Routledge
Taylor & Francis Group

INTERNATIONAL JOURNAL OF STRATEGIC PROPERTY MANAGEMENT
ISSN 1648-715X print / ISSN 1648-9179 online

2013 Volume 17(3): 278–304
doi: 10.3846/1648715X.2013.822032

REGIONAL HOUSE PRICE INDEX CONSTRUCTION – THE CASE OF SWEDEN

Lars-Erik ERICSON [1], Han-Suck SONG [2], Jakob WINSTRAND [3] and Mats WILHELMSSON [4] ✉

[1] *Valuguard Index Sweden AB, Sweden*
E-mail: lars-erik.ericson@valueguard.se
[2] *Department of Real Estate Economics, Royal Institute of Technology (KTH), Stockholm, Sweden*
E-mail: han-suck.song@abe.kth.se
[3] *Valuguard Index Sweden AB, Sweden*
E-mail: jacob.winstrand@valueguard.se
[4] *Institute of Urban and Housing Research (IBF), Uppsala University and Center for Banking and Finance, KTH, Sweden*
E-mail: mats.wilhelmsson@abe.kth.se

Received 18 January 2012; accepted 25 May 2012

ABSTRACT. The academic literature on the construction of regional house price indexes usually uses geographic areas whose boundaries are administratively drawn. However such administrative regions might not be optimal for the construction of regional price indexes. When producing housing price indexes, we often encounter problems with insufficient number of observations. One way to remedy this problem is to estimate a quarterly index instead of a monthly index. Another possible way to mitigate the thin markets problem is to construct indexes for geographically aggregated regions. However, the literature that discusses methods of dealing with the problem of thin markets and especially geographical aggregation is very rare. The goal of this paper is to construct a housing price index for a major part of Sweden, and to construct price index series for a number of regions. The number of regions, and how their boundaries should be created in order to construct reliable regional price indexes, is however an open question. We apply traditional hedonic methodology in order to estimate house price indexes for both predefined regions whose boundaries are based on a division of labor markets in Sweden, as well as a division of regions based on statistical cluster analysis. The results from this study suggest that regions should be clustered together based on regional price levels and/or price development as clustering variables. If only geographical proximity is used as clustering variable, our computations show that there is a high risk that we end up with some clusters having large standard errors, which in turn might result in inaccurate indexes.

KEYWORDS: Regional house prices; Hedonic price index; Cluster analysis; Aggregation

REFERENCE to this paper should be made as follows: Ericson, L.-E., Song, H.-S., Winstrand, J. and Wilhelmsson, M. (2013) Regional house price index construction – the case of Sweden, *International Journal of Strategic Property Management*, 17(3), pp. 278–304.

1. INTRODUCTION

When house price indexes are constructed by estimating hedonic price equations, having access to a large number of transactions is typically essential. Besides transaction prices, a rich set of attributes is needed in order to construct reliable house price indexes. In many cases, however, the number of available transactions on the housing market is less than

the amount which is desirable, which in turn might create problems for index constructors to, for instance, do statistical inference and economic analysis. The problem of how to construct price indexes when we face thin markets is therefore important to investigate. The literature about house price index construction with small sample sizes is not huge: Schwann (1998), McMillen (2003), Francke and Vos (2004), and Francke (2010), are examples of recent articles dealing with the problem of index construction in a thin markets environment.

One way to reduce the problem of thin markets is to aggregate smaller housing markets to larger housing markets, but it is not obvious how this aggregation should be carried out. In many cases there has been an arbitrary pooling of data across geography. To aggregate geographically adjacent areas may not be the best way to construct large housing market since different areas within a housing market can exhibit price evolutions that differ much. For instance, some housing markets have very distinct sub-markets. Furthermore, estimating a single price index for a whole region might not be a good solution, simply because such a method is based on implicit assumption that the aggregated price index has similar statistical properties as all the individual indexes in the sub-markets.

An interesting approach to create aggregated indexes is to combine housing markets that exhibit similar house price developments. Although this may seem like a good solution, there still is a problem of how to define what constitutes similar house price developments, and how to compare different regional housing markets.

