunately, such an approach is not possible using Land Registry data

Areal/Time unit	Median number of transactions	Percentage with no data	Percentage with no repeat sales
Ward, 5 day	1	43	69
Ward, 10 day	2	24	51
Ward, 20 day	4	9.8	32
Ward, 30 day	6	5.1	22
Ward, 90 day	18	0.44	5.8
LA, 5 day	25	1.7	3.6

Table 2.1: Volume of Land Registry data available, per space-time slice. Percentages are calculated as a proportion of slices from all areal units for which any data exists at all during the six year period of the study.

alone, as even the most basic variables necessary for a hedonic model (such as floor area) are not present in the data. While it would be possible to gather such data independently through a variety of means, to do so for all six million transactions in the data set would constitute another project in its own right so this approach has not been employed here.

2.4.3.3 Repeat price indices

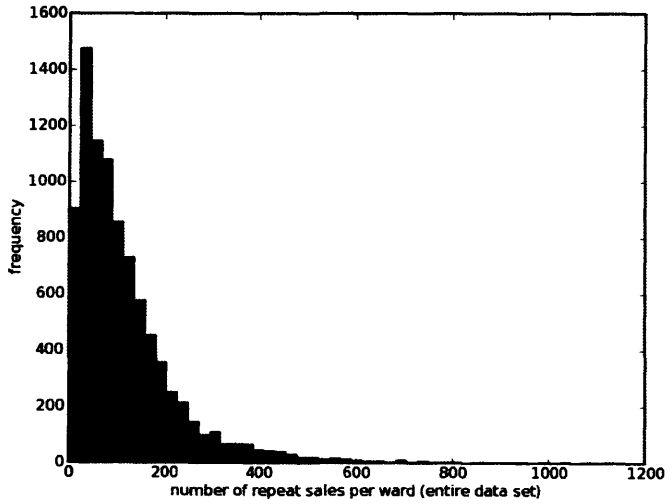


Figure 2.1: Histogram showing number of repeat sales per ward, over the entire data set

Another approach is to generate a *repeat price index* from the sales of houses which have sold more than once during the period under consideration (see Meese & Wallace (1997) or Meen (2001)). Thus, market growth can be estimated from the change in value of a single property between two subsequent sales of that property; regression is then used to combine this information from each property that sold more than once. It is assumed that the property has the same hedonic value each time it sells; obviously this is not always true due to changes to the property itself or to its local environment, but the assumption is nonetheless considered reasonable. It is certainly a better assumption than the one implicit in taking a simple average: that the shopping baskets of houses which sell each month have the same hedonic value. Meese does, however, note that repeat price indices "suffer from sample selection bias and inconstancy of implicit housing characteristic prices, and they are quite sensitive to small sample problems".

2.4.3.4 Discussion

As noted above, it was simply not possible to employ a hedonic index in this study, due to lack of data on housing attributes. Meese & Wallace (1997), which provides a useful comparison of the techniques described here, notes that missed attributes can cause major problems in hedonic index estimation.

Meanwhile, constructing a repeat price index entails throwing away five sixths of all the data available, which intuitively seems wasteful, particularly when attempting to study short term dynamics in small areas for which little data is available to start with. Figure 2.1 shows the distribution of repeat sales over wards. Table 2.1 shows the median number of transactions per ward per month as 6; the corresponding figure for repeat sales is 1. While no transactions whatsoever occur during 5.1% of ward-months, in the case of repeat sales, no data is available 22% of the time. While this may be a fairly accurate technique, then, for producing a price index in a high volume sales area, it is not sufficient for the study of low volume areas. Repeat price techniques are therefore not employed to generate price indices in this study, however they are used in section 2.4.5 for benchmarking purposes.

Meese concludes with the suggestion that "researchers interested in regional real estate cycles, or just a reliable estimate of the local trend in housing price movements can rely on the simple median sales price index". However, a tech-

nique is proposed in the next section which - without restricting the data used to repeat sales, or requiring detailed housing attribute data - can improve on this.

2.4.4 Methodology for aggregating data in time

This section proposes a new approach which makes use of census boundary data to approximate a hedonic model. For the purpose of the census, the UK is divided into approximately 220,000 *output areas* (henceforth OAs) which have been chosen partially with the goal of each OA representing a fairly consistent type of housing (Orford & Radcliffe 2007). To make use of this, it is explicitly assumed that all houses within the OA have exactly the same hedonic value. Like the other index generating assumptions mentioned above, this is of course not true - especially in rural areas where neighbouring properties tend to differ more than on a city street. However, as was the case with the repeat price index, it is perhaps better than the assumption used in taking a simple average (equation 2.3), that the monthly 'shopping basket' of properties available for sale always contain properties of exactly the same hedonic value. Assuming that OAs are internally homogeneous is preferable to assuming that larger areas are internally homogeneous.

Therefore, a price index is constructed for each OA using equation 2.3. We then compute a log relative index for each OA:

$$\log \text{relative index}_{OA}^{t} = \log \frac{index_{OA}^{t}}{index_{OA}^{0}} \tag{2.5}$$

where $index_{OA}^{t}$ is the value of the absolute index for district OA at time t, and $index_{OA}^{0}$ is the value of the absolute index for district OA at time zero (the first time slice in the study). This approximates a hedonic model because normalising each OA - i.e. studying only the price of housing relative to the average price in the OA at the start of the period under study - has the effect of removing the effect on prices of the average housing characteristics of that OA, at least for those housing characteristics whose relative value remains unchanged over the period of study.

Log price indices are not new, and are often appropriate for models of return on investment (Tsay 2002). In this case there are three reasons for converting to log indices:

	Relative return factor	Percentage profit	$\log_2$ relative return
A	0.5	-50	-1.00
B	1	0	0.00
C	2	100	1.00
D	2.1	110	1.07
E	11	1000	3.46
F	11.1	1010	3.47

Table 2.2: Thinking of financial returns on a logarithmic scale

1. Intuitively, human beings tend to think of the value of relative financial change in logarithmic terms. Table 2.2 gives some examples. Index A moves from 1 to 0.5 while C moves from 1 to 2: these movements are considered to be of equal magnitude but in opposite directions, and their log values of -1 and 1 respectively reflect this. B has a return exactly equal to the initial stake - no profit or loss - and hence a log value of 0. And the differences between lower returns are considered more significant than the differences between higher returns: the difference in log values of a 100% and 110% return (lines C and D) is 0.07, while the log difference between a 1000% and 1010% return (lines E and F) is 0.01.

2. Not only does our intuition match logarithmic valuing, but this also extends to the averaging processes. For example if lines A and C of table 2.2 correspond to two areas of a city, one doubling in value while the other halves in value, we would like to report no average change in value overall. With linear factors or percentage returns, the average of 0.5 and 2 is 1.25 which is not what we would like to report. Using a logarithmic scale however, the average of -1 and 1 is 0.

3. Logarithmic indices simplify the process of normalising returns over time. For example, calculating the annual equivalent of a 50% return over five years necessitates solving $(1+x)^5 = 1.5$, whereas with a logarithmic return, simple division can be used.

Finally the indices for OAs are combined into indices for larger areas using

an arithmetic mean:

$$\text{log relative index}^t_{AREA} = \frac{1}{|AREA|} \sum_{OA \in AREA} (\text{log relative index}^t_{OA}) \tag{2.6}$$

where $AREA$ is the set of census output areas OA in the area of concern, and $\text{log relative index}^t_{OA}$ is the value of the log relative price index for district OA at time t. It is then possible to convert from log indices back to normal relative indices:

$$\text{relative index}^t_{AREA} = \exp(\text{log relative index}^t_{AREA}) \tag{2.7}$$

It should be noted that mathematically, the logarithms and exponential can cancel, leaving a geometric mean over output area price indices:

$$\text{relative index}^t_{AREA} = \sqrt[|AREA|]{\prod_{OA \in AREA} \text{relative index}^t_{OA}} \tag{2.8}$$

However, for the reasons noted above, logarithmic indices are used for the rest of this study, so the calculation of indices ends with equation 2.6.

Note that it might be a cause for concern if the first timeslice in which transactions occur contains unusually high or low prices. However, the alternative to normalising by the first timeslice is normalising by an average of several slices, and this option has been discarded for two reasons. Firstly, it will be easier to comprehend relative price indices if they all start with a value of 1. Secondly, for the purposes of this study, the indices are mainly used in first derivative form - showing inter-interval price change rather than absolute price - so the actual initial value of the time series is irrelevant.

It would also be possible to smooth the resulting time series to remove 'noise' caused by the property basket effect at short time scales. This has not been done for three reasons. Firstly, it is not possible to tell signal from noise: how can we distinguish which changes in price are due to the random selection of properties which happen to be sold, as opposed to the underlying short term market changes which are the object of this study? Secondly, Tsay (2002) notes that smoothing of time series introduces false autocorrelation to the data. This would be of concern considering that the generated time series are to be used

in a cross-correlation analysis.[2] Finally, as with the spatially aggregated data, smoothing will be applied at a later date anyway - in visualisation, regression, or integration of a cross-correlation function - all of which amount to smoothing in one way or another.

2.4.4.1 Summary of index generating method

The indices are therefore constructed as follows:

1. leaseholds are removed from the index, as the property rent and duration of lease will have a significant impact on the transaction price. These factors are not included in the Land Registry data, so it is difficult to deduce much about underlying market value from any given leasehold transaction price.

2. regardless of the target spatial resolution of the index, transactions are assigned to census output areas (OAs).

3. for each timeslice and each OA, an average of all sale prices is taken to form an absolute index.

4. a relative price index is created from each absolute index, i.e. all index values are computed relative to the absolute index price of the first timeslice in which transactions occur

5. for OAs with few transactions, all time slices where no transactions occur are assigned the same index value as in the previous time slice (i.e. there is an assumption of no market change). All index values prior to the occurrence of the first transaction in the OA are also assigned a value of 1.0.

6. for each area unit of the target spatial resolution (either a census ward or Local/Unitary Authority) a price index is constructed which is the geometric mean of the indices of smaller units within it.

[2]Unfortunately, the process described here will itself generate some degree of false autocorrelation due to the technique used to estimate a price for areas with no data. However the author sees no reason to further exacerbate the problem!

2.4.5 Testing of index accuracy

The price indices developed in this chapter are to be used extensively in the remainder of the thesis, to produce visualisations (chapter 3), simple models of price growth (chapter 4) and more complex models based on inter-area time series correlation (chapters 5-6). They are therefore tested to ensure that they represent the raw transaction data in a reasonably accurate manner. Such tests can also be viewed to some extent as a test of the implicit ecological assumption inherent in indexing, that the price of each house within a certain area tends to behave in the same way as prices in the area on average.

To gauge the accuracy of the indexing process, therefore, some indices produced by it are compared with data derived from repeat sales: when a property has sold more than once during the period of the study, the rise in value of that property between subsequent sales is checked against the rise in the constructed index (which, for the purposes of the test, is constructed without using data from repeat sales; though these are included in the indices used in later chapters). As noted in section 2.4.3.3, repeat price indices - and hence all comparisons based on repeat sales - are subject to their own problems. However, they are used here only as a benchmark against which to compare other techniques, so it is their consistency rather than absolute accuracy which is of use. In particular, it is not necessary to define any spatial or temporal scale to examine a pair of repeat prices, so they are useful for comparing other indices in which the spatial or temporal granularity is varied. Additionally, any systematic difference between the behaviour of repeat price pairs and the behaviour of averaging indices - such as the fact that repeat sales suffer from sample selection bias, and their hedonic value tends to vary over time - is likely to affect the benchmarking of all averaging indices equally, rather than favouring one over another. In sum, benchmarking averaging indices against repeat price pairs may systematically overestimate the errors in the tested indices, however the relative differences in the benchmark scores of the tested indices are likely to be truly indicative of their relative performance.

For each agglomerated index, an Annual Mean Squared Error measure (AMSE) is defined, which gives an indication of the expected divergence of the

Parameter	Values tested
Time slice length (days)	5,10,15,20,25,30,60,90,120,150,180,210,240,270
Spatial scale	Local Authority, ward
Index construction method	Simple averaging, New technique with OA data

Table 2.3: Characteristics of indices tested

constructed index from repeat price data *per year*:

$$\text{AMSE} = \frac{1}{|\text{RPP}|} \sum_{t1,t2 \in \text{RPP}} \left(\frac{\log \frac{ind(t2)}{ind(t1)} - \log \frac{price(t2)}{price(t1)}}{\Delta T_{t1}^{t2}} \right)^2 \tag{2.9}$$

where RPP is the set of all repeat price pairs, $t1$ and $t2$ are a pair of transactions relating to the same dwelling, $price(t1)$ and $price(t2)$ are the prices of these transactions, $ind(t1)$ and $ind(t2)$ are the agglomerate index values for the same place and time as the transactions, and ΔT_{t1}^{t2} is the time difference between the transactions in years. Note that the set of all repeat price pairs can include individual prices more than once - for example if a given property sells three times at prices A, B and C, the price pairs AB, AC and BC are all incorporated in RPP and hence used to test the index. Again, log price indices are employed, hence, the actual expected divergence of the constructed index from a repeat price pair per year, is:

$$\text{AMSE expressed as ratio } = e^{\sqrt{AMSE}} \tag{2.10}$$

A total of 56 different agglomerate price indices - covering all possible combinations of the parameters in table 2.3 - are tested against all 1,035,097 repeat sales price pairs present in the data set. Each index is either a *simple average* (as described in equation 2.3) which does not use OA information, or is derived *using OA information* as described in equation 2.6. Indices are measured both at Local Authority (LA) and ward level; and the length of the time slice for aggregation is varied between 5 and 270 days to gain a picture of how index accuracy changes with increasing time resolution.

A pertinent question during testing is whether or not to include the test data (repeat price pairs) in the training data set (i.e. the data used to generate the index in the first place). Statistical convention would dictate that the two data

sets be kept separate, especially as in this case, when dealing with a short time slice and a small area, the only data present in the training set will be the test data itself - a single repeat price pair. On the other hand, *excluding* repeat price pairs from the training set risks including a systematic bias relating to the characteristics of housing which frequently changes ownership. The approach used is to present results without repeat sales included in the training data, although the AMSE values thus computed are assumed to be overestimations of the true error inherent in the indexing process.

2.4.6 Discussion

Results are shown in figure 2.2 and table 2.4.

2.4.6.1 Summary of results

Inspection of figure 2.2 and table 2.4 shows that the gains from the new indexing technique are small on long time scales (less than 3%) but large on shorter time scales: when using a time slice of five days, the index error is decreased by 20% at LA level and 51% at ward level.[3] Note that on short time scales, it is debatable whether or not the ward level price index is meaningful: table 2.1 shows that 43% of space-time slices contain no ward level data at all in the average month, therefore these data points will be filled in by copying values from previous months, introducing some degree of false autocorrelation to the data. However, if even one transaction exists within a space-time slice, then knowledge of the output area (and hence assumed housing type) to which it refers, endows that transaction with some degree of meaning. Therefore, if the time scale is increased to 20 days - making data available for over 90% of space-time slices at ward level - the new technique still gives accuracy gains of 34%.

One statistic not visible in the results as presented - due to their use of a mean squared error estimate - is the systematic bias in the agglomerate indices. In the case of the LA level 5 day index, for example, the Annual Mean Error expressed as a ratio is 0.92, as compared to the Annual Mean Squared Error which expressed as a ratio is 1.37. Thus on average, the index price tends to underestimate the gain

[3]These percentages are calculated from table 2.4, however note that as a divergence ratio of unity represents complete accuracy, percentage differences are calculated only on the component of the ratio exceeding 1, i.e. the percentage difference between ratios a and b is $100\frac{(a-1)-(b-1)}{a-1}$.

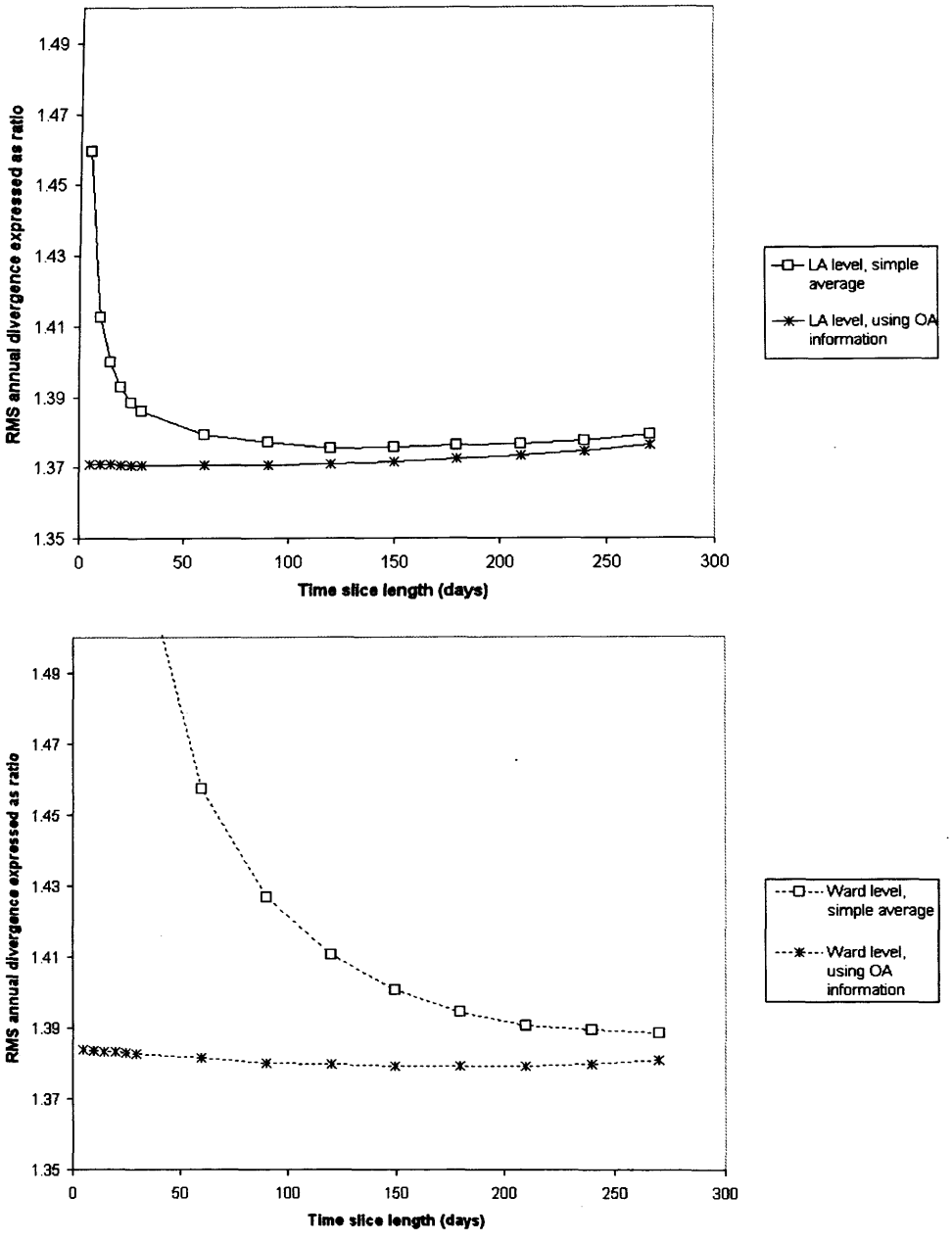


Figure 2.2: Plot of errors in new aggregated price index, as compared to errors from simple averaging index. The baseline chosen for comparison is repeat sale price pairs. AMSE is assumed to be a slight overestimation of true error, as the test data (repeat sales) are not included in indexing data.

Space/time scale	Divergence ratio (new index)	Divergence ratio (simple index)
LA level, 5 day (shortest tested/sensible)	1.37	1.46
Ward level, 5 day (shortest tested)	1.38	1.77
Ward level, 20 day (shortest sensible)	1.38	1.58
LA level, 60 day (optimal)	1.37	1.38
Ward level, 150 day (optimal)	1.38	1.40
LA level, 270 day (longest tested)	1.38	1.38
Ward level, 270 day (longest tested)	1.38	1.39

Table 2.4: Selection of results from testing generated indices against repeat price pairs

in value of multiply sold properties by a factor of $\frac{1}{0.92} = 1.09$. Comparison of this figure with the AMSE of 1.37 reveals that prices are not always underestimated, but that there is a slight tendency for underestimation to occur more frequently or more drastically than overestimation.

This could be a systematic error introduced by the fact that during the course of this study the market was generally rising, so perhaps in a falling market prices would be overestimated instead. However it is more likely to relate to the fundamental limits of repeat price accuracy (i.e. the fact that the hedonic value of properties tends to change, in particular to increase, between subsequent sales).

Determining the reason for this divergence with greater certainty would require comparison with a hedonic index.

2.4.6.2 Choice of optimal space/time aggregation units

It is interesting to ask which of the indices presented is the most accurate - i.e. given a property with a known price at some point in the past, which index can best be used to estimate its current price? This could be called the optimal index, and from figure 2.2 it is clear that optimal accuracy is obtained using the new indexing technique, at a Local Authority level, and with a 60 day time slice. However, despite this index being the most accurate index generated during this study, it is not necessarily the best to use in all cases.

Figure 2.2 shows that Local Authority level indices always give a better prediction of price gain than ward level indices. This is presumably because of the

greater number of properties present in the LA, which reduces the 'shopping basket' problem described in section 2.4.3.1; it must be the case that the advantages of eliminating that problem outweigh those gained by specifying a property's location more precisely when trying to estimate changes in its price. Notwithstanding, it is still relevant to study ward level dynamics, because the census data gives lots of interesting information about the constitution of wards and the relationships between them, and something can be gained from correlating this information with ward level price movements. Therefore, ward level indices are still employed as well as LA level indices. In the case of ward level indices, optimal accuracy is obtained with a 150 day time slice.

2.4.6.3 Conclusions

The new indexing techniques which use OA level information always exhibit better performance than their counterpart simple averaging indices. This is because, as expected, they mitigate the 'shopping basket' problem; and as expected this performance increase becomes much more significant at short time scales (figure 2.2), where the smaller number of properties used to construct the index becomes more susceptible to this problem.

It may be possible to further improve this process by using a weighted average based on the number of properties in each OA, rather than a simple average across OAs. Division into submarkets based on the housing type data provided by the Land Registry (*house* or *flat*) - thereby incorporating mix-adjusted index techniques - may also increase accuracy.

The results appear to indicate that time slices as small as 20 days can be used to study market dynamics at ward level, with little penalty to index accuracy. This may initially seem surprising given that housing transactions themselves usually take longer than this to complete, due to the legal complications of conveyancing. However, it is not clear how this time constraint should have an effect on short term market dynamics - in any case these are likely to be influenced by longer term macroeconomic trends, as well as local price fluctuations, so there is no reason to expect unstable behaviour over a shorter period *per se*. An interesting question to be addressed over the course of the thesis, is whether the short term dynamics thus revealed are simply the same as long term dynamics with extra measurement noise, or whether they follow patterns of their own.

One notable effect of relative, rather than absolute indexing is to remove the bias incurred by expensive properties. If sales in a given OA rise from around £1,000,000 to around £1,100,000, this will have the same effect on the index as another OA increasing its sale prices from £100,000 to £110,000. This is considered desirable for unearthing underlying market changes, however it should be noted that the indices do *not* directly reflect the total value invested in any given region of the spatial market.

For the remainder of this study, therefore, relative indices covering a variety of different time resolutions will be used to explore spatial market structure on different temporal and spatial scales. Absolute price indices will only be used when assessing absolute value, and the use of absolute indices will be restricted to long time scales to mitigate the 'shopping basket' problem. It can be assumed from testing the relative indices, on the other hand, that they are valid on all time scales greater than 20 days.

2.5 Temporal aggregation for Census data (or the lack thereof)

For the purposes of this study, time aggregation of census data is not needed because all census data used is taken from the year 2001 (the most recent full census of the UK). As such, the measurement is valid at one point in time only, and the statistics are assumed to remain static over subsequent years (2001-2006).

The reasons for this assumption are pragmatic. While there are some data available from the Census Bureau, and other sources, for the period from 2001 to 2006, they are not so comprehensive as the full 2001 census, and in any case considerable effort would be required to integrate them into comparable formats. The immediate aim of this study is to explore links between the full Land Registry data set, and the Census data set, at a level of detail which has not been conducted before; therefore it was decided to focus the study on this novelty rather than on incorporating smaller quantities of data from subsequent years.

It is realistic to expect that some statistics measured in the year 2001 are 'unstable' and will have changed during subsequent years, while others will have remained constant. However, note that the algorithms used in this study are designed to search for trends and correlations over the complete length of the

Land Registry data set. Therefore, as any trends discovered will apply to all years 2000-2006, there is a likelihood that any major trends will relate to statistics that remained fairly static over the entire 2001-2006 time period. Also, as any global trends should be interpreted on a case by case basis, later inspection of results provides a 'safety net' which may allow detection of those trends which, despite being stable themselves, relate to unstable statistics.

2.6 Summary

This chapter has described the data used during this study, and reviewed existing methods suitable for its aggregation in both space and time. Two spatial scales (those of the Local Authority and ward) have been chosen for further study; data will be aggregated to these scales by simple assignment.

A new technique has been proposed for aggregating house price data in time, which uses census boundary data to approximate a hedonic model. This has been shown to outperform a simple average, especially over short time scales. A set of indices has been produced which allows for the study of market dynamics for a number of different time scales between 20 and 270 days. Chapter 5 will discuss whether the short term price movements thus uncovered reveal any novel information relating to market dynamics.

Chapter 3

Visualisation of large datasets

"All you touch and all you see is all your life will ever be"

(Pink Floyd)

3.1 Introduction

From an early point in our scientific education, we are taught to double check both the outcomes of our measurements and the answers to our sums, if not by repeat calculation then by using common sense. If we deduce that the height of a mountain is 30cm, the speed of sound is 3km/h or that the age of the Earth is 6,000 years then we raise our eyebrows for a moment and then go back to see whether or not we made a mistake. Of course, doing so does not guarantee a correct answer, but it does catch a certain quantity of errors that would otherwise have slipped through the net.

The same applies to the data processing steps conducted in this thesis - it is vitally important to glance over both the inputs and outputs to each stage and check the figures for sanity. However, the quantity of data involved in both the measurement and calculation is somewhat larger than in the examples above. The Census data set is shipped out by the National Statistics office on a stack of CDs and DVDs almost as tall as the average desktop workstation, and there is no way that so many figures can be taken in with a quick glance. This chapter concerns visualisation techniques for distilling such large quantities of data into a form in which we can hope to quickly understand it. If that understanding is not precise and complete, then it should at least be adequate for the purpose of

broad comprehension.

As stated in chapter 1, a particular focus of this research is *interaction* datasets, such as the network of migration movements that extends over the UK. Before embarking on any detailed study, the author was curious as to what the data *looked* like. For example, one might ask the questions, where do people move to, and where do they move from? But printing out a list of the top 100 places for in- and out-migration respectively would not reveal anything about the nature of *interaction* between those places. Analysis techniques such as these only answer a specific and limited question, whereas what is really needed is to remove the metaphorical blinkers from our eyes and ask, 'What is the big picture?'

We could of course, in the above case, draw a flow map with arrows of varying thickness connecting every origin and destination. With the quantities of data involved however, the numerous arrows required would completely obscure the map and each other, unless some kind of thresholding process were applied to filter out all but the largest flows of migrants (as in Bertin 1984, page 350) or other methods taken to improve readability such as the computation of line densities (as in Rae 2009). This, however, would miss patterns in the smaller flows of migrants, which may well still be of great significance. This chapter introduces techniques which do not require thresholding of data to remove the smaller interactions. In fact, quite the reverse is true: the logarithm of data values is often computed, in order to 'tame' the larger data values and prevent them from obscuring the smaller ones.

This chapter deals with the development of techniques used to visualise the large data sets used in this study. These are used both as an end in themselves, to understand the contents of specific data sets in isolation, and also as a tool for understanding the output of computations in later chapters, such as data sets of cross-correlations and regression residuals. The remainder of the chapter is structured as follows. Section 3.2 surveys background literature on exploratory data analysis and visualisation. Section 3.3 discusses the research philosophy behind the visualisation techniques employed in this study. Sections 3.4 to 3.6 discuss different visualisation techniques used, and section 3.7 concludes.

3.2 Literature review: Exploratory data analysis and visualisation

The idea of Exploratory Data Analysis (EDA) is not new. One example of a work on the subject is Tukey (1977), which proposes the use of EDA for hypothesis generation: "Exploratory Data Analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step". Bertin (1984) also speaks of the use of graphics as a tool for thinking, for "augmenting [our] natural intelligence in the best possible way" and "finding the artificial memory that best supports our natural means of perception" (see also LeGates 2005).

Openshaw (1995) gives a good overview of the standard techniques used to visualise census data, for example nearest neighbour plots and scatterplot matrices. These are certainly helpful techniques, however they do suffer from the 'blinkers' problem to some extent in that they only ask specific questions, such as "to what degree is variable X spatially localised?". Also featured, however, is the more advanced technique of dimensionality reduction, both via principal component analysis and via projection pursuit. Either of these methods can reduce a complex dataset to a simpler one which preserves the key relationships between variables. Clearly this can be employed to understand greater quantities of data at once; indeed it is used in this chapter. However, while the standard use of dimensionality reduction is to assist in classifying data by simplifying a set of measured variables, its use in this thesis is to simplify geographical data (i.e. the positions of towns on the map), not measured social data. Instead, the simplification of the former assists in visualising the latter.

Much visualisation research seeks to make minor improvements to standard techniques. For example Cowell (2005) advocates displaying stream strength data as a literal 'stream' centred around the x-axis, Eick (1996) recommends displaying network data on the surface of a sphere rather than a plane, and Devaney (2005) employs head tracking to assist with the display of 3-d datasets. What these techniques have in common is the idea of presenting information in a more intuitive way, one which makes use of physical analogies to hijack the brain's ability to understand the physical world, and transfer that ability to the virtual one.

The manner in which Land Registry data is visualised in this thesis is novel.

Such data covers a range of both space and time, and the techniques for visualising data that extends into both space and time are surprisingly limited, despite the early suggestion by Hagerstrand (1975) that space-time diagrams should be borrowed from the world of physics, to help us understand both physical and sociological limitations on human behaviour. A field known as Temporal GIS has existed since at least Langran (1989), which focuses on temporal search and database operations more than visualisation, using the obvious two approaches to display time: either as an extra spatial dimension (e.g. Kraak 2001), or with an animated display (e.g. Claramunt et al. 2000). Alternative approaches have, however, existed: Blok (2000) for example creates a new computer language to describe temporal changes, while Cai et al. (2007) takes an interesting approach to monitoring marine algae growth over time: firstly, persistent features are identified and tracked as a single entity; secondly, periodic components are detected and then displayed on a simple frequency/strength plot. Changes over time are thereby reduced to a number of frequency components. In the field of house price analysis, Guerois & Le Goix (2009) uses discontinuity analysis to classify the change over time in boundaries between different regions of a spatial market.

Another key area of development is that of displaying interaction data. Bertin (1984) recommends some techniques applicable to this - using as a backdrop either the spatial domain (as is the case with flow maps) or the interaction domain (as is the case with interaction matrices). For the latter, the suggestion is to increasing comprehensibility by rearranging rows and columns to coalesce the larger data points into a familiar and easily recognisable shape - for example, a diagonal line or a triangle. Kwan (2000) uses 3d visualisations of the 'space time aquarium', sometimes normalised to a home-work axis, to show that people's spatial activities tend to congregate around their home, workplace or the route between. Kwan (ibid) also reduces timed flow data either to a single spatial dimension - the distance from home - much as is done with the UK Census distance-to-work data; or eliminates the time dimension entirely to produce activity density patterns. More recently, Yan & Thill (2009) use a self organising map to classify different types of interaction links according to their properties (in this case, the relevant data are air fares and airline market share for different routes). Likewise, Andrienko et al. (2007) uses clustering to group extensive car journey data for display on a comparatively simple map. Rae (2009) produces

UK interaction maps using flow thresholding, flow density and maps which distinguish between reciprocal and unique flows, while Cui et al. (2008) presents significant enhancements to traditional flow maps using on mesh-based warping of interaction lines. Wood et al. (2009) introduces *flow tress*, whereby each region of a spatial map is itself replaced by a miniature map of flow destinations from that region. Finally, Marble et al. (1997) takes an approach similar to that used in this chapter for the display of interaction data, albeit with a less sophisticated technique for ordering data points and somewhat bizarrely, making use of a 3d graphics where they are not needed.

While few techniques are specifically applicable to interaction or space-time data, some generic methods also exist for analysing any multidimensional dataset, and these may also be applied. Pryke & Beale (2005) gives a good overview. In general, multidimensional analysis focuses on looking for correlations between pairs of dimensions, which if they exist, represent frequently occurring patterns. Parallel Co-ordinate Plots (e.g. those on the *Geovista project website* n.d.) are useful here. Also, interesting techniques exist for displaying correlations once discovered, for example the 'spring network' plot (Ebbels et al. 2006, Pawlak 2005) in which strongly correlated variables are put more closely together on the page than weakly correlated ones. Alternatively, clustering can be used to group data points into a number of data-derived categories, a technique common in bioinformatics (Falkman 2001).

Finally it is important to mention Pixel methods (Keim 1996, Keim et al. 2001) which are used extensively in this research. These follow the principle of reducing each data entry to a single pixel, which is coloured according to the data value. Providing that the pixels are arranged in a comprehensible manner, a very large dataset can then be viewed entirely on one plot, and hitherto unsuspected patterns can be spotted.

3.3 Visualisation philosophy

This chapter started by stating that the datasets employed in this thesis are too large to comprehend at a glance. Actually however, the human brain can process an astonishing amount of data in a short quantity of time, so long as it is provided in the correct form. One such 'correct' form is that which the our brains were

primarily evolved to process: as a visual image.[1]

81 128 84 131 83 130 80 127 80 126 83 129 86 128 84 125 87 123 92 125 96 127 100 129 101 128 101 127 100
125 99 126 91 125 86 123 98 133 106 141 104 137 104 137 104 138 97 131 93 129 89 125 92 125 96 129 99 129
103 131 106 132 105 133 92 128 85 126 87 128 93 135 96 135 94 133 95 134 96 135 90 129 92 131 91 129 87
125 85 123 86 124 85 122 83 120 91 123 90 123 92 125 93 126 89 122 85 119 88 122 96 130 91 125 89 123 88
122 89 123 93 126 99 132 100 133 95 128 93 124 98 130 100 132 95 128 93 124 96 128 99 131 100 132 92 123
93 124 101 129 98 125 100 127 106 133 103 130 103 129 110 134 112 134 107 131 103 130 108 136 109 138
105 137 105 139 100 133 100 135 103 140 97 134 92 128 93 129 90 126 93 129 93 125 92 125 96 130 97 131 97
133 96 132 87 124 88 125 88 125 91 128 97 135 100 138 99 133 92 126 92 125 96 129 99 127 97 125 99 127
103 131 105 136 110 141 113 145 111 143 102 135 100 133 99 131 101 133 99 131 95 127 98 129 106 136 100
128 97 125 93 124 90 125 85 120 80 118 87 125 97 135 101 136 96 131 95 128 94 129 92 126 89 127 94 132 95
135 95 132 92 132 91 130 90 133 93 136 92 138 87 133 80 126 84 126 86 127 86 123 87 120 95 124 104 129
108 131 111 136 98 129 89 126 86 122 82 123 82 122 81 123 84 127 88 131 91 133 90 130 88 126 90 124 93
126 94 126 98 125 103 128 110 128 116 135 109 131 96 126 96 129 95 133 90 128 86 123 93 126 96 127 99 128
98 128 92 126 86 125 80 125 80 126 85 125 94 132 86 124 95 133 91 128 83 120 98 135 94 131 95 132 94 131
93 130 93 130 93 131 92 130 91 129 92 130 93 134 90 131 91 129 98 136 108 142 106 140 104 137 109 142 110
143 104 137 99 132 95 128 90 122 92 124 97 131 101 135 97 127 97 127 98 126 96 124 98 124 103 130 103 127
96 120 99 123 101 129 105 133 102 134 100 131 96 130 95 129 95 129 96 122 96 122 98 124 100 131 106 137
110 144 109 143 101 138 106 141 101 134 101 132 109 136 115 139 117 139 116 134 105 129 93 130 96 138 91
130 87 121 93 125 95 123 94 119 105 128 112 133 109 130 102 125 94 120 94 120 101 129 100 131 96 125 101
121 109 129 109 129 102 126 100 127 99 130 101 133 103 139 97 133 99 136 101 138 103 140 100 136 92 128
91 125 96 130 98 132 99 133 98 132 96 130 96 129 96 129 97 130 97 130 92 125 93 126 95 129 92 126 87 123
95 131 100 137 93 130 93 126 88 120 90 120 99 126 103 127 104 127 110 129 113 133 120 140 113 135 110
132 108 133 103 130 98 129 98 130 102 134 104 128 107 129 108 130 107 129 109 133 115 139 116 143 115
142 106 133 106 133 103 130 97 124 100 124 110 134 113 137 107 131 108 132 106 130 107 131 109 133 108
131 106 129 106 128 106 128 112 134 115 137 111 135 110 134 112 136 104 129 98 126 108 136 107 132 107
132 105 130 105 130 108 133 108 133 107 132 108 133 108 133 113 138 115 140 118 143 121 146 116 141 108
133 103 131 106 135 108 139 103 136 96 129 92 127 89 124 89 126 95 132 94 131 90 127 92 127 95 130 99 132
107 140 111 142 105 138 105 138 101 136 99 134 98 133 98 133 96 131 97 132 98 133 91 128 94 131 96 134 91
132 89 129 94 136 98 142 98 140 104 137 110 142 113 140 107 132 104 127 105 127 106 124 102 120 118 137
121 141 121 144 121 146 122 147 119 146 115 144 115 144 117 141 112 136 108 134 114 140 124 150 126 152
124 146 123 145 130 148 126 144 127 143 127 146 127 146 124 147 123 147 119 144 121 144 123 146 123 147
122 147 121 144 120 144 116 140 109 137 109 138 115 146 112 142 108 138 118 142 121 146 121 144 123 144

Figure 3.1: Extract from a large numeric dataset

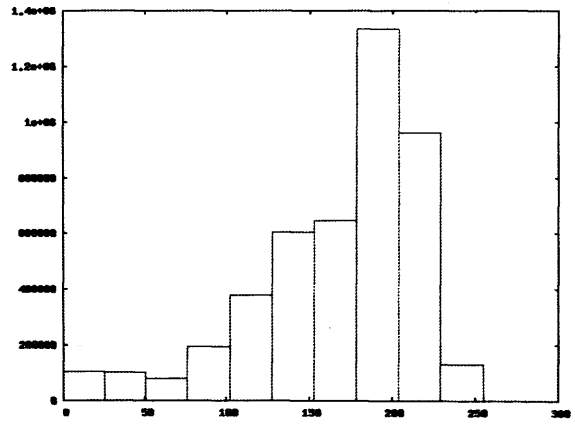


Figure 3.2: Histogram of number frequency in Figure 3.1 dataset

As an example, figure 3.1 presents an extract from a dataset containing approximately 10,000,000 points. Examination of this reveals that the data consists entirely of positive integers, and none can be observed which are greater than about 300. Is there any way in which we can understand them better? One way would be to plot a histogram of *frequency* against N, as in figure 3.2. This reveals that the range of $175 < N < 200$ contains by far the most frequent values of N, and that the numbers appear to roughly fit a skewed normal distribution.

[1]Sonification processes, during which data is converted to sound, are also an interesting area of research with similar aims; however they are not employed in this thesis.

Figure 3.3: Pixel based representation of Figure 3.1 dataset

Figure 3.4: Three possible errors in the data which are immediately obvious through visualisation.

Such analysis of the data set suffers from what was described in section 3.1 as the 'blinkers' problem: it addresses one specific question, and misses the so-called big picture. The big picture, in this case, is displayed in figure 3.3. On looking at this it is immediately obvious that the data was collected by some kind of camera on the summit of a mountain, and that the data consists of measurements of light reflected from three happy rock climbers. At a glance, the human brain has processed ten megabytes of data which previously seemed incomprehensible. It has identified certain areas as being composed of rock, sky, cloud or person by their texture (a task which for computer programs is very hard) and formed an internal model of the relationships between these areas.

This kind of visualisation is the ultimate aim of pixel based methods. Even if the colour value of every pixel is not exactly known, it is possible to gain an intuitive understanding of our datasets which may not have been possible by using more specific analysis. Also, it should be possible to spot certain types of error in the data which may have been missed by other techniques. Classes of error noticeable through this visualisation may be for example, fundamental errors in the ordering of the dataset, missing parts of the set, or spurious data points which don't seem to fit with their neighbours (each of which is illustrated in figure 3.4).

In a sense, pixel visualisations are an extension of the ASCII or numeric maps of data variables employed in the early days of computers; or indeed as the raw output from any modern computation. According to Burrough (1986), it was Edger M. Horwood who first made maps simply by printing grids of numeric data values onto paper; pixel visualisation simply makes this kind of output more readable.

3.4 Migration Pixel Matrix Plots

This section introduces the concept of pixel matrix plots, and describes their application to the visualisation of census interaction data. Section 3.4.1 describes the structure of the data, section 3.4.2 describes the methodology, and sections 3.4.3 and 3.4.4 discuss its application at Local Authority and ward level respectively. Section 3.4.5 provides an overall evaluation.

3.4.1 Structure of the data

Interaction datasets contain information not about every object, but about every *pair* of objects in the dataset. Thus the migration data sets shown here can each be said to be a function mapping the product space of geographical regions L to the natural numbers $\mathbb{N}_0$:

$$f : L \times L \mapsto \mathbb{N}_0 \tag{3.1}$$

The geographical regions themselves can be specified by a function from grid co-ordinates (latitude and longitude) to regions:

$$g : \mathbb{R} \times \mathbb{R} \mapsto L \tag{3.2}$$

so combining these, it can be seen that the aim is to visualise a four dimensional function:

$$\mathbb{R}^4 \mapsto \mathbb{N}_0 \tag{3.3}$$

3.4.2 Methodology

3.4.2.1 Pixelation

A simple example is given in figure 3.5. This could be considered as a migration table showing how many people moved between each pair of places a, b, c, d and e in a given year.

		From				
		a	b	c	d	e
To	a	5	6	0	1	0
	b	5	7	1	1	0
	c	1	2	3	2	1
	d	0	0	3	2	1
	e	0	1	1	1	1

Figure 3.5: Simple interaction dataset

Figure 3.6 illustrates how this would be displayed using a pixel-based method: each cell of the table is simply coloured in with an intensity proportional to the number of people it represents.[2] While the illustration is large, this would in

[2]Actually in later plots, to prevent smaller migrations being completely obscured by the

reality result in an image of 5×5 pixels, or less than 2×2 millimetres on a typical LCD display. Thus it is possible to display a much larger dataset on a computer monitor.