In order to reduce the problem of thin markets, we present in this paper different methods to aggregate housing markets to larger clusters using cluster analysis. Different clustering methods yield different regions, and henceforth different sets of price indexes. We therefore apply the Root Mean Squared Error (RMSE) as the out-of-sample measure in order to evaluate the different price indexes.

The disposition is as follow: a brief literature review is presented in the next section. Section 3 describes the methodology used. The empirical analysis is presented in Section 4. Section 5 concludes this paper.

2. LITERATURE REVIEW

As mentioned above, the literature that discusses house price index construction and the problem of thin markets is quite small. Schwann (1998) was one of the first who presented a method on how to tackle the problem. He defined thin markets as those that that have less than 30 sales per period (in his case per quarter). In order to estimate local house price indexes, temporal aggregation may be considered as a tool to increase reliability and accuracy of the index. Temporal aggregation is however not a practical or recommended solution, as Englund et al. (1999) have shown. Schwann (1998) proposed instead a method where earlier observations are added to current transactions and thereby increasing the number of observations. He used a data set from 1979 to 1992 in Vancouver with more than 60 000 observations. In order to evaluate his index method, he designed an experiment where he used smaller and smaller samples. In that way he could compare his method with the "true" price index using all observations. He used the root mean squared errors (RMSE), the average standard error, number of periods outside the confidence interval and turning point correctly identified. His conclusion is that the performance of this method is much better than a traditional hedonic price index.

Englund et al. (1999) analyzed whether temporal aggregation can be used in order to calculate local house price indexes. They concluded that time intervals should be as short as possible, that is, temporal disaggregation is considered to be most important. McMillen (2003) used locally weighted regressions on order to estimate reliable and accurate price indexes in submarkets that have very few transactions. The method is based on Fou-

rier expansion method and allowed him to estimate smooth house price indexes for 851 census tracts in the metropolitan area of Chicago. The main idea behind the method is to set up a regression model where observations from outside a census tract are used. Observations far away are down-weighted and more weights are placed on transactions in the census tracts. He used a data set consisting of almost 28 000 observations, repeat sales in his case, over a six year period in the beginning of 1990. Francke and Vos (2004) used a so-called hierarchical trend model (HTM) that is an extension of Schwann (1998). The HTM model allows the parameter to vary in space, time, and house type and thereby make it possible to estimate a price index for each segment of the market. One of the disadvantages with the HTM is that assumptions need to be made about the general trend and trend levels for sub-markets. They used about 30 000 observations over the period of 1985 to 1999 in Amsterdam and 21 000 in Breda, Holland. They studied standard deviation of each model in order to evaluate them. The results indicated that for small local housing markets, the HTM model seemed to be more accurate. A more recent article is Francke (2010) who applied repeat sales method to estimate indexes for thin markets. Costello et al. (2009) and Goh et al. (2012) compared different methods where the objective was to find the most accurate and robust price index in highly localized markets at frequent time intervals. They used the Mean Square Error (MSE) as the out-of-sample measure in order to evaluate their different price indexes. 75% of the observations were used in order to estimate the different models and the remaining 25% were saved to evaluate the performance. They used a data set of more than 500 000 observations from 1988 to 2005 in the city of Perth in Australia. The area is divided into 299 suburbs. They concluded that aggregation (both temporal and geographical) can be problematic and should not be done arbitrary. If it is done, the so-called hedonic imputation method shows better performance than the other methods used (longitudinal hedonic, repeat-sales method, hybrid and median approach).

What conclusions can be drawn from the literature? Different methods have been proposed in order to reduce the problem of thin markets. However the proposed methods typically put forward that different ways to make temporal and geographical aggregations. But such methods seem to come at a cost: they reduce accuracy and reliability of house prices indexes. In other words, there is an important trade-off between estimating indexes on aggregated regions in order to mitigate thin market problems, and to estimate price indexes on disaggregated levels in order to avoid problems of accuracy and reliability.

Below we will construct regional prices indexes based on geographical aggregation. However, we are not doing that in an arbitrary way, instead we test different methods based on cluster analysis. In this analysis, we are not only considering aggregating nearby areas, but also other ways to segment the regions, for instance by using price development and mean prices as clustering variables.