The catch is, that it would not be easy to comprehend such a dataset because the X and Y axes don't represent anything real. The positions of places a to e on the diagram are determined solely by their positions in the alphabet. It would be better if an ordering could be derived for the set of places which is more intuitive to somebody who knows the places concerned.

		From				
		a	b	c	d	e
To	a	5	6		1	
	b	5	7	1	1	
	c	1	2	3	2	1
	d			3	2	1
	e		1	1	1	1

Figure 3.6: Pixelation applied to the Figure 3.5 dataset

3.4.2.2 Selection of the ordering criterion

In order to display spatial interaction information on a 2-d plot in this manner, it is essential to reduce 2-d spatial information to a single dimension, in order to display spatial (2-d) information on each (1-d) axis of the table. The choice of ordering used on each axis will have a crucial effect on the information that will be interpretable, and indeed whether it will be possible to interpret any information whatsoever. The latter concern is aptly illustrated by the fact that in the days before digital television, premium channels on Sky TV were encrypted simply by changing the ordering of pixels along the X axis of each line of picture (*VideoCrypt* n.d.) and this was considered sufficient to prevent unauthorised access to the content! Likewise, the picture of UK migration could range from sharp to unintelligible depending on the ordering chosen for each axis. Inevitably, any such reduction from two dimensions to one, will not preserve spatial contiguity, as regions in 1-d space can have at most 2 neighbours, while the number of neighbours possible for an area in 2-d space is potentially unlimited.

larger ones, each pixel will be coloured with an intensity proportional to $log(n+1)$ rather than simply n, where n is the number of people.

A number of approaches are possible.

- Ordering from a single statistic, e.g. total migration, or population size. Figure 3.7 gives an example of this, where local authorities on the X and Y axes are ordered on the size of internal migration flows within the local authority. London, Birmingham and other large cities all appear merged into one bright area at the bottom right of the plot, while rural areas are situated more towards the top left. While it is conceivable that such a representation could be used to highlight certain elements of the social structure of a region, it is hard to discern much information about the UK from this plot, due to the lack of regional contiguity along the axes. Also, ordering based on a single statistic falls prey to the 'blinkered' approach discussed in the introduction to this chapter: while ordering based on population, for example, might reveal patterns related to population size and migration it might miss patterns related to other statistics. Therefore this approach is not used.

- The approaches suggested by (Bertin 1984, page 196) are either *diagonalisation* or *triangulation*. These are processes which use the data itself to determine an ordering: the rows and columns of the matrix are progressively swapped around until all of the larger data points appear in a coherent shape. In the case of diagonalisation, all large values are grouped along the diagonal, while in the case of triangulation the aim is to put all large values on the same side of the diagonal to form a triangle.

  This technique is commendable from the point of view of 'getting the big picture' - the ordering is not restricted to an arbitrary statistic so much as specifically rearranging it with the goal of producing a shape which is more easily comprehended by a human reader. However, it suffers from two drawbacks. Firstly, the process has the disadvantage that a different ordering will be employed for each of the different statistics displayed. For example, if one plot is produced detailing the number of migrants between each pair or locations, and a second plot is produced detailing the number of commuters, then these plots will likely each entail different orderings of the matrix, so not be directly comparable. Secondly, the technique was not designed for use with spatial information - Bertin applies it only to

non-geographic data such as sales clerks and sale items, or South American tribes and technological developments. In the case of migration, where each point along an axis represents a Local Authority or Ward situated in real geographic space, such an ordering would lose the inherent geographical information to some extent. As our intuitive understanding of the structure of a country seems to be based on geographical space, it seems a shame to discard this information - therefore a different approach is chosen which preserves it.

- The approach proposed in this thesis, therefore, consists of a hybrid method which both rearranges the matrix for greater comprehensibility, but also employs geographical data to fixate the graphic in real space. It thereby becomes, according to Bertin (1984), a *map*: a graphic where "the elements of a geographic component are arranged on a plane in the manner of their observed geographic order on the surface of the Earth" (page 285).

  Researchers familiar with a given area, for example the United Kingdom, have an intuition for which places are close together and which are further apart. For example, without much thinking, the author knows that Reading is near London, and Edinburgh is near Glasgow, but that there is a long distance between Bristol and Aberdeen. Therefore the ordering aimed for is one in which the following two criteria are fulfilled to the greatest extent possible:

  1. places which are physically close together on the map, will be close together in the ordering; likewise places which are far apart on the map will be far apart in the ordering
  2. places which are close together in the ordering, will be close together on the map; likewise places which are far apart in the ordering will be far apart on the map.

3.4.2.3 Methods for computing the desired ordering

Having chosen suitable criteria by which to order each of the axes of the pixel matrix plot, it is necessary to compute an ordering which fulfils these criteria. This task is not necessarily simple, as for a series of n geographical points there

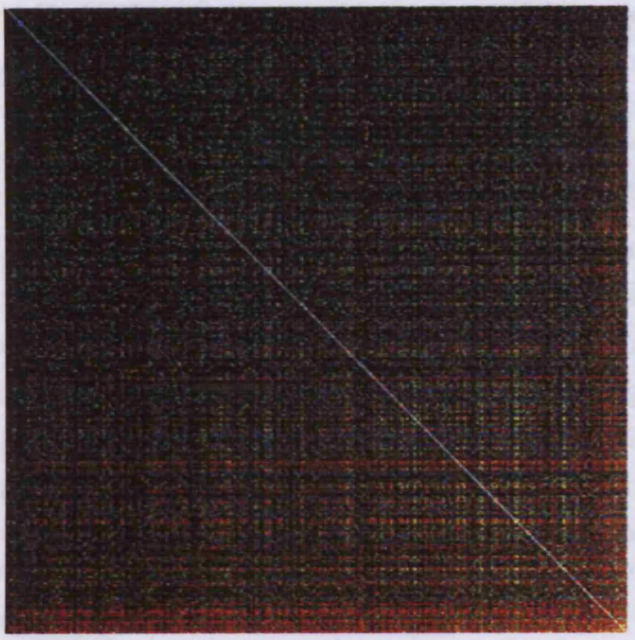


Figure 3.7: Local Authority level visualisation of UK internal migration, coloured by age as in figure 3.10, but ordered by size of internal migration flow rather than linearisation of geospace.

are $n!$ possible orderings to choose from. In the case of UK local authorities, this equates to about 10^{1000} different orderings, while in the case of wards in England and Wales there are a colossal 10^{31089} different alternatives to consider. Even the smaller of these numbers is far larger than the number of atoms in the known universe! Total enumeration of all possible orderings to find the 'best' (according to our criteria) is therefore impossible, and instead an algorithm must be chosen which somehow *approximates* a good ordering.

Three approaches are used:

1. The first approach used is complete linkage clustering and optimal ordering (CLO-OPT). This is an approach suggested in Guo & Gahegan (2006), in which a variety of different algorithms are investigated for their ability, on average, to fulfil the criteria described above. The best of the algorithms investigated, CLO-OPT, finds the shortest possible path which visits all of the geographical points, subject to a clustering constraint. In particular it outperforms the space filling curve techniques used by Marble et al. (1997).

   CLO-OPT solves a problem which is a minor variation on the Travelling

Salesman Problem (TSP) - the problem of finding the shortest path which links a set of points. The general case is NP-complete with time complexity $O(n!)$, where n is the number of points: there is no known way to find a shortest path (or optimal ordering) without considering all possible paths, which for more than a small number of points is computationally infeasible. CLO-OPT therefore differs from the standard TSP in that not all possible solutions are searched. Instead, the points are first *clustered* based on the geographic distance between them, such that (for the most part) towns and cities are recognised as single clusters. They are then optimally *ordered*, with the condition imposed that clusters of points must not be divided by the path (or ordering) chosen. Thus, any given cluster must remain in one continuous sequence in the final ordering. This approach has two advantages:

(a) the ordering derived will be more intuitively comprehensible to the researcher, as places they perceive as belonging in a group together (e.g. wards within the same city, or towns within a certain region) tend to be grouped together

(b) the problem is now computationally feasible. The time complexity of the Complete Linkage Clustering and Optimal Ordering algorithm is $O(n^3)$.

CLO-OPT was first developed in the field of bioinformatics (Bar-Joseph et al. 2003). On the set of approximately 400 local authority map points, it can compute an optimal ordering using a few seconds of time on an average desktop computer. Therefore, for the local authority level data set, only this technique is employed. However, for the 8850-point ward data set, this is not possible, and it is necessary to employ a 64-bit machine with around 8GB of RAM.

2. The second approach used is a novel variant, Hierarchical CLO-OPT (CLO-HIERA-OPT). Running CLO-OPT on the 8850-point ward data set discards information about administrative hierarchies, as each ward is situated within a Local Authority. As such information forms part of the intuitive geographical knowledge of planners, and indeed many local residents' mental models of space, this could well be useful in constructing a

more comprehensible ordering. Therefore, Hierarchical CLO-OPT ensures grouping together of all wards within the same LA.

This is achieved by first running CLO-OPT to produce an ordering for Local Authorities, then using the index from that ordering to define a position in three-dimensional space for each ward, such that

$$x_{ward} = \text{geographical x coordinate} \tag{3.4}$$

$$y_{ward} = \text{geographical y coordinate} \tag{3.5}$$

$$z_{ward} = kI_{LA(ward)} \tag{3.6}$$

where $I_{LA(ward)}$ is the index of the ward's containing Local Authority from the first-stage ordering, and k is a suitably large constant designed to ensure that grouping of wards within Local Authorities takes priority over grouping of wards within physical space. This permits the use of existing CLO-OPT software without modification.

Of course, whether or not political boundaries should be included in the visualisation is a debatable question. However it should be noted that as social policy, or at least spending, can differ between authorities, it might be expected that measured social data will reflect these differences and therefore grouping these data points according to administrative regions is not an illogical approach to visualisation.

3. In addition to the first two approaches, for the purpose of creating tools accessible to all, alternatives to CLO-OPT and CLO-HIERA-OPT were evaluated which allow ward level visualisations on more modest computing hardware. The first of these is a hierarchical complete linkage clustering with non-optimal ordering algorithm (CLO-HIERA-NONOPT), the non-hierarchical version of which is described in Guo & Gahegan (2006). Instead of a two-stage process in which points are first clustered and then ordered, in the non-optimal algorithm the points are clustered and ordered simultaneously. The clustering process works as before except that once a cluster is formed, its ends are automatically joined to their nearest neighbouring clusters. This produces non-optimal results, albeit using considerably less computing time and memory than the optimal algorithm.

temperature as function of time	$T(t) = 0.06f^t$
cooling factor	f = 0.99999
distance scalar	D = initial square path distance × 2
probability of accepting change	$exp(-(\text{pathlength\_after} - \text{pathlength\_before})/DT)$

Table 3.1: Parameters for Simulated Annealing.

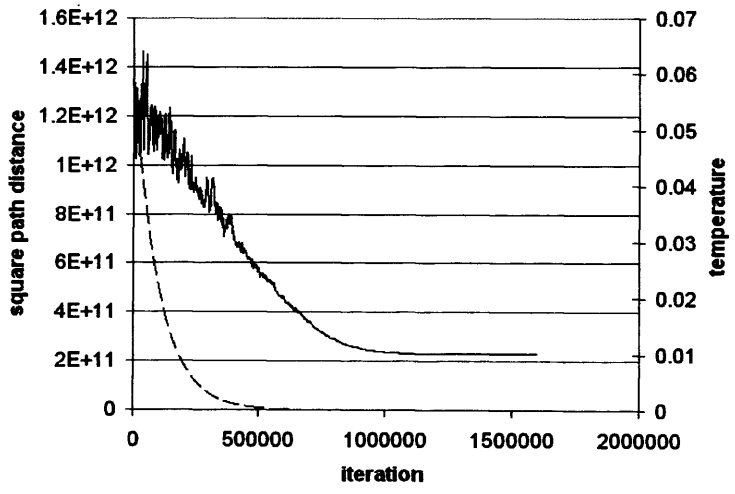


Figure 3.8: Plot of simulated annealing parameters (temperature and total squared path length) over annealing run.

4. Finally, and also with the aim of reducing computing power, a ward level simulated annealing approach is employed (CLO-HIERA-SA). Simulated annealing (see e.g. De Vicente et al. (2003)) works by repeatedly making random changes to the linear ordering, and accepting these changes *if they do not degrade the quality of the ordering E by more than a certain amount ΔE*. The quality of each possible ordering is measured, in this case, by the squared path length of the ordering:

$$E = \sum_{l \in LINKS} LENGTH_l^2 \tag{3.7}$$

The quantity ΔE reduces from a high number to zero over time, so that in the end changes are only accepted if they actually improve the ordering. The process is analogous to the physical process of a liquid freezing to become a solid; with ΔE representing the energy available for reconfiguring atoms into a different state, related to the temperature T of the system. Provided that T decreases slowly enough, a low energy configuration of atoms will be found; in the case of the ordering algorithm this equates to a good linear ordering i.e. a short overall path length.

Table 3.1 summarises the parameters used for simulated annealing, while figure 3.8 shows a plot of how temperature and overall path length change during the annealing run.

3.4.2.4 Software to enable display of the visualisations

This section discusses considerations in the creation of software to display pixel matrix plots.

1. Before a large pixel plot can be displayed, the place names on the axes must be labelled. However, as each column of the matrix is only one pixel wide, and each row only one pixel high, there is no room for text labelling of every place displayed. As it would be possible to label only key places, the approach taken was instead to display the plots via an interactive computer interface which, as the mouse is moved over the pixel plot, indicates on a map which geographical points are being viewed. This was implemented in

Javascript, to allow easy dissemination via the world wide web. An example is displayed in figure 3.9.

2. Pixel matrix plots derived from Local Authority level have a size of approximately 400x400 pixels, similar plots for Ward level data have dimensions of 8850x8850 pixels. The latter does not fit on the average computer screen, so it was necessary to build a visualisation tool which allowed for zooming in on the data. Owing to the greater computing requirements of this task, the tool was implemented in Java. Some screenshots are shown in figure 3.14.

3. A formula for selecting the colour of each pixel based on the data it represents must be chosen - especially in the case of the zooming tool where each pixel may represent more than one data point. In the case of migration and commuting plots, each pixel was set to the log sum of all points represented by it, normalised according to the average across the extent of the current display (rather than the entire data set):

$$\text{pixel\_value} = \frac{log(\sum_{pixel} \text{data\_values})}{6 \times log(\sum_{display} \sum_{pixel} \text{data\_values})} \tag{3.8}$$

This has the consequence that bright pixels - representing high data values - tend to 'clip' (i.e. above a certain data value the maximum brightness is reached so no contrast is distinguishable). In practice, this artifact is rare and hence not usually a concern.

Colour was used to convey further information on some of the plots: the red, green and blue channels allowing for simultaneous display of three data sets. Alternatively, in the case of plots of market correlations - which can encompass both positive and negative values - the positive and negative data was summed separately to generate information for the red and blue colour channels respectively.

3.4.3 Discussion of LA level migration plot

The Pixel Matrix Plot was found to be a useful tool for obtaining an overview of a complex data set. This section is dedicated to the discussion of an example

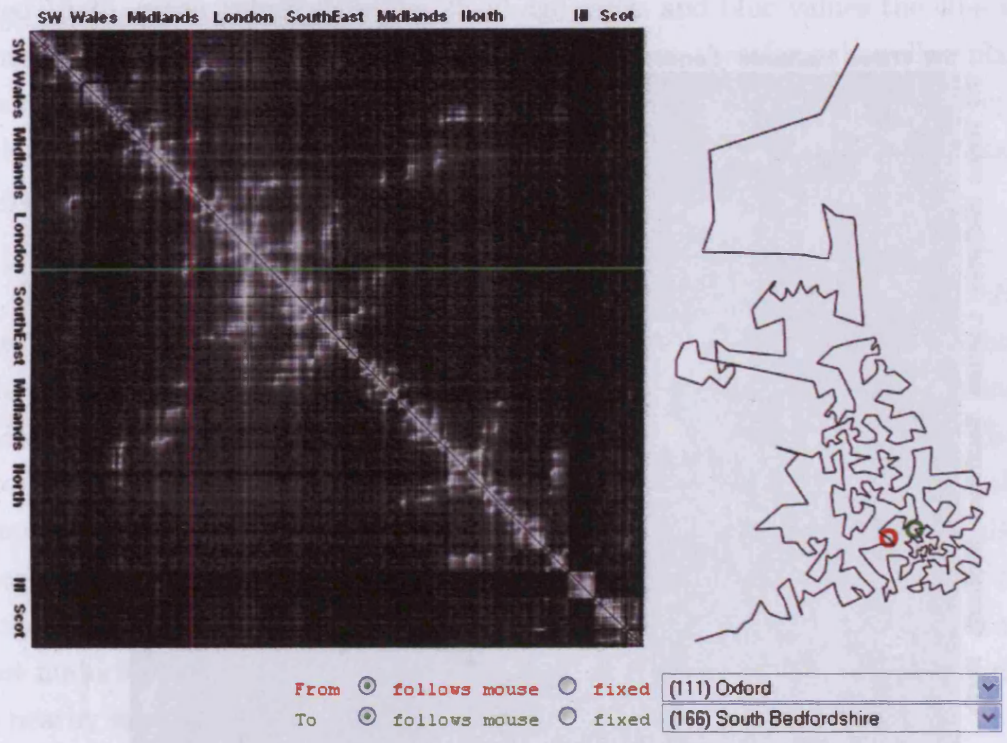


Figure 3.9: Visualisation of UK commuting flows between all local and unitary authorities in the UK, ordered with the CLO-OPT algorithm (the only algorithm used at LA level). The line formed by linearising all the geographic points is shown on the right hand side. As the mouse is moved over the pixel plot on the left hand side, red and green markers move over the map to show origin and destination points respectively. On the pixel plot, origin is represented by x-axis position, and destination by y-axis position. Axis labels are not present in the visualisation software though have been added to the figure for clarity. The bright diagonal line indicates that most commuting takes place on a local basis.

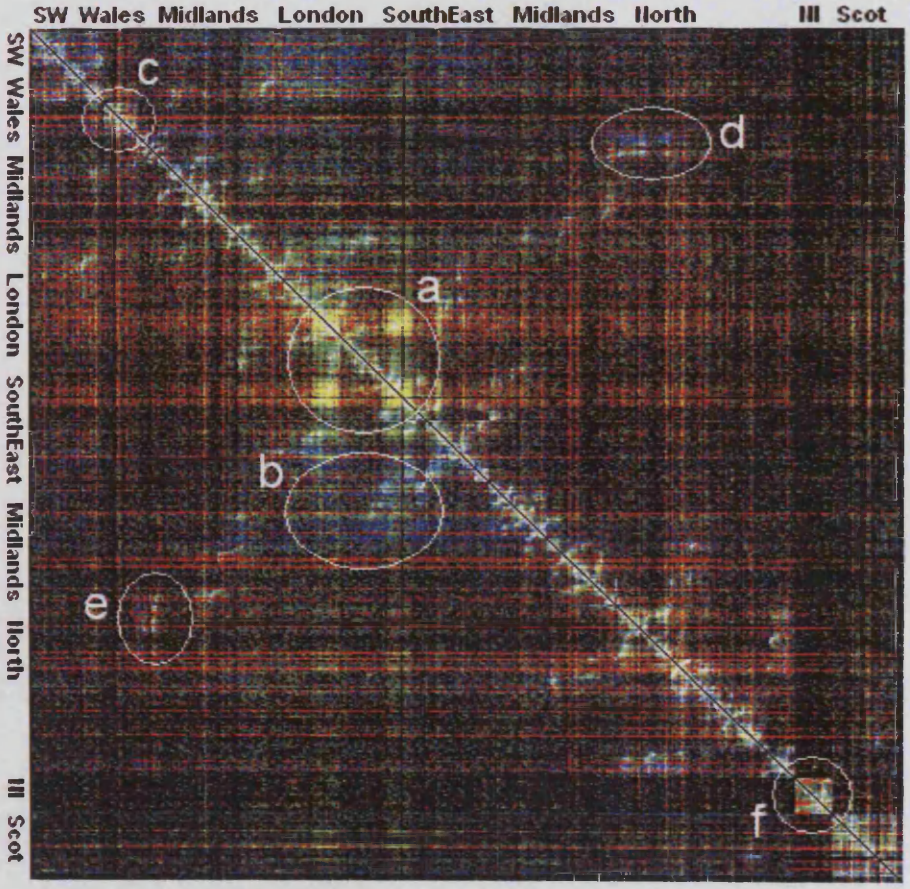


Figure 3.10: Local Authority level visualisation of UK internal migration, coloured by age. Various features have been labelled - see section 3.4.3 for explanation.

(shown in figure 3.10), which shows internal migration for the 2001 census year, between all local authorities in the UK. The plot has additionally been coloured by age - the red, green and blue components of each pixel represent the following age bands, which were chosen following a principal component analysis of the age composition of moving groups: red pixel values show the component of migrants aged 16-25, green values show the 25-40 age range and blue values the 40+ age range. (Due to the nature of the colour mixing process, therefore, the yellow pixels seen on much of the plot indicate a mixture of 16-40 year olds, light blue/cyan colours indicate a predominance of people over the age of 25, and white pixels show an even combination of all three age bands).

Several patterns in this plot are worthy of discussion.

The first point to note is the black diagonal line of pixels running right across the image, and the tendency of other pixels near this diagonal to be brighter. The black diagonal has been marked in afterwards to aid with interpretation - each point on it would represent migration from a local authority to itself. These pixels would naturally be shaded very near to white, as much migration takes place within local authorities; however they have been changed to black to allow easy identification of the diagonal. Given this feature, it is easy to see that the diagonal is surrounded by further bands of white pixels. This illustrates that the vast majority of migrations in the UK are localised - i.e. people tend to migrate to nearby areas.

Various other features have been labelled, to give an idea of the kind of patterns which can be spotted. These are discussed in the following sections.

3.4.3.1 London

The four yellow squares of feature *a* represent London. Unfortunately, the algorithm fails to group all of Greater London into one cluster; instead it is divided into two regions interspersed with some of the local authorities to the north (see figure 3.9 for a map of the ordering). Hence, London appears as $2 \times 2 = 4$ squares on the diagram, because on each axis it is represented by two different regions.

A lifecycle migration pattern is visible with respect to London. Thick orange lines, extending horizontally outwards from region *a*, indicate a flow of younger people from all over the country migrating into London. Fainter green lines extending vertically out of region *a* indicate the middle-aged leaving London;

while the strong blue patch below (marked *b*) shows an older population migrating from London to Norfolk and surrounding areas.

The City of London is clearly visible as a black cross centred in the lower right yellow square. This is because it has little residential population and therefore very few migrations to and from the City occur.

3.4.3.2 Northern Ireland

The feature marked *f* represents Northern Ireland, easily identifiable because of its strong internal structure (a bright square of migration movements) but having little interaction with the rest of the UK (the black bars extending horizontally and vertically from it).

3.4.3.3 Visible evidence of known migration models

Aside from the diagonal bright region, the rest of the image contains several bright horizontal and vertical lines representing certain local authorities. Two such lines cross in the area marked *c* (which happens to be Cardiff). These lines tend to intersect on the diagonal, as large centres of population tend to have both high in- and out-migration.

However, where bright horizontal and vertical lines cross *away from* the diagonal, even brighter pixels can be found at the intersection of the two lines. This indicates (roughly speaking) that for longer distance movements between two points X and Y, the number of people moving between X and Y is proportional both to the *total* number of migrants leaving X and to the *total* number of migrants arriving at Y, regardless of their actual destinations and origins.

Overall, the fact that the diagram appears to consist of a bright diagonal band, augmented with a series of horizontal and vertical lines, tells us that the migration data could broadly fit some kind of gravity model

$$migrants_{XY} \propto \frac{population_X \times population_Y}{f(distance_{XY})} \tag{3.9}$$

except in this case that we are not directly observing the populations of X and Y, instead we are basing our mental model on the crude assumptions that

$$population_X \propto out\_migrants_X \tag{3.10}$$

and

$$population_Y \propto in\_migrants_Y \tag{3.11}$$

so the model we see in the diagram is therefore

$$migrants_{XY} \propto \frac{out\_migrants_X \times in\_migrants_Y}{f(distance_{XY})} \tag{3.12}$$

Thus at shorter distances, the denominator of the fraction dominates and a high level of migration is seen, while at longer distances the numerator dominates and migration becomes more strongly related to the origin and destination populations.

It should be noted that this observation is not presented as concrete evidence; rather it is presented as a *hint* given to us by the visualisation. It would equally be possible, for example, to look at the visualisation and intuit the existence of a 2-state heterogeneous migration model: in this case, one type of migration (perhaps motivated by wanting a change of residence) would take place locally, explaining the bright diagonal band; meanwhile another type of migration (perhaps motivated by a change of employment) would not be influenced much by distance but rather follow the multiplicative part of the gravity model:

$$long\_distance\_migrants_{XY} \propto population_X \times population_Y \tag{3.13}$$

If we wished to test these hypotheses properly, we would of course have to compute the models numerically. However, it is clear that the visualisation is of use in generating the two alternative hypotheses in the first place: it has given us some ideas about the data which we are now free to go and rigorously test. As it happens, such models are investigated in the existing literature e.g. Dennett & Stillwell (2008).

3.4.3.4 Polycentricity in London

A further interesting facet of this visualisation is that it has been possible to see evidence of urban polycentricity in London. Hall (2001) notes that London, rather than having a single centre which exceeds all other parts of the region in its provision of products and services, is "now the centre of a system of some 30-40 centres within a 150km radius". Each of these multiple centres can be defined as

Algorithm	Square path length (Gm^2)	Crossings
CLO-OPT	208	390
CLO-HIERA-SA	230	870
CLO-HIERA-OPT	293	1180
CLO-HIERA-NONOPT	1210	9390

Table 3.2: Ordering metrics for ward level (8850 point) linearisation algorithms.

such because in some specialist area of provision (for example financial services, or legal services) they are not superseded by anywhere else. The centre of expertise for each service category is in a different geographical location, rather than all such centres coinciding in one key location such as the City of London. In a more numerically-focused study, Taylor et al. (2006) shows that the entire south-east can be said to be polycentric, the principal nodes being London, Reading and Southamption. This is achieved via a lengthy measurement process involving the identification of links between regional branches of law firms.

The pixel matrix plot visualisation arguably also shows that Greater London is polycentric in terms of migration movements, because no discernible internal structure is visible within the yellow squares of region *a*. This is in contrast to other areas in the UK, for example region *c* which represents South Wales, with Cardiff clearly being a monocentric keystone of interaction for the region.

Again, this hypothesis would require numerical testing in order to confirm it rigorously; however again the visualisation has hinted at the existence of a pattern in the first place.

3.4.4 Discussion of Ward level migration plots

The two visualisations presented so far have concerned commuting and migration flows for the 426 local authorities of Great Britain. However, much of the analysis in later chapters is conducted on the 8850 census *wards* of England and Wales. As interaction data relates to every *pair* of locations, rather than simply to every location, the size of the data set grows with the *square* of the number of locations. Thus the ward data set is over 400 times larger than the local authority data set, and different problems are encountered both in producing and interpreting the visualisation. The issues with producing the visualisation stem from the computational task of calculating an optimal ordering; as discussed in section

Optimal ordering

Non-optimal hierarchical ordering

Optimal hierarchical ordering

Simulated anneal
approximate optimal
hierarchical ordering

Figure 3.11: Map of UK wards ordered by four different linearisation algorithms. The wards are linked by a single line, coloured according to the spectrum from red through to violet.

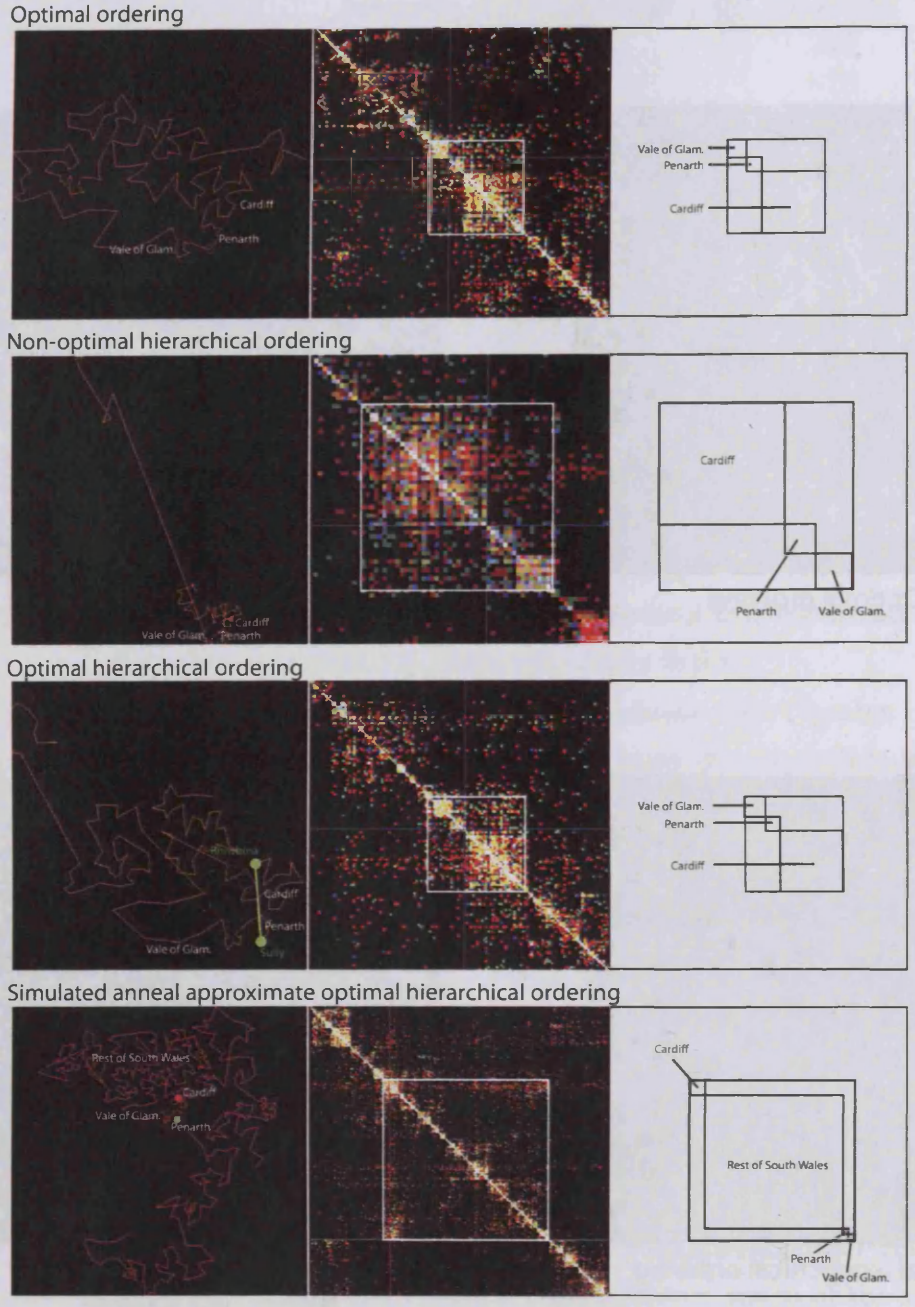


Figure 3.12: UK-wide Ward-level migration visualisations, accompanied by maps showing linearisation, 'zoomed in' to Cardiff area, for four different ordering algorithms. Place annotations are added manually, and the sketch figures to the right illustrate the relative positions of Cardiff, Penarth and the rest of the Vale of Glamorgan in each plot. The Penarth region is defined here as the wards of Sully, St. Augustine's, Plymouth, Stanwell, Dinas Powys, Cornerswell and Llandough, while the rest of the Vale of Glamorgan is defined as Gibbonsdown, Cadoc, Court, Buttrils, Castleland, Baruc, Dyfan, Rhoose and Illtyd.

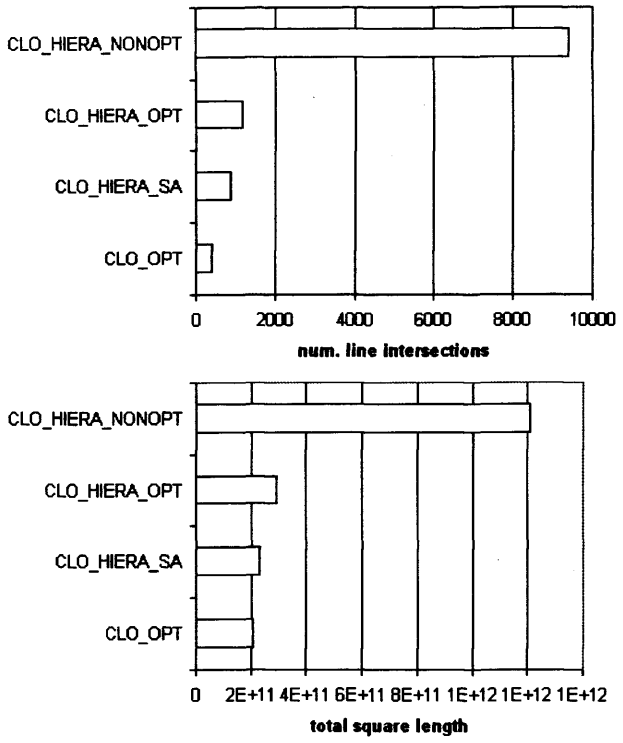


Figure 3.13: Ordering metrics for ward level (8850 point) linearisation algorithms.

Algorithm	Time requirements	Space requirements
CLO-OPT	3 hours (Java code)	8 Gb
CLO-HIERA-SA	25 hours (Python code)	minimal
CLO-HIERA-OPT	3 hours (Java code)	8 Gb
CLO-HIERA-NONOPT	negligible	minimal

Table 3.3: Computational requirements for ward level (8850 point) linearisation algorithms. Time requirements for CLO-HIERA-SA and CLO-HIERA-NONOPT refer to the linearising, rather than clustering phases, while time requirements for CLO-OPT and CLO-HIERA-OPT refer to both clustering and linearising phases. The author considers there to be considerable scope for time-optimisation of the Python-coded CLO-HIERA-SA algorithm, giving it the potential to compete with the run time of CLO-OPT if such work is undertaken.

3.4.2.3. However, there are also problems with the *interpretation* of the larger Pixel Matrix Plots.

Figure 3.11 shows orderings of all wards in England and Wales, produced by each of the algorithms discussed in section 3.4.2.3. The maps are represented by a single continuous line illustrating the ordering, coloured through all hues in the spectrum from red to violet to give a general idea of its course through 2d space without having to study the details.

Some general statistics by which to compare the algorithms are given in figure 3.13 and tables 3.2 and 3.3. The metric of total path length is what the ordering algorithm seeks to minimise; however, the number of intersections between lines (points at which the ordering crosses itself) is also given as this is thought to influence comprehensibility of the map somewhat. Note that complete solution of the travelling salesman problem would give an ordering with no intersections whatsoever, as any intersection of lines is an indication that the path may be further shortened (for example, in the case of a square ABCD, by visiting points in the order ABCD instead of ACBD). Notwithstanding, even the optimal algorithms presented here display some degree of path crossing due to the constraint that clusters must be contiguous in the ordering. According to both metrics, the outright optimal algorithm - CLO-OPT - performs best, followed by the hierarchical algorithms, in order from better to worse: CLO-HIERA-SA, CLO-HIERA-OPT, CLO-HIERA-NONOPT.

It seems counterintuitive that the hierarchical simulated annealing algorithm should outperform the 'optimal' hierarchical algorithm. This is due to the fact that Local Authorities in the CLO-HIERA-OPT result are constrained to a preassigned order, whereas with CLO-HIERA-SA the Local Authority and Ward orderings are optimised concurrently. One would therefore expect an enhancement of the CLO-HIERA-OPT algorithm coded in a similar manner to outperform simulated annealing.

In contrast to the numerical findings, the author found orderings produced by CLO-HIERA-OPT to be by far the most comprehensible. This is because

1. the intuitive hierarchy imposed by grouping Local Authorities was thought to be of benefit, thereby causing the algorithm to outperform CLO-OPT;

2. the orderings produced by simulated annealing, despite having shorter path length and fewer crossing points, tended to be less intuitive, and

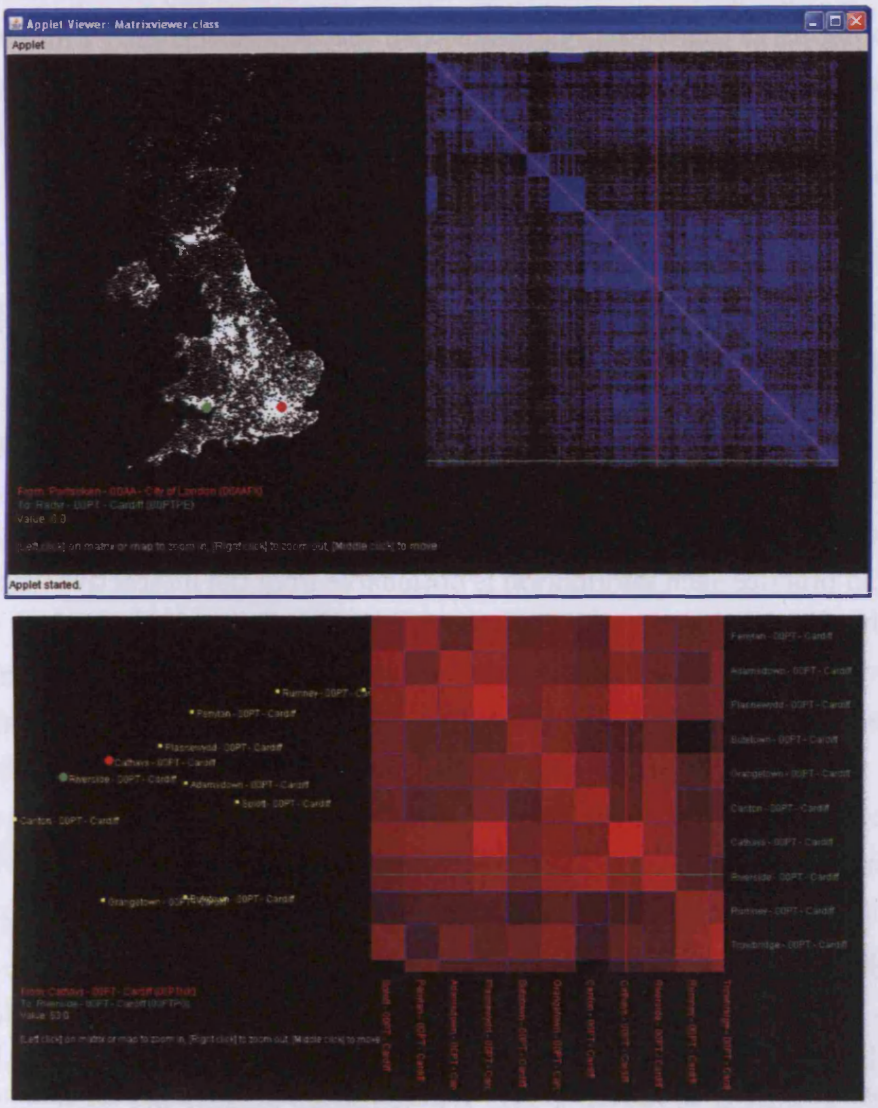


Figure 3.14: Screenshots of software showing Ward level visualisation of UK internal migration, at wide and close zoom levels. Text to the left of the display shows names of the currently selected origin and destination, together with the relevant value(s) from the interaction matrix. For close levels of zoom (shown in the lower figure), place names are added to both the map and matrix.

3. the orderings produced by non-optimal hierarchical ordering were mostly unreadable.

Point 2 above highlights the need for further research on what constitutes an 'intuitive' ordering as this is clearly at odds with the metrics currently presented. For the purposes of present discussion, an example of a problematic ordering derived by simulated annealing - one which despite a short path length and few crossing points, is harder to read than the alternatives - is presented in section 3.4.4.1.

3.4.4.1 Study of Cardiff and surroundings via ward level migration plot

Figure 3.12 shows visualisations of migration in Cardiff and the Vale of Glamorgan derived from each of the four ward-level ordering algorithms. In each case, the visualisations have been manually annotated to clarify the meaning of each section of the plot; in practice such information is obtainable from the interactive display interface.

By way of a case study, the relationship between Cardiff, the Penarth area and the remainder of the Vale of Glamorgan is considered. In all cases, internal interaction for each of the three areas is indicated by a bright square of pixels situated along the diagonal of the pixel matrix plot; therefore it is possible to say that each area is to some extent a self-contained entity. The interaction between the different areas, however, is perceived differently depending on the ordering algorithm used.

- with optimal ordering, the presence of a fainter square of bright pixels encapsulating the brighter squares representing the three areas, indicates that while Penarth, Cardiff and the Vale of Glamorgan can be considered as three independent entities, they also form a single loosely connected unit. Penarth, the central of the three squares, appears to be equally connected to both Cardiff and the Vale.

- with non-optimal hierarchical ordering, the loose grouping of the three areas is also visible, however a sudden jump in the ordering to Penarth from the opposite side of Cardiff displaces pixels representing local Cardiff-Penarth

movements away from the corners of the bright squares. This causes Penarth to appear more connected to the remainder of the Vale of Glamorgan than it is to Cardiff.

- a very similar situation exists with optimal hierarchical ordering. As the map for this ordering is displayed at a greater level of zoom, the offending sudden jump in the ordering (in this case from Rhiwbina to Sully) is annotated on the map.

- with simulated annealing, a different perspective is shown entirely. The ordering has separated Cardiff from Penarth and the Vale, placing most of the rest of South Wales in between. The loose grouping of the three areas is not so immediately apparent, though it can be deduced from the bright groups of pixels situated well away from the diagonal axis. It is almost impossible, however, to compare the relative strengths of the Penarth-Cardiff and Penarth-Vale of Glamorgan links.

The variety of 'pictures' of Cardiff area migration presented by the different algorithms highlights one limitation of pixel matrix plots, as it is undesirable for the information presented to vary so drastically with the precise details of the ordering algorithm chosen.

3.4.4.2 Alternatives to zoomable pixel matrix plots

The approach suggested by Guo (2007) is to first aggregate the data into a more manageable number of regions, before applying the clustering and linearisation algorithm. However, the zoom tool, at this stage, is a fun (if not easy) method for exploring the data set, and is consistent with the principle of not altering the underlying data but instead rearranging it into a more comprehensible format. Also, the zoom facility naturally allows analysis at any level of spatial aggregation chosen by the viewer, rather than just the top level.