3. METHOD

The estimation of regional price index series is done for two types of geographical divisions of regions. First we estimate price indexes for already existing regions. The Swedish Agency for Economic and Regional Growth (Tillväxtverket) has created these regions, whose boundaries reflect functional labor market regions. Some regions are further divided into sub-regions. We use these sub-regions when possible in this paper. Henceforth the regions are simply referred to as “FA-(sub)regions” or simply “regions”. We could have used single municipalities as smallest geographical unit in this first step, but the number of observations would not be sufficient in order to estimate hedonic price equations in most of the 290 municipalities in Sweden.

Secondly, we estimate price indexes for clusters of regions that we create with cluster analysis. Therefore the price index estimation

and the following evaluation procedure involve a large number of steps. In order to help the reader to follow our procedures we first give a short overall summary of the main steps of the estimation procedure. Thereafter we explain the estimation procedure in more detail. We also present some key numbers.

Our analysis is based on following main steps:

- *Initial step* (*step 0*): Collection of data on single-family house transactions and on FA-(sub)regions.
- *Step 1*: Estimation of annual hedonic price index series for each region and preparing a dataset with descriptive statistics for the regions – based on a sample of 90% of the data.
- *Step 2*: Applying cluster analysis in order to create different sets of homogenous groups of housing sub-markets – the clustered regions.
- *Step 3*: Estimation of hedonic price index series for the different sets of clustered regions, and comparison and performance evaluation of the different price index series.

Below we explain in more detail the different sub-procedures involved in the steps above.

Step 0: Collection of data on single-family house transactions and labor market regions

The data on single-family house transactions comes from a unique database provided by Valueguard Index Sweden AB. The database contains about 70% of all house sales in Sweden from 2005 to 2010. The database has been constructed by merging data from real estate agents and the official property register. In total, 209 126 observations are included in this dataset. For each transaction, following variables are observed: transaction price, contract date, a number of quality and size variables (living area, number of rooms, lot area, semi-detached, detached, quality index, building year), and a number of location variables (*X*- and y- coordinates, sea front, sea view, value areas for taxation purposes, urban, municipality).

As mentioned above, the functional labor market regions that we estimate price index series of are based on the Swedish Agency for Economic and Regional Growth's FA-subregions. Sweden is divided into 72 FA-regions, and some FA-regions are further divided into a number of FA-subregions. The total number of FA-regions and FA-subregions amount to 93. The FA-regions have been constructed for analytic purposes. The idea is that each region shares the same labor-market. Many regions consist of a city and its surrounding areas.

However, some regions do not have sufficient number of transactions to make the estimation of (yearly) price index series based on regression analysis possible. We define the criteria to be used when determining whether a FA-(sub)region contains enough number of transactions as follows: the minimum amount of required house sales per FA-(sub)region and year must be at least 83 (at least 500 observations over six years). This cut-off criterion is slightly higher than e.g. Geltner (1997). Given this criterion, 66 of the 93 regions are considered to have enough number of transactions. This means that 27 FA-(sub)regions are not included in the following analysis, representing 2.2% of the transactions.

Step 1: Estimation of annual hedonic price index for each labor market region

In this step we estimate yearly hedonic house price indexes for each of the 66 FA-(sub) regions that are considered to have enough number of observations.

A hedonic equation is a regression of prices against attributes that determine these prices and time. The regression coefficients are interpreted as estimates of the implicit (hedonic) prices of these attributes, and hence, the willingness-to-pay for the attribute in question (see Rosen, 1974). The method has a long tradition. Recent articles are, for example, Song and Wilhelmsson (2010) and Ceccato and Wilhelmsson (2011). Following the literature, the hedonic price equation is equal to

$$Y_{i,t} = \beta_0 + X_{i,t}\beta_1 + TD_t\beta_{2,t} + \varepsilon_{i,t} \quad i = 1, \ldots, N \text{ and } t = 1, \ldots, T, \tag{1}$$

where: Y denotes the dependent variable transaction price (normally in log form); β_1 is a vector of parameters (regression coefficients) associated with exogenous explanatory variables, X. The stochastic term e is assumed to have a constant variance and to by normally distributed. Usually we implicitly assume that all relevant attributes are included in the matrix X: in other words, no omitted variable bias problem exists. We can decompose X into, for example, structural apartment and property attributes, as well as neighborhood attributes. The variable TD with subscript t is a dummy variable for each period and equals one for period t and zero otherwise. The number of observations is denoted by N, and T denotes the number of time periods.

The two major approaches measuring hedonic price indexes are the time dummy approach and the so-called hedonic imputation approach (see Diewert et al., 2009). Song and Wilhelmsson (2010) is an example of the former and Gouriéroux and Laferrère (2009) is an example of the latter. Here we are utilizing a time dummy approach. The main difference between the methods is that the hedonic imputation allows all estimated parameters to change over time while the time dummy method assume that the parameters are constant over time. One way to overcome the problem of unstable parameters over time is to use moving window regression as in Song and Wilhelmsson (2010). In their article a hedonic time dummy approach is compared to a moving window time dummy approach with a window span of one year. Their conclusion is that there is no difference in estimated parameters concerning the time dummies. However, the general conclusion seems to be that the hedonic imputation method is preferable if the parameters are unstable over time (see e.g. Berndt and Rappaport, 2001 and Pakes, 2003 besides the article referred to above). Diewert et al. (2009) concludes with the following statement: "favor HI [hedonic imputation] methods unless degrees of freedom are very limited". In our case the degrees of freedom are very limited. Our overall objective is to estimate hedonic price indexes on market that are very thin. Consequently, the hedonic imputation method is not an approach that can be utilized. The main advantages with the time dummy approach are that the degree of freedom is preserved and that the methods minimize the influence of outliers (see Diewert et al., 2009). Hence, hedonic the time dummy approach is used in this study.

Spatial dependency is a problem that is more or less always present in this type of hedonic models. In order to minimize the problem of spatial dependency, we are including a number of different variables such as submarket dummies, coordinates and distance to the city. Coordinates have earlier been used in e.g. Wilhelmsson (2009) and Galster et al. (2004) In order to reduce spatial.

However, it is an empirical question whether spatial dependency creates biases in the coefficients concerning the price index. Song and Wilhelmsson (2010), using the same data, found that it did not. We have also estimated spatial autoregressive model (SAR) and spatial error model (SEM) following Anselin (1988) and Wilhelmsson (2002). We are using inverse distance as spatial weight matrix.

Since the out-of-sample forecast evaluation below requires some proportion of the observations to be saved, we choose to set aside ten percent – randomly chosen – of the historical data to be reserved for out-of-sample testing. In other words, the hedonic price index estimations will be based on random sample of 90% of the transactions (that is, 90% of 209 126 transactions from 2005 to 2010).

The dependent variable is transaction price based on contract dates. The explanatory variables consist of size variables (living area, secondary area, number of rooms, lot size), type of house variables (semi-detached, detached), and standard and location variables (quality index, building year, municipality, sea front, sea view, urban, and X- and Y- coordinates).

A dummy variable is created for each municipality. The quality index is defined by tax authorities in order to appraise the properties for taxation purposes. It is a composite of 25

questions concerning different quality aspects such as construction materials and amenities. Each question gives a number of points (2.5 points on average, but some questions can give as much as 11 points). One additional unit of quality can refer to very different things: for example, the existence of a car port or that the house has a new roof. Information whether the single-family house is semi-detached or detached are included as an attribute. Based on the building year variable, we construct a number of dummy variables that reflect different building periods since the beginning of the twentieth century (see Song and Wilhelmsson, 2010, for further information).

Step 2: Identifying homogenous groups of housing sub-markets with cluster analysis

Cluster analysis has been used in earlier research as a tool for constructing housing sub-markets (see for example Wilhelmsson, 2004). Here we will use it as a tool to aggregate smaller housing markets into larger homogenous housing markets – the clusters. The smaller housing markets in a specific cluster are supposed to share many characteristics, such as proximity to each other, price development, and/or price level.

Variables used in the cluster analysis. We have chosen to use number of different variables and combinations of these as clustering variables.