3.4.5 Evaluation of Pixel Matrix Plots

It is interesting to compare the features of pixel matrix plots to the ideal properties of graphics as suggested in Bertin (1984). The latter proposes that graphics should be *efficient*, a property which is defined as such:

> "If, in order to obtain a correct and complete answer to a given question, all other things being equal, one construction requires a shorter observation time than another construction, we can say that it is more efficient for this question". (Bertin 1984, page 139)

Arguably, as they take some practice and training to read, pixel matrix migration plots are not particularly efficient. However, if a researcher is already well trained in their use then they may be a suitable technique for answering the open-ended question, 'What are the primary features of this data set?'.

Bertin suggest further guidelines for the construction of graphics: that the presenter should aim "to represent the information in a single image" and "to simplify the image without reducing the number of correspondences" (ibid, page 171). These aims are both fulfilled by migration pixel matrix plots; indeed the second aim is fulfilled far better than in the case of flow maps (the conventional means of presenting migration information) as small correspondences need not be discarded in order to create a meaningful image.

As pixel matrix plots were developed primarily to assist with the other research presented in this study, it is valid for the time being to evaluate their usefulness purely with respect to the other tasks undertaken in the thesis.

3.4.5.1 Advantages

Using pixel matrix plots assisted with gaining a broad overview of the data handled. In the case of Local Authority level migration, they reveal a broad two-level structure of local and global migrations. In the case of Ward level migration, they highlight above all that the data is sparse, with few long-distance migrations between individual pairs of wards. This hints that long-distance ward level migration might not be of great relevance in determining property market movements, as there isn't very much of it! In fact, the findings to be presented in chapter 5 confirm this.

In certain instances, the use of pixel matrix plots enabled the detection of software bugs. For example, in one case, migrations originating or ending *outside* the UK had not been removed from a data sets, a processing step which was required for the task at hand. This was immediately obvious when looking at the visualisation - the high volume of migrations showed up as a pair of extremely bright horizontal and vertical lines, with ward codes of 8888 and 9999 - those

used to denote external migration by the census office. In another case, a list of wards from the entire UK was accidentally being used to index housing market data which related only to England and Wales. When visualising the output, it was clear that large sections of the plot were blank, because no data were present for Scotland or Northern Ireland.

Therefore, these plots are considered to be a useful tool both for creating and debugging systems which make use of large interaction datasets.

3.4.5.2 Disadvantages

One drawback of the pixel matrix plot is the learning curve associated with starting to use it. When presenting data to colleagues, they will often take some time to learn to interpret the output, which is far less intuitive to read than, for example, a flow map. However, it could be argued that for certain users, the time taken to learn how to read a pixel matrix plot is well worth the effort.

Another drawback is that some features shown on the plots are artifacts of the algorithm used to produce it rather than inherent properties of the data set. An example of this is the regions *d* and *e* marked in figure 3.10. These clusters of bright points far away from the diagonal would seem to imply a significant non-local migration movement between two specific regions of the country; a movement which furthermore does not fit the models proposed for longer distance migration in section 3.4.3.3. The viewer may be tempted to conclude that there is a specific reason for this movement, perhaps related to reasons such as the economic rise or downturn of a particular region. This, however, is not really the case: the two regions in question are in fact Liverpool and North Wales, which are situated close to one another on a map of the UK. Thus, the migration between them can be explained purely by their proximity; nothing out of the ordinary is occurring. The problem is that the linearising process applied to the UK has split up two regions which are geographically close together: examination of the map in figure 3.9 reveals that the ordering used first visits North Wales, then wanders around the Midlands, the South East, East Anglia and even South Yorkshire before returning to Liverpool.

The converse problem to this - the fact that some migration flows may be obscured by an unfortunate ordering of the matrix - was apparent when studying ward level migration in the Cardiff region. While it seems possible to deduce

broad hierarchies of connections from the visualisation, the viewer should be careful if trying to gauge the relative importance of such connections by visual methods alone.

Such limitations are unavoidable side-effects of the linearisation process. The algorithms used are reducing two dimensional data to a single dimension and as such, can never perfectly replicate the properties that the data possessed in its original form. These limitations therefore highlight above all the need for the user to be aware of the linearisation process and its possible consequences.

3.4.5.3 Future improvements

It is possible to envisage a number of incremental improvements which would enhance comprehensibility of such a visualisation system in future.

- User-controlled restriction of origin/destination areas to those of interest, and thresholding flows to levels of interest, are both mentioned in Marble et al. (1997) and would certainly be of use.

- The ability to search by name for places of interest in the interaction matrix has obvious benefits.

- It would be helpful to leave choice of data reduction formula for the zoom tool to the user, for example allowing them to select whether the average, maximum or total of all contained points is used to determine the colour of a pixel. In the case of the Property Market Correlation Pixel Matrix Plots discussed in section 3.6, it is useful to implement more complex algorithms, for example assigning the maximum positive data value to the red level of a pixel at the same time as assigning the maximally negative data value to the blue level of a pixel.

- When choosing colours for the linearised map, it would be useful to specify whether this is done so globally (thus ensuring continuity between zoom levels) or locally (thus displaying greater contrast within each level).

- Likewise, when colouring the interaction matrix, it is also useful to control whether the data is normalised globally (thus ensuring continuity between zoom levels) or adjusted locally for the section of the matrix currently displayed (thus displaying greater contrast within each level).

- When studying origins and destinations far apart, it would be useful to split the interactive map into two panes, showing both the origin and destination zones in greater detail.

- It would also be useful to interactively colour areas in the map to show migration patterns (inward or outward flows) for the place currently under the mouse pointer - using the same colour scheme as in the matrix.

- Finally, due to the possibility of unfortunate matrix orderings having a greater impact on the display of some flows than on others, it would be helpful to allow the user to reorder the matrix at will, to assist with producing plots for specific areas of interest.

3.5 Property Market Time Series Pixel Matrix Plots

This section describes the generation of property market time series plots, for visualising changes in market prices distributed over both space and time.

3.5.1 Structure of the data

In contrast to the interaction data displayed above, each time series data point relates only to one geographical region L, but now it also relates to a time slice, which is indexed by a natural number $\mathbb{N}_0$. Thus the data set can be seen as a function h from the product space of geographical regions and time, to the space of all possible prices expressed as real numbers:

$$h : L \times \mathbb{N}_0 \mapsto \mathbb{R} \tag{3.14}$$

As before, regions are defined by the grid coordinates of points which fall inside them:

$$g : \mathbb{R} \times \mathbb{R} \mapsto L \tag{3.15}$$

thus a property market time series can be seen as a three dimensional function from geo-coordinates and time to some financial quantity:

$$\mathbb{R}^2 \times \mathbb{N}_0 \mapsto \mathbb{R} \tag{3.16}$$

3.5.2 Methodology

The geographical locations are reduced from a 2-dimensional to a 1-dimensional data space, exactly as in section 3.4.2.3. A plot can then be produced with 'location' on the x-axis, and 'time' on the y-axis; where as before, the location axis has the following properties:

1. locations which are close together in real space, will be close together on the axis; likewise locations which are far apart in real space will be far apart on the axis;

2. locations which are close together on the axis, will be close together in real space; likewise locations which are far apart on the axis will be far apart in real space.

In all cases, the logarithm of data values $log(d + 1)$ is taken so as to 'tame' larger data points and prevent them from obscuring any structure present at smaller levels of interaction.

3.5.3 Discussion

Figure 3.15 shows the result of applying this process to the Land Registry data. In this case, absolute average prices are plotted for each Local Authority. It is immediately apparent that the plot does not change much as time progresses (moving down the page) - the 'data', in this case, appears to be constructed of continuous vertical lines. Consequently it must be concluded that changes in house prices over time for each area are far exceeded by the disparity between different areas. The only tangible pattern apparent from this plot is that London and the South East exhibit higher house prices on average than the rest of England and Wales; the brightest vertical band being in the region of Westminster.

There is a second reason why the trend of prices generally rising over time is not easily visible in figure 3.15. The human eye is not sensitive to gradual colour

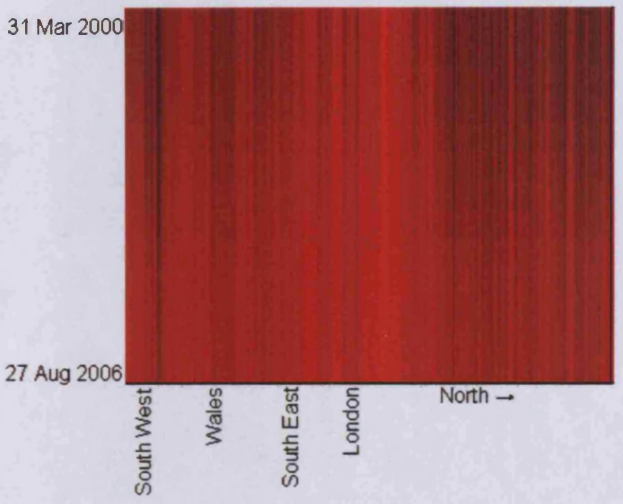


Figure 3.15: Visualisation of average Local/Unitary Authority house prices from 2000 to 2006. Each time slice represents 180 days. Relatively little structure is seen in terms of change over time, because such differences are obscured by the initial disparity between prices in different regions.

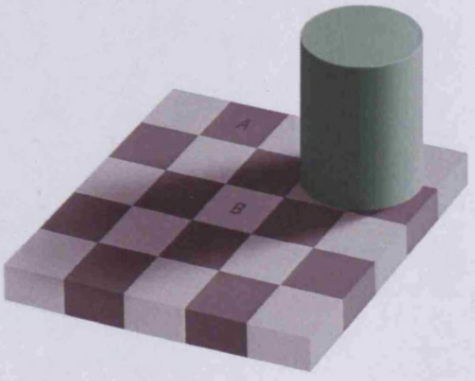


Figure 3.16: Illusion demonstrating problem with visualisation in figure 3.15. The human eye is bad at judging absolute colour values; demonstrated here by the fact that square A and square B are shaded exactly the same colour. Image courtesy of Edward H. Adelson and Wikimedia commons.

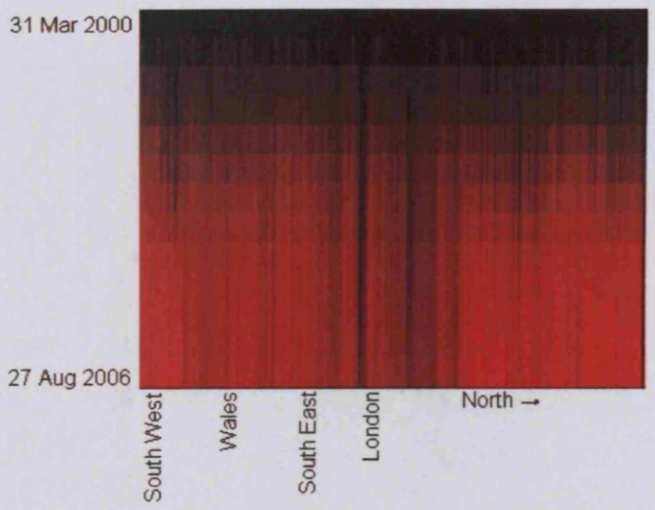


Figure 3.17: Visualisation of Local/Unitary Authority price indices from 2000 to 2006. Each index is relative, i.e. it starts with a value of 1, so pixel brightness shows local prices relative to prices for the same area in the year 2000. Each time slice represents 180 days.

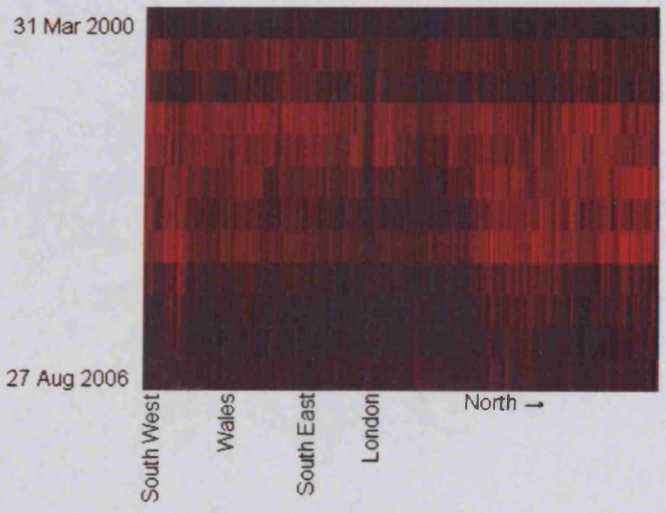


Figure 3.18: Visualisation of Local/Unitary Authority price change from 2000 to 2006. Pixel brightness shows price *change* i.e. this is the first derivative in time of the plot in figure 3.17. Each time slice represents 180 days.

Figure 3.19: Map of UK Wards with the greatest relative increase in price between 2000 and 2006. Wards with < 500 transactions are not counted.

gradients, nor good at judging absolute as opposed to relative colour. Figure 3.16 shows a trick where the eye is deceived into thinking identical colours are different; conversely in the absolute time series plot, differing colours (at the top and bottom of the plot) are interpreted as being the same.

This suggests that a different method is needed for visualisation, which is presented in figure 3.17. In this case, prices for each Local Authority are normalised to remove the disparity between different areas, by dividing all indices by their own value at the beginning of the period studied. Thus, instead of visualising absolute average prices, each price index now begins at 1.0. It is now apparent that, having initially started with higher house prices, London and the South East do not grow as much in price over the time period under consideration. Thus, the gap between London and the periphery is seen to be narrowing.

Finally, growth can be visualised directly by taking the derivative of the data in figure 3.17. The result is given in figure 3.18. This shows more clearly the time slices in which the most growth occurred. In the 180 day period beginning in October 2001, growth occurs fairly uniformly across the country (except in some areas of Central London); outside of the South East, significant growth is sustained until September 2004. In September 2003 another significant wave of growth occurs nationwide, including the South East although it is not as strong here as elsewhere.

For comparison, figure 3.19 displays in a more traditional manner, the wards in which the most growth occurred between 2000 and 2006. However, as this is plotted on a conventional map of the UK it cannot (unlike figure 3.18) show in which years, as well as in which wards, the greatest growth occurred.

It is clear that further interactive improvements to the time series visualisation are possible. These would mostly be of a similar nature to those discussed for the migration matrix visualisation tools in section 3.4.5.3.

3.6 Property Market Correlation Pixel Matrix Plots

Another visualisation technique which is used later in this study is the *Property Market Correlation Plot.* This is a means of viewing the Land Registry data set not as a set of geo-coded time series, but as interaction data - defined, as with

migration, as relationships between pairs of geographical regions. For any two regions A and B, the interaction data in question is a *market correlation* measure which provides an estimate of the extent to which the property market at place A is driven by that at place B, i.e. the extent to which, based on existing data, a price change at B is likely to be followed by a similar price change at A. The derivation of market correlation will be described in chapter 5; however a brief overview of the resulting visualisation is given here.

3.6.1 Structure of the data

The market correlation data sets shown here can each be said to be a function mapping the product space of geographical regions L to the real numbers $\mathbb{R}$:

$$f : L \times L \mapsto \mathbb{R} \tag{3.17}$$

The geographical regions themselves can be specified by a function from grid co-ordinates (latitude and longitude) to regions:

$$g : \mathbb{R} \times \mathbb{R} \mapsto L \tag{3.18}$$

so combining these, it can be seen that the aim is to visualise a four dimensional function:

$$\mathbb{R}^4 \mapsto \mathbb{R} \tag{3.19}$$

The geographical regions are assigned to a linear ordering as described in section 3.4.2.3.

3.6.2 Discussion

Figure 3.20 shows the result. This will not be discussed at this stage, except to say that the large blocks of red situated on the diagonal axis indicate a region of strongly correlated price movement. Thus, as already identified above, is it noticeable that London and the South East have followed a different pattern of development between 2000-2006 than most of the rest of England and Wales.

This and similar visualisations, including display of regression predictions and residuals, were found to be of considerable use in developing the *market*

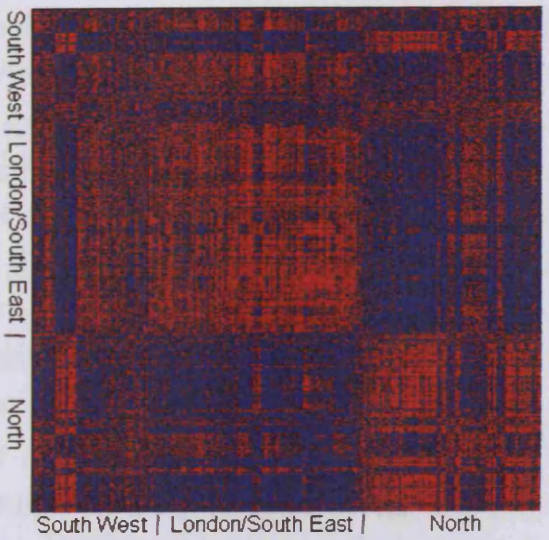


Figure 3.20: Interaction plot of market correlation between all pairs of Local Authorities in England and Wales. Red indicates above average correlation; blue indicates below average correlation.

correlation measure displayed here, however this discussion will be postponed to chapter 5.

3.7 Conclusions

3.7.1 Novelty

A number of visualisation techniques have been presented in this chapter; all of which are based on the idea of reducing 2-d spatial data to a single dimension through linearisation. This has allowed for the display of spatial migration and correlation matrices as well as price change data over time.

To the best of the author's knowledge, this is the first time such a technique has been used to visualise the spatial distribution of price change over time. While Marble et al. (1997) has used similar techniques to visualise US migration data, this study increases the scale and complexity of that work, by application of the CLO-OPT algorithm rather than a semi-contiguous space filling curve. Guo (2007) uses the same process to display simulation outputs, although stays clear of actual demographic data. Also, the technique has never been applied to correlation data before (further details on correlation will be described in chapter

5).

Moreover, the work presented in this chapter improves on all of the above (i) by providing a zoomable tool, (ii) by using political hierarchies to increase intuitive comprehensibility of the orderings, and (iii) by investigating alternative techniques which can be used if computing resources are limited.

Overall, these techniques have proven to be useful both for gaining a general impression of the contents of a large data set, and also for 'sanity checking' of model inputs and results, hypothesis generation, algorithm comparison and program debugging. Several deductions about the nature of migration patterns derived from visualisation have already been presented in this chapter; however it will also be seen in chapter 5 that two significant hypotheses of this thesis are generated through visualisation of property market cross-correlations. Their usefulness notwithstanding, the limitations of these visualisations have also been investigated, and suggestions for overcoming these limitations have been put forward.

3.7.2 Closing comments on social construction

A final advantage of the linearisation-visualisation approach has become evident when presenting this work to colleagues. In our current research climate, numerical techniques in the social sciences have come under criticism for being too 'inhuman': ignoring the complexity of social construction, and making too many assumptions about human behaviour. Visualisation methods such as those presented in this chapter are one possible answer to such criticism, because rather than employing a computer to reduce data to a small set of figures - making many assumptions about meaning and validity along the way - pixel visualisation can present a researcher with a large data set *in its entirety.* Arguably in the case of Pixel Matrix Plots of migration and Time Series Plots of house prices, the data has not been changed at all - instead, a plot of raw data has merely been formatted for easier comprehension. Also, while the nature of the data *has* been changed in the case of Property Market Correlation Plots (they are a derivative of the time series data) - it *has not been overly simplified,* merely transferred from the time domain to a correlation domain. Some information is lost in the latter process, but it is in no way comparable to the data loss inherent in (for example)

reducing the data to a map of price increase as in figure 3.19[3].

The lack of data reduction has the fundamental advantage of *requiring fewer assumptions* about the data. Making assumptions, both explicit and implicit, is inevitable when carrying out any scientific task whatsoever; however there is always the risk that these assumptions will be wrong, especially in the case of implicit assumptions - of which (by their nature) the researcher may not even be aware. Therefore, requiring fewer assumptions overall means that there is less likely to be a false assumption hidden among our set of implicit axioms.

This state of affairs allows the researcher, then, to dynamically make their own assumptions about the data, form their own mental models and make their own conclusions about its meaning, rather than having these assumption, models and conclusions forced upon them by the statistical techniques such as aggregation or averaging. Social construction need not be ignored, because the visualised data is simply another input to the human mind, which is where the real deduction takes place.

[3]It should be pointed out that while the *data* have not been altered, ordering it in different ways reveals different *information*; so it is possible that patterns may be present in the data which are missed by the approach chosen here. However, presenting the data in their entirety rather than in simplified form increases, rather than decreases the number of patterns that could conceivable be spotted; therefore it can only be a good thing if taken in conjunction with more traditional approaches.

Chapter 4

Development of an exploratory housing market regression model (and some basic results)

"But I'm just saying
I don't think you're special
I mean... I think you're special
But
You fall within a bell curve" (Tim Minchin)

4.1 Introduction

From the moment they awake to the moment they sleep, the majority of human beings use vision to assist with every possible task - whether eating breakfast, solving equations or avoiding being hit by a bus. The author recalls attending a presentation course where the vast majority of those present (all of whom were scientists) purported to be 'visual thinkers' - a condition even more common among the scientific community, according to the course lecturer, than in the other groups he had taught. And yet in matters of research, we do not trust vision alone - a sentiment expressed well by physicist William Kelvin: "When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it your knowledge is of a meagre and unsatisfactory kind". As children we were taught to join the num-

bered dots to create a picture. As scientists using a visualisation-based workflow, we must learn to do the opposite: number the dots in the picture to create a narrative.

Regression was created primarily as a tool to perform exactly this task, and as such is usually used not for *exploratory* but for *confirmatory* data analysis. However, the use of regression to confirm patterns seen in a visualisation would contradict one of the principles of statistical hypothesis testing: that the hypothesis should be decided upon before the data set is examined. Therefore, the techniques presented in this chapter are instead presented in an exploratory spirit. Firstly, they are used to complement the visualisations of chapter 3 by allowing for the logical discussion of patterns first seen during the visualisation process; also however, they enable the detection of additional patterns missed by the visualisation, and provide the bridge between the two data sets studied - the link between the house price and census data. Thus, they extend the exploratory nature of visualisation into a non-visual, symbolic realm.

Discussing regression at this point means that the chapters in this thesis are presented out of sequence, as the cross-correlation analysis of chapter 5 was in fact performed before this one. However, as the regression model will be used in both chapters 5 and 6, it is logical to discuss it beforehand.

This chapter therefore deals with the development of a generalised exploratory regression model, based on explanatory data relevant to the housing market, which will be used during the cross-correlation analyses of chaters 5 and 6. The model is tested on the simpler problem of predicting average prices on a ward level during the year 2001. The remainder of the chapter is structured as follows: section 4.2 discusses the development of a statistical model, and section 4.3 discusses the choice of explanatory variables. Sections 4.4 and 4.5 present the results of a test analysis used to validate the statistical framework, and section 4.6 concludes.

4.2 Development of a statistical model

4.2.1 Choice of statistical technique

Although the title of this chapter focuses on regression, it should be noted that regression is by no means the only way to achieve the aim of analysing the patterns

seen in visualisations, and searching for patterns not already spotted. This section therefore discusses the other options available, and explains why regression has been chosen.

A statistical technique is needed which fulfils three criteria:

1. Computational feasibility. While the data sets presented in this chapter are only moderate in size (a set of house prices and about 90 census variables for each of 8,850 wards), those to be used in chapter 5 are interaction sets, and therefore encompass a larger set of variables - up to 1000 - for each possible pair of wards (78 million pairs). Whatever analysis is used must be computationally feasible.

2. Reliability. For a tool used to assist in analysing a large data set for the first time, as is the case in this study, complex analyses are rejected in favour of simpler approaches. Note that the techniques presented in this chapter are not themselves the main topic of the thesis; instead they are used to examine the output of other, more novel processes. As such, it is essential to use a tool which can be relied upon to give accurate results, as these will in turn be used to evaluate other tools. There is no reason not to undertake a complex analysis at a later date, but without the simpler results for comparison, it would be impossible to tell whether or not the complex approach is actually better.

   Thus, Occam's razor is applicable to the selection of this technique - the principle that "entities must not be multiplied beyond necessity" (or in the original Latin, *entia non sunt multiplicanda praeter necessitatem*). It may be, however, that the simplest approaches are computationally infeasible, so the approach taken may be better summarised in Einstein's maxim: "Make everything as simple as possible, but not simpler."

3. The final criterion is explanatory power. In particular, it is desired to chose a technique which can reveal complex interactions between variables in addition to calculating basic correlations.

		Ability to speak Latin				
		High	Fair	Moderate	Low	None
Tendency to quote in Latin	High	1	1	0	1	0
	Fair	1	1	0	6	0
	Moderate	3	2	1	1	0
	Low	1	0	0	1	2
	None	0	0	0	6	18

Table 4.1: 2-d illustration of a data cube, for a fictitious survey of PhD students

4.2.1.1 Rule based data mining

A logical starting point for analysis of large quantities of data is rule based data mining. The term 'data mining' has gained popularity of late, mainly in the world of retail, due to the desire of marketing departments to squeeze every last drop of commercially valuable information out of customer transaction databases! Han & Kamber (2006) gives a good overview of principles. Data cube techniques will not be employed in this study, however, they are described here in some detail because they inspired the choice of the alternative methods used, and because they illustrate the fundamental problems of multidimensional analysis.

The first stage of data mining usually involves dividing data points into different buckets based on their parameters; together, these buckets form a multidimensional *data cube*, with the number of dimensions equalling the number of parameters being studied. Table 4.1 shows the process in two dimensions, for a fictitious survey of the tendency of PhD students to use Latin quotations (such as the one above) in their theses. Note that the data cube contains far more information than would a simple correlation coefficient fitted between the two variables. For example, in table 4.1 we see that most (5/6) students with a high level of fluency in Latin have at least a moderate tendency to use Latin quotations, while most (18/20) students with no knowledge of Latin do not use Latin quotations at all. Fitting a coefficient of correlation would reveal the obvious trend that greater fluency in Latin corresponds to greater likelihood to use Latin quotations. However, this misses the fact that a considerable number (8/15) of low-ability Latin speakers have a moderate to high tendency to use it in their work - an anomaly revealed by the data cube (and, the author fears, applicable to the quotation from the previous section!).

Rule based mining processes improve on correlation analysis by searching for rules based on a support / confidence / lift framework. So for a rule of the form 'X implies Y' ($X \Rightarrow Y$),

- Support indicates the proportion of the data set to which the rule applies, $P(X)$.
- Confidence indicates the certainty of the rule, i.e. the probability of the rule being true for an arbitrary item to which it applies: $P(Y|X)$.
- Lift is a measure of the extra knowledge represented by the rule, and is calculated as the confidence of the rule relative to the confidence of making such a prediction without using the rule: $P(Y|X)/P(Y)$.

So for the data set above, some examples of rules might be:

- Students with low ability to speak Latin will have fair tendency to quote in Latin

  Support = 6/46 (13%), Confidence = 6/15 (40%), Lift = $\frac{6/15}{8/46} = 2.30$
- Students with low ability to speak Latin will have high tendency to quote in Latin

  Support = 1/46 (2%), Confidence = 1/15 (7%), Lift = $\frac{1/15}{3/46} = 1.02$
- Students with high ability to speak Latin will have moderate tendency to quote in Latin

  Support = 3/46 (7%), Confidence = 3/6 (50%), Lift = $\frac{3/6}{7/46} = 3.29$

Such rules are then filtered (for example, discarding rules with low support, confidence or lift, depending on the requirements of the researcher) and merged if they refer to adjacent sections of the data cube. The end result should be a set of rules which effectively describe the data set. It would be hoped that the end result of data mining the above table would be

- a rule showing the tendency of students with no fluency in Latin to avoid using Latin
- a rule showing the tendency of students with fluency in Latin to use Latin

- a rule showing the anomaly whereby some students with a little knowledge of Latin use it in excess of their ability.

Thus, complex *interactions* between variables, beyond simple correlation, can be automatically captured.

4.2.1.2 The curse of dimensionality

The technique illustrated above is easy enough to understand and compute when applied to two variables. However, the computational complexity increases exponentially with the number of dimensions. If, for example, we have five categories per variable (continuous variables are usually divided into discrete intervals), then for n variables we will require 5^n buckets, where each bucket represents a unique combination of categories from each variable, that is to say, it is a single box in the data cube. Even for the 90 variables used in this chapter, an estimated 10^{53} gigabytes of buckets would be required - and one of the analyses in chapter 5, which uses 1000 derived variables, would require 10^{689} 'gigabuckets'.

As the count of data items (78 million) is vastly smaller than 10^{689}, most of these buckets would be empty, so the data cube is *sparse* and the patently ridiculous amount of storage just quoted would not be required. Still, however, the mining process would need to check for the existence of data in each of these buckets - a task which remains computationally infeasible.

Several approaches are commonly used to mitigate this problem (Han & Kamber 2006):

1. Limited rule based mining. It is possible, instead of searching for interactions between *all* possible combinations of variables, to search for interactions between (for example) every *pair* of variables. This would mean performing up to 1000^2 computations using a simple 5×5 grid such as that in table 4.1 - a task which is feasible. However, more complex interactions would be missed.

2. Heuristic rule based mining. 'Intelligent' algorithms can be used to guide the selection of further rules to seek, based on rules already found.

3. Clustering. This is a process whereby points in the data set are repeatedly combined with near neighbours to produce *clusters*. Typically, the user

must specify how many clusters they wish to see, and the clustering algorithm will vary the degree of aggregation accordingly. Once similar points have been grouped together, the clusters can be described by rules: clustering the data of table 4.1 would hopefully produce the same final three rules as rule-based data mining.

Clustering takes advantage of the sparseness of the data cube, because the complexity of the algorithm scales polynomially with the number of points p, as $O(p^2)$ - instead of exponentially with the number of dimensions, as $O(2^n)$.

The three approaches listed above all have disadvantages. In the first case, any interaction which is a product of more than two variables will be missed. In the second and third, any form of interaction could potentially be missed, depending on whether or not the heuristic search or clustering process finds it. This is not to say that these processes are not worthwhile, and should not be employed in future! However, for the time being, the curse of dimensionality has forced the introduction of greater complexity to the data mining framework, and this now violates the principle of Occam's razor stated in section 4.2.1, as another, simpler technique can be used to process the data: Multivariate Linear Regression.

4.2.1.3 Multivariate Linear Regression

Regression is the process of drawing a best-fit line through a range of data points. While it misses many of the subtleties of interaction revealed by the data cube, the previous section demonstrated that data cube analysis is not possible for the large number of variables present in this study, without employing more complex algorithms. In light of this, linear regression becomes a reasonable first step for understanding the data set. While limited rule based mining is capable of detecting *any* interaction between any *pair* of variables, multivariate regression is capable of picking out any *linear* interaction between *all* of the variables. Crucially, when multiple variables all appear to have an effect on the target variable, linear regression can tell us which one best models the target phenomenon. This is a useful property in a study involving detailed sociodemographic data in which many variables will exhibit correlations with one another, but many such correlations will be far better explained by a smaller subset of variable interactions.

The trade-off made by linear regression is that by restricting the search for interactions to linear rules, it is not necessary to restrict the subset of variable combinations that are searched.[1]

Multivariate linear regression also has the advantage of being a well-understood and frequently used approach in the biological, social and economic sciences. It therefore seems reasonable to conduct a multivariate regression as a benchmark against which other techniques can be compared in future.

In keeping with the visualisation philosophy presented in chapter 3, numerical techniques are not relied upon in entirety. In addition to computing regression coefficients, scatterplots are also viewed of each regression dimension versus the target variable. Due to the large quantities of data points involved, these are shown as *scatter density plots* rather than simple scatterplots, allowing for the viewing of detail where points on a scatterplot become too dense for meaningful analysis. In chapter 5, some instances will be presented where viewing of these scatterplots allowed for the discovery of data features missed by the linear regression.

Finally, in keeping with the principle of reliability, a correlation coefficient can be computed between each explanatory variable and the target, because this is a process simpler than that of linear regression. The correlation coefficient suffers from the disadvantage of not recognising multi-variable interactions, nor being able to identify which of a set of strongly correlated variables best models the target phenomenon, so it is best viewed as a simpler yet inferior analysis conducted at the same time. However, it also provides a valuable fall-back, and a point of reference in case any of the assumptions upon which linear regression rests happen to be violated.

Section 4.2.2 discusses the assumptions inherent in linear regression and the steps taken to verify them.

Assumption	Consequence of violation	How checked
Data can be explained by linear relationship	Model is mis-specified	Assumed false (though scatter-plots of explanatory vs target variables still used)
Explanatory variables are free of error	Parameters may be underestimated	Assumed false
Mean residual is zero	Model is mis-specified, possibly variables left out	Value computed and checked
Errors are not spatially correlated	Model is mis-specified	Residuals mapped, Moran I test
Errors are homeoscedastic	Model is mis-specified *and* confidence intervals may be wrong	Plot of residuals vs prediction
Errors are approximately normally distributed	Confidence intervals may be wrong	Histogram plot of residuals vs Gaussian curve
Explanatory variables not exactly collinear	Model unsolvable (multiple solutions)	Assumed true as PCA used
Explanatory variables not seriously collinear	Parameter estimates unreliable	Assumed true as PCA used

Table 4.2: Assumptions of linear regression

4.2.2 Assumptions of Linear Regression

Table 4.2 lists the assumptions inherent in performing a linear regression analysis. The treatment of these is discussed below. In some cases it is permissible to violate the assumptions; in other cases they must be rigorously checked. However it should be noted, that as the data is divided into a 'training' and 'test' set, the effects of violating any regression assumptions are minimised, as checking of errors from a test set will highlight the presence of any bad regression model where more complex analysis of residuals may have failed to do so.

4.2.2.1 Linearity

As its name suggests, linear regression assumes a linear relationship between the explanatory and target variables - but, in the complex real world of market interactions, it is unlikely that this will hold! Three justifications are presented for violating this assumption. Firstly, the consequences are not dire: nonlinearity merely means that the model thus derived may not be the most accurate possible. As the models will be tested for accuracy (both by checking residuals and errors) it will be possible to decide experimentally whether they are accurate enough for the task at hand. Secondly, regression is being used not as the final analysis, but instead to complement the visualisation process. It allows us to codify and enumerate the contents of the visualisations, so we need not rely on it in totality. Finally, for reasons of computational feasibility, we have chosen to violate the assumption of linearity in order to study multivariate interactions in a very large data set. Without taking such a step, this study would not be possible.

4.2.2.2 Accuracy of explanatory variables

It is not assumed that the measurement of the explanatory variables is free of error. Even if the census were completely accurate on matters such a population count, it should be noted that population would have changed over the six year period of the study. However, the consequence of violating this assumption is

[1]It should be noted that other nonlinear multivariate regression techniques exist, for example, Stepwise regression or Multivariate Adaptive Regression Splines (MARS). The algorithms for these are comparable to those of heuristic rule based mining: regression terms are repeatedly added and removed based on a heuristic guess as to their ultimate usefulness. Therefore, while these techniques may be valuable, they are rejected in this study for the same reason that heuristic data mining is rejected:- the more reliable approach should be tried first.

usually that coefficients will be 'diluted' - that the effect of explanatory variables will be underestimated. For the reasons described in section 4.2.2.1 this is not considered a major problem. Also, it is usually considered sufficient to assume that the errors in the explanatory variables are insignificant compared to the errors in the target variable. When the target variable is a measure derived from market fluctuations (as will be the case in chapters 5 and 6), this will almost certainly be true.

Note that this assumption is not necessary if an errors-in-variables model is used. However, such models are more complex and more computationally expensive than simple least squares regression. Computational cost is a major concern for the datasets which will be analysed in chapter 5, so errors-in-variables models are not used. Also, as the large data sets in this study have necessitated the coding of custom software, implementation complexity is also a concern, because it is hard to prove the correctness of complex statistical software without extended trials.

4.2.2.3 Residuals have mean of zero and are not spatially correlated

Violation of either of these assumptions implies a mis-specified model, which could be improved by inclusion of an extra variable. Again, for the reasons described in section 4.2.2.1 this is not considered serious. However, testing for violation of the assumptions is undertaken nonetheless; through spatial mapping of the residuals and application of the Moran I test. The latter is performed as described in Cliff & Ord (1981), and for speed of computation, using a 20km-threshold weight matrix of the same type as in e.g. Mella-Marzuez & Chasco-Yrigoyen (2006), and a permutation approach to generate inference statistics.

4.2.2.4 Homeoscedasticity of errors

If the errors are not homeoscedastic - that is, errors tend to be greater in certain parts of parameter space than in others - this can imply a mis-specified model, and also may mean that confidence intervals calculated for parameters are inaccurate. The latter consequence is considered serious, as the calculated confidence intervals are used to decide on whether or not regression findings are significant. After Mather (1976), this is checked via scatterplots of residuals vs predictions. It is also considered prudent to check plots of residuals vs explanatory variables;

however this is not necessary as plots of target vs explanatory variables are being produced as part of the visualisation methodology (section 4.2.1.3) and any heteroscedasticity should be visible in these - the plot would take the form of a wedge or triangle shape rather than an uncorrelated scattering.

4.2.2.5 Approximate normal distribution of errors

The assumption of normality of errors is likewise necessary for accurate computation of confidence intervals, and is therefore important. If the other assumptions of regression are fulfilled then for a sufficiently large data set, central limit theorem dictates that normality of errors will hold as the errors are the sum of a number of small independent variables (Mather 1976). However, as not all of the above assumptions are fulfilled, histogram plots of residuals are made and compared with normal probability density functions with the same location and scale to check for approximate normality.

4.2.2.6 Collinearity

In the case of exact collinearity of explanatory variables, linear regression is an unsolvable problem as infinitely many valid combinations of parameter will exist. In the case of major (but not exact) collinearity, parameter estimates are still unreliable; ordinary least squares regression may derive unusually large parameter values which happen to cancel for the purposes of the data at hand. This problem is a form of *overfitting* - whereby it is possible, by choosing ridiculous parameter values, to represent the training data set more accurately, albeit at the expense of destroying the predictive power of the regression when applied to data outside of the original set.

Both of the problems of collinearity are avoided by use of Principal Component Analysis for dimensionality reduction.

4.2.3 Use of Principal Component Analysis (PCA)

Principal component analysis is performed using the mdp toolkit in Python (Zito et al. 2009). This section discusses the need for PCA, interpretation of results and choice of an appropriate number of dimensions.

4.2.3.1 Need for dimensionality reduction

Reducing the number of dimensions is necessary for three reasons:

1. Elimination of multicollinearity. As discussed in section 4.2.2.6, this causes technical problems with parameter estimation during regression.

2. Automatic variable selection. Part of the purpose of multivariate linear regression is to be able to automatically analyse a large number of variables and discover which are the most relevant to the phenomenon at hand. With so much data available, it is preferable to be able to do this rather than specifically choose a small subset of variables for analysis. In general, the approach has been taken of including a large number of variables so it is necessary to be told automatically which are the most important. Linear regression can do this, although with large numbers of variables, the computed confidence intervals for each coefficient become large so the results become largely meaningless if dimensionality reduction is not also used.

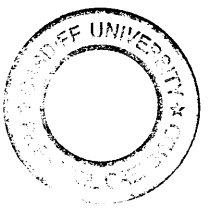
3. Ease of computation. The computational complexity of linear regression increases linearly with the number of dimensions, so the reduction of the data set to 40 dimensions from (in some cases) up to 1000 initial variables has a significant effect on the time taken to compute a regression.

4.2.3.2 Interpretation of results

PCA can be used to achieve dimensionality reduction by finding a number of orthogonal linear combinations of the input variables - called *components* - from which most of the variance in the original data set can be reconstructed, while discarding components which account for little of the variance in the input data. Regression is then performed on the output of the PCA process. A disadvantage of this is that when regression coefficients are computed, they relate to PCA components rather than directly relating to input variables, which can make interpretation of the results difficult. Therefore in this study, a form of reverse transform is applied to deduce parameters for each variable. This transform is described below.

The PCA regression model predicts a relationship of the form

$$y = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + ... \quad (4.1)$$



where y is the target variable, β_n are the component coefficients and C_n are the PCA components, each defined thus:

$$C_n = \gamma_{(n,1)}x_1 + \gamma_{(n,2)}x_2 + \gamma_{(n,3)}x_3 + ... \tag{4.2}$$

where x_p are the input variables and $\gamma_{(n,p)}$ is the PCA coefficient for component n and input variable x_p. Substituting (4.2) into (4.1):

$$\begin{aligned} y = \beta_0 \; &+ \; \beta_1(\gamma_{(1,1)}x_1 + \gamma_{(1,2)}x_2 + \gamma_{(1,3)}x_3 + ...) \\ &+ \; \beta_2(\gamma_{(2,1)}x_1 + \gamma_{(2,2)}x_2 + \gamma_{(2,3)}x_3 + ...) \\ &+ \; ... \end{aligned} \tag{4.3}$$

Multiplying and collecting terms;

$$\begin{aligned} y = \beta_0 \; &+ \; (\beta_1\gamma_{(1,1)} + \beta_2\gamma_{(2,1)} + ...)x_1 \\ &+ \; (\beta_1\gamma_{(1,2)} + \beta_2\gamma_{(2,2)} + ...)x_2 \\ &+ \; ... \end{aligned} \tag{4.4}$$

The terms in brackets represent the contribution of each input variable x_n towards the target variable y, thus it can be said that:

$$contribution(x_p) = \beta_1\gamma_{(1,p)} + \beta_2\gamma_{(2,p)} + ... \tag{4.5}$$

As PCA is merely a rotation and scaling of parameter space, the coefficients $\gamma_{n,p}$ are free from error. Hence, as β_n are assumed normally distributed with standard deviations σ_{β_n}, the standard deviation of $contribution(x_p)$ can be assumed to be

$$\sigma_{contribution(x_p)} = \sqrt{(\sigma_{\beta_1}\gamma_{(1,p)})^2 + (\sigma\beta_2\gamma_{(2,p)})^2 + ...} \tag{4.6}$$

as the variance of a sum of normally distributed variables can be calculated by the formula $\sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2}$. This enables the calculation of confidence intervals for each explanatory variable. Figure 4.1 illustrates the entire process.