Each clustering variable and each combination of these corresponds to a specific clustering method. The first cluster analysis method (C1) uses the average annual price development over time (2005–2010) in the 66 FA-(sub) regions. The average price changes are determined by the hedonic price index estimations conducted in step 1 above. The second method (C2) uses the mean price level over the period. Method three (C3) uses the price development pattern over time by using a structural break every second year. The fourth cluster analysis method (C4) uses distance between FA-(sub) regions. The distance between the housing markets have been estimated from the average coordinates of the transactions in the housing market. Method number five (C5) uses both geographical proximity and price development. Finally, method number six (C6) combines of all the above clustering variables. As a reference, we also perform a seventh cluster analysis (C7) based on a variable containing random numbers. All cluster analysis are weighting the housing labor markets by size where size is measured by the number of transactions (observations).

Transformation, normalization and weighting of the variables. The size of the clustering variables naturally varies substantially between different housing submarkets. Furthermore, different types of cluster variables are used simultaneously in the cluster analyses below. In order to make the variables comparable, the following steps are taken.

Meanprice

The variable mean price is first transformed to logarithmic price. This means that the difference between two regions with mean price of 200 000 SEK and 400 000 SEK will have the same importance as the difference between 2 MSEK and 4 MSEK. The logarithmic price is also standardized with its standard deviation, in order to make it comparable with other variables.

Price development

The price development captures the average annual price changes. The price development is transformed to logarithmic scale and normalized by dividing it with its own standard deviation.

Price development pattern

Even if two regions exhibit the same total price development from 2005 to 2010, there might be large differences in the annual price developments. That is why we construct variables to measure the price development pattern during the period. The pattern is defined as four variables measuring the price development over two years, that is (price index year Z) / (price index year Z-2) for Z is 2007, 2008, 2009 and 2010. Each variable is transformed to logarithmic scale.

These variables are not individually standardized in order to avoid a situation in which we underestimate the importance of periods with large price developments, and overestimate the importance of stable periods, which would be counter-intuitive.

Coordinates

The coordinates are based on the Swedish RT90-system, which measures distance in meters. X-coordinates represent the north-south direction and *Y*-coordinates represent the east-west direction. Sweden is an oblong country, and the standard deviation of the *X*-coordinates is almost twice as big as the standard deviation of the y-coordinates. If they were normalized with their own standard deviations, one meter in north-south direction would have less bearing than one meter in east-west direction. Because of this, we standardize the coordinates y dividing them with the same number when used together with other variables.

The coordinates are not normally distributed; instead they exhibit very fat tails. If we would just standardize them with their standard deviation, the coordinates would be more important than other variables in a cluster analysis.

In order to find a good weighting of coordinates, we have tried different weights and analyzed the resulting outcome (maps). We have found that dividing the coordinates with 200 000 seems to give the coordinates a well-balanced importance in the cluster analysis. This could of course be tested further.

Note that there are two variables that measure the geographical position (the *X*- and *Y*-coordinates), which is important when coordinates are included in clustering models with other variables.

Relative weights between different types of variables

We also need to consider the relative weights between the variables because we use different number of variables for measuring similarity.

- Geographical location is measured by two variables, *X* and *Y*.
- Mean price is measured by one variable and is therefore multiplied by 2.
- Price development is measured by one variable and is therefore multiplied by 2.
- Price development pattern is measured with four variables. However, these are not standardized. They have rather low standard deviations and after tests, we find that multiplying all variables by 2 is most appropriate. However, further tests should be undertaken. This weighting is only used when all clustering variables are used simultaneously.

Weighting of regions in the cluster analysis. The objective is to create clusters which have large enough number of observations needed for constructing reliable indices based on hedonic regression analysis. However, it may require several clusters to reproduce the different market characteristics among the many housing markets. But, segmenting the market in too many small clusters creates problems with thin markets. Thus, there is a trade-off between identifying large enough clusters and to identify as many clusters as needed to reflect the heterogeneity among the housing submarkets.

With the clustering procedure we are using, it is not possible to influence how small or how large the clusters are going to be, which can result in very small or very large clusters. For instance, an "outlier-region" with a low number of sold houses might become its own cluster. On the other hand, several large cities can be assigned into one very large cluster. In this case the statistical clustering should ideally allow a region with many observations to become more "viscous" to avoid that such regions too easy become assigned into "too large" clusters, that is to say, we would not like the three metropolitan areas (Stockholm, Göteborg and Malmö) to become one cluster. If we could assign weights to regions with many observations, we would solve this problem. Unfortunately, we cannot perform the weighting directly using a standard clustering procedure. Therefore, we have implemented, into the cluster analysis, a method where we create "cop-

ies" of the regions, where the number of copies that are created are based on the number of sales in the region. Thus, each region will be represented by n copies, where n is defined as the number of sales. As a result, the cluster analysis will first create "clusters" with only the duplicate observations because they have the same values on all cluster variables. This is equivalent as giving them a weight, and it is no longer likely that the largest cities will be put together in the same cluster. Smaller regions on the other hand will more easily be joined with other regions.

Cluster procedures and similarity/dissimilarity measures. There are many methods of clustering the data (see e.g. Mooi and Sarstedt, 2011). We have evaluated a number of different methods, but of course this could be further investigated.

The cluster procedures we apply are *k-means and k-medians*, which assign each point to the cluster whose center is nearest. The first clusters are randomly chosen. The k-means procedure is a relatively simple clustering procedure that is suitable for large data sets, which we also have. We use both Euclidean distance and Canberra distance in order to find similarities between price indexes or geographical proximity. However, we found very small differences between the cluster procedures and the similarity measures. As a result, and in order to simplify the presentation, we only use K-means and Euclidean distance.

We iterate the number of clusters by starting by estimating two clusters and then three, four and so on. The iteration stops when the number of observations in the smallest cluster is on average below 60 observations each month and/or below 30 observations in an individual month (that is in line with Geltner, 1997; Schwann, 1998). Furthermore, we remove – if possible – the three largest regions from their respective clusters in a second step. If for instance Stockholm is assigned into the same cluster as other large regions, Stockholm will be removed. This will only happen if the cluster initially contains at least 15 000 observations.

If not necessary, we want to avoid clustering the largest regions with other regions since they have enough observations to constitute their own clusters. This kind of "post clustering" could probably be investigated further, for instance by performing new cluster analyses on the resulting clusters from the first cluster analysis, in order to divide the largest clusters into smaller ones.

Step 3: Comparison and performance evaluation of the different regional price index series

Since each clustering method (see C1 to C7 above) generates different set of clusters, the corresponding price indexes will also be different. Thus, it is important to compare and evaluate the performance of the different clustering methods for price index construction purposes. However the problem is to define a natural choice of benchmark against which the different clustering methods can be compared. In this paper, we use an out-of-sample prediction measure utilized by Costello et al. (2009) and Goh et al. (2012). They estimate Root Mean Squared Error (RMSE) on a sub-sample of 25%. However, we choose to set aside 10% of the observations for the out-of-sample test. The remaining 90% of the observations will be used to estimate the hedonic price indexes. In order to compensate for only using a subsample of size 10% we do a simulation with 100 out-of-sample replications (sampling with replacement). Then we use the mean figures of the 100 replications.

We have found that the results from the clustering vary a lot between the replications. The same method can result in different number of clusters, and the resulting regression models will sometimes be relatively good and sometimes relatively bad. By using a random sample of 90% of the data, we obtain somewhat different values on the variables used in the cluster analyses. These small variations might explain why the results from the cluster analysis differ.

4. EMPIRICAL ANALYSIS

Data source

We use sales-data from real estate agencies. The dataset contains variables that originate from the sales-process, for instance to create advertisements and publishing them on the Internet. All sales have a contract date. This is a big advantage compared to using data from when the transaction is recorded in official registers, which is often done several months after that the buyers and sellers have agreed upon the sales price.

In order to get more information about the sold houses, this dataset has been combined with data from the Swedish Real Estate register. From that register we get variables like see view, lot size and a quality index used for taxation purposes. Some descriptive statistics can be found in the part where the hedonic regression is described. The dataset has been provided to us by Valueguard Index Sweden AB.

Regions

The number of transactions in different regions varies a lot. There is a close connection between the population in the region and the number of transactions. In Figure 1, number of transaction per month is plotted against the population rank.

The relationship between population and number of transactions per month is positive. The largest four labor markets have all more than 100 observations per month in average. However, it is only 11 labor markets that have more than 60 observations per month on average. In Figure 2, number of observations per 100 000 inhabitants is shown.