Such unpacking of regression results for PCA components, into the total contribution of each input variable to the model prediction, is a novel and unusual method of displaying coefficients. It is employed to assist in interpreting results

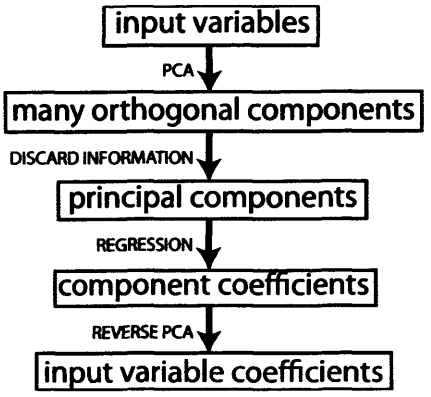


Figure 4.1: Illustration of the combined PCA and regression process

from the regressions which in chapter 5, will contain on the order of 1000 variables. However, care is still needed with the interpretation of such results, which in the case of variables which are not strongly represented in any of the PCA components selected for regression, may not accurately reflect the importance of those variables.

4.2.3.3 Choice of dimensionality

The problem remains of choosing an appropriate number of dimensions to represent the data. The primary reason for dimensionality reduction, as discussed in section 4.2.2.6, is to eliminate problems of overfitting. Figure 4.2 illustrates this. The Land Registry and Census data sets are randomly divided into two subsets, one for 'training' and the other for test purposes, and a test regression of house price inflation against a set of approximately 90 census variables is conducted, using only the training set. Two separate quantities are plotted:

- mean square residual - a measure of goodness of fit of the regression to the training set, and
- mean square error - a measure of the accuracy of predictions from the training set when applied to the test set.

It can be seen that if more than 75 dimensions are used, the model is overfitted and errors (computed from checking against the test set) are huge. This is likely to be because of multicollinearity in population variables - e.g. the sum of the variables defining population by age is likely to equal the sum of

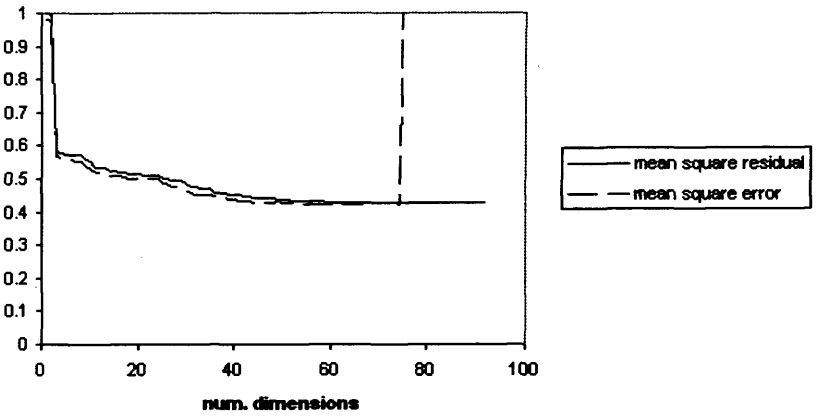


Figure 4.2: Plot of mean square residual, and mean square error vs number of dimensions used, for regression of 2000-2006 growth. Residual/error values are normalised such that a value of unity is equivalent to modelling the entire population by computing the mean target variable. Note that after dimension 75, the model falls victim to multicollinearity and as a result is over-fitted: although the mean square residual continues to fall, the mean square error suddenly rises off the chart, eventually reaching a staggering maximum of about 370,000.

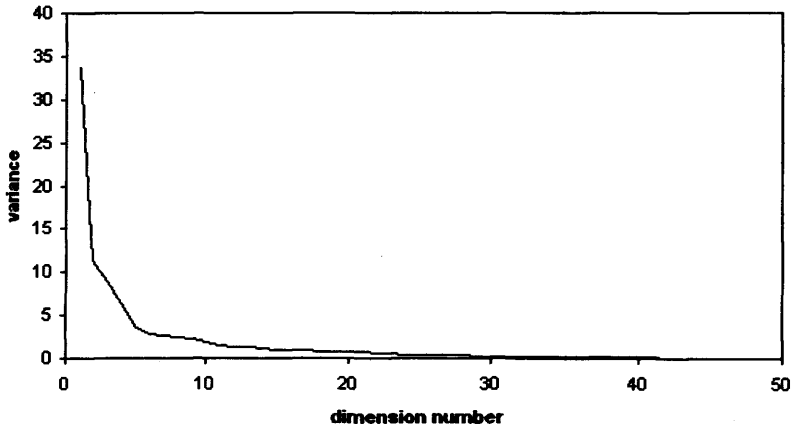


Figure 4.3: Plot of dimension variance for regression of 2000-2006 growth

variables defining population by social class, which is also likely to equal the total population. However, if fewer than 75 dimensions are used, the mean square error is similar to the mean square residual - i.e. the regression model makes accurate predictions about points not in the original set.

Computing a regression for each possible number of dimensions is too expensive a process to use throughout this study, so has only been conducted here to put an approximate upper bound on the number of dimensions which can be used. More commonly, the number of dimensions is chosen by use of a variance plot as in figure 4.3, which shows the amount of variability in the data accounted for by the inclusion of each extra dimension. Here, it can be seen that when using even as many as forty dimensions, the amount of unaccounted-for variability is negligible. As this is well clear of the 75-dimension limit at which overfitting occurs, it can be seen that the exact choice of number of dimensions is not important - any of a wide range of dimensionalities could be used. Therefore, 40 dimensions are chosen as safely representing most of the variability in the data (even leaving room for the extra dimensionality of the data sets expected to be encountered in chapter 5) while staying well clear of the problems caused by overfitting.

It should be noted that PCA is based on the assumption that dimensions accounting for little variability in the explanatory data set will also account for little variability in the target variable. This is not usually considered to be a problem, if explanatory variables are sensibly chosen. In this case, the specific assumption is that major differences in the housing market are caused by major, rather than minor, sociodemographic variation. This was shown to be the case in the graph of figure 4.2, because beyond a certain point, adding extra dimensions did not increase the goodness-of-fit of the model.

4.2.4 Development of software

4.2.4.1 System Architecture

Owing to the large size of the datasets to be studied in chapter 5, it was necessary to hand-code a linear regression algorithm which does not load the entire data set into memory, but instead makes multiple passes through the source files on disk in order to accumulate the required parameters. This was conducted in Python, using the pylab interface to the open-source numpy and matplotlib libraries to

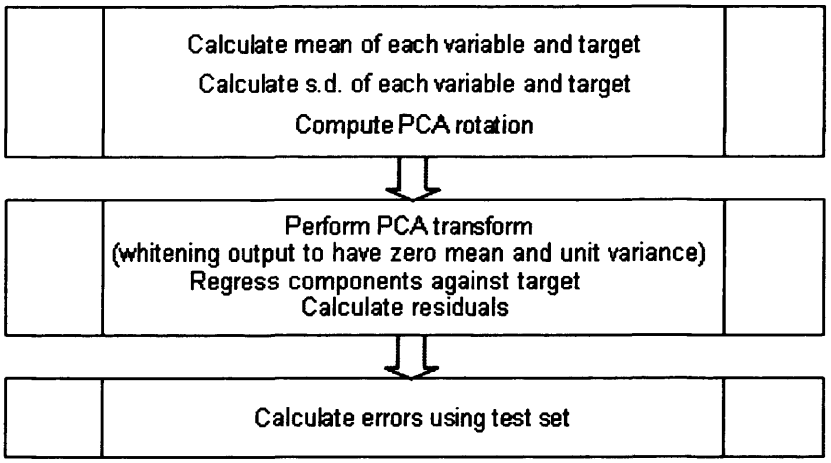


Figure 4.4: Serial passes through the data set made by the regression engine.

assist with linear algebra and graph display functions. Custom modifications were made to the PCA whitening mdp node, in order to perform PCA based on normalised input variables, therefore ensuring that components represent the relative importance of different variables within the data set rather than simply being a reflection of units of measurement. Thus, normalisation (to zero mean and unit variance) is undertaken both before and after the PCA stage. Figure 4.4 shows the passes through the data set made by the regression engine.

The software is engineered according to the 'small-tools' philosophy embodied in Unix. In other words, each program is run from the command line and executes one specific task. The largest program is the regression engine, but several smaller programs are used to split census data tables, normalise, sort and take logarithms of data before regression, and calculate the Moran I statistic after regression. Shell scripting, or makefiles can then be used to combine multiple programs to carry out more complex tasks.

The data is held in plain text comma separated value format, with only one variable per file. This allows for easy specification of the variables to include in any regression, simply by specifying the set of input files on the command line. However, as variable files must be read in parallel, it is important that each input file is pre-sorted before regression. The sorting order used is the same as in the linearised visualisation of chapter 3.

Variables are stored in a directory structure which separates ward from LA

data, thus allowing quick selection of all data from a given level. A sparse format is used - meaning that rather than storing a literal zero to disk when a variable has a value of zero, no data at all is stored. This saves considerable space when storing the migration interaction datasets.

The software is run on the Advanced Research Computing at Cardiff (ARCCA) facility. This allows for parallel execution of several regression runs concurrently, with no loss of speed. It is also possible to parallelise the PCA regression itself; however this level of speed was not necessary for the initial regressions on geographical space.

4.2.4.2 Validation

As is a problem with any software which crunches a large quantity of data into a small set of figures, it is difficult to know whether those figures are correct, or whether a bug in the software has caused it to present the wrong answer. Both the regression and Moran-I software was validated by means of a number of small test cases. A larger scale - though less formal - test of the software is also conducted in section 4.5.2 by comparing the results of a regression on 2001 house prices, with the qualitative predictions of a number of known models of the property market.

Before conducting a regression, however, it is necessary to select the set of explanatory variables.

4.3 Choice of variables

It is likely that explanatory models of the housing market have existed ever since humankind first started assigning financial value to pieces of land, and almost any such model can be used to suggest a choice of variables for a regression framework. Therefore, rather than committing to any specific model, variables from several common models are combined to produce a dataset which should be meaningful to housing market analysis. In this section, a list of variables is constructed, with reference to the models which inspired their selection.

As the analysis employed is spatial rather than temporal, variables which are known to have an impact on the housing market, but which are spatially homogeneous (such as the overall interest rate) have been discarded.

In the case of many variables discussed, there is a choice to be made between relative and absolute variables. For example, if considering the effect of the 16-25 year old population, do we express this as an absolute count of individuals or instead as a proportion of the total population of all ages? The approach taken is to use both: this is in the spirit of using a regression model which can cope with a vast amount of data and pick out only the important factors. Obviously, adding both relative and absolute statistics introduces serious multicollinearity to the data; however as PCA is used to reduce dimensionality, this is not a problem.

As is standard practice (Openshaw 1995), the logarithms of all absolute financial and population variables have been taken in order to remove outliers. To avoid taking a logarithm of zero, this has been achieved with the formula $log(x + 1)$.

Also, as discussed in chapter 2, variables are included on two spatial scales simultaneously: both on the level of the individual ward, and of the containing Local Authority. This could be considered a crude form of multi level model; while the target variable is not split into regional and local components, the explanatory variables are supplied in such a manner. Thus if, for example, Local Authority level statistics are found to be better determinants of local price change than local statistics, then the regression model should detect this.

The models giving rise to choice of variables are discussed below.

4.3.1 Life cycle models

Meen (2001), Murphy & Muellbauer (1993) and Orford (1999) all refer to life cycle models of migration. Obviously various different 'life cycles' are possible depending on the individual, but an example might be that of a worker growing up in the countryside, seeking their fortune in a central urban area, later migrating to the suburbs to raise a family, and finally retiring once more to the countryside. Such cycles of migration are apparent in the visualisations of section 3.4.3.1.

If properties in different regions are typically purchased by inhabitants of different ages, it is reasonable to expect that the economic situation of buyers and sellers will vary according to the age distribution of people in the region. Therefore a detailed breakdown of local population by age is included in the regression dataset, using the following variables:

- number of children under the age of 16

- number of 16-25 year olds
- number of 26-35 year olds
- number of 36-45 year olds
- number of 46-55 year olds
- number of 56-65 year olds
- number of 66-79 year olds
- number of people aged 80 or above

These are taken from census table UV004.

4.3.2 Bid rent theory

Bid rent theory was first introduced by Von Thunen in 1826, though both Orford (1999) and Meen (2001) note its ongoing relevance to housing market structure. When applied to residential property, the implication is that home buyers make a choice of housing location based on a trade off between commuting costs and land rent. This suggests that local travel-to-work distances will have an impact on the housing market. The following variables are therefore included:

- population with travel to work distance under 2km
- population with travel to work distance 2km - 5km
- population with travel to work distance 5km - 10km
- population with travel to work distance 10km - 20km
- population with travel to work distance 20km - 30km
- population with travel to work distance over 30km
- population with no fixed place of work
- population working outside the UK
- population working offshore

These are taken from census table UV035.

4.3.3 The Ripple Effect

Many sources refer to a so-called "Ripple effect" (see Meen 2001, for an overview) whereby price movements in the UK market appear to start in London and the South East, and the rest of the country follows. However, the cause of this is debatable. Is it because the rest of the country turns to London to set their

expectations on the underlying state of the market? Or is it because certain socioeconomic variables differ between the North and South - such as the degree of property speculation, or the average Loan-To-Value ratio (and hence sensitivity of the market to changes in the interest rate)? In the latter case, if the variables are measurable, they should be used in a regression model, but if not they may cause *spatial coefficient heterogeneity*, whereby deduced regression coefficients appear to vary depending on the region from which they were derived.

A simplistic approach is adopted in this study; a single variable is included:

- distance to London

on the assumption that London and surrounding areas will behave differently to everything else. Note that it may seem somewhat odd to include an explicit measure of space in this spatial model, when it would be more enlightening to look at other spatial variables for explanatory power (therefore answering the question 'why do different areas behave differently?' rather than 'do different areas behave differently?'). However, as this study is an exploration of data rather than an attempt at detailed model construction, selection of the variable with most explanatory power is left to the regression engine - and distance to London is included in the variable set.

Distance to London is calculated from the UKBorders dataset, and measured from the centroid of each areal unit.

4.3.4 Speculative models

In an environment where home buyers increasingly engage in property speculation with their own residence - relying, for example, on an increase in the price of their house to fund a retirement plan - speculative models are of relevance. Cameron et al. (2005) and Meen (2001) note that buyers may be attracted by anticipated capital gains on a property, but are also tempered by considerations of affordability. The following variable is therefore included:

- average 2001 house price

This is derived from the Land Registry dataset, and is used when regressing for a target variable of relative growth (but not, obviously, when regressing for a target variable of average 2001 house price)!

4.3.5 Models of supply and demand

Cameron et al. (2005) and Meen (2001) note that supply and demand enter the housing market in two ways. Firstly, potential inhabitants can choose an area in which to live based on a risk-reward calculation; namely trading off the average income in an area against the chances of gaining employment there. Secondly, and more directly, the quantity of available housing stock will have an impact on its price. The following variables are therefore included:

- log average weekly income
- total population
- housing stock: number of household spaces
- housing stock: number of occupied household spaces
- housing stock: number of unoccupied household spaces
- housing stock: quantity of second residence/holiday accommodation
- housing stock: number of vacant household spaces
- employment: number of people employed
- employment: number of people unemployed
- employment: number of people categorised as 'other'

Population, housing stock and employment data is taken from census tables UV004, UV053 and UV028 respectively. Average weekly income data is available separately from the Census office.

4.3.6 Submarket-based models

Orford (1999) states that "the principle of stratification of a housing market into subsets is widely recognised in the valuation literature" (page 79). These subsets can be based on spatial division, or type of property. In the former case, our model is already explicitly spatial; however in the latter case we can extend it by including data on the types of property inhabited by households (household data) and types of dwellings, using the following variables:

- num. households
- num. households in an unshared dwelling
- num. households in bungalow
- num. households in detached house

- num. households in semi-detached house
- num. households in terraced housing (including end terrace)
- num. households in flat, maisonette or apartment
- num. households in a purpose-built block of flats
- num. households in part of a converted or shared house
- num. households in a commercial building
- num. households in a caravan or other mobile or temporary structure
- num. households in a shared dwelling
- num. dwellings
- num. dwellings unshared
- num. dwellings shared

The household and dwelling data are found in census tables UV056 and UV055 respectively. An extra variable is included to help distinguish between urban and rural areas, in case different submarkets exist for each:

- population density (from census table UV002).

4.3.7 Models of market turnover

Meen (2001) notes that there is a high level of correlation, both in the US and UK, between price movements and transaction frequency. Theoretically, in an efficient market this should not be the case; however in the real world they "exhibit strong autocorrelation ... [although] there is little evidence that either transactions affect prices or vice versa" (page 20). In other words, it is not yet known why this correlation exists. In any case, the following variable is included:

- number of transactions between the years 2000 and 2006

This is derived from the Land Registry dataset.

4.3.8 Models linked to socio-economic class

Variables relating to the population distribution over National Statistics Socio-economic Classifications (NS-SeC) - social class - are also included. Class often appears as a factor in property market behaviour, for example

- Murphy & Muellbauer (1993), states that "high relative house prices in the South East encouraged migration of skilled and professional workers from the South East". In this case, of certain economic conditions have had an impact only on certain classes of worker.

- Meen (2001) mentions the effect of labour mobility: "in the short run, prices differ because labour is immobile, but the differences are gradually eroded over time through migration". If labour mobility has an impact on the market, then it is reasonable to expect that the class of participants has an effect also as differing classes tend to have differing mobility.

- a clustering-based study of the property market in Paris (Guerois & Le Goix 2009) indicates that social class has a significant effect on prices, and that both of these variables are linked to other neighbourhood characteristics.

In short, economic factors affecting specific types of professions might generate separate submarkets for different social classes. The following variables are therefore included:

- num. people classed as Higher Managerial And Professional
- num. people classed as Lower Managerial And Professional
- num. people classed as Intermediate
- num. people classed as Small Employers
- num. people classed as Lower Supervisory And Technical
- num. people classed as Semi-Routine
- num. people classed as Routine
- num. people classed as Never Worked, Long Term Unemployed or Not Classified

These are found in census table UV031.

4.4 Preliminary results

The analyses presented in this chapter are not considered to be thorough. Results from two preliminary regressions are given, these are presented not as novel findings, but as a means of testing and validating the regression engine before it is used in anger in chapters 5 and 6.

The test regressions have target variables of (i) average 2001 house price, and (ii) average relative growth, respectively. For the purpose of this study, price growth is defined as the average price in the final fifth, divided by the average price in the first fifth of the time span under consideration, using the relative price indices developed in chapter 2. As almost no areas fell in price during the period of study, this amounts to measuring a (smoothed) gradient of growth over the entire time series.

4.4.1 Visualisation

As stated in section 4.2.1.3, scatterplots of principal components against the target are used in conjunction with regression to check for any details which have been missed. All scatterplots showed either a clear correlation (in a form that would be picked up by regression analysis) or no correlation; also none showed any evidence of heteroscedasticity. Figure 4.5 displays an example of a correlated principal component, and an uncorrelated principal component, for the 2001 price regression.

4.4.2 Regression diagnostics

4.4.2.1 Diagnostics for the 2001 house price regression

To test for heteroscedasticity, a scatter density plot of residual versus prediction is given in figure 4.6. No significant correlation is visible, thus this test does not indicate the presence of heteroscedasticity. Figure 4.7 shows a histogram of the residuals; these are seen to be approximately normally distributed, allowing for valid computation of confidence intervals on regression parameter estimates.

The Moran I statistic for spatial autocorrelation in the residuals is 0.050. This is significant (with the 99% confidence level being 0.003) though not large (the theoretical maximum is approximately 0.5). The residuals have been mapped in figure 4.8. If our aim were to create the most accurate model possible of house prices in the UK, we would of course deduce that there is likely to be a spatial variable missing from our model, and the residuals map would provide us with a hint as to what that variable might be. However, as the purpose of the regression is not to provide a complete explanatory model, but rather to

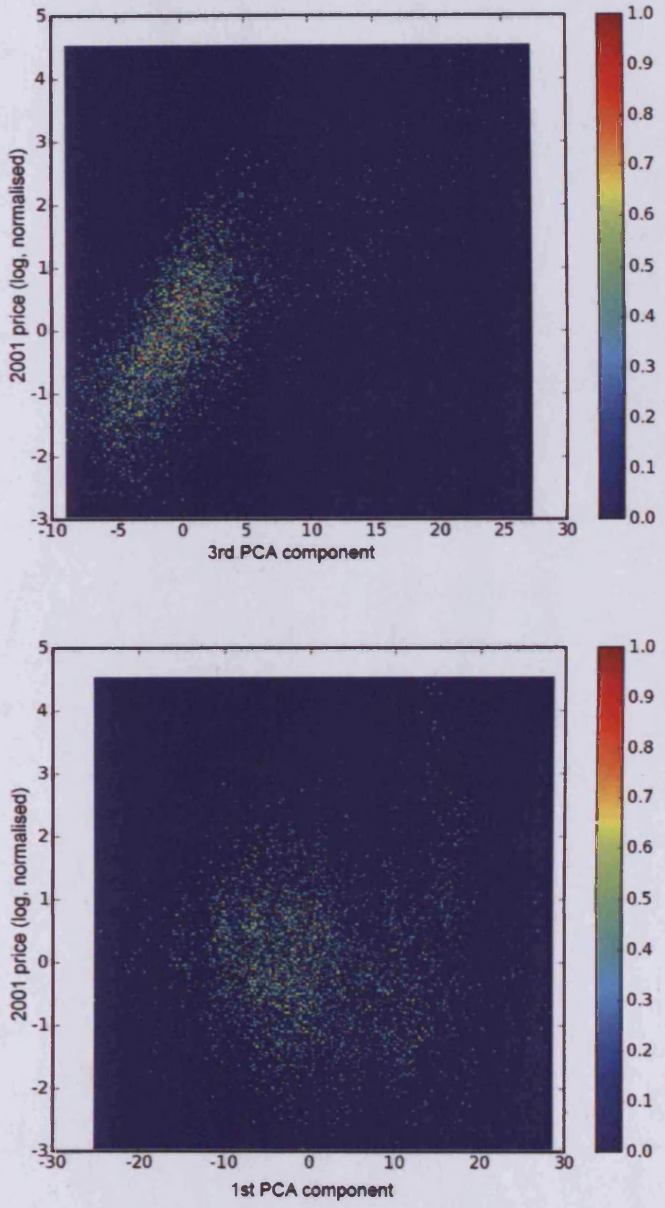


Figure 4.5: Plot of typical correlated and uncorrelated components vs the target variable. All such scatterplots were checked for interesting features. In this case, the first component correlates strongly to absolute numbers of dwellings/people, and the third with high income professions and shared dwellings.

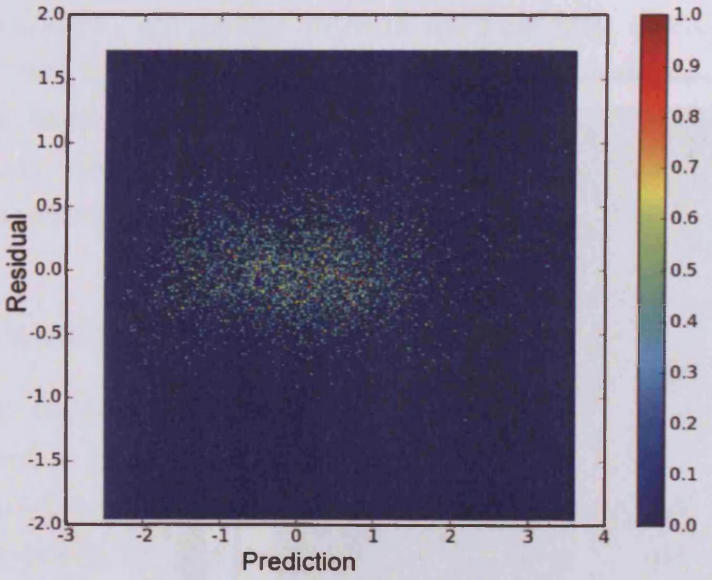


Figure 4.6: Plot of residuals vs predictions for 2001 house price regression

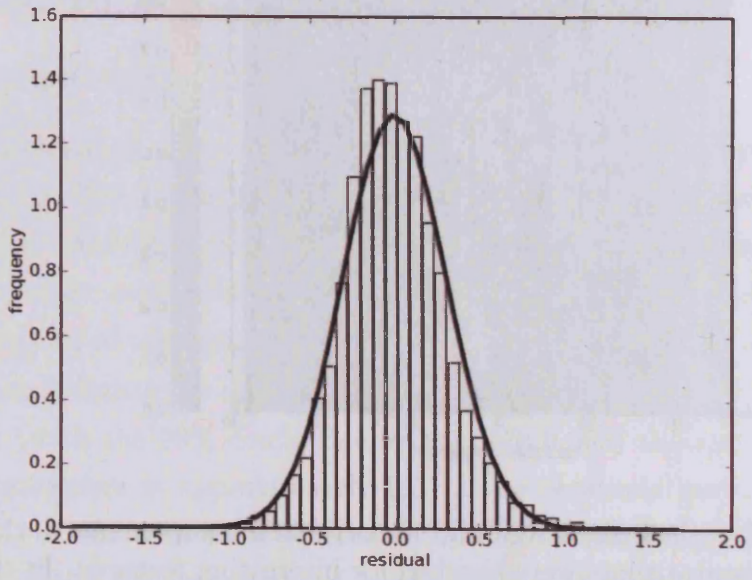


Figure 4.7: Histogram plot of residuals for 2001 house price regression; bold line shows normal distribution curve

Figure 4.8: Map of residuals for 2001 house price regression. The residuals have been divided into quintiles, coloured (in order from positive to negative): red, pink, white, light blue, dark blue. Spatial autocorrelation is strongly visible.

Regression	Mean square residual	Mean square error
2001 prices	0.095	0.107
Price growth	0.385	0.365

Table 4.3: Measures of goodness-of-fit for the test regressions. The data set is divided into a training and test set; the mean square residual is goodness-of-fit to the training set; mean square error is goodness-of-fit to the test set. The test set is employed to ensure that the data has not been over-fitted, as explained in section 4.2.3.3.

complete a preliminary exploration of the data, no further variables are added and the model is used as it stands.

4.4.2.2 Diagnostics for the house price growth regression

In the case of the price growth regression, a plot of residuals versus predictions shows no correlation, and a histogram plot of residuals shows approximately normal distribution. As these are similar in appearance to figures 4.6 and 4.7, neither are reproduced. The Moran I statistic is again significant (at 0.030, with the 99% confidence level being 0.002) but not large.

4.4.3 Tables of parameters

Having discussed the validity of the results, they can be presented. Tables 4.4 and 4.5 list the top 40 variables for each regression, in descending order of the magnitude of their regression coefficient. Where explanatory variables have been taken from census data, they have been named in a uniform manner to clarify data sources.

- all names for ward level data take the form *UVxxx-yzz-VariableName*, where *xxx* indicates the univariate census table number; *y* indicates whether this is a normalised statistic (*N*), expressed as a proportion of population, or absolute (*A*) expressed as a direct count; and *zz* is a table column index assigned for the purpose of this study.

- all names for Local Authority level data take the form *UVxxx-LA-yzz-VariableName*, where *xxx*, *y* and *zz* are defined as above.

Table 4.3 gives the mean square residuals and errors for the regressions.

Variable Name	Mean	Std. Dev	Reg. Coeff	99% Conf	Corr. Coeff
logweeklyincome	2.7	0.12	0.11	0.0061	0.79
UV031-N00-HigherManageria	0.036	0.019	0.11	0.0054	0.8
distance to London.log	5.1	0.43	-0.086	0.0064	-0.63
UV031-N02-Intermediate	0.038	0.01	-0.084	0.0091	0.15
UV031-N04-LowerSupervisor	0.031	0.0088	-0.068	0.0059	-0.65
UV031-N05-SemiRoutine	0.048	0.013	-0.065	0.0036	-0.73
UV056-N05-Terracedincludi	0.025	0.016	-0.06	0.0079	-0.48
UV056-LA-N07-Inapurposebu	0.011	0.0085	0.059	0.0051	0.38
UV031-A00-HigherManageria	2.4	0.37	0.057	0.0023	0.47
UV031-LA-N00-HigherManage	0.035	0.013	0.054	0.0041	0.71
UV031-N01-LowerManagerial	0.077	0.02	0.053	0.0036	0.73
ward numtransactions	7.4e+02	5.6e+02	-0.05	0.0064	-0.17
UV028-N01-Unemployed	0.013	0.0065	-0.05	0.0051	-0.52
UV004-N05-46to55	0.03	0.0052	0.049	0.0092	0.25
UV035-LA-N06-NoFixedPlace	0.019	0.0043	0.046	0.0064	0.46
UV031-LA-N02-Intermediate	0.038	0.0066	-0.045	0.0079	0.28
UV053-LA-N04-Vacanthouseh	0.0067	0.0022	-0.045	0.0086	-0.42
UV056-LA-N04-Semidetatche	0.033	0.008	-0.041	0.0088	-0.42
UV056-LA-N06-Flatmaisonet	0.016	0.012	0.039	0.0031	0.4
UV056-LA-N02-Houseorbunga	0.083	0.011	-0.039	0.0033	-0.41
UV056-LA-A07-Inapurposebu	3.7	0.41	0.039	0.0032	0.15
UV031-N06-Routine	0.038	0.016	-0.039	0.0039	-0.76
UV004-N03-26to35	0.029	0.008	-0.038	0.0063	-0.022
UV056-N04-Semidetatched	0.033	0.014	0.038	0.0091	-0.28
UV031-A04-LowerSupervisor	2.4	0.32	-0.035	0.0026	-0.36
UV031-LA-N07-NeverWorkedL	0.11	0.016	-0.035	0.0041	-0.44
UV035-N06-NoFixedPlaceOfW	0.02	0.0061	0.034	0.0083	0.38
UV028-LA-N00-Employed.log	0.21	0.015	0.033	0.0038	0.49
UV053-A03-Secondresidence	0.87	0.53	0.033	0.0057	0.29
UV028-LA-N02-Other.log	0.13	0.015	-0.033	0.0041	-0.48
UV031-A06-Routine.log	2.4	0.37	-0.032	0.0019	-0.48
UV031-A05-SemiRoutine.log	2.6	0.33	-0.031	0.0017	-0.36
UV031-LA-N01-LowerManager	0.074	0.012	0.029	0.0031	0.72
UV056-A08-Partofaconverte	1.4	0.66	0.029	0.0048	0.18
UV056-LA-A06-Flatmaisonet	3.8	0.4	0.029	0.0026	0.17
UV004-LA-N08-Over79.log	0.0096	0.0023	0.028	0.006	0.041
UV056-A05-Terracedincludi	2.6	0.5	-0.027	0.0047	-0.33
UV004-LA-N03-26to35.log	0.029	0.0048	0.026	0.0049	0.25
UV056-LA-A04-Semidetatche	4.2	0.27	-0.026	0.0041	-0.34
UV056-LA-A03-Detatched.lo	4.1	0.27	-0.026	0.0048	-0.11

Table 4.4: Top 40 determinants of 2001 house prices. Means and standard deviations of explanatory variables are given to add context to their normalised coefficients, and data for reconstruction.

Variable Name	Mean	Std. Dev	Reg. Coeff	99% Conf	Corr. Coeff
distance to London.log	5.1	0.43	0.072	0.011	0.63
log 2001 prices	5	0.3	-0.072	0.016	-0.63
UV056-LA-N07-Inapurposebu	0.011	0.0087	-0.058	0.01	-0.33
logweeklyincome	2.7	0.14	-0.057	0.017	-0.51
UV031-N00-HigherManageria	0.036	0.019	-0.051	0.0097	-0.62
UV028-LA-N00-Employed.log	0.21	0.015	-0.05	0.0076	-0.51
UV028-LA-N02-Other.log	0.13	0.015	0.049	0.0081	0.52
UV031-LA-N00-HigherManage	0.035	0.013	-0.048	0.0083	-0.66
UV031-LA-N07-NeverWorkedL	0.11	0.016	0.048	0.0083	0.47
UV053-LA-N04-Vacanthouseh	0.0067	0.0022	0.048	0.018	0.45
UV056-LA-N04-Semidetatche	0.033	0.0081	0.041	0.018	0.32
UV031-LA-N01-LowerManager	0.074	0.012	-0.041	0.0063	-0.67
UV056-LA-A07-Inapurposebu	3.7	0.41	-0.039	0.0064	-0.17
UV004-LA-N03-26to35.log	0.029	0.0049	-0.038	0.0092	-0.29
UV004-LA-N05-46to55.log	0.029	0.0026	0.038	0.01	0.046
UV056-LA-N06-Flatmaisonet	0.016	0.012	-0.037	0.0062	-0.33
UV056-LA-N02-Houseorbunga	0.083	0.011	0.035	0.0066	0.33
UV056-N05-Terracedincludi	0.025	0.016	0.035	0.016	0.3
UV028-LA-N01-Unemployed.l	0.013	0.0046	0.035	0.0086	0.39
UV056-A10-Caravanorotherm	0.59	0.57	0.033	0.028	-0.092
UV028-N01-Unemployed	0.013	0.0065	0.033	0.0095	0.4
UV056-LA-A06-Flatmaisonet	3.8	0.4	-0.031	0.0054	-0.16
UV004-LA-N06-56to65.log	0.023	0.003	0.03	0.0084	0.22
UV053-LA-A04-Vacanthouseh	3.2	0.27	0.029	0.0089	0.29
UV031-A00-HigherManageria	2.4	0.38	-0.029	0.004	-0.43
UV035-LA-N06-NoFixedPlace	0.019	0.0043	-0.027	0.011	-0.37
UV002-LA-A00-PopulationDe	2.5	0.35	-0.027	0.0094	-0.055
UV056-N00-ALLHOUSEHOLDS	0.1	0.00058	-0.026	0.015	-0.13
UV056-LA-N01-Inanunshared	0.097	0.00019	-0.025	0.011	0.077
UV056-N08-Partofaconverte	0.0032	0.0061	0.025	0.019	-0.18
UV031-LA-A00-HigherManage	3.8	0.29	-0.025	0.0046	-0.36
UV053-A03-Secondresidence	0.87	0.53	-0.024	0.012	-0.11
UV056-LA-A04-Semidetatche	4.2	0.28	0.024	0.0081	0.23
UV031-N05-SemiRoutine	0.048	0.013	0.024	0.0075	0.49
UV053-LA-A03-Secondreside	2.4	0.48	-0.024	0.0083	-0.012
UV004-N04-36to45	0.031	0.0039	0.024	0.012	-0.23
UV031-LA-N02-Intermediate	0.038	0.0066	-0.023	0.014	-0.39
UV031-N02-Intermediate	0.038	0.01	0.022	0.018	-0.28
ward numtransactions	7.4e+02	5.6e+02	0.022	0.012	0.069
UV031-LA-N06-Routine.log	0.038	0.0098	0.022	0.0081	0.58

Table 4.5: Top 40 determinants of house price growth 2000-2006. Means and standard deviations of explanatory variables are given to add context to their normalised coefficients, and data for reconstruction.

4.5 Discussion

This section is divided into three. Subsection 4.5.1 summarises results. Subsection 4.5.2 evaluates these with respect to existing models of market behaviour, and subsection 4.5.4 discusses the possibility of there being important variables missing from the model.

4.5.1 Summary of results

The primary advantage of the regression models presented here is their ability to identify the most important explanatory variables from a very large set of candidates. Such variables can be found at the top of tables 4.4 and 4.5 as these are ordered in decreasing order of coefficient magnitude. However, it is also possible to summarise these tables more succinctly by listing the order in which different *categories* of variables first appear. In the case of 2001 house prices, these are

1. income
2. social class
3. distance to London
4. housing type
5. market turnover
6. employment level
7. age distribution[2]
8. housing stock

For subsequent (2000-2006) price growth the order of importance for variable categories is

1. distance to London
2. 2001 price
3. housing type
4. income
5. social class
6. employment level

[2]Though only one age variable appears at this level, therefore, this result may not be stable in the face of adding more explanatory variables.

7. housing stock
8. age distribution

As these summaries cannot encapsulate the full complexity of the regression output, the author recommends that they are only used as a rough guide to the real interactions between variables. They are still, however, considered useful for basic comprehension of the data set.

4.5.2 Evaluation of 2001 price regression with respect to housing market models

In this section, the results will be discussed with reference to the models which inspired choice of variables, in the same order as presented in section 4.3. It can be seen that, in the majority of cases, coefficients behave as expected. Most of the coefficients derived are significant at the 99% level, however to shorten the discussion, only those greater than the average coefficient magnitude (0.01) will be considered. Coefficients below this level, while statistically significant, can be considered of diminished importance compared to other factors in the regression.

4.5.2.1 Lifecycle models

The premise of life cycle models is that there is a tendency for people of different ages tend to live in different areas, and also for people of different ages to have differing financial means. One would therefore expect areas populated by more affluent age groups to exhibit higher house prices. The patterns exhibited by the regression coefficients are discussed below.

- Negative effects on price for the 56-65 and 66-79 age band

This might be explained as indicating the presence of couples whose children have left home and are fully financially independent, and retirees, many of whom choose to 'downscale' at this stage in life, freeing some property capital to cover living expenses.

- Positive effects on price for the 80+ age band

This, by definition, will be positively correlated with the quantity of residents with greater than average longevity - this is in turn known to be correlated

with better financial means so it is not surprising that housing in areas with a significant population in the oldest age band is more expensive.

- Mixed effects on price (positive at Ward level, negative at LA level) for the 16-25 age band
- Mixed effects on price (positive at Ward level, negative at LA level) for the 46-55 age band
- Mixed effects on price (negative at Ward level, positive at LA level) for the 26-35 age band.

These are harder to explain, at least without increasing the level of wild speculation! It is not uncommon, however, for mixed effects to occur in a multi-level model. One explanation for the first of these results may be that the 16-25 age band are not usually high earners, so may have a negative effect on prices at LA level; however at Ward level the presence of many 16-25 year-olds might be indicative of a strong rental market, thus locally increasing property prices. The latter result, conversely, may relate to first time buyers of houses who will typically chose to locate in cheaper areas; however their greater affluence overall contributes positively to average prices in the Local Authority.

These results are not presented as definitive, so much as a being plausible validation of the regression model.

4.5.2.2 Bid rent theory

Bid rent theory predicts that, as labourers make a trade-off between land rent and commuting costs when choosing a location in which to live, locations where the average travel-to-work distance is small will command greater land rents. At Ward level, this is reflected in some of the regression parameters, with a positive coefficient of 0.017 for people travelling less than 2km to work. However, mixed effects are also noted for most travel to work distances, particularly the $> 30km$ band, which has an unexpected positive effect at ward level as well as the expected negative effect at LA level. The reason for this may be understood by studying the 'working outside the UK' category, which has a universally positive effect on house prices. It seems likely that on an international level, bid rent theories no longer apply, while the presence of a population which works abroad may be indicative of higher earners, whose commuting costs - if these are actually paid for

by the worker rather than employer - are insignificant compared to their wages. Similar reasoning could probably be applied to the case of those travelling over 30km to work.

4.5.2.3 The Ripple Effect

In a period of overall market growth, the ripple effect predicts that the log distance to London will have a large negative impact on house price. This prediction is borne out in the regression coefficient, which has a value of -0.09.

4.5.2.4 Supply and demand

The theory of supply and demand applied to the mobile labour market suggests that high incomes and employment levels will have a positive impact on house prices. The theory of supply and demand applied to housing stock suggests that high levels of housing availability will have a negative impact on prices.

As expected, large positive coefficients (> 0.01) are noted for the average income and employment level. Large negative coefficients are noted for unemployment and vacant household spaces. Additionally, the quantity of second residence/holiday accommodation has a large positive effect, presumably because of the increased demand for housing in areas considered pleasant for recreation.

4.5.2.5 Submarket-based models

Submarket models suggest that different types of accommodation will be sought by different categories of people, so there is no reason why the markets for these different types should be completely correlated. In the case of absolute house prices, a more direct effect is noticeable - that different categories of accommodation tend to have different absolute hedonic values. Thus at Ward level, bungalows, terraces and commercial buildings have a negative effect on prices, while semi- or detached housing and flats have a positive impact on prices. Bizarrely, shared housing and caravans/temporary structures also have a positive effect on prices. The former may be indicative of a strong rental market while the latter may be indicative of rural areas. Population density was seen to have a negative effect on prices at a local level, perhaps because high density is indicative of poor local housing quality.

The relation of Local Authority level housing type variables to house price is unclear.

4.5.2.6 Models of market turnover

The number of transactions between the years 2000 and 2006 was seen to be negatively correlated with 2001 house prices. Using this as an explanatory variable is of course reversing causality somewhat! However, the result is easily explained. The general trend during these years was for the cheaper areas of the market to grow more - either because they were affordable, or because they were behind London and the South East (the prices of which had already risen) in a ripple pattern of growth.[3] As growth, in turn, is correlated with high numbers of transactions[4] we therefore would expect high numbers of transactions to correlate with low 2001 house prices.

4.5.2.7 Effect of social class

As predicted in section 4.3.8, the socio-economic classification of residents has a direct impact on prices, with strong positive effects noted for managerial and professional classes, and negative effects for intermediate, lower supervisory, technical, semi-routine and routine classes. Mixed effects were noted for small employers.

4.5.3 Evaluation of 2000-2006 price growth regression with respect to housing market models

The regression on price growth between the years 2000-2006 requires far less discussion than the regression on 2001 house prices. This is because by far the largest coefficients estimated are for the distance to London (which had a positive effect on price growth) and the 2001 house price (which had a negative effect on price growth). Both of these are reconcilable with the ripple effect model: namely that in the tail end of a rising phase of the market, properties in London had already increased in price before commencement of the study, while properties

[3]The price growth regression gives a coefficient of -0.07 relating 2001 prices to subsequent growth.

[4]This is backed up by a coefficient of 0.02 in the price growth regression.

elsewhere had yet to exhibit a similar gain in value. Thus during this period, low priced houses far from London gained more in value than any elsewhere.

4.5.4 Missing Variables

The presence of spatial autocorrelation in the residuals points to the conclusion that at least one important spatial variable has been missed from the model. Examination of the residuals map hints that this variable may be related to the urban/rural divide. While population density has been included as a variable, perhaps this is not sufficient, and failing to include more detailed information on this division has perhaps resulted in the regression process attempting to compensate in other ways. For example, in the case of population travelling 2-5km to work (measured at ward level), a significant positive coefficient has been assigned to the absolute number of people, while a significant negative coefficient has been assigned to the corresponding relative proportion of the total population. Also in many cases, variables are shown to have a positive effect at one spatial scale and a negative effect at another.

It is almost certain that simple regression is not the optimal technique for making accurate predictions of the market, and a model which allows for coefficient heterogeneity between urban and rural areas may perform better. However, it is not the purpose of this study to produce the best possible model, so much as to conduct a preliminary exploration of the data, for which the model presented - in light of the low mean square error - is considered to be adequate.

4.6 Conclusions

In this chapter, one of the simplest possible approaches for understanding a large data set has been implemented. Data mining has been rejected in favour of a hybrid visualisation-regression technique. Regression assumptions are checked, and PCA is used to reduce the dimensionality of the data, thus allowing for input of a large number of explanatory variables. Custom software is necessary in anticipation of the size of the data sets to be analysed in the upcoming chapters. Based on existing market models, a set of variables has been selected for study.

Overall, despite evidence of missing variables, the regression framework is found to be effective. For the 2001 house price model, the normalised mean

square residual - the error in the average prediction made by the model - is 0.095, which is low; by contrast, using a global mean price to model housing market variations would result in a mean square residual of 1.0. Likewise for the price growth model the mean square residual is 0.385. It is clear that in each case the regression analysis, while not completely modelling the data, has accounted for a large proportion - in fact, the majority - of variation in the target variable.

The model has allowed us to identify the most important out of a range of candidate variables. Additionally, evidence for many of the existing models of housing market behaviour has been found in the output from the model. This would suggest that the regression framework is a valid approach.

Having used this framework to identify some of the more relevant variables, it may be fruitful in future to return to data mining techniques to carry out a more detailed analysis of the way in which these variables interact. In particular, such techniques may help to untangle the causes of the mixed effects on house price noted for certain classes of variable, such as population age, social class and travel-to-work distances. However in sum, the framework developed here is considered sufficient for a preliminary exploration of the data, and will be extended and applied in anger to the much larger data sets encountered in chapter 5.

Chapter 5

Development of a general, exploratory, interactive and reactive housing market model

5.1 Introduction

This chapter presents perhaps the primary novel piece of work in the thesis: a combined analysis of three data sets at an unprecedented level of detail.

Returning to the mountaineering analogy of the introduction, this is the crux of the mountain of data, where the most complex manoeuvres must be conducted, using both innovative statistical technique and raw computing power. In chapter 2 the mountain's easy lower slopes were climbed through aggregation of data. In chapters 3 and 4 two subsidiary peaks of visualisation and regression were climbed - not so important in and of themselves, so much as for the skills thereby developed which will be applied to the main summit. But now it is time to address the main challenge of the ascent: the joint exploratory analysis of two large spatial time series and network interaction data sets, those of the Land Registry and Census Office.

The route taken in this analysis is chosen so as to fulfil three criteria:

- both the Land Registry and Census data must be used, on a fine scale - to take maximum advantage of the available data,
- the Census Interaction data must be used - because detailed comparison of

this with Land Registry data is unprecedented, and

- the Census Interaction data should not be used in isolation, as that would not be indicative of its overall importance in explaining house price interactions, which must somehow be compared to the importance of the Census Area Statistics.

It is hoped that any significant findings from this exploration will be novel. Because detailed study of the interaction data in this manner is unprecedented, a guiding principle of this study is that *if there are interactive effects in the housing market, however small or localised, it is the aim of the analysis to find them.* This principle is of relevance to many of the decisions taken on methodology.

The remainder of this chapter is structured as follows. Section 5.2 reviews previous work in this area. Section 5.3 discusses the choice of an appropriate methodology for analysis of the interaction domain. Section 5.4 applies this to an exploratory regression analysis of the data, however visualisation remains an important component of the study and is therefore employed in section 5.5. Section 5.6 concludes.

5.2 Literature review

5.2.1 Spatial housing market models

A suitably broad starting point for models of spatial processes is that of Cliff & Ord (1981):

> "When we develop a model for a spatial process, we must always ask whether the levels of the process at two (neighbouring) sites reflect interaction (between the sites) or reaction to some other variable. The case is rarely open and shut."

Cliff and Ord go on to define interaction through spatial auto-correlation models, and reaction through regression models, culminating in the presentation of combined models which are capable of reflecting both kinds of process, with parameters fitted by regression.

In the case of housing markets, such reactive and interactive models are both present. Meen (2001) provides a good overview. In general, regional (and hence spatial) housing market models fall into one of three categories:

1. Reactive models. Probably the most common type of model, these deal with the response of regions to changes in macroeconomic variables such as housing stock or average income (as were discussed in chapter 4), or to the response of prices to local variables such as neighbourhood quality, as is the case with hedonic modelling (discussed in chapter 2).

2. Interactive models. These deal with the correlation between markets in different regions, and the propagation of price changes between them. Such models can either be parametrised, e.g. with distance decay or spatial autocorrelation terms, or alternatively - if there is enough data - the data itself can be used to directly estimate the values in the spatial interaction matrix. This second type of model is rare, but the work presented in this chapter belongs in this category, therefore such models are therefore discussed in more detail in section 5.2.2.

3. Mixed models. These combine both reactive and interactive models, and are far more common than pure interactive models. However, interaction between regions is typically incorporated in a somewhat constrained manner, using the first (parametrised) form of interaction model noted in (2). An example is the spatial cross-section model of (Anselin 1988, in Meen, 2001):

$$y = \rho W y + \beta x + \epsilon \tag{5.1}$$

where y is the vector of regional prices, W is a spatial weights matrix, x is a vector of regression variables, ρ and β are coefficients to be determined and ϵ is the error term. However, the model's weakness is that W must be specified directly - based for example on spatial contiguity or distance decay - rather than deduced from the data. While this has the advantage of providing strong explanatory power if the model fits the data well, it is likewise limited to explaining interactions only of the form explicitly specified.

As directly estimated interactive models are of greatest relevance to understanding interaction data, they are discussed in more detail in section 5.2.2.

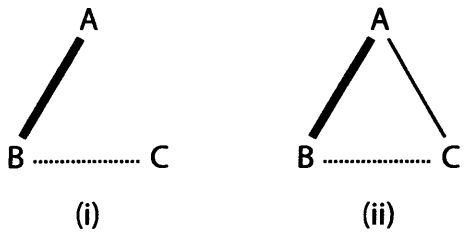


Figure 5.1: Illustration of transitive cross-correlation for three time series A, B and C. If A, B and C are linked by weak and strong links as shown in (i), then VAR analysis for series A will deduce the correct structure. However, Granger or Cross-correlation analysis will deduce the structure in (ii) whereby a link is deduced between A and C because of the transitive correlation between A-B-C. While series A and C are indeed correlated, this misses the point that such a correlation is better explained by the mutual link to B than by a direct link.

5.2.2 Spatial housing market models with directly estimated interaction matrices

5.2.2.1 Estimation techniques

A number of techniques have been used to directly estimate the form of interaction matrices, when sufficient data is available.

- Seemingly Unrelated Regression (SUR) can be used to separately estimate reactive regression models for each region, but look for correlations in the residuals across regions (Meen 2001, page 166). This essentially treats the interaction matrix as a form of error analysis to be studied after conducting a regression, although it does allow for the inclusion of spatial coefficient heterogeneity.

- Cross-correlation, as used in Giussani & Hadjimatheou (1991) and Shi et al. (2009), looks for simple correlation between the current growth of a region and past growth values of all other regions.

- Granger causality tests, as used in Giussani & Hadjimatheou (1991), Shi et al. (2009), Worthington & Higgs (2003) and MacDonald & Taylor (1993) are a more sophisticated form of correlation analysis. A regression is performed relating current growth in a region to past growth in the same

region and each other region separately. This has two advantages, firstly that by modelling autocorrelation is it is possible to find out the relative importance of auto- and cross-correlation; and secondly that it is easier to estimate statistical significance of the resulting coefficients.

- Vector Auto-regression (VAR), as used in Pollakowski & Ray (1997), Shi et al. (2009) and Worthington & Higgs (2003) is the multivariate extension of Granger causality tests. For each time series, a multivariate regression is conducted relating that series to both its own past values, and the past values of all other time series. It is thus possible to separate the cross-correlations which have greater explanatory power, from those which appear simply because of transitive cross-correlation. Figure 5.1 illustrates the difference between Granger tests and VAR.

- Johansen tests, as used in MacDonald & Taylor (1993) and Shi et al. (2009) are the most sophisticated, as they can search for all combinations of co-integrating time series simultaneously. Time series are said to be co-integrated if a linear combination of them can be made which is stationary. For example, for two near-identical time series A and B, the linear combination $A - B$ would always be near zero so the time series could be said to be co-integrated. Results from Johansen tests are therefore presented not as a matrix but as a set of sets of time series.

5.2.2.2 Explanatory power

While all of the above studies uncover some form of interaction structure between regions, little attempt is made to extend the model to explain why that interaction exists. Instead, the results are interpreted on a qualitative basis. Pollakowski and May's US study, for example, notes that

> "although many [regional] cross lags are significant, neither a spatial pattern nor any other discernible pattern is evident"

and that

> "From inspection [of New York districts], it does appear that there

> may be some sort of spatial diffusion pattern ... [though] a statistical inference clearly cannot be reliably made."

Shi et al. (2009) notes that in New Zealand, correlation is measurable within but not between regions, and thence argues that the apparent correlation is a reactive phenomenon - caused by differing response to external economic factors rather than by spatial proximity.

In the case of the British literature, most of the studies are targeted at answering a question often posed in the wider literature, namely: "Is there a ripple effect between London and the other regions?" (Holmes & Grimes 2005). Giussani, having concluded that the effect exists, mentions a variety of possible reasons, including the possibility that regions react differently to housing stock, income etc (essentially reactive models as described in section 5.2.1). Correlations are also noted between acceleration in inter-region migration and price change, albeit aggregated to a very course level. Worthington & Higgs (2003) concludes that the South East drives the rest of the market, but offers little explanation for why this may be the case; meanwhile Macdonald and Taylor openly admit that their deduced interaction structure remains unexplained. Meen meanwhile, puts the UK ripple effect down to spatial coefficient heterogeneity, with respect to responsiveness to incomes and interest rates, concluding that

> "The spatial pattern that we observe has little to do with spatial movements *between* regions through migration, for example, but relates to adjustments *within* regions" (emphasis in original).

However, Meen also speculates that in other cases, differing economic conditions, information dissemination through property searches and equity transfer (the latter two being interaction-related) could also cause inter-regional price differences. This may be supported by Cameron et al. (2005) which, although not using an interaction model, aggregates in- and out-migration on a regional level and finds "strong housing market effects consistent between inflow and outflow equations".

Cameron also cautions, however, against "drawing policy inferences from one set of relationships in a complicated web". Indeed, overall it would seem that the order of cause and effect, when applied to the interacting systems of migration, house prices, economic conditions and housing market structure, is almost

never clear. What is therefore sought in the study is correlation, not causation - however, regression analysis does make it possible to decide which correlations explain the structure of the data better than others.

Section 5.3 discusses a technique for exploring the HMLR and Census data in a manner which can incorporate much of the above. In particular, not only are interactions directly estimated, but in section 5.4, a quantitative attempt is made to explain them, in terms of both interaction and reaction.

5.3 Interaction domain methodology: cross-correlation and other comparisons

5.3.1 Choice of analysis domain

To recap the criteria for choosing an analysis technique, it is desired to compare the currently underutilised Census interaction data to the Land Registry (HMLR) data, but also including Census Area Statistics (CAS), so the relative explanatory capabilities of each data set can be directly assessed. This is largely to be conducted at Census ward level (8850 areal units) although data aggregated to local/unitary authorities (of which there are 376 in England and Wales) will also be employed.

Referring back to the thesis map of figure 1.2, the Land Registry data is now in the form of per-area house price time series, the area statistics are a two-dimensional geographical data set, and the interaction data is a four-dimensional set (linking 2-d origins to 2-d destinations). As such, to compare these is to compare 'apples and oranges' - they must first be converted to a common format before they can be analysed. Once this has been done, the data sets can be regressed against one another in the manner described in chapter 4.

To make best use of the interaction data, therefore, the HMLR and CAS data sets are converted into an interaction format. All subsequent analysis is conducted purely in the interaction domain. The remainder of this section is structured as follows. Section 5.3.1.1 explains the conversion of HMLR data to an interaction format, while section 5.3.1.3 deals with the conversion of CAS to an interaction format. Section 5.3.1.4 further discusses, and presents justification for, the choice of analysis domain.

5.3.1.1 Converting time series data to a network interaction format using cross-correlation

The literature presented in section 5.2.2 presents several techniques for converting multivariate time series data into an interaction matrix. These are discussed in turn.

- Seemingly Unrelated Regression can be used to perform regressions in a reactive or mixed model for each time series separately, then study the error terms for possible interactions. This would be a fairly complex method to use here as it would divide the explanatory (regression) phase into two - it would first be necessary to regress with a reactive model, then regress again for interactive components. This complexity seems unwarranted for an initial analysis.

- Johansen and Vector Auto-regression (VAR) techniques are difficult to use here due to the ratio of time series slices (i.e., 21 slices of 100 days over 6 years) to locations (i.e., 8850 census wards). The problem is clearest when considering VAR: for each census ward, we must estimate at least 8850 regression parameters to determine the influence of each other ward, on the ward under consideration. To do so from only 21 data points in the time series would lead to confidence intervals on the parameters which are wide enough to be useless. As many wards exhibit similar behaviour, the regression would be prone to serious multicollinearity. This problem may be solvable with principal component analysis as in chapter 4. However to do so would add considerable complexity to the analysis.

  It is also the case that 8850 regions are a lot to use in a VAR framework - Pollakowski & Ray (1997) only uses 9 census regions in one test, and 5 districts of New York in another. Likewise (Meen 2001, page 174) notes that even "nine census divisions are a lot to use in a single Johansen co-integrating system".

- Granger causality tests. As Granger tests only consider pairwise relationships between time series (and not multivariate relationships), they would be computationally feasible.

- Cross-correlation is computationally simpler than all the above techniques. It is therefore chosen for conversion of the HMLR data to an interaction format.

For the purposes of this study, cross-correlation between two time series A and B is defined as a normalised cross-correlation function,

$$(A \star B)_x = \sum_{t=0}^{T} \frac{(A_t - \bar{A})(B_{t+x} - \bar{B})}{\sigma_A \sigma_B} \tag{5.2}$$

where $(A \star B)_x$ is the value of the correlation function for time offset x, A_t and B_t are the values of time series A and B respectively at time t, $\bar{A}$, $\bar{B}$, σ_A and σ_B are the mean and standard deviation of each time series, and T is the length of the original time series.

However, this time series is only studied for $x > 0$, that is to say, for the region in which the present value of B is correlated with past values of A, and not *vice versa*. Thus it is a measure of the extent to which 'A causes B' - though note that this is Granger causality, not true causality as both series could be responding to the same external cause.

Section 5.3.2.1 will discuss the typical properties of this series, with respect to its highest peak, the time offset of this peak and the sum of the series. The sum of the series is defined as

$$\text{corr\_sum}_A^B = \sum_{x=wmin}^{wmax} (A \star B)_x \tag{5.3}$$

where *wmin* and *wmax* are the minimum and maximum time offsets defining the window over which a search is conducted for correlations. Thus, correlation sums are not symmetric:

$$\text{corr\_sum}_A^B \neq \text{corr\_sum}_B^A \tag{5.4}$$

except in the trivial case where $wmin = wmax = 0$. Note that in addition to these windowing parameters, the time resolution of the original series has a major effect on the computed cross-correlations. Thus while it is possible to interpret window parameters as the time scales over which correlations are sought, e.g.:

$$\text{max\_corr\_offset\_searched\_(days)} = \text{time\_slice\_length} \times wmax \tag{5.5}$$

it is not necessarily the case that the correlations for parameters that result in the same correlation window (measured in days) will be identical. Therefore the parameters used to define a cross-correlation metric are:

- wmin (measured in time slices)
- wmax (measured in time slices)
- time slice length (measured in days).

It should finally be noted that before cross-correlations are computed, all time series are first differentiated, therefore all cross-correlations discussed hereafter relate to price *movements* rather than absolute prices.

Computation of cross-correlations for 8850 wards can be achieved in a few hours on modest computing hardware (a single 3GHz Itanium core). There is probably further scope for code optimisations to speed up such computation.

5.3.1.2 Converting time series data to a network interaction format using simple division

The format conversion described in section 5.3.1.1, then, allows the discussion of data trends in terms of linked housing markets. For example, by comparing HMLR with interaction data, the model could answer the question "does a flow of migrants between areas A and B cause the housing market at A to affect the market at B?". However, by using this technique, the opportunity to answer a simpler class of interaction-related questions is missed, such as "do people generally move upmarket?" Therefore, a second technique is also used to convert HMLR data to an interaction format:

$$\text{log price ratio}_A^B = \begin{cases} log\left(\frac{\text{average 2001 sale price}_A}{\text{average 2001 sale price}_B}\right) & \text{if both prices are defined} \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

This is a simple log ratio and will therefore be positive if the prices at A are higher than at B, and negative if the prices at B are higher than at A. As A is the origin and B the destination, positive values of this measure indicate a move downmarket, while negative values indicate a move upmarket.

An additional advantage of studying this simpler interaction structure is that, price ratios being better understood than market correlations, it provides a means of testing and validating the regression engine on more familiar data.

5.3.1.3 Converting Census Area Statistics to a network interaction format

Converting Land Registry data to an interaction format allows for direct comparison of housing market interactions with census interaction data. However, it is well known that many other factors influence the housing market, and it is wise to assume that the influence of these other factors will be greater than those of the interactions captured by the census.

Therefore, the Census aggregate statistics are also converted into an interaction format, allowing analysis of housing market interactions with respect to all the categories of variables discussed in chapter 4:- population age distribution, travel to work distances, distance to London, average 2001 house price, housing stock, incomes, employment, housing type and social class.

This conversion is achieved by including three interaction variables for each area statistic variable X. As all interactions are defined between pairs of places A and B, the following are included for the interaction between A and B:

1. FROM-X, the value of X at origin, X_A
2. TO-X, value of X at destination, X_B
3. CHANGESQ-X, square difference in value of X between origin and destination, $(X_B - X_A)^2$.

Thus, it is possible in principle to deduce results such as, "there is a tendency for areas with a similar proportion of skilled professions in the resident population, to have correlated housing markets." It is also possible to directly estimate the relative importance of these effects over the importance of interaction effects such as migration.

5.3.1.4 Discussion

The methods presented above merit further discussion. The justification presented for their choice is that they fulfil the aims of the study, by combining

the three incompatible data sets to produce a unified analysis. As such, they are different to existing techniques which either (a) parameterise spatial interaction or (b) directly estimate the interaction matrix. Instead, they take these techniques further by first directly estimating the matrix, and then analysing it - which may hint at methods of parameterising the interaction in future. Another advantage of their use, is that they are also sufficiently general that they may find application in other fields of study. As discussed in chapter 4, general techniques are preferred for an exploratory data analysis as they do not, so far as possible, confine the study to a search for evidence of a specific model.

It should be noted that there is considerable scope for criticising the techniques on grounds of information theory: they appear to greatly expand the information originally entered into the system. Taking a 21-point time series for 8850 wards (a total of 185,850 data points) and using it to compute 8850 $\times$ 8850 pairwise cross-correlations (a total of 78,322,500 data points) is expanding the data set by a factor of 421. Meanwhile, expanding an aggregate statistic defined for 8850 wards to three sets of interaction statistics as described in section 5.3.1.3 expands the data by a factor of 26,550.

The justification for such expansion is that it could be seen as a process analogous to data decompression: converting a small, high-entropy file into a larger, low-entropy file, which may take more space but crucially, *is easier to understand.* It is noted, however, that as the entropy of the data has decreased, there is likely to be some kind of spatial autocorrelation present in the four-dimensional interaction data space. This correlation may or may not take the form of correlation with neighbouring points (however we chose to define those), but it is safe to assume that it could violate the regression assumptions of section 4.2.2. Fortunately, the penalty for this is not severe: if spatial autocorrelation is present, then the regression does not become invalid, it is simply the case that perhaps there is a better model which can be used instead. Therefore in addition to the regression analysis, section 5.5 presents a visualisation methodology which is employed to check for the possibility of using alternative models and take appropriate action.

5.3.2 Validation of cross-correlation data

5.3.2.1 Explanatory power of cross-correlation with respect to reconstructing time series

To validate the interpretation of cross-correlation metrics as a meaningful statistic prior to their use in a regression analysis of the housing market, an attempt is made to use them to reconstruct the market growth patterns from 2000-2006, i.e. to reproduce the first derivative of the original 8850 ward timeseries (as was shown in figure 3.18 back in chapter 3). To prevent exponential divergence of the reconstructed time series from the original, this is performed using series which are normalised over all wards at each time step. Thus, overall market growth and volatility (defined here as the variance in growth over all wards for a given time slice) is considered exogenous to the system, because a pure market cross-correlation model cannot be expected to predict them.

The algorithm for reconstruction is as follows:

```
1. compute cross-correlation sums (corr_sum) from source data
     as specified in equation 5.3
     using parameters wmin=1, wmax=2, time slice length = 100 days
2. initialise prediction matrix to store predictions
     number of rows = number of wards
     number of columns = time series length
3. copy initial time series values from source data to first column
     of prediction matrix
4. for each time step t
     4.a. for each ward w
          4.a.1 set prediction matrix [w,t] as
```

$$\text{prediction}_t^w = \sum_{w' \neq w} \text{corr\_sum}_w^{w'} \times \text{prediction}_{t-1}^{w'} \tag{5.7}$$

```
     4.b. normalise predictions over all wards such that mean is 0
     4.c. normalise predictions over all wards such that variance is 1
5. compare prediction matrix with source data
```

(In equation 5.7, prediction_t^w is the predicted growth for ward w at time t, $\text{corr\_sum}_w^{w'}$ is the sum of the cross-correlation function between wards w and w'

and $\text{prediction}_{t-1}^{w'}$ is the prediction for the previous time step. Note that the autocorrelation term $\text{xcorr}_{ward}^{ward}$ is excluded from the summation. Thus intuitively this equation is an attempt to predict the growth of each ward by summing the growth of all other wards in the previous time step, weighted by the inter-ward correlation matrix). The entire algorithm is run for five separate sets of initial conditions, representing five different simulation start points spread evenly through the time span of the data.

For comparison, the time series are also reconstructed using only autocorrelation information. Thus, equation 5.7 is replaced with

$$\text{prediction}_t^{ward} = \text{corr\_sum}_{ward}^{ward} \times \text{prediction}_{t-1}^{ward} \tag{5.8}$$

The results of the cross-correlation based reconstruction are presented in figure 5.2, while the results of the autocorrelation-based reconstruction are presented in figure 5.3. The RMS divergence of cross-correlation based reconstruction, for the first time slice, is 0.53; rising over a six-year period to approximately 0.9. This compares favourably to predictions made by assuming homogeneity of growth (i.e. assuming all time series are equal to the national average time series, which for normalised series will always have an RMS divergence of 1.00) and also to predictions based on autocorrelation alone (which have first time slice RMS of 1.12, rising rapidly to 1.4).

It is not clear why predictions based on autocorrelation alone behave so badly - worse, on average, than a prediction based on market average. One hypothesis for this might be that such predictions are based on a single data point in the cross-correlation matrix and are therefore far more susceptible to noise in the data than other predictions. As they are only used as a benchmark, however, the quality of autocorrelation predictions is not of concern here. Meanwhile, it is encouraging to note that cross-correlation based predictions show significant improvement over homogeneous predictions in the short term (1 year), and slight improvement in the long term (4 or more years).

5.3.2.2 Stability of cross-correlations over time

It is pertinent to ask whether the structure of cross-correlations revealed in this analysis are a permanent feature of the market, or merely a transient reflection of its current state. Dividing the time series into two halves, and comparing cross-

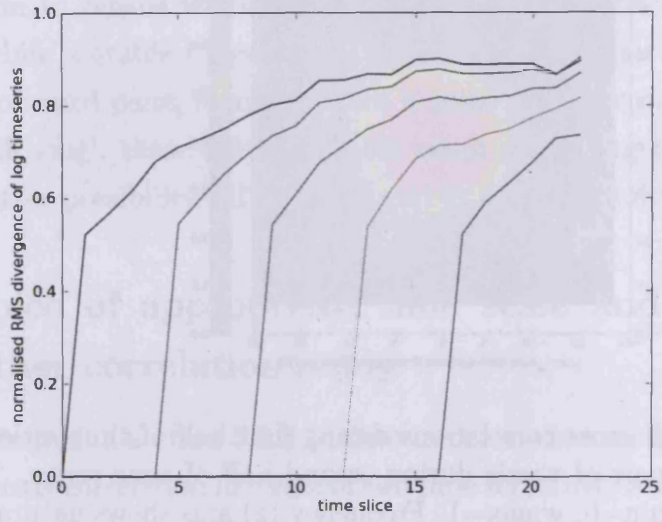


Figure 5.2: Plot of RMS divergence of predicted from actual time series, based on cross-correlation data with time slice of 100 days, wmin=1, wmax=2. Results from five different simulation runs are shown as separate curves, with start times evenly spaced over the 2000-2006 time span of the data set.

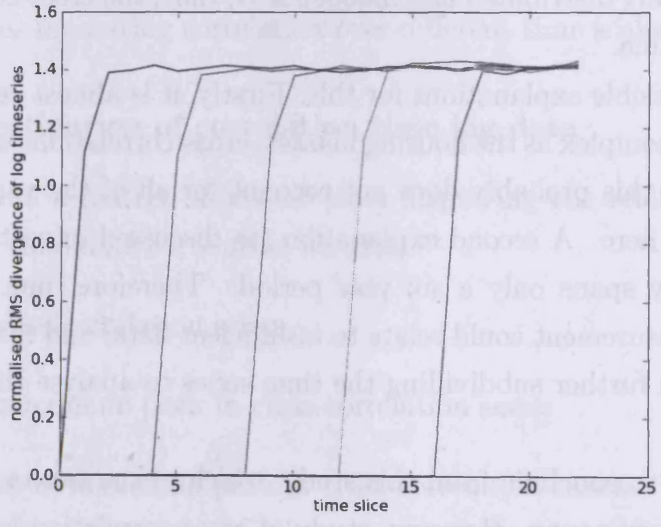


Figure 5.3: Plot of RMS divergence of predicted from actual time series, based on auto-correlation data with time slice of 100 days, wmin=1, wmax=2. Results from five different simulation runs are shown as separate curves, with start times evenly spaced over the 2000-2006 time span of the data set.

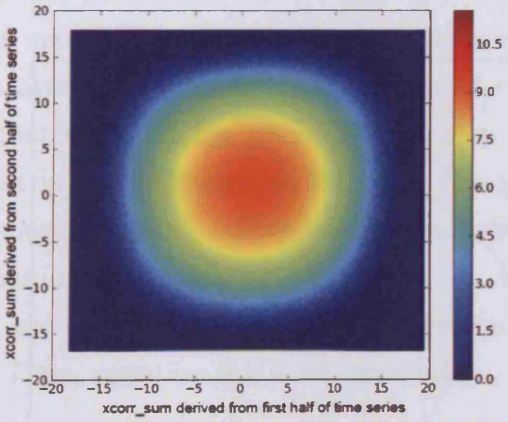


Figure 5.4: Plot of all cross-correlations during first half of time series, vs their values for the same pair of wards during second half of time series. Time slice length of 90 days, wmin=0, wmax=1. Frequency (z) axis shows natural logarithmic frequency count $ln(f + 1)$.

correlation derived from each half, answers this question succinctly. Figure 5.4 shows a scatterplot of cross-correlation sums derived from the years 2000-2003, versus the same correlation sums from the years 2003-2006. The circular shape of this plot indicates that cross-correlations derived from each time period are approximately normally distributed and independent; thus, the cross-correlations are not stable over time.

There are two possible explanations for this. Firstly, it is almost certain that in a social system as complex as the housing market, cross-correlations *will* change over time. However, this probably does not account for all of the variability in measurement shown here. A second explanation (as discussed in section 1.3) is that the entire study spans only a six year period. Therefore, much error in cross-correlation measurement could relate to insufficient data, and this will only be exacerbated when further subdividing the time series to analyse the stability of cross-correlations.

It is not possible to conclude from this study, which of the two explanations holds more explanatory power. However, study of cross-correlation is still valid for a number of reasons. Firstly, it is useful to explore methodologies for handling data sets such as these, even if the results appear unstable in this case. Secondly, even though the results are unstable, they may well be a good measure of the internal state of the market during the period under study. This is evidenced

both by their usefulness in reconstructing time series (section 5.3.2.1) and by their correlation to census statistics (which will be discussed in section 5.4.3). And finally, while a stable cross-correlation relationship may not exist for the vast majority of ward pairs, it may be that a small minority *do* exhibit a stable relationship, although these are not visible when displaying the data set as a whole. This latter possibility will be discussed in section 5.4.3.1.

5.3.3 Choice of appropriate time scale and metrics for further correlation study

All results presented in section 5.3.2.1 relate to correlation metrics calculated by summing the cross-correlation function of two time series over a window of a single time step. It is therefore pertinent to ask whether using longer time slices would lead to better reconstruction of the original time series. However, if extending the time span of the cross-correlation metric, it is also pertinent to ask whether any better alternatives exist than simple summation of the cross-correlation signal (as in section 5.3.1.1), such as the inter-series time difference at which maximum correlation occurs, and the magnitude of that maximum correlation.

Section 5.3.3.1 discusses use of alternative correlation metrics, and section 5.3.3.2 discusses measuring correlation over different time scales.

5.3.3.1 Investigation of correlation time lag data

Figure 5.5 shows a matrix of scatter plots displaying the relationship between three different candidate correlation metrics:

- sum of cross-correlation series
- value of maximum peak in cross-correlation series
- time offset of maximum peak in cross-correlation series.

The most important of these plots is that of *sum* vs *maximum*, the shape of which may be explained as follows:

- The x-axis reflection is caused by the fact that the value with greatest magnitude may be either positive or negative.

- The tails to the right of the upper cluster of points, and to the left of the lower cluster of points, show that for a small subset of cross-correlation functions, the presence of a single large maximum value has some effect on the overall sum.

- The vast concentration of points in the centre of the clusters, and their approximate y-axis symmetry (discounting the tails) reveals that there is little overall correlation between *maximum* and *sum*.

Similar analysis may be conducted for the plots of *time of maximum peak* vs *sum* and *maximum*, in all cases showing little in the way of interesting correlation. The lack of such correlations points toward the conclusion that the cross-correlation functions themselves are too noisy for the extraction of more detailed metrics, as in the absence of noise it would be expected that at least *maximum* should correlate with *sum* (as would be the case for e.g. a bell curve). Such detailed metrics are therefore discarded.

5.3.3.2 Investigation of different correlation time scales

Bearing in mind that the cross-correlation functions are too noisy for detailed temporal analysis, it is unclear how the additional temporal information gained from longer cross-correlation functions could be used to increase the accuracy of predictions. Therefore, in investigating different correlation time-scales, the reconstruction algorithm remains the same as in section 5.3.2.1. Figure 5.6 shows the result of reconstructing the original time series based on longer-term (up to 800-day) correlations.

The divergence graph for 200-day correlations initially rises to a higher value than was the case with 100-day correlations, however divergence in the longer term increases more slowly. Thus, the results display decreasing performance of the reconstruction when longer time scales are used to predict short-term movements, and increasing performance when longer time scales are used to predict longer term movements. This is presumably because short-term and long-term correlations differ (so it would be logical to expect short term measurements to be best for short term predictions, and so on) - however it may also be the case that predictions from long-term data are less susceptible to short term noise.

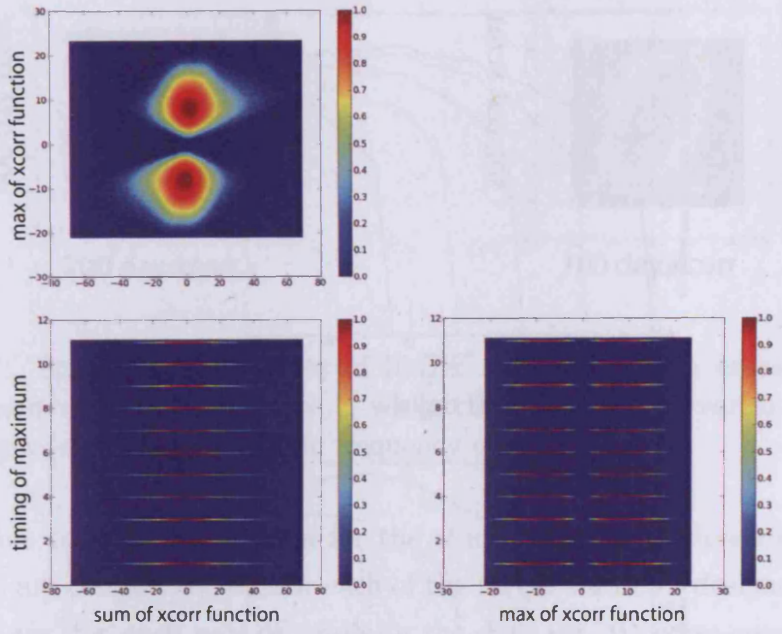


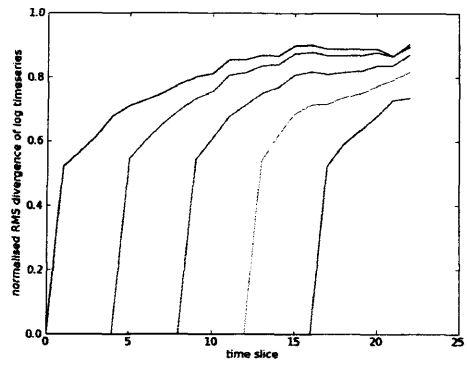
Figure 5.5: Scatter plots of sum/maximum/time of maximum peak for cross-correlation functions of each pair of wards. Frequency (z) axis shows natural logarithmic frequency count $ln(f + 1)$.

For completeness, figure 5.7 presents scatterplots of shorter vs longer term market correlations. The generally uncorrelated shape of the plots shows that short term correlation only dimly reflects long term correlation, and vice versa.

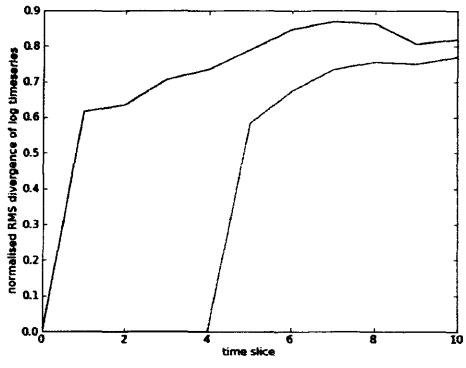
For the regression analysis, both 100- and 200-day correlations were used. The 200-day analysis is presented as a stronger relationship to the census data is visible in the results.

5.4 Regression analysis

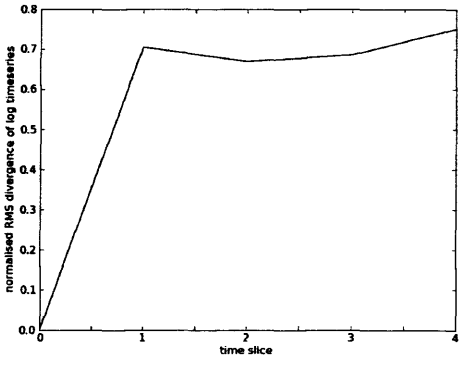
As stated in the opening section of this chapter, the purpose of this study is both to make use of interaction data to gain new perspectives on the UK housing market, and also to rate its relative importance to housing market structure as compared to the importance of more standard spatial data sets. Now that all data has been converted to a directly comparable network interaction format,



100 days



200 days



400 days

Figure 5.6: Reconstruction of time series based on long-term cross-correlation. In each case, $wmin = 0$ and $wmax = 1$ while the time slice length is varied. First time-slice errors are approx. 0.5 for 100 days, 0.6 for 200 days, 0.7 for 400 days and 1.0 for 800 days (not shown).

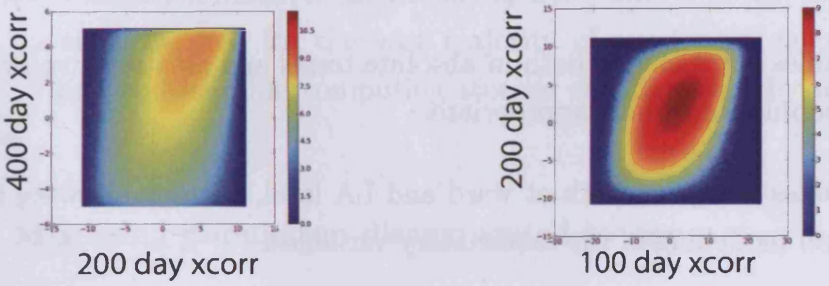


Figure 5.7: Scatter density plots of longer vs shorter term cross-correlation. Again, $wmin = 0$ and $wmax = 1$, while time slice length varies. Frequency (z) axis shows natural logarithmic frequency count $ln(f + 1)$.

it is possible to fulfil the criteria for the study using regression analysis. Two regressions are conducted, one for each of the target variables designed in section 5.3: these are, for each pair of wards in the data set, (i) cross-correlation, and (ii) log price ratios, formed by simple division of relative growth.

It should be remembered that the role of regression in this study is not, as is usually the case, to provide a comprehensive model of the data accompanied by a substantive interpretation of the meaning of that model. Instead, regression has been selected as the most appropriate tool for providing an exploratory analysis of the contents of the data set. However, it is impossible to entirely separate the processes of exploration and interpretation, because in the current context - where the aim is to develop analytical methods - those methods must be evaluated, and in evaluation, can only be considered useful if there is a meaningful interpretation of their outcomes. Therefore, an outline sketch of how the results may be interpreted is provided alongside the raw results.

The remainder of the section is structured as follows. Section 5.4.1 deals with the choice of explanatory variables. Section 5.4.2 discusses computational issues, and regression results for cross-correlation and log price ratios are presented in sections 5.4.3 and 5.4.4 respectively.

5.4.1 Choice of explanatory variables

When applying regression to the outcome of the cross-correlation analysis, it is again necessary to chose explanatory variables. Almost all considerations are the

same as those covered in chapter 4, therefore the reader is referred back to that chapter for more detail on the following points:

1. All variables are included both in absolute terms and as a relative proportion of population, where appropriate.

2. All variables are given both at ward and LA level, thereby allowing crude multi-level modelling of the explanatory variables.

3. All variables relating to population and finance are used in the form $log(x+1)$ to 'tame' outliers.

4. Principal Component Analysis is applied to reduce approximately 1000 explanatory variables to 40 components, thus eliminating problems of collinearity. It is arguable that a higher number of components should be used for the more complex interaction data; however due to software and compute time limitations the number has been kept at 40 as was the case in previous regressions.

5. Finally, a back-transform of the estimated regression coefficients is used to display results in terms of explanatory variables rather than components.

As noted in the introduction to this chapter, spatial models can be classified as *interactive* or *reactive*. The same classification applies to explanatory regression variables.

- The reactive variables used are identical to the set used in chapter 4, falling into the categories of population age distribution, travel to work distances, distance to London, average 2001 house price, housing stock, incomes, employment, housing type and social class. Chapter 4 should be referred to for a full discussion of these variables. As described in section 5.3.1.3, reactive variables are converted to an interaction format, and named as either FROM, TO or CHANGESQ variables depending on whether they relate to the origin or destination area, or the squared difference between them.

- The interaction variables used are described in sections 5.4.1.1 to 5.4.1.5 below.

All interaction variables relate to each of 78,332,500 *pairs* of census wards, however, the interaction matrices are usually sparse - that is to say, most entries are equal to zero, because for the vast majority of ward pairs no migrations occurred. Thus, considerable computing storage can be saved by not storing zero-values.

5.4.1.1 Migrating population disaggregated by age

Dennett & Stillwell (2008) provides a comprehensive district-level analysis of UK migration, noting (among other things) age as a key determinant of migration behaviour. It is reasonable to expect that migrants differing in behaviour - i.e., migrating for different reasons - will have differing effects on the housing market.

The following variables are therefore included:

- total number of migrants
- number of migrants aged under 16
- number of migrants aged 16 to 24
- number of migrants aged 25 to 34
- number of migrants aged 35 to 44
- number of migrants aged 45 to 59
- number of migrants aged 60 to 64
- number of migrants aged 65 to 74
- number of migrants aged 75 to 110

Data on the age of migrants is derived from census interaction table MG201.

5.4.1.2 Migrating population disaggregated by social class

Dennett & Stillwell (2008) also notes that socio-demographic characteristics of origins and destinations have major influence on the behaviour of migrants. Characteristics of origins and destinations are already included in the regression model employed here, using the variables described in chapter 4 and the expansion technique described in section 5.3.1.3. However, it also seems pertinent to include information on the class of migrants.

The following variables are therefore included:

- number of Total Wholly moving households

- number of Total Other moving groups
- number of Large employers and higher managerial occupations (Wholly moving households)
- number of Large employers and higher managerial occupations (Other moving groups)
- number of Higher professional occupations (Wholly moving households)
- number of Higher professional occupations (Other moving groups)
- number of Lower managerial and professional occupations (Wholly moving households)
- number of Lower managerial and professional occupations (Other moving groups)
- number of Intermediate occupations (Wholly moving households)
- number of Intermediate occupations (Other moving groups)
- number of Small employers and own account workers (Wholly moving households)
- number of Small employers and own account workers (Other moving groups)
- number of Lower supervisory and technical occupations (Wholly moving households)
- number of Lower supervisory and technical occupations (Other moving groups)
- number of Semi routine occupations (Wholly moving households)
- number of Semi routine occupations (Other moving groups)
- number of Routine occupations (Wholly moving households)
- number of Routine occupations (Other moving groups)
- number of Never worked and long term unemployed (Wholly moving households)
- number of Never worked and long term unemployed (Other moving groups)
- number of Full time student (Wholly moving households)
- number of Full time student (Other moving groups)
- number of Not classifiable for other reasons (Wholly moving households)
- number of Not classifiable for other reasons (Other moving groups)

Data on the social class of migrants is derived from census interaction table MG204.

5.4.1.3 Migrating households disaggregated by tenure type

It is reasonable to expect that migrants between regions will have differing impact on inter-regional market linkage depending on whether they rent, buy or occupy social housing. It is certainly already known that housing type affects mobility and hence migration behaviour: Boyle (1993) finds in a UK study that owner-occupier migrants are more restricted by distance than other types, the population of the South enjoys greater overall mobility, and that council tenants tend to migrate less for a variety of social and structural reasons.

The following variables are therefore included:

- number of owner occupied households migrating (whole household)
- number of owner occupied households migrating (partial household)
- number of social rented households migrating (whole household)
- number of social rented households migrating (partial household)
- number of private rented households migrating (whole household)
- number of private rented households migrating (partial household)

Migrating household tenure type data is derived from census interaction table MG205.

5.4.1.4 Commuting flows

Bidrent theory, as discussed in chapter 4, suggests that working households may locate based on a trade-off between commuting and land rent costs. Therefore, if it is feasible to commute between two places A and B, it is reasonable to expect that workers at A can chose to live either at A or B, provided land rents at A are not too high.

For this reason, the following variable is included:

- total number of commuters

This figure is derived from the census travel-to-work data.

5.4.1.5 Inter-ward distances

Finally, a key (perhaps definitive) component of a spatial analysis is the study of any links between spatial proximity and correlation! Therefore, a distance metric

is included, not only because of it's known connection to migration behaviour (Dennett & Stillwell 2008) but because it defines this study:

- inter-ward distance

Inter-ward distance is measured from ward centroids and derived from the UKBORDERS dataset. In the case of Local Authorities, an average distance between all pairs of wards across the pair of LAs is used.

5.4.2 Computation

The data set used in this regression is large: 78 million data points, for each of

- approx. 40 interaction variables, duplicated for absolute (direct counting)/relative (as a proportion of population) measures, and duplicated again for LA/ward measures = 160 variables
- approx. 50 reactive variables, each used three times as described in section 5.3.1.3, duplicated for absolute/relative measures and duplicated again for LA/ward measures = 600 variables

to create a total of approximately 760 explanatory variables. In order to complete regression on the entire data set within a reasonable time frame, it was necessary to parallelise some of the steps described in chapter 4, notably

- calculation of the correlation matrix for principal component analysis, and
- producing scatter plots of all explanatory components vs the target.

The data expansion phase described in section 5.3.1.3, and the ward to Local Authority lookup and subsequent retrieval of LA-level variables was also parallelised. Fortunately the mdp toolkit provides resources with which to parallelise data generation and PCA transformation. For the visualisation process, each plot is produced independently therefore parallelisation is relatively simple.

The regressions each took approximately 48 hours to run on a 16-core 3GHz Itanium machine. There is probably much scope for code optimisation which would reduce this runtime considerably (however for research purposes it is considered a better use of time to attend to other tasks while the machine continues computing!). The final version of the code also used 32Gb of memory. Such

memory use is unnecessary as the data set can be processed from disk in a linear manner, however loading the output of the PCA dimension reduction into available memory allows a memory/runtime tradeoff to be made, reducing run times.

5.4.3 Regression results for cross-correlations

This section presents results for regression of cross-correlation data against census interaction and aggregate statistics. The time slice length is 200 days, with $wmin = 0$ and $wmax = 2$ time slices, therefore all correlated movements must occur within 400 days of each other in order to be detected.

5.4.3.1 Cross-correlation regression diagnostics

The regression of cross-correlation against census variables had a mean square residual of 0.88, indicating that the model does not fit the data well. However, it should be remembered that this regression is an attempt to explain market movements, which in an efficient market should be entirely unpredictable. Therefore, the ability to explain even as much as 12% in the variation in correlation is a useful finding - even though few would argue that the UK property market is efficient! Furthermore, the mean square error (when regression was applied to the test set) was also 0.88, indicating that the findings of the regression do not lose any predictive power when applied to other data from the same time period; this fact helps to validate the meaningfulness of the derived parameters.

A histogram plot of residuals (figure 5.8) shows that they are normally distributed. A plot of residual against prediction (figure 5.9) shows no correlation for the majority of data points, but considerable evidence of heteroscedasticity for a subset of outlying data. The slope of the plot indicates that high correlation predictions are too high, while low correlation predictions are too low: this may be explained by nonlinearity at the extreme ends of the data range, however without a fuller investigation it is impossible to be certain. The Moran I statistic is not computed for the residuals, as to do so for a data set of this size would require implementation of a more complex algorithm than that used in chapter 4; moreover, a visualisation of the residuals themselves (to be shown at the end of this chapter, in figure 5.20) shows that massive spatial autocorrelation is present

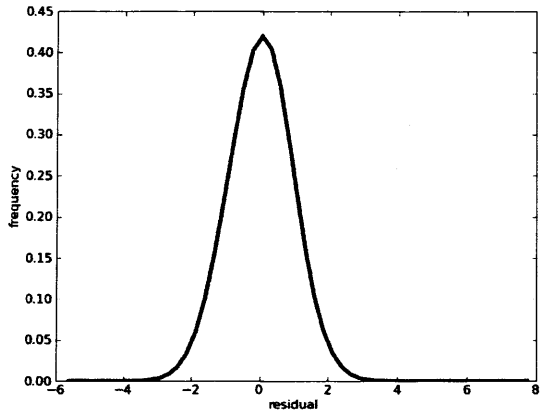


Figure 5.8: Histogram plot of residuals from cross-correlation regression.

in any case, indicating that exploratory analysis has only begun to scratch the surface of what may be explainable in the data set. The similarity of this figure to the target variable of cross-correlation itself may be partially explainable by the nonlinearities shown in the plot of residual against prediction.

As visualisation is considered a crucial part of this research, scatterplots of all PCA components versus the target variable were created and checked for interesting behaviour. The most correlated component is displayed in figure 5.10; the most significant variables represented by this component relate to absolute numbers of residents and dwellings, dwelling type and social class for the origin ward (with positive correlation) and destination wards (with negative correlation). Thus, component describes a signed difference between the wards in terms of these variables.

5.4.3.2 Discussion of cross-correlation regression results

The results of the cross-correlation regression can be summarised (as with the house price regression of chapter 4) by an approximate description of the list of explanatory coefficients, ranked in order of magnitude. The coefficients themselves are shown in table 5.1. For the definition of FROM, TO and CHANGESQ variables see section 5.3.1.3.

- The first four coefficients relate to migration at ward level. Note, however, that the signs of these coefficients differ: it is not the total level of migration

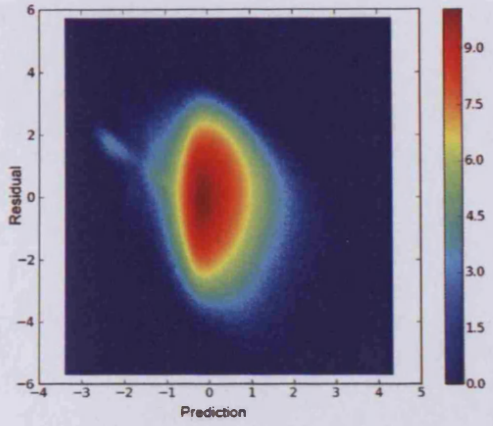


Figure 5.9: Plot of residuals against corr_sum values predicted by regression. Frequency (z) axis shows natural logarithmic frequency count $ln(f + 1)$.

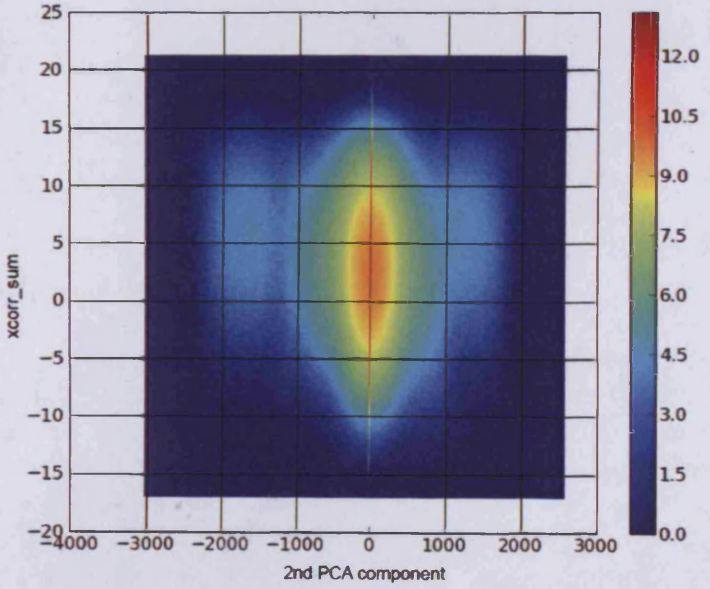


Figure 5.10: Plot of cross-correlation against most correlated component for cross-correlation regression. Frequency (z) axis shows natural logarithmic frequency count $ln(f + 1)$.

Variable Name	Reg. Coeff	99% Conf
MG204-N01-TotalOthermovinggroups.log.wardindex.sor	0.0017	2.1e-05
MG204-N00-TotalWhollymovinghouseholds.log.wardinde	-0.0016	1.8e-05
MG204-A00-TotalWhollymovinghouseholds.log.wardinde	-0.0014	1.6e-05
MG204-A01-TotalOthermovinggroups.log.wardindex.sor	0.0014	1.8e-05
TO-UV031-LA-N03-SmallEmployers.log	-0.0012	1.3e-05
TO-UV055-LA-A02-Shared.log	-0.0012	1.1e-05
TO-UV035-LA-A03-10k20k.log	-0.0012	1.1e-05
TO-UV028-LA-A00-Employed.log	-0.0011	1.1e-05
TO-UV031-LA-A05-SemiRoutine.log	-0.0011	1.1e-05
TO-UV056-LA-N07-Inapurposebuiltblockofflats.log	-0.0011	1.2e-05
TO-UV035-N04-20k30k	-0.0011	1.1e-05
TO-UV004-LA-A07-66to79.log	-0.0011	1e-05
FROM-UV056-N03-Detatched	0.0011	1.1e-05
MG205-N05-PrivateRentedOther.log.wardindex.sorted	0.0011	1.3e-05
TO-UV035-A03-10k20k.log	-0.0011	1e-05
TO-UV056-N03-Detatched	-0.0011	1e-05
FROM-UV031-LA-N03-SmallEmployers.log	0.0011	1.3e-05
FROM-UV035-N04-20k30k	0.0011	1.1e-05
TO-UV035-LA-N08-WorkingOffshore.log	-0.0011	1e-05
TO-UV035-LA-N03-10k20k.log	-0.001	1e-05
TO-UV053-N04-Vacanthouseholdspace	-0.001	1.1e-05
TO-UV056-LA-A03-Detatched.log	-0.001	1.1e-05
TO-UV035-LA-N00-Under2k.log	-0.001	1.1e-05
TO-UV056-N08-Partofaconvertedorsharedhouse	-0.001	1.1e-05
FROM-UV055-LA-A02-Shared.log	0.001	1.1e-05
TO-UV031-LA-A01-LowerManagerialAndProfessional.log	-0.001	1.1e-05
TO-UV056-LA-A00-ALLHOUSEHOLDS.log	-0.001	1.1e-05
FROM-UV035-LA-A03-10k20k.log	0.001	1.1e-05
TO-UV056-A11-Inashareddwelling.log	-0.001	1.2e-05
TO-UV031-N04-LowerSupervisoryAndTechnical	-0.001	1e-05
TO-UV056-A06-Flatmaisonetteorapartment.log	-0.001	1.1e-05
TO-UV035-N05-Over30k	-0.001	1e-05
FROM-UV004-LA-A07-66to79.log	0.001	1e-05
FROM-UV056-LA-N07-Inapurposebuiltblockofflats.log	0.00099	1.2e-05
FROM-UV056-A11-Inashareddwelling.log	0.00098	1.2e-05
MG204-N07-Lowermanagerialandprofessionaloccupation	0.00098	1.2e-05
FROM-UV035-LA-N03-10k20k.log	0.00098	1e-05
FROM-UV053-N04-Vacanthouseholdspace	0.00098	1.1e-05
FROM-UV031-LA-A05-SemiRoutine.log	0.00097	1.1e-05
TO-UV004-LA-N01-Under16.log	-0.00097	9e-06
TO-UV004-A07-66to79.log	-0.00096	1e-05
FROM-UV028-LA-A00-Employed.log	0.00096	1.1e-05
TO-UV031-LA-A04-LowerSupervisoryAndTechnical.log	-0.00096	9e-06

Table 5.1: Most significant parameters (those with greatest magnitude) from the ward-level cross-correlation regression.

which forms an indicator of housing market connectedness, so much as its type, both in absolute numbers and as a proportion of migrants. Wholly moving households tend to indicate less correlation between areas, while other moving groups indicate a more connected market. Coefficient magnitudes range from 0.0017 to 0.0014; although these appear small, they are at least five times larger than the average coefficient magnitude from the regression, which is 0.00025. Also, the 99% confidence intervals are small in comparison to the coefficient magnitudes.

The presence of these variables appears to be due mainly to the 29th principal component, which appears to distinguish primarily between wholly moving households and other moving groups, and is the 5th most important component for determining market correlation. Private renting and lower managerial/professional occupations are also noted to correlate with the 'other moving groups' category, thus, further migration variables relating to these categories appear prominent in the results.

Local authority level migration does not appear until 91st place, with a coefficient magnitude of 0.00074. This hints that the usefulness of migration as an indicator of market linkage might be limited to local movements, as long-distance inter-ward migration is fairly sparse.

- In 5th to 261st position, a large number of FROM- and TO- type variables appear, generated from census tables UV031 (social class), UV035 (travel to work distance), UV055 (dwelling occupancy), UV056 (dwelling type), UV004 (population age), UV053 (housing stock) and UV028 (economic activity). Coefficient magnitudes range from 0.0012 to 0.00017.

  It can therefore be said that the individual wards' characteristics have a large part to play in whether they drive, or are driven by, the market of other wards. The importance of FROM- and TO- type variables relative to CHANGESQ- type variables, suggests that for a significant proportion of pairs of differing wards A and B, it can be said that A affects B more than B affects A. This could be interpreted as a kind of ripple effect, though not necessarily in the form popularised in the literature. The question of which wards tend to more strongly drive others will be explored in more detail in chapter 6.

It is noticeable that variables from both ward and LA level are present; also both absolute and relative measures are significant (that is, both direct head/dwelling counts, and relative fractions of population/housing stock by type). Little pattern is discernible as to which sub-types of variable the regression model has 'chosen' to represent the underlying statistics; however as an exception to this, it is noticeable that housing stock tends to have a greater effect at ward level, while short commuting distances have a greater effect at LA level - the latter possibly being indicative of urban centres.

- In 262nd place, the first CHANGESQ (square change) variable appears with a coefficient of 0.00018, suggesting that the unsigned square difference between area characteristics is not nearly so important as their signed difference in determining market correlation. That is to say, knowing that areas A and B differ in terms of a certain statistic is not of great use compared to knowing which one possesses the higher value of that statistic. Thus, differing housing submarkets defined by the characteristics of areas are not actually independent; instead, certain types of submarket appear to drive certain other types.

- In 390th place, the interaction variable relating to the absolute distance between wards appears ($\beta = 0.000095$). Paradoxically, the insignificance of physical distance in determining ward level market correlation would be a significant finding! However, such a low coefficient may relate in part to the fact that migration interaction data is a better indicator of the proximity of areas than straight line distance, which fails particularly in cases where road network distance is much greater than a straight line, e.g. for towns situated on opposite sides of a river estuary. Still, it is shown that greater inter-ward distance tends to reduce market cross-correlation.

  To reassure the reader that the "CHANGESQ-distance-to-London" variable is not being used as a proxy for inter-ward distance, it is noted that this variable is also assigned only a small parameter (-0.000059, in 487th place).

  Overall however, this finding should be taken with the caveat that inter-ward distance, being a single variable, accounts for only a small proportion of variability in the combined set of interaction data, and is therefore not strongly or independently represented in any of the 40 PCA components

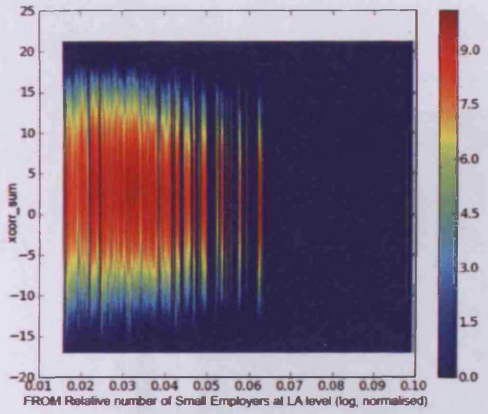


Figure 5.11: Plot of cross-correlation metric (Y) against a significantly correlated 'FROM' variable (X) - difference between relative number of small employers at local authority level. Vertical lines appear due to grouping of 8850 wards into 376 local authorities; meaning that the variable can only take one of 376 discrete values. Frequency (z) axis shows natural logarithmic frequency count $ln(f + 1)$.

used for regression.[1] Thus, it may be the case that the limitations of PCA regression preclude meaningful discussion of coefficients with a magnitude so small as this.

As a key component of this research is visualisation, scatterplots of all input variables versus the target variable were produced. Figure 5.11 shows an example for a significant FROM variable, while figures 5.12 and 5.13 show the relationship of market correlation to total migration flow, and inter-ward distance respectively.

The latter two plots merit further discussion. It should firstly be noted that both variables show a clear correlation with property market connectedness. However, the low coefficients assigned to them by the regression engine imply that this correlation is better explained by characteristics of individual areas, and by the type of migration flow, than by total migration flows or inter-ward distances.

Secondly it should be noted that the migration versus market correlation plot of figure 5.12 displays a 'tail' of points about which more information is known: if total migration flows exceed 100 people, then it is almost certain that the areas concerned exhibit above average correlation. Such a relationship could be further analysed through the data mining techniques discussed in chapter 4, using

[1] Out of the 40 regression components used, distance appears most strongly in number 22, which relates mainly to social class and dwelling type.

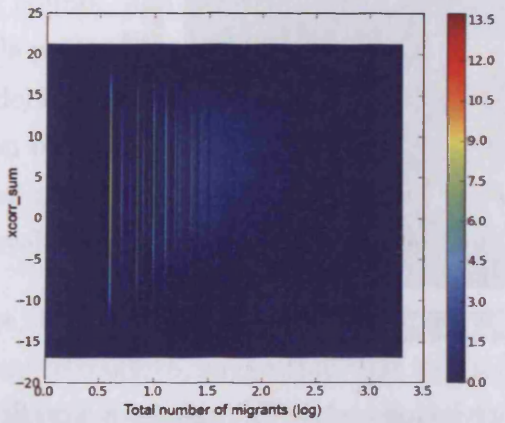


Figure 5.12: Plot of cross-correlation metric (Y) against total migration flow (X). Vertical lines appear due to the discrete nature of migration flows, with a gap where $0 < log(migration) < 0.5$, i.e. for migration values of 1 or 2, caused by small value adjustment carried out by the census office to preserve anonymity. Frequency (z) axis shows natural logarithmic frequency counts $ln(f + 1)$.

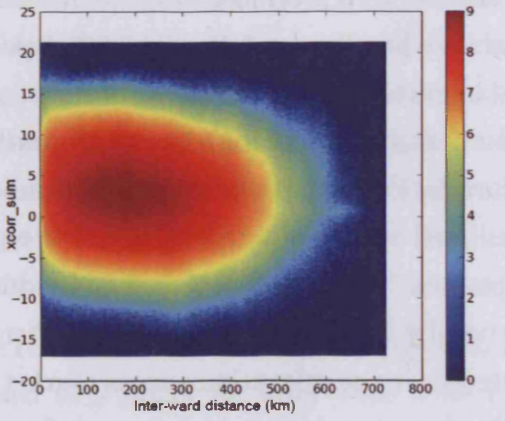


Figure 5.13: Plot of cross-correlation metric (Y) against inter-ward distance in metres (X). Frequency (z) axis shows natural logarithmic frequency counts $ln(f + 1)$.

Figure 5.14: Map of 'tail' interactions from figure 5.12: highly correlated wards linked by strong migration flows. Note that not all wards shown on this map are correlated with one another - they are most likely each correlated only with their close neighbours.

a support/confidence/lift framework. However, this finding is not considered to be particularly meaningful for two reasons:

1. all that has been identified in the migration plot 'tail' is a small subset of highly linked areas. To illustrate, the areas represented by the tail of figure 5.12 are plotted on a map in figure 5.14. The map simply identifies a certain quantity of urban areas, within each of which the housing market is tightly integrated.

2. in any case, as mentioned above, characteristics of origins and destinations and the type of migration flow, explain market correlation far better than total migration flows.

Several of the scatter plots of components against the target variable exhibit similar 'tails' - figure 5.10, for example, exhibits two such features. These could, in future, be further explored through data mining; such work certainly has the potential to produce more accurate regression models.

5.4.4 Regression results for log price ratios

This section presents the results for the regression of log price ratios against census statistics. While this is a somewhat redundant study - as a price ratio is determined from two individual prices, and the variance of these has already been well explained in chapter 4 - the analysis is still of value, firstly because by producing sensible results relating to better understood systems, the properties of the regression engine can be further studied and validated; and secondly because it allows a study to be made of the links between interaction data (specifically, migration) and house price difference.

5.4.4.1 Log price ratio regression diagnostics

The mean square residual in the log price ratio regression is 0.25, indicating a much better goodness-of-fit than for market cross-correlation data. The mean square error against a test data set is likewise 0.25, confirming the applicability of the derived parameters to other data recorded during the same area and time period.

A histogram plot of residuals is not shown as it appears very similar to figure 5.8, exhibiting a normal distribution. Figure 5.15 gives a plot of residuals against predicted target value for each regression point, which shows a number of features of interest:

- The majority of the data points show a very slight positive correlation between residual and prediction, indicating small nonlinearities in the data unaccounted for by the model - high predictions turn out to be not quite as high as the actual target value, while low predictions turn out not to be quite as low as the actual target.

- A backward diagonal line is clearly visible spanning the entire plot. This is caused by the price ratio being defined as zero if either the origin or destination exhibits no transactions from which to measure the price. In the case where $target = 0$, then $residual = target - prediction = -prediction$. Arguably for these points, the prediction provides better valuation than the data itself, which contains no transactions.

- Overall, the plot is approximately symmetrical, which is reassuring as price ratios should appear in matched pairs a/b and b/a.

As visualisation is considered a crucial part of this research, scatterplots of all PCA components versus the target variable were created and checked for interesting behaviour. The most correlated component is displayed in figure 5.16; the most significant variables in this component relate to differences in social class, age, commute distance and dwelling type.

5.4.4.2 Discussion of log price ratio regression results

Results for the log price ratio regression are shown in table 5.2.

No single category of variable appears to be more important than any other in the log price ratio regression. For the most important 40 variables, coefficient magnitudes range from 0.003 to 0.001, as compared to a mean coefficient magnitude of 0.0004.

It is notable that FROM and TO variables appear in matched pairs, e.g. the FROM and TO coefficients for the same variable will tend to have opposing signs. This is to be expected as any statistic which increases the origin-destination ratio

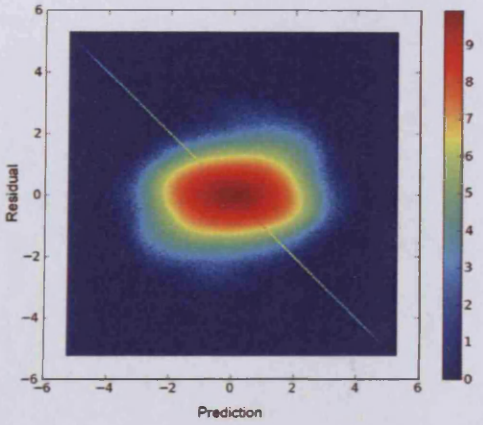


Figure 5.15: Plot of residuals against log price ratios predicted by regression. Frequency (z) axis shows natural logarithmic frequency count $ln(f + 1)$.

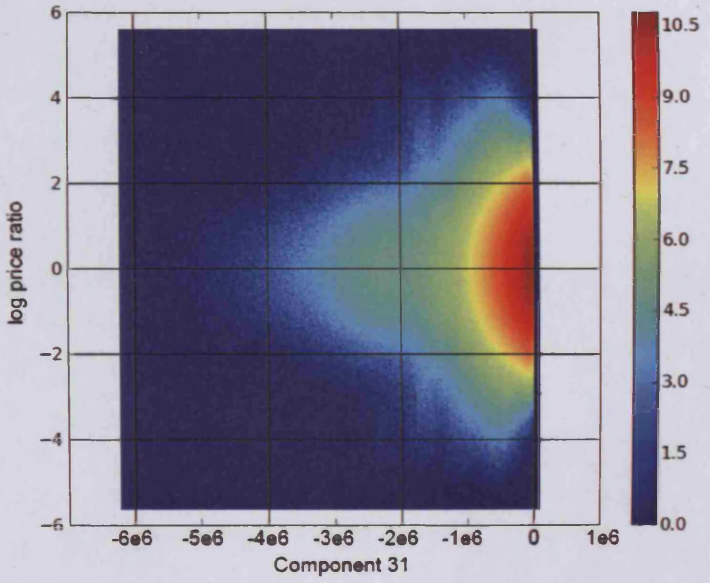


Figure 5.16: Plot of most correlated component for log price ratio regression. Frequency (z) axis shows natural logarithmic frequency count $ln(f + 1)$.

Variable Name	Reg. Coeff	99% Conf
FROM-UV031-N06-Routine	-0.0032	4e-06
TO-UV031-N06-Routine	0.003	4e-06
FROM-UV031-LA-A01-LowerManagerialAndProfessional.l	-0.0027	4e-06
TO-UV031-LA-A01-LowerManagerialAndProfessional.log	0.0026	3e-06
MG204-N00-TotalWhollymovinghouseholds.log.wardinde	-0.0026	9e-06
MG204-A00-TotalWhollymovinghouseholds.log.wardinde	-0.0026	1e-05
TO-UV056-N04-Semidetatched	-0.0024	6e-06
FROM-UV004-A06-56to65.log	0.0024	4e-06
FROM-UV056-N04-Semidetatched	0.0023	6e-06
TO-UV004-A06-56to65.log	-0.0022	4e-06
FROM-UV031-LA-N00-HigherManagerialAndProfessional.	0.0022	5e-06
TO-UV031-LA-N00-HigherManagerialAndProfessional.lo	-0.0021	5e-06
FROM-UV056-A09-Inacommercialbuilding.log	0.002	4e-06
FROM-distance to London.log	-0.002	5e-06
TO-UV056-A09-Inacommercialbuilding.log	-0.002	4e-06
TO-distance to London.log	0.0019	5e-06
FROM-UV035-LA-N01-2k5k.log	0.0018	5e-06
MG205-A00-OwnerOccupiedWholeHousehold.log.wardinde	-0.0018	7e-06
TO-UV035-LA-N01-2k5k.log	-0.0018	5e-06
MG204-A06-Lowermanagerialandprofessionaloccupation	-0.0017	6e-06
TO-UV004-LA-A01-Under16.log	-0.0017	5e-06
FROM-UV004-LA-A01-Under16.log	0.0016	5e-06
FROM-UV056-LA-N10-Caravanorothermobileortemporarys	0.0016	3e-06
TO-UV056-LA-N10-Caravanorothermobileortemporarystr	-0.0016	3e-06
FROM-UV035-LA-N06-NoFixedPlaceOfWork.log	-0.0016	6e-06
FROM-UV056-A10-Caravanorothermobileortemporarystru	-0.0016	4e-06
FROM-UV035-LA-A03-10k20k.log	-0.0016	6e-06
TO-UV056-A10-Caravanorothermobileortemporarystruct	0.0016	4e-06
MG205-A04-PrivateRentedWholeHousehold.log.wardinde	-0.0016	7e-06
MG205-N00-OwnerOccupiedWholeHousehold.log.wardinde	-0.0015	5e-06
TO-UV035-LA-N06-NoFixedPlaceOfWork.log	0.0015	6e-06
MG204-N06-Lowermanagerialandprofessionaloccupation	-0.0015	5e-06
FROM-UV004-LA-A03-26to35.log	0.0015	4e-06
FROM-UV031-N03-SmallEmployers	0.0015	4e-06
TO-UV004-LA-A03-26to35.log	-0.0014	4e-06
TO-UV031-N04-LowerSupervisoryAndTechnical	-0.0014	6e-06
TO-UV004-A05-46to55.log	-0.0014	3e-06
FROM-UV004-A05-46to55.log	0.0014	3e-06
MG205-N04-PrivateRentedWholeHousehold.log.wardinde	-0.0013	5e-06
TO-UV035-LA-A03-10k20k.log	0.0013	6e-06
TO-UV031-N03-SmallEmployers	-0.0013	4e-06
MG201-A01-Under16.log.wardindex.sorted	-0.0013	7e-06
MG201-A04-35to44.log.wardindex.sorted	-0.0013	7e-06

Table 5.2: Most significant parameters (those with greatest magnitude) from the log price ratio regression.

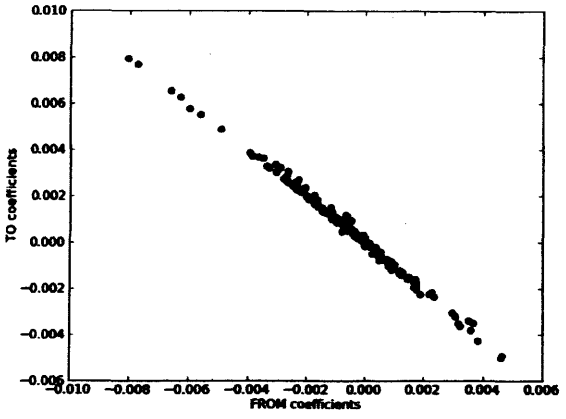


Figure 5.17: Plot of derived coefficients for FROM- and TO- pairs of variables in the log price ratio regression.

will also decrease the destination-origin ratio. Figure 5.17 illustrates this correlation by plotting FROM- coefficients against their TO- counterparts. By the same reasoning, CHANGESQ coefficients should theoretically always equal zero, otherwise their effect will be to increase the accuracy of one origin-destination prediction at the expense of decreasing the accuracy of another. The largest value assigned to any such coefficient is 0.00017, which is 18 times smaller than the maximum coefficient, but still half the size of the average coefficient magnitude. This finding has implications for the accuracy of the market cross-correlation regression: showing that parameters derived from the 760-variable regression engine, while generally indicative of the underlying trends, may still be susceptible to some small degree of error. However, the clear linear plot of figure 5.17 helps to validate the assumption that the regression engine is performing the task expected of it.

Migration variables are also important, with coefficients up to 0.002. In general - as was the case with the cross-correlation regression - ward level migration variables are assigned larger coefficients than local authority level migration variables; possibly because they provide valuable information on lifecycle-type moves within a region, which can be used to deduce the structure of the local housing market.

5.4.4.3 Analysis of the relation of migration to price differences via simple correlation

While the study of regression coefficients allows for estimation of price differentials, as noted in section 5.4.4, this is a redundant prediction useful only for validation of the regression engine itself. In terms of general interest, the question of which variables best predict price differential between two areas is a slightly odd question to ask, when it is already well known which variables determine price for each area individually. It is more interesting, therefore, to study the link between price differential and migration not in terms of regression coefficients but in terms of simple correlation. These statistics are displayed in table 5.3. To pick out a few salient features of the data;

- the top upmarket movers are owner-occupied whole households and all wholly moving households, while the top downmarket movers are the 'other moving groups' (partial households) in the private rental market, higher professional and lower managerial/professional occupations.
- all age bands tend to move upmarket except for 16-25, 26-34 and 35-44 year olds, and those over the age of 75.
- commuting usually takes place from a lower- to higher-value area.

Note that as this calculation is performed on a per-interaction basis, not a per-person or per-household basis. Thus the figures do not necessarily reflect the relative likelihood of each individual from that category to move up or downmarket, instead they represent the tendency of up- and downmarket price differentials to be coincident with moving groups of the types listed. Also, as the data does not represent the value of individual houses but is captured at ward level, interpretation in terms of people moving up/downmarket is likely to be inaccurate in wards with diverse housing types, i.e., particularly in rural areas.

5.5 Visualisation

While scatterplot visualisation of principal components versus the target variable, and each input variable versus the target variable is undertaken in sections 5.4.3 and 5.4.4, to date the new data generated in this chapter - the cross-correlation

Variable Name	Coeff
MGcommute.log	-0.0072
MG205-A00-OwnerOccupiedWholeHousehold.log	-0.0032
MG205-A05-PrivateRentedOther.log	0.0028
MG204-A00-TotalWhollymovinghouseholds.log	-0.0026
MG204-A05-HigherprofessionaloccupationsOthermovinggroups.log	0.0026
MG204-A07-LowermanagerialandprofessionaloccupationsOthermovi	0.0022
MG204-A21-FulltimestudentOthermovinggroups.log	-0.0021
MG204-A20-FulltimestudentWhollymovinghouseholds.log	-0.0018
MG204-A01-TotalOthermovinggroups.log	0.0015
MG201-A05-45to59.log	-0.0014
MG204-A22-NotclassifiableforotherreasonsWhollymovinghousehol	-0.0012
MG201-A06-60to64.log	-0.0012
MG204-A14-SemiroutineoccupationsWhollymovinghouseholds.log	-0.0012
MG204-A09-IntermediateoccupationsOthermovinggroups.log	0.0012
MG204-A06-LowermanagerialandprofessionaloccupationsWhollymov	-0.0011
MG204-A08-IntermediateoccupationsWhollymovinghouseholds.log	-0.0011
MG204-A03-LargeemployersandhighermanagerialoccupationsOtherm	0.0011
MG201-A07-65to74.log	-0.00091
MG204-A16-RoutineoccupationsWhollymovinghouseholds.log	-0.00087
MG201-A00-AllPeople.log	-0.00085
MG205-A03-SocialRentedOther.log	-0.00067
MG205-A02-SocialRentedWholeHousehold.log	-0.00066
MG204-A18-NeverworkedandlongtermunemployedWhollymovinghouseh	-0.00065
MG204-A12-LowersupervisoryandtechnicaloccupationsWhollymovin	-0.00046
MG201-A01-Under16.log	-0.00043
MG204-A10-SmallemployersandownaccountworkersWhollymovinghous	-0.00038
MG204-A13-LowersupervisoryandtechnicaloccupationsOthermoving	-0.00034
MG204-A02-LargeemployersandhighermanagerialoccupationsWholly	0.00034
MG205-A01-OwnerOccupiedOther.log	-0.00034
MG204-A19-NeverworkedandlongtermunemployedOthermovinggroups.	-0.00033
MG201-A02-16to24.log	0.00033
MG204-A17-RoutineoccupationsOthermovinggroups.log	-0.0003
MG204-A11-SmallemployersandownaccountworkersOthermovinggroup	-0.00024
MG201-A08-75to110.log	0.00022
MG201-A03-25to34.log	0.0002
MG201-A04-35to44.log	-0.00013
MG204-A23-NotclassifiableforotherreasonsOthermovinggroups.lo	-0.00011
MG204-A15-SemiroutineoccupationsOthermovinggroups.log	-8.9e-05
MG205-A04-PrivateRentedWholeHousehold.log	-7.4e-05
MG204-A04-HigherprofessionaloccupationsWhollymovinghousehold	-8e-06

Table 5.3: Correlation coefficients for ward level interaction variables versus log price ratio. Negative values indicate a move upmarket.

matrix between all pairs of wards - has not been visualised. Such visualisation can now be performed using the techniques developed in chapter 3, and is shown in figures 5.18 and 5.19. Regression residuals are also visualised in figure 5.20, though these are not discussed further at this stage - the possibility of extended analysis of residuals is dealt with in chapter 7.

Examination of figures 5.18 and 5.19 shows that wards and Local Authorities both exhibit strong inter-correlation. However, while LAs overall tend to be strongly positively or negatively correlated (relative to the average correlation level), the correlation properties of individual wards within them are not determined by the correlation of the LA alone. In other words, an LA with positive correlation to an external location X may still contain some wards with strong negative correlation to X. This points to the need, in future studies of correlation, for true multi-level modelling whereby the variance of the target variable (correlation) is split into broader and finer spatial levels which are then treated separately. There is even a remote possibility that correlation matrices exhibit fractal behaviour, down to the physical limit of an individual property. If this is the case, it may be worth computing their fractal dimension - although exactly what such a computation would reveal about the system overall would be difficult to answer unless a variety of different markets were studied.

Two features of the cross-correlation plots are particularly noticeable.

- the existence of large red and blue blocks - more visible in the LA level plot (fig. 5.19) but also visible for wards (fig. 5.18) - indicate that large spatially contiguous regions have above average intra-region time series correlation. Meanwhile, the inter-region correlation appears to be below average.

- the prevalence of horizontal and vertical lines. This agrees with the finding of section 5.4.3.2 that characteristics of origins and destinations, taken in isolation, are important in determining patterns of cross-correlation.

The existence of these two features points toward the potential usefulness of two simpler techniques which can be used to disentangle the complexity of property market interactions. In the former case, the presence of large, spatially contiguous regions exhibiting similar market behaviour could be uncovered by a *cluster analysis*. In the latter case, the presence of horizontal and vertical lines suggest that it is worth studying what causes a ward to *drive* the rest of the

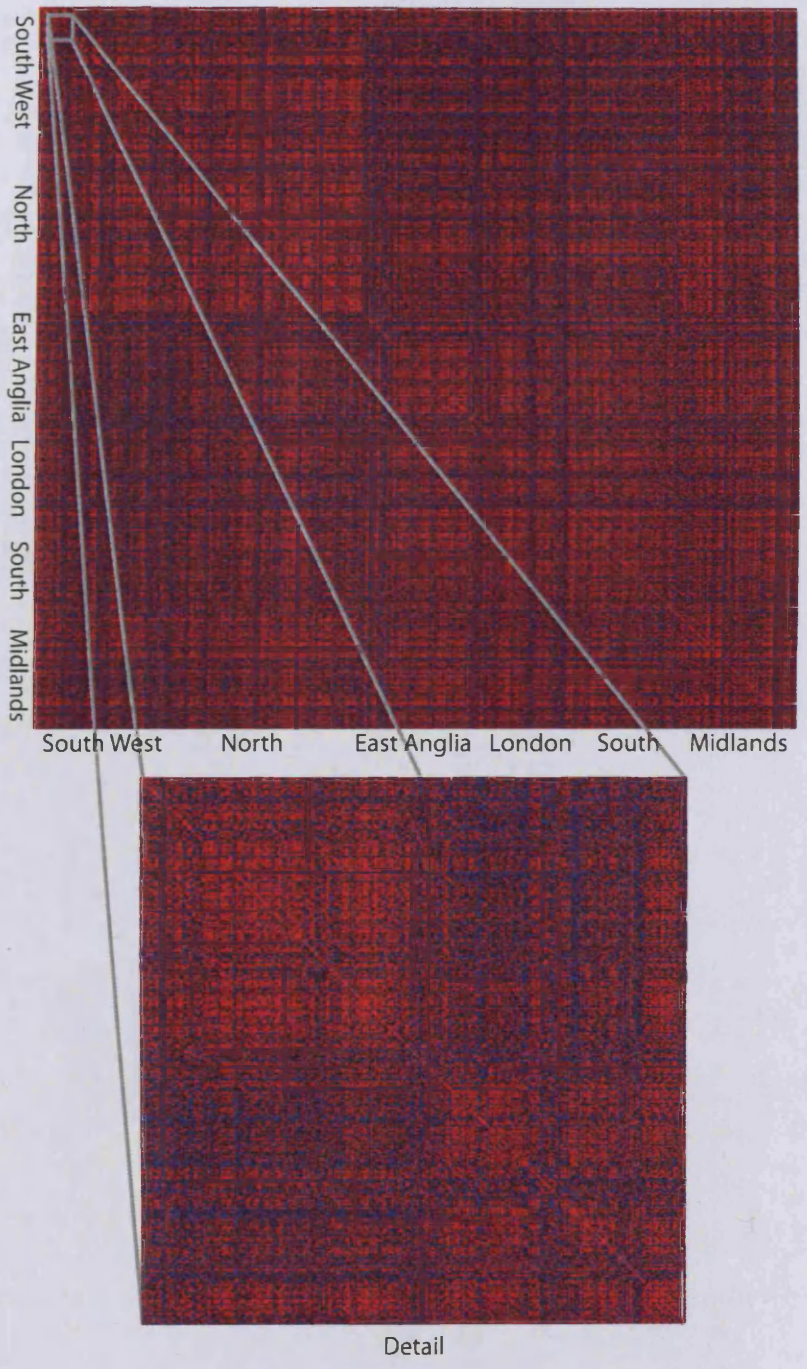


Figure 5.18: Pixel matrix plot of ward cross-correlations, 200 day time slice, $wmin = wmax = 0$. Axis ordering is hierarchical non-optimal - for further explanation of the display technique, see chapter 3. Red indicates above average correlation; blue indicates below average correlation.

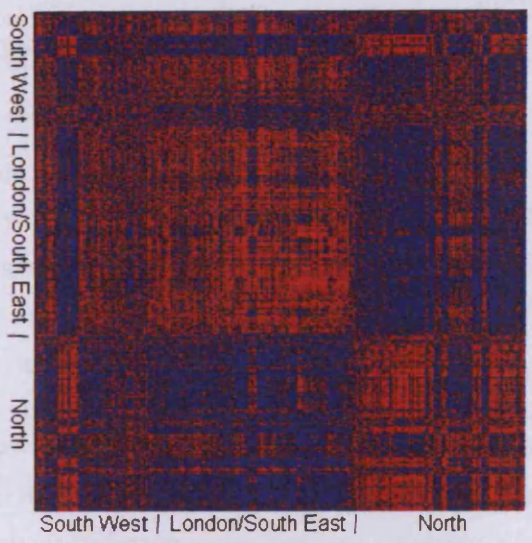


Figure 5.19: Pixel matrix plot of local authority cross-correlations. Axis order is optimal. Red indicates above average correlation; blue indicates below average correlation.

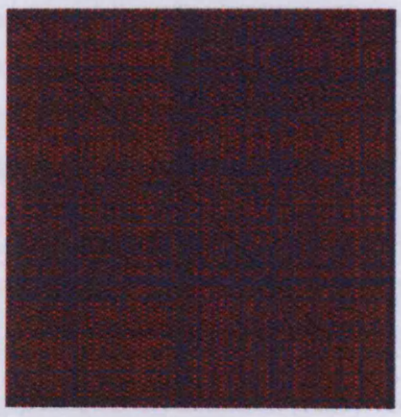


Figure 5.20: Pixel matrix plot of cross-correlation regression residuals, zoomed out to small scale using nearest neighbour interpolation. Axis ordering is hierarchical non-optimal. Red indicates positive residual, blue indicates negative. Massive spatial autocorrelation is visible, demonstrating that the model presented has only scratched the surface of what is explainable in this data. Note the similarity of the pattern of residuals to the cross-correlation itself shown in figure 5.18.

market (indicated by a strong red vertical line) or be *driven* by it (indicated by a strong red horizontal line).

Such analyses will therefore be conducted in the remaining chapter of this thesis.

5.6 Conclusion

5.6.1 Summary of key findings

The key findings of this chapter can be summarised as follows.

- analysis of the cross-correlation between each pair of wards is a valuable, if computationally expensive way to study fine grained spatial movements in the housing property market. Derived cross-correlation data can go some way toward reconstructing the original market time series, thereby validating its meaning.

- property market cross-correlation appears to be a multi-level phenomenon, therefore further analysis of it should be conducted using multilevel modelling

- it is hard to deduce what constitutes a good indicator of cross-correlation, as only 12% of the variance in correlation has been explained by this study. However, inasmuch as it is possible to make predictions:

  - at ward level, the composition, rather than total size, of migration flows has a significant effect
  - the characteristics of each area are also important - social class, travel to work distance, dwelling occupancy, dwelling type, population age, housing stock and economic activity
  - the total size of migration and commuting flows, along with physical distances, are relatively unimportant in determining market linkage.

  The importance of origin and destination characteristics is concomitant with the explanation for the UK ripple effect presented in Meen (2001), namely that the seeming propagation of prices from one region to another is

in fact caused by locally differing responses to external economic conditions rather than direct interaction effects. This is also reflected in the conclusions of Giussani & Hadjimatheou (1991) in the UK, and mirrored by Shi et al. (2009) in New Zealand.

The importance of the composition of migration flows, on the other hand, would appear to contradict the above studies. However it should be noted that the current analysis is conducted on a finer spatial scale than any of the cited literature, therefore different effects may be expected to apply. In Cameron et al. (2005), a correlation between migration and price propagation is noted, although only on a very broad spatial scale.

- alternatively, it is possible to explain cross-correlation by simpler methods, either
  - as an effect of the existence of separate clusters of areas, which are internally more correlated but externally less correlated;
  - as an effect of the existence of certain areas which drive the market, and other areas which are more susceptible to their influence.

The final analysis chapter of this thesis will address alternative explanations of cross-correlation, through both clustering and a driving/driven wards analysis.

5.6.2 Limitations

One crucial limitation of the study is the time period over which data was available. As all findings relate to the years 2000-2006, they should be interpreted as a reflection of the state of the market during those years as opposed to some kind of universal invariant. The years studied represent the latter phase of an upswing of the market - thus, a logical extension of the analysis would be to study all phases of the property market cycle as and when appropriate data becomes available. It would also be worthwhile to study data spanning several cycles.

A second limitation relates to the use of PCA to reduce the explanatory variables to 40 dimensions. This is the same number chosen for the purely reactive regression modelling of chapter 4 - despite the increased complexity of interaction data; the reason for this relates to compute time and software limitations. In an ideal situation, it would be possible to perform the computation for every possible

number of dimensions between 1 and the number of explanatory variables (760); thereby precisely determining the point at which collinearity and overfitting occur in the model and thus preventing it without any loss of accuracy. Without such a process, it is impossible to verify a crucial assumption of PCA based regression: that a low-variance component currently excluded from analysis, does not display significant correlation with the target variable. Checking this assumption may be possible in future with increased compute times and better-optimised code.

5.6.3 Novelty

The work presented in this chapter constitutes the first combined study of house price data, census aggregate statistics and census interaction data at a fine spatial and temporal level (UK census wards with 200-day time slices).

It is also the first interaction study with explanatory power, in the sense of offering parameters which begin to predict interactions between wards.

The reconstruction of time series based on cross-correlation alone is believed to be novel. While of limited use in isolation - other than to validate the use of cross-correlation as a metric for further study - such reconstruction techniques may be of use as an incremental improvement to existing predictive regional price models.

Chapter 6

Alternative analyses of housing market interactions

6.1 Introduction

If this thesis is to be seen as a first ascent of a mountain of data, then the summit was reached in chapter 5. A significant quantity of computing power was used to perform a regression analysis which revealed some trends in the data. However, the summit was not reached in the dark, nor in a white-out: visualisations of the data were also produced. The pictures generated reveal two alternative routes up the mountain, both steeped in assumptions not present in the original ascent, but justified on the grounds that we now know that they also lead to the summit. It is also noted that the alternative routes require far less computing power.

These alternative routes - that is, hypotheses which can be tested by alternative analyses - are as follows:

- large, spatially contiguous blocks of wards exhibit similar price time series behaviour, therefore it is reasonable to model the system as being composed from a small number of 'clusters'
- certain wards display a consistent tendency to drive the market or be driven by it, therefore it is reasonable to model the system in terms of these wards.

This chapter therefore focuses firstly on identifying the clusters, driving and driven wards in the market, and secondly returning to regression to identify what,

over the six year period, causes a ward to belong to a particular cluster, to drive the market or be driven by it. Section 6.2 deals with the clustering analysis, while section 6.3 deals with driving/driven analysis. Section 6.4 concludes.

6.2 Cluster analysis of house prices

Like many of the techniques in this thesis, cluster analysis of data has been taking place for a long time. As even the simplest artificial neural networks are capable of performing clustering (Kohonen 1982), it is arguable that the process has been taking place since organisms with neurons first evolved during the Cambrian explosion (for which it would be appropriate to cite Earth's Oceans, 500 million years ago). Modern recognition of the technique, however, dates back to the 1950s (Steinhaus 1957), leading to the development of the now widely-used k-means algorithm (MacQueen 1967). This algorithm, given a number of data points and a distance metric between them, divides the data into a pre-specified number of clusters as follows:

1. create n new *centroid* points at random (these centroids will eventually represent the centres of the n clusters)

2. repeat the following steps:

   (a) assign each data point to the cluster represented by the closest centroid

   (b) update the position of each centroid to be the mean of the points in its cluster

   (c) compute *distortion* (the mean square distance of each data point from the centroid of its assigned cluster - which should reduce at each step),

   until no further reduction in distortion (greater than a certain iterative threshold) takes place.

K-means clustering has been used in numerous financial studies; however for two applications in spatial property markets see Goetzmann & Wachter (1995), which classifies the economies of US cities based on rental prices and vacancies, and Bourassa et al. (1999) which uses k-means clustering as part of a complex process for identifying housing submarkets in Melbourne.

A full, fine-grained cluster analysis of the UK housing market could doubtless constitute a whole thesis in its own right. Instead, then, a terse analysis is presented here: firstly to explore the hypothesis generated in chapter 5 that large spatially contiguous blocks of wards exist with similar market behaviour, and secondly to provide a brief alternative exploration of the data.

6.2.1 Methodology

6.2.1.1 Data

Clustering is undertaken on two alternative sets of time series data, thereby grouping together wards whose time series display similar behaviour. The data sets are:

1. relative ward price indices, as defined in chapter 2 - whereby an index is generated for each ward which starts at 1.0 in the year 2000
2. relative ward price indices, converted to a log series, normalised so that the maximum of each index is 1, and then re-converted to linear series. Hence each series starts with 1.0 and never exceeds 2.72 (e).

The second data set is included because clustering data set (1), as will be seen, divides wards mainly according to total growth over the years 2000-2006 - essentially the *height* of the time series graph. Removing this information allows for further division according to the *shape* of that graph.

6.2.1.2 Algorithm

Clustering is implemented using python's built-in k-means routines defined in the *numpy* library. A list of algorithm parameters is given in table 6.1. Clustering solutions are evaluated in terms of *distortion*, which in the case of UK wards is defined as

$$distortion = \sum_{ward \in wards} \sum_{t=0}^{t=T} (centroid_t^{ward} - data_t^{ward})^2 \qquad (6.1)$$

where T is the time series length, $data^{ward}$ is the time series for each $ward$ and $centroid^{ward}$ is the time series of the assigned centroid of the cluster for the same

number of clusters tried	$n = 1 - 30$
cluster solution chosen for each n	best of 20 runs
iterative threshold for each run	10^{-5}
length of time series	12 slices
time slice length	200 days

Table 6.1: Parameters for the k-means clustering algorithm. The iterative threshold is the criterion by which computation of a single clustering run is considered complete, if altering cluster centroids does not decrease distortion by more than this threshold. Each solution is derived by choosing the best results from 20 runs.

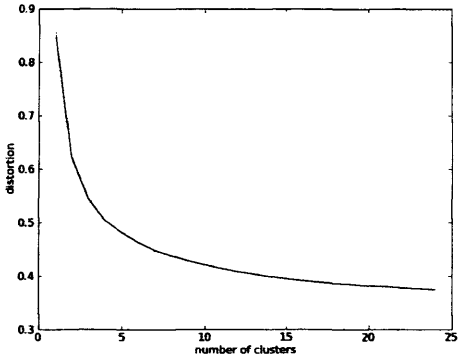
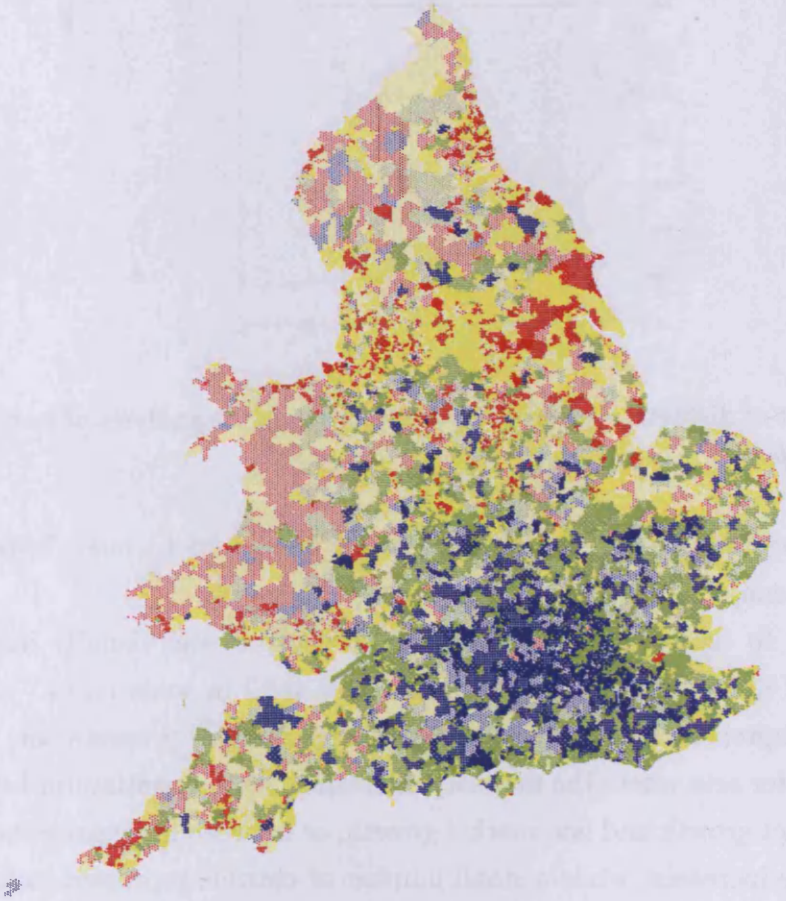


Figure 6.1: Plot of distortion against number of clusters for analysis of normalised time series.

ward. In other words, distortion is the mean square error caused by assuming that each ward behaves exactly like all other wards in the same cluster. In the case of a single cluster system, this single cluster will be the system-wide mean, and *distortion* therefore becomes the variance.

6.2.1.3 Choosing the number of clusters

Choosing the number of clusters is a similar process to that of choosing the output dimensionality of principal component analysis (as discussed in section 4.2.3.3). The criteria for choice are slightly different however: with PCA, the aim is to eliminate collinearity and maximise explanatory power of the regression, which led to a choice of 40 dimensions, however with clustering the requirement is to make a good trade-off between distortion and comprehensibility of results. In each case a graph of distortion versus number of clusters is plotted, and in all cases it will be seen that even modelling the system by as few as two clusters



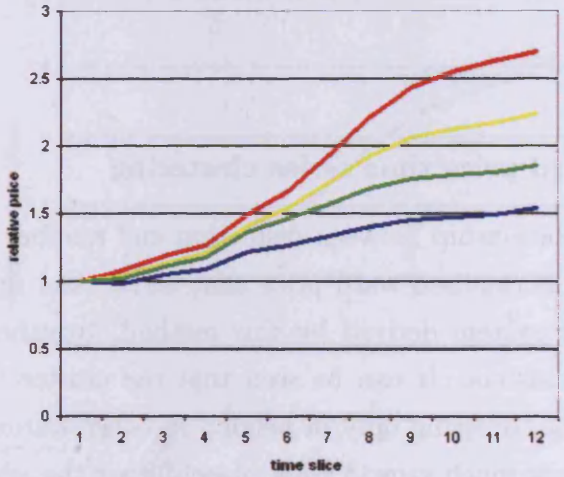


Figure 6.2: 4-cluster map of relative time series. The four colours represent time series as shown in the graph beneath. Saturation level represents residuals, the median residual being 0.44. Thus, bold colours indicate locations where the model fits well, and paler colours indicate a looser fit.

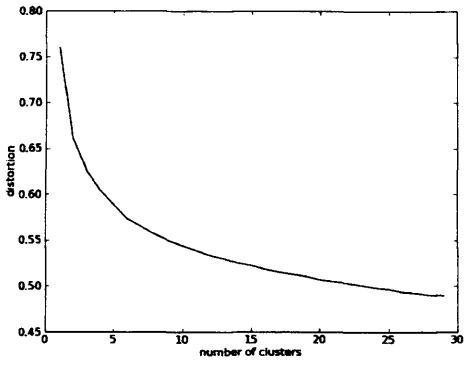


Figure 6.3: Plot of distortion against number of clusters for analysis of normalised time series *shapes*.

leads to a significant reduction in distortion as compared to one cluster (the system-wide mean).

In addition to the analysis of distortion, the output was visually inspected over a range of different cluster counts - as many as 12 in some cases - and the most easily comprehensible results were selected by hand for presentation. These tend to be cluster sets where the majority of clusters form a continuum between, say, early market growth and late market growth, or between large price increases and small price increases; while a small number of clusters represent exceptions to the continuum.

6.2.2 Results

6.2.2.1 Relative ward price time series clustering

Figure 6.1 shows the relationship between distortion and number of clusters for the initial clustering of normalised ward price time series, and figure 6.2 shows a map of a four-cluster system derived by this method, together with a time series plot of the four clusters. It can be seen that the cluster time series are almost identical in shape, differing only in height: in other words, this method has divided wards on how much growth took place during the years 2000-2006. A clear spatial pattern is visible, with the core (London and the central South) tending to exhibit the least growth while the periphery (Wales and the North) exhibits the most. The residuals are greater for some rural areas, as illustrated by paler colours.

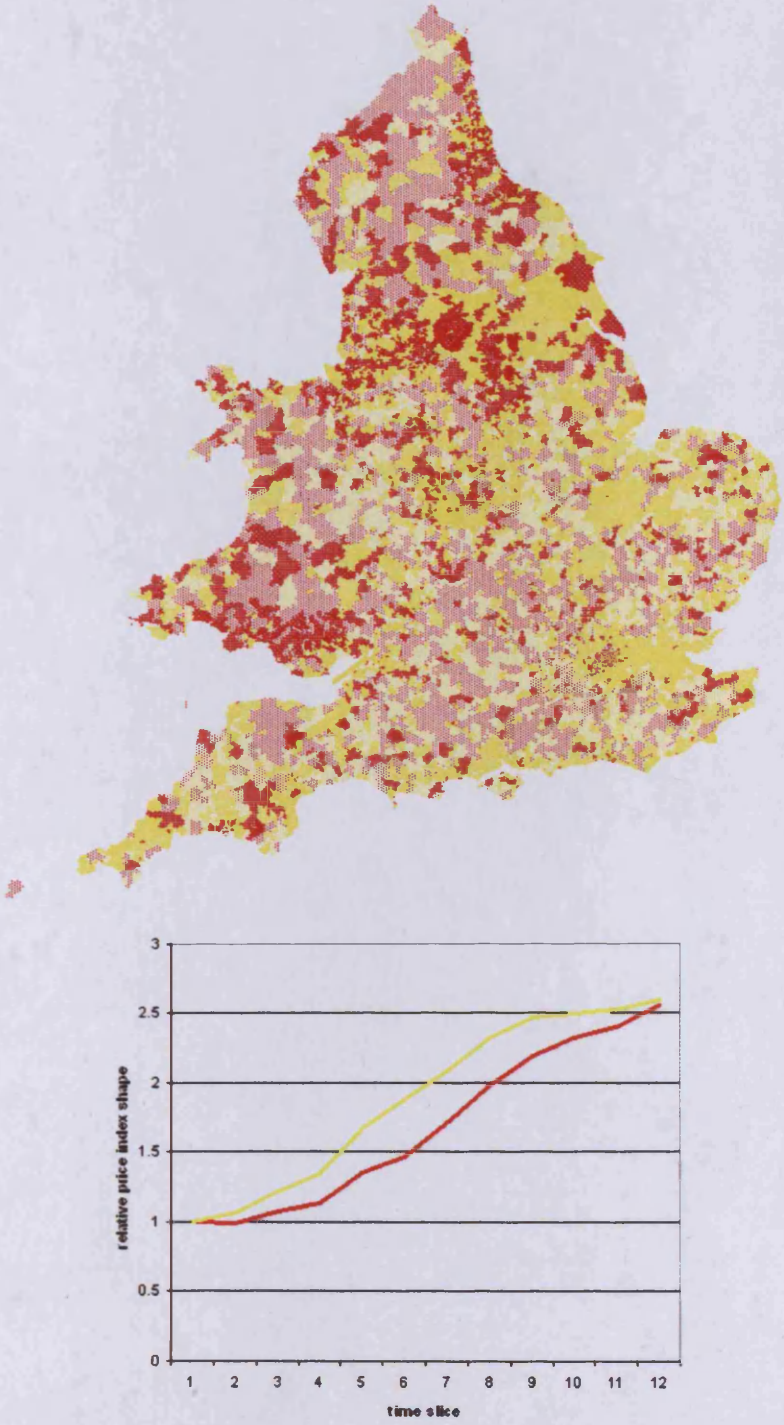


Figure 6.4: 2-cluster map of relative time series *shape*, i.e, time series are normalised to have identical minima and maxima. The two colours represent time series as shown in the graph beneath. Saturation level represents residuals, the mean residual being 0.56. Thus, bold colours indicate locations where the model fits well, and paler colours indicate a looser fit.

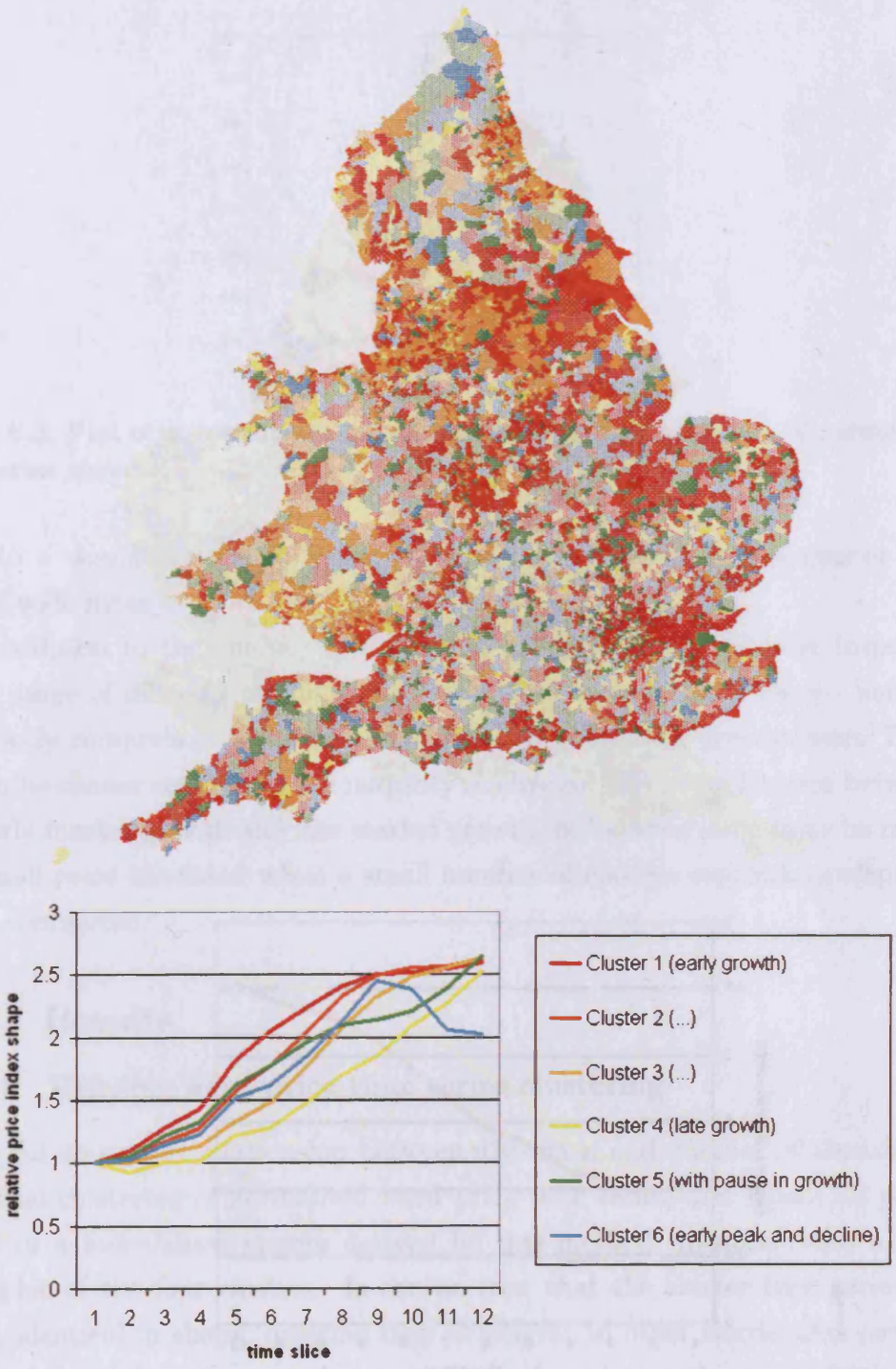


Figure 6.5: 6-cluster map of relative time series *shape*, i.e, time series are normalised to have identical minima and maxima. The six colours represent time series as shown in the graph beneath. Saturation level represents residuals, the mean residual being 0.49. Thus, bold colours indicate locations where the model fits well, and paler colours indicate a looser fit.

For a numerical analysis of this result, the reader should refer back to table 4.5 in chapter 4, in which overall growth from 2000-2006 was analysed. As may be suspected from examination of the map, spatial co-ordinates (namely the distance to London) turn out to have a large amount of explanatory power with respect to price growth. However, regional prices from the year 2001 are equally important in this respect, and a number of variables relating to socio-demographics of the population follow shortly after. The question of which class of variable constitutes the underlying *cause* - spatial or sociodemographic - is left to an economic geographer. A data driven study can only demonstrate explanatory power, not causality.

6.2.2.2 Time series shape clustering with two clusters

As initial clustering of time series divided them only on overall growth or height, and the factors affecting this were already explained in chapter 4, a second clustering analysis was undertaken with height information removed, allowing for division of time series on their shape.

Figure 6.3 shows the relationship between distortion and number of clusters for normalised ward price time series shapes. The simplest fruitful analysis is of a 2-cluster system, and is shown in figure 6.4. The distinction between wards revealed by the algorithm is in the timing of growth: in general it appears that southern and urban areas had a tendency to grow earlier than rural and northern/Welsh areas. Northern/Welsh urban and southern rural areas are divided in their characteristics, and as before, rural areas tend to exhibit greater residuals i.e. they do not fit the cluster pattern so well. Overall, the pattern is consistent with the hypothesis that the years 2000-2006 represented the tail end of a ripple growth phase, whereby most of the growth in the core had already taken place while the periphery had yet to follow suit.

Table 6.2 shows the results of a regression on cluster membership. The regression target is a dummy variable set to 1 for those places which increased in value earlier (the yellow cluster in fig. 6.4) or 0 for those which increased later (the red cluster). As can be seen from the map and regression, locations near to London had a tendency to lead the market, although this is better explained by many of the following characteristics: class (areas inhabited by people with intermediate occupations tending to exhibit earlier growth, while areas with high unem-

ployment were later); age (the 36-45 age band indicating a later growth); housing characteristics (detached and shared housing areas exhibiting earlier growth, while terraced and semi-detached areas follow) and housing turnover (more transactions implied earlier growth). While both ward- and LA-level characteristics are important, it seems that relative characteristics are slightly more strongly represented than absolutes, indicating that the composition of an area may be of slightly greater importance than its population count. It should not be forgotten, though, that population density itself is a key factor. Predictions made by the regression were good, with a 75% accuracy rate.[1]

6.2.2.3 Time series shape clustering with six clusters

For systems with 3-5 clusters, similar behaviour to the two-cluster analysis is displayed: the derived clusters divide wards on the differing timing of their growth. A 6-cluster analysis is the simplest result which exhibits qualitatively different behaviour, and is therefore presented in figure 6.5. Clusters 1-4 (shaded with colours red through yellow) provide the same timing information as the 2-cluster analysis, albeit with finer resolution. Cluster 5 (shown in green) exhibits initial growth matching cluster 2, but which later hits a plateau before growing again. Most interesting however is cluster 6 (shown in light blue) which reaches an early peak in 2005 and actually begins a decline.

It should be noted that most areas classified as belonging to cluster 6 exhibit large residuals, i.e. the fit is only approximate. The results of a regression on the membership of cluster 6 is provided in table 6.3. The regression target is a dummy variable set to 1 for areas within cluster six, or 0 otherwise. Examination of the (relatively few) positive coefficients in this regression suggests that such areas are rural, high priced and occupied by higher managerial professions with high income. Predictive power of this regression was poor, however, with a mean square error on the dummy variable of 82%.

6.2.3 Discussion

Overall, and perhaps unsurprisingly given its long history of use in such problems, clustering of time series has proven to provide a useful perspective on house price

[1] To avoid misrepresenting Bayesian statistics: False negatives rate is 9%, the false positive rate 16%, and actual membership of 1st cluster 62%.

Variable Name	Mean	Std. Dev	Reg. Coeff	99% Conf
UV031-N02-Intermediate	0.038	0.01	0.11	0.025
UV004-N04-36to45	0.031	0.0039	-0.07	0.017
UV056-A03-Detatched.log	2.6	0.37	0.061	0.016
UV035-LA-N06-NoFixedPlace	0.019	0.0043	0.061	0.016
UV004-LA-N04-36to45.log	0.031	0.002	-0.059	0.015
UV031-LA-N02-Intermediate	0.038	0.0066	0.056	0.019
UV053-LA-N04-Vacanthouseh	0.0067	0.0022	-0.054	0.024
ward numtransactions	7.4e+02	5.6e+02	0.05	0.018
UV002-A00-PopulationDensi	1	0.58	0.049	0.014
UV004-N02-16to25	0.024	0.0096	0.047	0.021
UV056-LA-N03-Detatched.lo	0.03	0.013	0.047	0.014
UV028-N01-Unemployed	0.013	0.0065	-0.046	0.013
UV053-N04-Vacanthousehold	0.0067	0.0042	-0.046	0.024
UV053-N03-Secondresidence	0.0022	0.0056	0.045	0.022
UV031-N04-LowerSupervisor	0.031	0.0088	0.044	0.017
UV056-N03-Detatched	0.032	0.022	0.044	0.013
UV004-N05-46to55	0.03	0.0052	-0.043	0.024
UV056-LA-A03-Detatched.lo	4.1	0.28	0.042	0.013
UV031-A02-Intermediate.lo	2.5	0.36	0.04	0.008
UV031-LA-N00-HigherManage	0.035	0.013	-0.038	0.011
UV031-N00-HigherManageria	0.036	0.019	-0.037	0.014
UV004-LA-N05-46to55.log	0.029	0.0026	-0.036	0.014
UV053-LA-A04-Vacanthouseh	3.2	0.27	-0.035	0.012
UV056-LA-A11-Inashareddwe	1.9	0.58	0.035	0.016
UV031-N06-Routine	0.038	0.016	-0.035	0.011
UV053-A04-Vacanthousehold	1.8	0.37	-0.034	0.014
UV004-N08-Over79	0.0096	0.0042	0.034	0.019
UV031-LA-N05-SemiRoutine.	0.048	0.0074	0.033	0.012
UV056-LA-N05-Terracedincl	0.025	0.0095	-0.032	0.018
UV056-LA-N04-Semidetatche	0.033	0.0081	-0.032	0.024
UV031-LA-N04-LowerSupervi	0.031	0.0055	0.032	0.016
UV056-N04-Semidetatched	0.033	0.014	-0.031	0.024
UV002-LA-A00-PopulationDe	2.5	0.35	0.031	0.014
UV004-LA-N01-Under16.log	0.041	0.0031	0.03	0.023
UV055-A02-Shared.log	0.24	0.42	0.03	0.025
UV028-N00-Employed	0.23	0.027	0.03	0.0097
UV004-LA-N08-Over79.log	0.0096	0.0023	0.03	0.016
UV031-N03-SmallEmployers	0.035	0.017	-0.028	0.017
UV031-A04-LowerSupervisor	2.4	0.33	0.027	0.0071
UV056-A11-Inashareddwelli	0.43	0.61	0.027	0.022
distance to London.log	5.1	0.43	-0.027	0.015

Table 6.2: Top determinants of earlier growth 2000-2006, from regression on 2-cluster time series shape. Variables for which the coefficient is less than its 99% confidence interval have been omitted.

Variable Name	Mean	Std. Dev	Reg. Coeff	99% Conf
UV031-N02-Intermediate	0.038	0.01	-0.04	0.027
UV031-N03-SmallEmployers	0.035	0.017	0.037	0.018
log 2001 prices	5	0.3	0.033	0.017
UV002-A00-PopulationDensi	1	0.58	-0.031	0.016
UV053-A03-Secondresidence	0.87	0.53	0.03	0.018
UV031-N04-LowerSupervisor	0.031	0.0088	-0.03	0.018
logweeklyincome	2.7	0.14	0.028	0.019
average la income	2.7	0.13	0.027	0.02
UV035-N04-20k30k	0.026	0.018	-0.027	0.026
UV035-A04-20k30k.log	2	0.36	-0.025	0.017
UV004-N08-Over79	0.0096	0.0042	-0.024	0.02
UV031-N05-SemiRoutine	0.048	0.013	-0.021	0.011
UV031-N00-HigherManageria	0.036	0.019	0.021	0.015
UV002-LA-A00-PopulationDe	2.5	0.35	-0.02	0.016
UV031-A04-LowerSupervisor	2.4	0.33	-0.018	0.0077
UV031-A02-Intermediate.lo	2.5	0.36	-0.017	0.0087
UV031-A03-SmallEmployers.	2.4	0.25	0.017	0.012
UV053-N02-Unoccupiedhouse	0.0089	0.0073	0.016	0.012
UV053-N00-ALLHOUSEHOLDSPA	0.19	0.003	-0.016	0.012
UV053-N01-Occupiedhouseho	0.18	0.0092	-0.015	0.012
UV031-A05-SemiRoutine.log	2.6	0.34	-0.015	0.005
UV004-A08-Over79.log	2.3	0.33	-0.014	0.0094
UV053-A02-Unoccupiedhouse	1.8	0.36	0.013	0.0099
UV031-A06-Routine.log	2.4	0.37	-0.011	0.0057
UV035-LA-A00-Under2k.log	4.2	0.19	0.01	0.0079
LA-log 2001 prices	5	0.22	0.0099	0.0091
UV004-A02-16to25.log	2.7	0.37	-0.0092	0.0079
UV004-A03-26to35.log	2.8	0.37	-0.0086	0.0048
UV031-A01-LowerManagerial	2.8	0.32	-0.0078	0.0043
UV053-A01-Occupiedhouseho	3.3	0.3	-0.0071	0.0028
UV004-A00-AllPeople.log	3.7	0.3	-0.0066	0.0029
UV004-A07-66to79.log	2.7	0.29	-0.0066	0.0064
UV056-A01-Inanunshareddwe	3.3	0.3	-0.0066	0.0029
UV055-A00-ALLDWELLINGS.lo	3.3	0.3	-0.0066	0.0029
UV055-A01-Unshared.log	3.3	0.3	-0.0066	0.0029
UV028-A00-Employed.log	3.3	0.3	-0.0066	0.0035
UV056-A00-ALLHOUSEHOLDS.l	3.3	0.3	-0.0064	0.0029
UV053-A00-ALLHOUSEHOLDSPA	3.3	0.3	-0.0064	0.003

Table 6.3: Top determinants of 6th cluster membership, i.e. an early peak followed by decline. Variables for which the coefficient is less than its 99% confidence interval have been omitted.

data. While the behaviour thus revealed is not as rich as that discovered through cross-correlation analysis, it is simpler to understand, and for all cases presented here, accounts for a large proportion (around 50%) of the variability in time series data. Moreover, as the work presented in chapter 5 has shown that the characteristics of individual areas have a much greater influence on house prices than interaction between areas, the inevitable loss of interaction information necessary to conduct such a simple clustering analysis can be considered justified.

In terms of direct results, it has been demonstrated that core areas of the UK exhibited earlier (though lesser) growth during the years 2000-2006, while peripheral areas exhibited later (and greater) growth. Certain rural areas had already peaked and commenced a decline by the start of 2006. Expressing results in terms of a core-periphery relationship begs the question of what constitutes core and periphery. While the simple spatial measure of distance-to-London was shown to have some explanatory power, socio-economic and demographic variables were shown to be of greater relevance in determining cluster membership.

6.3 Driving/driven analysis of house prices

The second analysis presented in this chapter is that of classifying ward interactions in terms of which wards tend to drive the market overall, and which wards are more susceptible to being influenced by movements in the wider market. This is based on the observation that the visualisation of cross-correlation in figure 5.18 exhibits a large quantity of continuous horizontal and vertical lines, indicating that such driving and driven places do exist. It would seem logical to derive a metric for such places in order to further study the phenomenon.

6.3.1 Methodology

Breaking down a full interaction matrix into the corresponding influence of sources and destinations has many characteristics in common with calibration of an unconstrained spatial interaction (SI) model (Fotheringham & O'Kelly 1989). Such a process involves assigning parameters such as attractiveness/repulsiveness coefficients (or in the current case, driving/driven scores) to each location such that the model explains the observed pattern of interaction as closely as possible.

It is, however, worth noting the differences between this correlation analysis and traditional SI modelling.

- SI models were developed to handle flows of *people*, in particular with applications of migration, commuting and shopping behaviour. It is thus one of their assumptions that people make a conscious choice concerning their behaviour. This is not true of property market correlations, which are the sum total of many factors rather than being under the control of a limited number of individuals. Different census wards do not consciously choose to correlate their price behaviour with others!

- SI models will typically include some kind of distance decay function, which (as was shown in chapter 5) is not particularly relevant in the case of house price interactions. Such a function is therefore excluded from the current analysis.

- It is also the case with many migration based interaction models, that a fixed sum of total migrations can be assumed - whereas in the current case of house price correlations it is certainly not true that total system correlation is limited, so such a constraint cannot be applied.

On the other hand, one relevant consideration of SI models is the concept of *entropy* and *maximum likelihood estimation* (ibid.) whereby the location coefficients are assigned in such a way as to be as robust as possible with respect to small changes in the observed data. It is thus hoped that the estimated model will still be relevant for making predictions as well as explaining current observation.

For a preliminary exploration, a simpler method than this is used - although the study of residuals from this method (to be presented in section 6.3.2) hints at potential future improvements. As horizontal and vertical lines are clearly visible in the plots of inter-ward correlations presented in section 5.5, the chosen method is to sum the columns of the ward interaction table to produce 'driving' scores for each ward, and to sum the rows to produce 'driven' scores. In section 6.3.3 this approach will be shown to produce results sufficiently robust for analysis, inasmuch as the resulting parameters exhibit strong correlation with census characteristics.

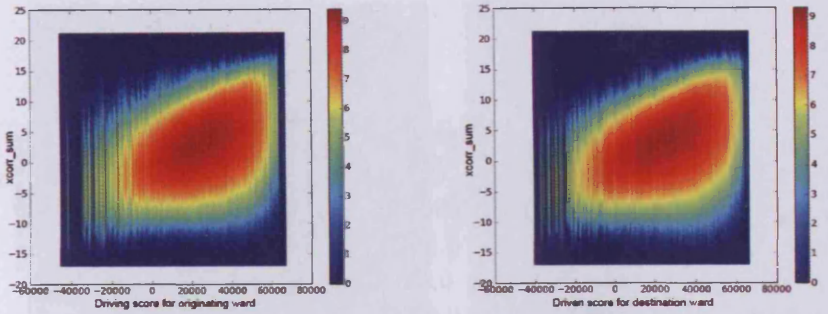


Figure 6.6: The relationship between the cross-correlation and driving/driven frameworks, shown as (i) scatter density plot of cross-correlation against driving score of origin; (ii) scatter density plot of cross-correlation against driven score of destination.

Thus, the 'driving' characteristic is defined as

$$driving_{ward} = \sum_{other\_ward \in wards} \text{corr\_sum}_{ward}^{other\_ward} \tag{6.2}$$

while the 'driven' characteristic is defined as

$$driven_{ward} = \sum_{other\_ward \in wards} \text{corr\_sum}_{other\_ward}^{ward} \tag{6.3}$$

where corr_sum is the sum of a ward cross-correlation function as defined in equation 5.3.

6.3.2 Results

In order to provide an indication of the extent to which the cross-correlation matrix can be explained by driving and driven scores, scatterplots of inter-ward correlation against the driving score for the origin, and the driven score for the destination, are given in figure 6.6. A clear correlation is visible in each case, however it is noted that as discussed in chapter 5, the cross-correlation data is very noisy so the observed correlation is weakened as a result. One advantage of summing the cross-correlation data to produce driving and driven metrics is, indeed, a reduction is this level of noise.

Figure 6.7 shows details from the cross-correlation matrix after driving and driven characteristics have been removed from the data. This could be considered

Variable Name	Mean	Std. Dev	Reg. Coeff	99% Conf
UV031-N02-Intermediate	0.038	0.01	0.14	0.022
ward numtransactions	7.4e+02	5.6e+02	0.085	0.016
UV053-N04-Vacanthousehold	0.0067	0.0042	-0.077	0.022
UV031-N00-HigherManageria	0.036	0.019	-0.063	0.013
UV053-N03-Secondresidence	0.0022	0.0057	0.062	0.02
UV031-N04-LowerSupervisor	0.031	0.0088	0.061	0.015
UV004-N05-46to55	0.03	0.0052	-0.06	0.022
UV031-LA-N02-Intermediate	0.038	0.0066	0.056	0.018
UV004-N04-36to45	0.031	0.0039	-0.051	0.016
UV031-A02-Intermediate.lo	2.5	0.36	0.051	0.0072

Table 6.4: Top ten predictors of 'driving' wards along with regression coefficients. Average residual for the whole regression was 0.61.

Variable Name	Mean	Std. Dev	Reg. Coeff	99% Conf
UV031-N02-Intermediate	0.038	0.01	0.093	0.023
ward numtransactions	7.4e+02	5.6e+02	0.068	0.016
UV053-N04-Vacanthousehold	0.0067	0.0042	-0.064	0.022
log 2001 prices	5	0.32	-0.06	0.015
UV031-N00-HigherManageria	0.036	0.019	-0.058	0.013
UV053-N03-Secondresidence	0.0022	0.0057	0.054	0.02
UV002-A00-PopulationDensi	1	0.58	0.044	0.013
UV031-N04-LowerSupervisor	0.031	0.0088	0.042	0.015
UV004-LA-N01-Under16.log	0.041	0.0031	0.039	0.021
UV031-A02-Intermediate.lo	2.5	0.36	0.038	0.0073

Table 6.5: Top ten predictors of 'driven' wards along with regression coefficients. Average residual for the whole regression was 0.62.

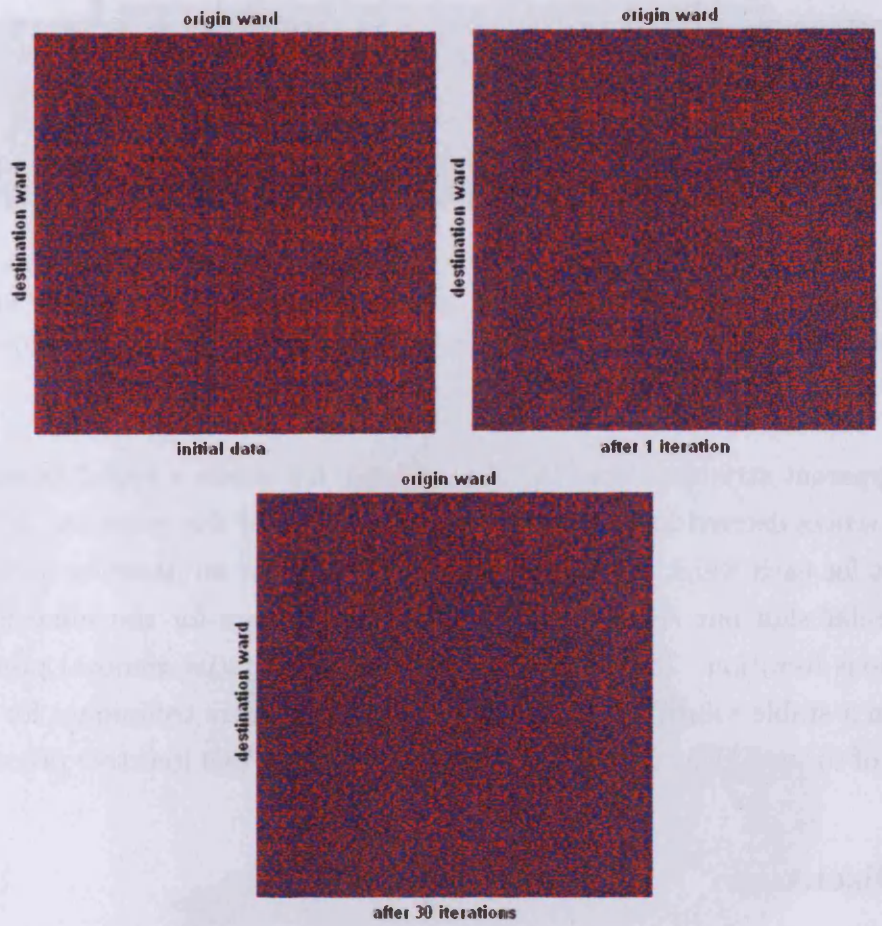


Figure 6.7: Detail from inter-ward cross-correlation residuals as driving/driven characteristics are iteratively removed. Note how apparent structure decreases with the number of iterations.

as a kind of residual analysis. Removal of driving and driven characteristics is achieved by subtracting the average of the origin's driving, and destination's driven scores for each cross-correlation, i.e.

$$residual_A^B = \text{corr\_sum}_A^B - \frac{driving_A + driven_B}{2} \tag{6.4}$$

Note that a driving/driven structure, characterised by horizontal and vertical lines, is still apparent after the initial removal of these characteristics. This is due to the fact that removal must necessarily be a compromise between eliminating both the driving and driven characteristics - the horizontal and vertical lines. Therefore, the process is iteratively repeated, and in this way succeeds in

Figure 6.8: Display of mean 'driving' scores through five successive iterations of driving/driven analysis. The tendency of each data point to be of the same sign though smaller magnitude than the same point for the preceding iteration indicates a stable convergence of the algorithm.

removing apparent structure from the data. Figure 6.8 shows a visualisation of the driving scores derived at each of the first five stages of the iteration. It can be seen that for each ward, the driving score computed for an iteration appears to be of similar sign but smaller magnitude than the score for the same ward in the previous iteration. This would suggest that the iterative removal process converges on a stable solution. It may be the case that future techniques for the assignment of driving/driven parameters can make use of this iterative process.

6.3.3 Discussion

Maps of the driving/driven scores of each ward are given in figures 6.9 and 6.10. Visual inspection shows that broadly similar spatial patterns are exhibited for both measures: urban areas tend to be both more strongly driving and driven, particularly in the North, while rural areas tend to score less strongly on both fronts. Analysis of the stability of driving/driven scores over time, in a similar manner to that used for cross-correlation characteristics in chapter 5, reveals that little correlation exists between driving or driven characteristics of the same place during different time spans. The usual disclaimer applies, therefore, that these metrics appear to represent the actual state of the market during the period studied, rather than revealing a universal invariant in market behaviour.

It is noteworthy, however, that approximately 40% of the variability in driving/driven scores is explainable by regression against census aggregate statistics. This compares favourably to a figure of only 10% for the proportion of cross-correlation variability explained by the census data. It might therefore be suggested that the driving/driven scores, being based on a larger quantity of

Figure 6.9: Market driving areas in the UK. Lighter shade indicates higher driving score.

Figure 6.10: Market driven areas in the UK. Lighter shade indicates higher driven score.

aggregated data, are less prone to noise than cross-correlation. Moreover, it is shown that despite the instability over time of driving/driven scores, these metrics do correlate to something *real* and this goes a long way towards validating them as a worthwhile object of study.

Results of the regression are given in tables 6.4 and 6.5. Overall, it can be seen that for both driving and driven characteristics, the presence of a high proportion of intermediate and lower supervisory occupations, and second residences, as well as a high market turnover correlate positively with high driving/driven scores. Negative correlations are noted for vacant households and higher managerial and professional occupations. Note the prevalence of ward-level relative variables, showing the importance of the local composition of an area in determining its level of driving/driven-ness.

6.4 Conclusions

In sum, this chapter has shown that the property market cross-correlation behaviour discovered in chapter 5 can be more simply, though quite efficiently explained (i) through a clustering framework and (ii) through a driving/driven wards framework. It is thus a useful result that the complex market behaviour initially defined in a 78-million-point interaction space, can mostly be approximated in an 8850-point geographical space. Furthermore, it is argued that these derived representations of market interaction are meaningful, because a large proportion of their variability can be explained by regression against census statistics. The fact that clusters and driving/driven scores correlate with real data should be taken as an indication of their 'reality'.

Unfortunately, driving/driven characteristics were shown to be unstable over time, inasmuch as scores derived from the years 2000-2002 did not correlate with scores derived in the years 2003-2006. There are three possible reasons why this might be the case:

1. the scores are always unstable over time;

2. the scores are stable, but noisy, therefore data covering a longer time span is required for their accurate estimation;

3. the scores are stable, but only for a small subset of wards studied.

A short pilot study investigating hypothesis (3) showed that while there are indeed some wards which exhibit stable driving/driven scores over the time period studied, there is no way of telling whether or not this is purely due to coincidence: no method was discovered for predicting which wards will be stable, other than directly measuring their tendency to change over time. The explanation then, is assumed to be either (1) or (2), and further investigation must be left to a future study as it is not possible to tell which of these is true without acquiring digitised data spanning a much larger time period.

6.4.1 Combined models as an avenue for future exploration

A final suggestion for future investigation is the question of how to combine the clustering and driving/driven models into a single, unified framework. It is also relevant to ask, after this framework has been applied, whether the residuals correlate to any interaction characteristics. The latter question could be investigated using the techniques of chapter 5, as it is possible that once the most significant determinants of cross-correlation have thus been removed, a faint pattern of inter-ward interaction may remain. However, as noted above, consideration of such a combined model should probably be postponed until data covering a greater time-span is available, as this study remains limited by the lack of such data and the ensuing inability to tell whether there are any timescales for which derived correlation statistics are in fact stable.

Chapter 7

Conclusions

7.1 Summary of contributions

The aim of this thesis was to conduct exploratory analysis of a new data set: a set comprising nearly 90% of all housing transactions recorded by the UK Land Registry between mid 2000 and mid 2006. This data was analysed in tandem with census data from the year 2001, including both aggregate and interaction statistics. The extensive coverage of these data sets, combined with their complexity (in the case of the interaction data) and fine grained spatial scale, meant that their analysis - conducted at ward level - entailed contributions both to our knowledge of the data, and to the techniques of exploratory analysis themselves.

The key contributions are:

- **An improved technique for forming price indices from transaction data**, in which relative indices are first created for output areas, which are later aggregated to the areal units desired (section 2.4). This has the advantage of incorporating hedonic information (from the definition of the output area) into the index, when such information may be otherwise unavailable.

- **Application of complete linkage clustering and linearisation algorithms, for the first time, to visualise census and Land Registry data** (chapter 3).

- **Incremental improvements to these techniques** during the course of their application to the data sets (which are larger than those for which they were designed). Alternative algorithms based on hierarchical optimal

clustering, and on simulated annealing, were investigated, and an interactive zooming tool was created. Comparison of the information revealed by the different algorithms revealed limitations on the size of the data sets to which they can usefully be applied (chapter 3).

- **A novel method for unification of data in the interaction domain, to allow comparison of price time series, interaction data and aggregate census statistics.** This method computes cross-correlations of time series and regresses them (i) against interaction data and (ii) against values of each aggregate statistic from the origin and destination area, their difference and square difference (section 5.3.1). Principal component analysis is used to combat the inevitable collinearity that results from this technique. Such unified analysis of data, combined with visualisation, has led to:

  ⇒ **potential improvements to time series prediction**, as regional cross-correlation based predictions are shown to significantly outperform those based on auto-correlation or a market average (section 5.3.2.1);

  ⇒ **confirmation of an existing hypotheses** that spatial patterns of diffusion in housing markets, and particularly the UK housing market, are often caused by reactive rather than interactive processes (section 5.4.3.2);

  ⇒ **an additional finding** that the composition of migration flows is an important indicator of market cross-correlation on a fine spatial scale (section 5.4.3.2);

  ⇒ **provision of a novel justification for the use of clustering approaches**, as visualisations of the cross-correlation matrix suggest a clustered structure (section 5.5) which is also explicitly confirmed, and each cluster linked to explanatory variables (section 6.2);

  ⇒ **development of new theories concerning market-driving and market-driven areas**, as visualisations of the cross-correlation matrix also suggest a driving/driven structure (section 5.5);

  ⇒ **detection of such areas, and some suggestions as to their causes** by row- and columnwise summation of the cross-correlation

matrix to produce driving/driven metrics, which are regressed against census aggregate statistics (section 6.3);

⇒ **overall increased understanding of the inherently noisy nature of housing market cross-correlations** - to be summarised in section 7.2.3.

There is an apparent incongruency in the cross-correlation analysis which is worth pointing out at this stage. Time series reconstruction (as conducted in section 5.2) was only possible if the global trend of mean market value was first removed from the data. Conversely, removal of this trend *decreased* the strength of results from the regression analysis of cross-correlation, and these results are therefore presented with the trend left in place. Thus, time series reconstruction using cross-correlation data with the trend removed, is used to justify analysis of cross-correlation data with the trend present.

This inconsistency is not considered a problem, as other grounds exist for justifying the cross-correlation regression: namely, that the results themselves show a clear link between census and Land Registry data, and also lead to a fruitful market-driving/market-driven analysis. The necessity of trend removal before time series reconstruction is probably best seen as a limitation only of the reconstruction technique, relating in no small part to its consideration of reactive mechanisms as exogenous to the system. As such, the reconstruction technique - despite its limitations - still supports the hypothesis that cross-correlation data (whether or not global trends have been removed) is a meaningful target for analysis.

The remainder of this chapter will discuss the limitations of these findings (section 7.2) and their implications for the wider field of research in which they are situated (section 7.3). Section 7.4 concludes.

7.2 Limitations and future work

It is often the case that the limitations of existing research are a useful source of inspiration for proposals relating to the extension of that research. Therefore, this section represents a combined discussion of both limitations and possible continuation projects.

There are two exceptions - in the present case - to this duality of limitation and opportunity. One is the endless potential for inventing new visualisation techniques, and combining them with existing ones. Animations, 3-d displays, further interactive features and indeed methods from any of the literature reviewed in chapter 3 could potentially be combined with the techniques used here to produce ever more sophisticated visualisation tools.

The other extension not directly related to a limitation of the study, is the obvious idea of applying similar methods to other data sets, of which many exist not only in housing (the housing markets of other countries, for example); but also in the wider social sciences, and indeed beyond that, in fields such as biology and informatics. The sharing of techniques between disciplines can hardly fail to benefit both sides.

The remainder of this section discusses issues that are both limitations of this study, and possible ideas for future investigation.

7.2.1 Choice of target variable

It is notable that the analysis was conducted with all target variables derived from Land Registry data, and most explanatory variables derived from census data. This is an acceptable approach when the census data is already well-understood, and it is the Land Registry information which requires more detailed investigation. However, if applying these techniques to other data sets, it may be more appropriate either to directly analyse the output of PCA, or to repeat the regression with every variable, in turn, as the target. Additional techniques such as spring networks (Ebbels et al. 2006, Pawlak 2005) may be necessary to visualise the output of such an analysis, which would provide a very general overview of the interrelationship of all variables in a complex data set.

7.2.2 Left out variables

"Garbage in, garbage out" has become a universal mantra relating to computer systems, however in the case of regression, failing to input the relevant can often be a more serious problem than including the irrelevant. Undoubtedly there must be measurable quantities which have an impact on housing markets, but which have not been included in this study, which for the most part has settled on

census statistics.

Future work could always, therefore, aim to include more sources of information - the system has been designed to be as general as possible, thus it should not be too hard to include additional data sets. However, there is one in particular which springs to mind. A large body of literature exists geared towards computing abstract qualities of interaction networks which often correlate with real world statistics (the reader is referred back to section 1.2.3 for a brief review). As network data - in terms of distance, migration and road linkage is readily available - it would make sense to reduce this to such metrics (for example, polycentricity and spatial locality of settlement interactions) and include the resulting data in the analyses.

7.2.3 Inherent noisiness of ward level cross-correlations

A key inference drawn from this study is that the ward level cross-correlations computed were inherently very noisy. This cannot be conclusively demonstrated, as to do so would require a measure of the underlying 'signal' of ward interactions which cross-correlation imperfectly attempts to measure. However, the following evidence points towards this conclusion:

1. the complete lack of consistency between identical ward pair correlations derived from different timespans (section 5.3.2.2),

2. the inability of the regression analysis to explain more than 10% of the variability in cross-correlation (section 5.4.3.1),

3. the fact that the quantity of data is expanded by a factor of around 400 when computing cross-correlations from time series, when no new information has been generated (section 5.3.1.4),

4. the relative success of time series predictions based on cross-correlation when compared to predictions based on autocorrelation alone. If cross-correlations are noisy then we would expect this to be the case, as the former type of prediction averages over more data points thereby reducing noise to some extent (section 5.2),

5. the fact that repeat-sale house price predictions based on local authority information outperform those based on ward information, which may indicate that ward level price indices are themselves noisy (section 2.4.6.2).

This limitation of noise may well apply not only to the study of housing market cross-correlations, but also to any methodology that involves expansion of data in a similar manner. In this case, due to the presence of noise, it has been impossible to say to what extent cross-correlations naturally change over time (and are thus accurate for the time period and time scales studied) as compared to the extent that they are imperfectly measured. It would appear that the latter is strongly true, though the possibility of change over time remains to be investigated.

For the analysis of Land Registry data, four avenues are suggested for mitigation of noise-related problems in future research.

- One is to aggregate the data to larger areas in order to reduce noise. An obvious candidate is the Local Authority level; also, such analysis will help to answer the question of whether migration plays a significant part in market correlation except on a fine spatial scale. The result may well be a much more accurate predictive regression; however there is a caveat that despite the noise in the data, some variables were still found to have a meaningful effect on cross-correlation at ward level (section 5.4.3.2) and moreover, almost all of the top characteristics of driving/driven areas are determined at ward level (section 6.3.3). Therefore is is probable that information will be lost, as well as gained, by such a change in spatial scale.

- An alternative possibility is to work on reducing the noise in cross-correlation data. It may be the case that, as the indexing techniques of chapter 2 allow for meaningful predictions to be made on a shorter time scale, it is better to employ price indices with shorter time slices in order to provide more data points with which to compute meaningful correlations. This is in contrast to the work presented in this thesis, which used time slice lengths of around 100-200 days based on an 'optimum' length estimated for predictive accuracy in section 2.4.6.2. It is nonetheless conceivable that a small decrease in predictive accuracy, combined with a large increase in temporal precision, could allow more meaningful analysis of the relative shapes of different time series. Thus, in the terms of section 5.3.1.1,

cross-correlation could be computed with [time_slice_length = 20, wmax = 10] instead of [time_slice_length = 100, wmax = 2] - even though both computations give an overall correlation window length of 200 days.

- Seemingly in contradiction to the suggestion of aggregating data into larger areas, it is noted that in many cases, output areas will not accurately capture the housing attributes of individual properties. This ecological fallacy applies especially in rural areas. Therefore, a heterogenous analysis based on different submarkets, or perhaps even incorporating individual property data, may also reduce what currently appears to be noise in the data, but is in reality caused by individual variation.

- The final way to reduce noise in the cross-correlation data is simply to obtain more house price data. This will be discussed in section 7.2.4.

7.2.4 Time span of the data

An ever-present concern throughout this research has been the limited time span of the data, from the years 2000-2006. This means that despite studying one of the largest sets of housing transactions to date, the scope of the study is necessarily limited to the nature of housing interactions during the tail end of a nationwide increase in the overall market that was occurring during these years. It also means that any correlations, or good predictive relationships discovered in the data *do not reflect a universal characteristic of the UK property market* but are better seen as a measure of its state during those years. As well as mitigating these concerns, the acquisition of more data to analyse may assist in reducing the noise of cross-correlations as discussed in the previous section (7.2.3).

Since commencement of the study, the average house price has dropped by as much as 30% in some cases, to pre-2001 levels. Inclusion of even these past three years of data would greatly broaden the scope, as it would allow for the study of both a rising and falling market. Extending the time span of the data in the other direction is not so easy, as records before the year 2000 were not digitised at time of creation. However, it seems that an ongoing drive exists to put some of this data into electronic form. As was seen in section 7.2.3, it is perhaps the case that aggregating data to local authority (rather than ward) level would be a more fruitful avenue of study. Therefore, increasing availability

of historic housing data, even if aggregated to coarser units of time and space than used here, may yet contribute to our understanding of housing markets.

7.2.5 Study of residuals

A standard component of regression analysis is to study not only the estimated coefficients, but also to study the regression residuals for any kind of pattern missed by the regression. While this has been undertaken to a limited extent (such as checking for heteroscedasticity of errors in the plots of residual versus prediction, figure 5.9), further analysis has been limited by the sheer amount of noise in the data as discussed in section 7.2.3. In the absence of this problem, study of residuals would likely provide hints as to how to improve existing models, and indeed it is possible to envisage an incremental development cycle whereby the residuals for each model in turn are used to inspire its successor. This possibility will be discussed in more detail in section 7.3.

7.3 Broader implications: on scientific method in the era of data-driven science

The work presented in this thesis is situated within the field of *exploratory data analysis* - a collection of techniques used to familiarise oneself with a data set before proceeding to *confirmatory data analysis*, in which hypotheses are formally tested. The complex visualisations employed are simply an extension of such techniques to ever-larger and more complex data sets. Visualisation has shown itself to be of use not only for exploratory analysis, but also for debugging software, and in this age of the world wide web, for quick dissemination through online tools. Visualisation based research workflows are now commonplace in a wide range of fields - for example climate science (Kehrer et al. 2008), intelligence (Thomas & Cook 2006) and biodiversity (Kelling et al. 2009).

It should be noted, however, that the regression coefficients presented are also in the spirit of exploratory analysis, effectively used to 'visualise' facets of the data which are not so easily represented by a graphic. Such analyses should not be confused with the confirmation of any hypothesis - which is the more traditional use for regression! The only exception to this is where hypotheses have previously

been proposed by other researchers: for example, the assertion of Meen (2001) that differing reaction to external influences is responsible for apparent interactive price propagation in the UK, which chapter 5 serves to test. In formal terms, use of statistical hypothesis testing requires that the hypothesis be decided on before the test is conducted, else the data is not really an independent test and does not tell us anything about the truth of the hypothesis.

It is notable that this consensus on the requirements of scientific method was developed against a background of research studies in the physical and biological sciences, where the data collected represented only a very small sample of the underlying population. However, it may in fact be possible to waive this requirement depending on the scope of the study. For example, if one is interested only in the England and Wales housing freehold market from 2000-2006, the data set employed here is almost entirely complete. Two possible interpretations are therefore suggested for this work:

1. If the reader is interested only in the area studied during the period of the study, then the estimated coefficients can be taken as entirely representative, and any hypothesis supported by these results can be taken as confirmed - even if such hypotheses were generated by reading the results in the first place.[1]

2. If the reader is interested in the behaviour of housing markets in multiple countries, or in multiple time periods including the past or future, then the estimated coefficients can only be considered a limited sample of the underlying population. Any hypothesis supported by the results is therefore only a suggestion which is subject to further testing on other data before confirmation can occur.

A related feature of all the regression analyses conducted in this thesis is that the estimated parameters were tested on a subset of the data not used in the estimation process. Again, this should not be confused with the confirmation of any hypothesis, because the test data set is sampled within the same time and space constraints as the training set (e.g. England and Wales, 2000-2006)

[1]It would be wise in so doing, however, to be mindful of the limitations of principal component analysis, and if greater certainty is required it would be prudent to conduct a further regression of one's own, dispensing with PCA and using only the variables of interest.

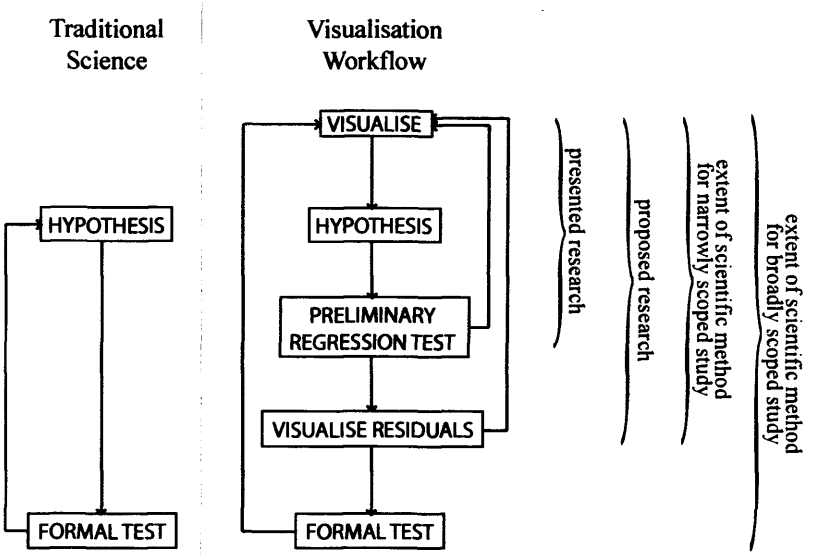


Figure 7.1: Illustration of the visualisation-based workflow.

and therefore cannot be expected to be representative of any data outside of this scope. However, it is nonetheless hoped that such a process of testing increases the likelihood of any generated hypothesis surviving the test process on other data sets in the future.

An illustration of the overall workflow is given in figure 7.1. The new workflow (and the proposed extension of studying residuals mentioned in the previous section 7.2.5) can be considered as fitting in with traditional scientific method; however, exactly which part of the scientific cycle has been conducted depends, as noted above, on the stated scope of the study.

Whereas traditional method consists a repeated iteration of hypothesis and test, the new workflow starts with visualisation and hypothesis generation, followed by a regression test of the hypothesis, *which is performed on the same data used to create that hypothesis.* At this stage, it is possible to reject the hypothesis altogether; however if successful, it is possible to study the residuals and further refine the hypothesis. Overall, two 'inner loops' have been added to the scientific process before the traditional hypothesis test is reached. These inner loops have rapid execution times, inasmuch as they allow for the quick development of complex hypotheses that have a fair chance of standing up to wider confirmation; this is the primary contribution of the visualisation based workflow. It should be remembered however, that if we are interested not in the particulars of the data

but in developing a broad scientific understanding of the world around us, the workflow presented stops short of the final confirmatory test.

7.4 Closing remarks - or, the return to base camp

And there it is: the mountain of data was climbed, the summit reached, the clouds parted and the view was recorded. It is often said that the most dangerous part of a mountaineering trip is the descent, and perhaps that can be likened to the conclusions of a research project. Just as we must be careful, once the summit has been reached, not to slip on the way down, in science we must also be careful not to slip when interpreting the wider meaning of our work.

Much as expected, a little knowledge was gained of the mountain; much more was gained, however, in terms of mountaineering expertise itself. In the 1950s, the ascent of Everest was considered justified purely in terms of the existence of the challenge itself, but it should be noted that such acceptance of the merits of mountaineering has not always been widespread in society at large. When Michel-Gabriel Paccard and Jacques Balmat first summited Mont Blanc in 1786, for example, the trip was justified to the wider public on the grounds of a barometer reading taken from the summit. In this day and age however, ascents of difficult peaks are accepted as worthwhile in their own right; likewise, the prevailing scientific opinion is (and always was) that methodological contributions are a valid and essential part of progress in research.

Bibliography

Andersson, C., Frenken, K. & Hellervik, A. (2006), 'A complex network approach to urban growth', *Environment and Planning A* **38**(10), 1941–1964.

Andrienko, G., Andrienko, N. & Wrobel, S. (2007), 'Visual analytics tools for analysis of movement data', *SIGKDD Explor. Newsl.* **9**(2), 38–46.

Anselin, A. (1988), *Spatial Econometrics: Methods and Models*, Kluwer, Dordrecht, Netherlands.

Axelrod, R. (2006), A guide for newcomers to agent-based modeling in the social sciences, *in* L. Tesfatsion & K. L. Judd, eds, 'Handbook of computational economics, volume 2: Agent-based computational economics', North-Holland, Amsterdam, p. 1647.

Axtell, R., Epstein, J. & Young, H. (2000), The emergence of classes in a multi-agent bargaining model, Papers 9, Brookings Institution - Working Papers. available at http://ideas.repec.org/p/fth/brooki/9.html.

Bar-Joseph, Z., Demaine, E. D., Gifford, D. K., Srebro, N., Hamel, A. M. & Jaakkola, T. S. (2003), 'K-ary clustering with optimal leaf ordering for gene expression data', *Bioinformatics* **19**(9), 1070–1078.

Barrat, A., Barthelemy, M., Pastor-Satorras, R. & Vespignani, A. (2004), 'The architecture of complex weighted networks', *Proceedings of The National Academy of Sciences* **101**(11), 3747–3752.

Batty, M. (2001), 'Editorial: Agent-based pedestrian modelling', *Environment and Planning B: Planning and Design* **28**(3), 321–326.

Batty, M. (2006), 'Rank clocks', *Nature* **444**, 592.

Berger, T., Schreinemachers, P. & Woelcke, J. (2006), 'Multi-agent simulation for the targeting of development policies in less-favored areas', *Agricultural Systems* **88**(1), 28–43.

Bertin, J. (1984), *Semiology of Graphics*, University of Wisconsin Press.

Blok, C. (2000), Monitoring change: Characteristics of dynamic geo-spatial phenomena for visual exploration, *in* 'Spatial Cognition II (LNAI 1849)', Springer-Verlag, Berlin, p. 16.

Bourassa, S. C., Hamelink, F., Hoesli, M. & MacGregor, B. D. (1999), 'Defining housing submarkets', *Journal of Housing Economics* **8**(2), 160.

Boyle, P. (1993), 'Modelling the relationship between tenure and migration in England and Wales', *Transactions of the Institute of British Geographers, New Series* **18**(3), 359–376.

Burrough, P. A. (1986), *Principles of Geographical information Systems for Land Resources Assessment*, Oxford University Press.

Cai, Y., Stumpf, R., Wynne, T., Tomlinson, M., Chung, D., Boutonnier, X., Ihmig, M., Franco, R. & Bauernfeind, N. (2007), 'Visual transformation for interactive spatiotemporal data mining', *Knowledge and Information Systems* **13**(2), 119.

Cameron, G., Muellbauer, J. & Murphy, A. (2005), 'Migration within England and Wales and the housing market', *Economic Outlook* **29**(3), 9.

Claramunt, C., Jiang, B. & Bargiela, A. (2000), 'A new framework for the integration, analysis and visualisation of urban traffic data within geographic information systems', *Transportation Research C* **8**(1), 167.

Cliff, A. D. & Ord, J. K. (1981), *Spatial Processes, Models and Applications*, Pion, London.

Cliff, D. (2003), 'Explorations in evolutionary design of online auction market mechanisms', *Journal of Electronic Commerce Research and Applciations* **2**(2), 162–175.

Cliff, D. & Miller, G. (2006), 'Visualizing coevolution with CIAO plots', *Artificial Life* **12**(2), 1–4.

Congdon, P. (2006), 'A model for geographical variation in health and total life expectancy', *Demographic Research* **14**, 157–178.

Cowan, R., Jonard, N. & Zimmermann, J.-B. (2006), 'Evolving networks of inventors', *Journal of Evolutionary Economics* **16**(1), 166–174.

Cowell, A. J. (2005), Scientific discovery within data streams, *in* 'LNAI3345: Ambient Intelligence for Scientific Discovery', Springer, Berlin, p. 66.

Cui, W., Zhou, H., Qu, H., Wong, P. & Li, X. (2008), 'Geometry-based edge clustering for graph visualization', *Transactions on Visualization and Computer Graphics* **14**(6), 1227.

Day, R. H. (1993), Nonlinear dynamics and evolutionary economics, *in* R. H. Day & P. Chen, eds, 'Nonlinear dynamics and evolutionary economics', Oxford University Press.

De Montis, A., Barthelemy, M., Chessa, A. & Vespignani, A. (2007), 'The structure of inter-urban traffic: A weighted network analysis', *Environment and Planning B* **34**(5), 905.

De Vicente, J., Lanchares, J. & Hermida, R. (2003), 'Placement by thermodynamic simulated annealing', *Physics Letters A* **317**(5-6), 415.

Dennett, A. & Stillwell, J. (2008), Internal migration in Great Britain - a district level analysis using 2001 census data, Technical report, School of Geography, University of Leeds.

Devaney, J. E. (2005), Science at the speed of thought, *in* 'LNAI3345: Ambient Intelligence for Scientific Discovery', Springer, Berlin.

Development of a migration model (2002), Technical report, Office of the Deputy Prime Minister, London.

Drummond, B. & Cauty, J. (1988), *The Manual: How to Have a Number 1 the Easy Way*, Ellipsis, London.

Eaton, B. C. & Lipsey, R. G. (1975), 'The principle of minimum differentiation reconsidered: Some new developments in the theory of spatial competition', *The Review of Economic Studies* **42**(1), 27–49.

Ebbels, T. M. D., Buxton, B. F. & Jones, D. T. (2006), 'Springscape: visualisation of microarray and contextual bioinformatic data using spring embedding and an 'information landscape'', *Bioinformatics* **22**(14), e99.

Eick, S. G. (1996), 'Aspects of network visualization', *IEEE Computer Graphics and Applications* **16**(2), 69–72.

Epstein, J. M. & Axtell, R. (1996), *Growing artificial societies: social science from the bottom up*, The Brookings Institution, Washington, DC, USA.

Falkman, G. (2001), 'Information visualisation in clinical odontology: multidimensional analysis and interactive data exploration', *Artificial Intelligence in Medicine* **22**(2), 133.

Faust, K., Entwisle, B., Rindfuss, R. R., Walsh, S. J. & Sawangdee, Y. (2000), 'Spatial arrangement of social and economic networks among villages in Nang Rong district, Thailand', *Social Networks* **21**(4), 311.

Fotheringham, A. S. & O'Kelly, M. E. (1989), *Spatial Interaction Models: formulations and applications*, Kluwer, London.

Gabszewics, J. J. & Thisse, J.-F. (1992), Location, *in* R. J. Aumann & S. Hart, eds, 'Hand Book of Game Theory with Economic Applications, Vol. I', North Holland, Amsterdam, chapter 9.

Geovista project website (n.d.).
URL: *http://www.geovistastudio.psu.edu/jsp/publications.jsp*

Giussani, B. & Hadjimatheou, G. (1991), 'Modeling regional house prices in the United Kingdom', *Papers in Regional Science* **70**(2), 201.

Goetzmann, W. N. & Wachter, S. M. (1995), 'Clustering methods for real estate portfolios', *Real Estate Economics* **23**, 271.

Guerois, M. & Le Goix, R. (2009), 'La dynamique spatio-temporelle des prix immobiliers a differentes echelles : le cas des appartements anciens a Paris (1990-2003)', *Cybergeo: Systemes, Modelisation, Geostatistiques* **470**.

Guo, D. (2007), 'Visual analytics of spatial interaction patterns for pandemic decision support', *International Journal of Geographical Information Science* **21**(8), 859.

Guo, D. & Gahegan, M. (2006), 'Spatial ordering and encoding for geographic data mining and visualisation', *Journal of Intelligent Information Systems* **27**(3), 243–266.

Hagerstrand, T. (1975), Space, time and human conditions., *in* A. e. a. Karlqvist, ed., 'Dynamic allocation of urban space', Saxon House Lexington.

Hall, P. (2001), 'Christaller for a global age: Redrawing the urban hierarchy', *Stadt und Region: Dynamik von Lebenswelten, Tagungsbericht und wissenschaftliche Abhandlungen* **53**, 110.
URL: *http://www.lboro.ac.uk/gawc/rb/rb59.html*

Han, J. & Kamber, M. (2006), *Data Mining: concepts and techniques*, Morgan Kaufmann, San Francisco.

Heikkila, E. J. & Wang, Y. (2006), 'Fujita and Ogawa revisited: An agent-based modeling approach', *Submitted to Regional Science and Urban Economics* .

Holmes, M. & Grimes, A. (2005), Is there long-run convergence of regional house prices in the uk?, Technical report, Motu Economic and Public Policy Research, Wellington, New Zealand. working paper 05-11.

Hotelling, H. (1929), 'Stability in competition', *Economic Journal* **39**(153), 41–57.

Irmen, A. & Thisse, J.-F. (1998), 'Competition in multi-characteristics spaces: Hotelling was almost right', *Journal of economic theory* **78**(1), 76–102.

Isaaks, E. & Srivastava, M. (1989), *Applied Geostatics*, Oxford University Press, New York.

Jiang, B. (2004), 'Topological analysis of urban street networks', *Environment and Planning B: Planning and Design* **31**(1), 151–162.

Kehrer, J., Ladstadter, F., Muigg, P., Doleisch, H., Steiner, A. & Hauser, H. (2008), 'Hypothesis generation in climate research with interactive visual data exploration', *IEEE Transactions on Visualization and Computer Graphics* **14**(6), 1579–1586.

Keim, D. A. (1996), 'Pixel-oriented database visualizations', *SIGMOD record* **25**(4), 35–39.

Keim, D., Hao, M. C., Ladisch, J., Hsu, M. & Dayal, U. (2001), 'Pixel bar charts: a new technique for visualizing large multi-attribute data sets without aggregation', *Information Visualization* **1**.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G. & Hooker, G. (2009), 'Data-intensive science: A new paradigm for biodiversity studies', *BioScience* **59**(7), 613–620.

Kohonen, T. (1982), 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics* **43**(1), 59.

Kollman, K. & Page, S. E. (2006), Computational methods and models of politics, *in* L. Tesfatsion & K. L. Judd, eds, 'Handbook of computational economics, volume 2: Agent-based computational economics', North-Holland, Amsterdam, p. 1433.

Kraak, M. (2001), 'Visualize Overijssel's past, interactive animations on the WWW', *Kartografisch Tijdschrift* **27**(4), 5.

Kreps, D. M. (1990), *Game theory and economic modelling*, Clarendon Press, Oxford.

Kwan, M.-P. (2000), 'Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set', *Transportation Research Part C* **8**, 185.

Lai, S.-K. (2006), 'A spatial garbage-can model', *Environment and Planning B: Planning and Design* **33**(1), 141–156.

Lake, M. (2001), 'The use of pedestrian modelling in archaeology, with an example from the study of cultural learning', *Environment and Planning B: Planning and Design* **28**(3), 361–383.

Langran, G. (1989), 'A review of temporal database research and its use in GIS applications', *International Journal of Geographical Information Systems* **3**(3), 215–232.

Latora, V. & Marchiori, M. (2003), 'Economic small-world behaviour in weighted networks', *The European Physical Journal B* **32**(2), 249–263.

Lebaron, B. (2006), Agent-based computational finance, *in* L. Tesfatsion & K. L. Judd, eds, 'Handbook of computational economics, volume 2: Agent-based computational economics', North-Holland, Amsterdam, p. 1187.

LeGates, R. (2005), *Think Globally, Act Regionally*, ESRI Press, Redlands, California.

Li, X. (2005), *Advanced GIS Modelling Techniques and Applications*, China Education and Culture Publishing Company.

MacDonald, R. & Taylor, M. P. (1993), 'Regional house prices in Britain: long-run relationships and short-run dynamics', *Scottish Journal of Political Economy* **40**(1), 43.

MacQueen, J. (1967), 'Some methods for classification and analysis of multivariate observations.', Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967).

Manley, D., Flowerdew, R. & Steel, D. (2006), 'Scales, levels and processes: Studying spatial patterns of British census variables', *Computers, Environment and Urban Systems* **30**(2), 143–160.

Marble, D., Guo, Z., Liu, L. & Saunders, J. (1997), Recent advances in the exploratory analysis of interregional flows in space and time, *in* Z. Kemp, ed., 'Innovations in GIS 4', Taylor and Francis, London, p. 75.

Masuda, N., Miwa, H. & Konno, N. (2005), 'Geographical threshold graphs with small-world and scale-free properties', *Physical Review E* **71**(3 part 2A).

Mather, P. M. (1976), *Computational methods of multivariate analysis in physical geography*, Wiley, London.

McCann, P. (2001), *Urban and Regional Economics*, Oxford University Press, Great Clarendon Street, Oxford.

Meen, G. (2001), *Modelling spatial housing markets: theory, analysis and policy*, Kluwer, Dordrecht, Netherlands.

Meese, R. A. & Wallace, N. E. (1997), 'The construction of residential house price indices: a comparison of repeat-sales, hedonic-regression and hybrid approaches', *Journal of Real Estate Finance and Economics* **14**(1,2), 51.

Mella-Marzuez, J. M. & Chasco-Yrigoyen, C. (2006), Urban growth and territorial dynamics: a spatial-econometric analysis of Spain, *in* A. Reggiani & P. Nijkamp, eds, 'Spatial Dynamics, Networks and Modelling', Edward Elgar, Cheltenham, UK, p. 325.

Murphy, A. & Muellbauer, J. (1993), Explaining regional house prices in the UK, Technical report, Department of Economics, University College Dublin. Working Paper WP94/21.

Oceans, E. (500,000,000 B.C.), *The Cambrian Explosion (cataclysmic evolutionary event leading to the creation of life as we know it today).*

Openshaw, S., ed. (1995), *Census Users's Handbook*, Geoinformation International.

Orford, S. (1999), *Valuing the built environment: GIS and house price analysis*, Ashgate, Aldershot, England.

Orford, S. & Radcliffe, J. (2007), 'Modelling UK residential dwelling types using OS Mastermap data: A comparison to the 2001 census', *Computers, Environment and Urban Systems* **31**(2), 206.

Ormerod, P. (2005), *Why Most Things Fail: Evolution, Extinction and Economics*, Faber and Faber, London.

Page, S. (1999), 'On the emergence of cities', *Journal of Urban Economics* **45**(1), 184–208.

Pawlak, Z. (2005), 'Flow graphs and intelligent data analysis', *Fundamenta Informaticae* **64**(1-4), 369.

Pollakowski, H. O. & Ray, T. S. (1997), 'Housing price diffusion patterns at different aggregation levels: an examination of housing market efficiency', *Journal of Housing Research* **8**, 107.

Pontius Jr, R. G. et al. (2007), 'Comparing input, output, and validation maps for several models of land change', *Annals of Regional Science* **42**(1), 11.

Propper, C., Jones, K., Bolster, A., Burgess, S., Johnston, R. & Sarker, R. (2005), 'Local neighbourhood and mental health: Evidence from the UK', *Social Science and Medicine* **61**(10), 2065–2083.

Pryke, A. & Beale, R. (2005), Interactive comprehensible data mining, *in* 'LNAI3345: Ambient Intelligence for Scientific Discovery', Springer, Berlin, p. 48.

Rae, A. (2009), 'From spatial interaction data to spatial interaction information? geovisualisation and spatial structures of migration from the 2001 census', *Computers, Environment and Urban Systems* **33**(3), 161. doi:10.1016/j.compenvurbsys.2009.01.007.

Rosen, S. (1974), 'Hedonic price and implicit markets: Product differentiation in pure competition', *Journal of Political Economy* **82**(1), 34.

Sandberg, M. (2007), 'The evolution of IT innovations in Swedish organizations: a Darwinian critique of 'Lamarckian' institutional economics', *Journal of Evolutionary Economics* **17**(1), 1–23.

Schelling, T. C. (1971), 'Dynamic models of segregation', *Journal of Mathematical Sociology* **1**, 143–186.

Schwoon, M. (2006), 'Simulating the adoption of fuel cell vehicles', *Journal of Evolutionary Economics* **16**(4), 435–472.

Shi, S., Young, M. & Hargreaves, B. (2009), 'The ripple effect of local house price movements in New Zealand', *Journal of Property Research* **26**(1), 1.

Silveira, J. J., Espindola, A. L. & Penna, T. (2006), 'An agent-based model to rural-urban migration analysis', *Physica A: Statistical Mechanics and its Applications* **364**, 445–456.

Slater, P. B. (1976), 'A multiterminal network-flow analysis of an unadjusted Spanish interprovincial migration table', *Environment and Planning A* **8**(8), 875–878.

Steinhaus, H. (1957), 'Sur la division des corps materiels en parties', *Bull. Acad. Pol. Sci., Cl. III* **4**, 801–804.

Stillwell, J. (2008), 'Inter-regional migration modelling: A review and assessment (forthcoming book chapter)'.

Stillwell, J. & Duke-Williams, O. (2007), 'Understanding the 2001 UK census migration and commuting data: the effect of small cell adjustment and problems of comparison with 1991', *Journal of the Royal Statistical Society A* **170**(2), 425.

Strogatz, S. H. (1994), *Nonlinear Dynamics and Chaos*, Perseus Books Publishing, LLC, Washington.

Tassier, T. & Menczer, F. (2001), 'Emerging small-world referral networks in evolutionary labour markets', *IEEE Transactions on Evolutionary Computation* **5**(5), 482.

Taylor, P., Evans, D. & Pain, K. (2006), Applications of the inter-locking network model to mega-city regions: Measuring polycentricity within and beyond city-regions, Technical Report 201 (A), GaWC Research Bulletin.
URL: *http://www.lboro.ac.uk/gawc/rb/rb201.html*

Tesfatsion, L. (1997), How economists can get ALife, *in* W. B. Arthur, S. Durlauf & D. Lane, eds, 'The Economy as an Evolving Complex System, II', Westview Press, Washington, pp. 533–563.

Thomas, J. & Cook, K. (2006), 'A visual analytics agenda', *Computer Graphics and Applications, IEEE* **26**(1), 10–13.

Titheridge, H. & Hall, P. (2006), 'Changing travel to work patterns in South East England', *Journal of Transport Geography* **14**(1), 60.

Tsay, R. S. (2002), *Analysis of financial time series*, Wiley-Interscience, Malden, USA.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison Wesley.

UK Census Geography description web page (n.d.).
URL: *http://www.statistics.gov.uk/geography/census_geog.asp*

VideoCrypt (n.d.).
URL: *http://en.wikipedia.org/wiki/VideoCrypt*

Vragovic, I., Louis, E. & Diaz-Guilera, A. (2005), 'Efficiency of informational transfer in regular and complex networks', *Physical Review E* **71**.

Watts, D. J. & Strogatz, S. H. (1998), 'Collective dynamics of 'small-world' networks.', *Nature* **393**(6684), 440–442.
URL: *http://dx.doi.org/10.1038/30918*

Webster, C. (2001), 'Contractual agreements and neighbourhood evolution', *Planning and Markets* **4**(1), 7.

Webster, C. & Lai, L. W.-C. (2003), *Property Rights, Planning and Markets*, Edward Elgar Publishing, Cheltenham, UK.

Whalley, J. & Zhang, S. (2007), 'A numerical simulation analysis of Hukou labour mobility restrictions in China', *Journal of Development Economics* **83**, 392–410.

Witt, U. (1992), 'Evolutionary concepts in economics', *Eastern Economic Journal* **18**(4), 405.

Witt, U. (1993), Emergence and dissemination of innovations: some principles of evolutionary economics, *in* R. H. Day & P. Chen, eds, 'Nonlinear dynamics and evolutionary economics', Oxford University Press, USA.

Wood, J., Dykes, J., Slingsby, A. & Radburn, R. (2009), Flow trees for exploring spatial trajectories, *in* D. Fairbairn, ed., 'Proceedings of the GIS Research UK 17th Annual Conference, University of Durham, Durham, UK', pp. 229–234.

Worthington, A. & Higgs, H. (2003), 'Comovement in UK regional property markets: a multi-variate cointegration analysis', *Journal of Property Investment and Finance* **21**(4), 326.

Wu, F. (2005), Introduction - urban simulation, *in* D. Atkinson, Foody & Wu, eds, 'GeoDynamics', CRC Press, Boca Raton, Florida, pp. 205–214.

Xu, Y. (2006), 'The behaviour of the exchange rate in the genetic algorithm with agents having long memory', *Journal of Evolutionary Economics* **16**(3), 279–297.

Yan, J. & Thill, J.-C. (2009), 'Visual data mining in spatial interaction analysis with self-organizing maps', *Environment and Planning B* **36**(3), 466.

Zito, T., Wilbert, N., Wiskott, L. & Berkes, P. (2009), 'Modular toolkit for data processing (mdp): a python data processing frame work', *Front. Neuroinform.* **2**(8).
URL: *http://mdp-toolkit.sourceforge.net*