On average, over the period, there are around 2 500 observations per 100 000 inhabitants. However the variation around the average is big, especially for the smaller labor markets. Some of the smallest labor markets have not that many transactions but related to the population, the number of sales is large. That is to say, it is not obvious that all small labor markets have few observations as the number of observations per inhabitants can be large. Some regions have more owner occupied houses than others, and in regions with low prices, many sales are not reported because they are not sold by real estate agents. More information about the regions can be found in the Appendix.

Descriptive statistics for the dataset

Table 1 presents descriptive statistics concerning the observed attributes that are used in the hedonic price equation. As earlier described we are including a number of different attributes in order to control for all variation across time and space. Descriptive statistics concerning the municipality-dummies and the time dummies are not presented in the Table.

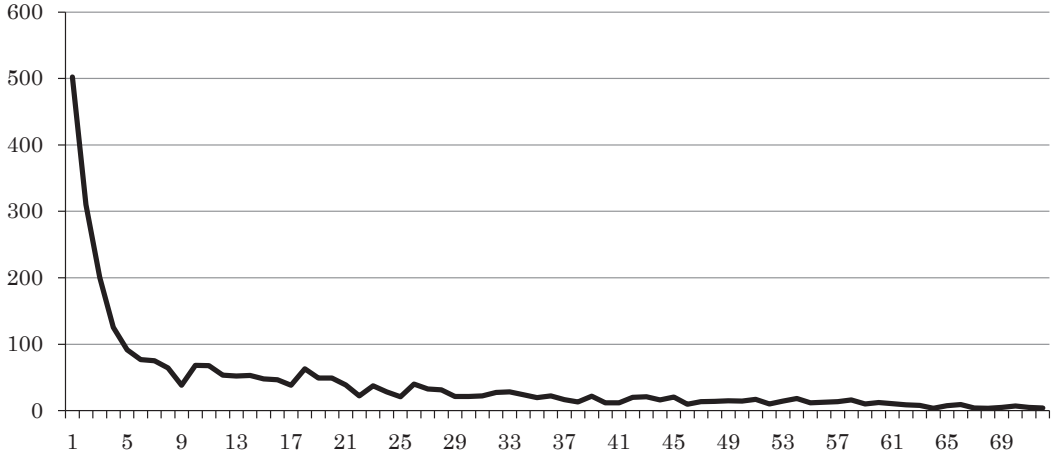

Figure 1. Number of transactions per month

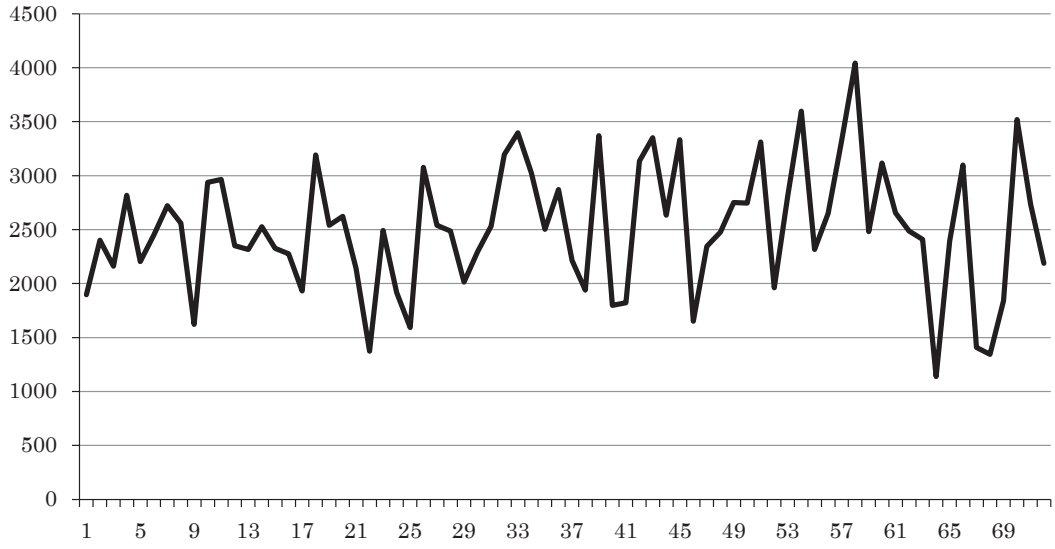

Figure 2. Number of transactions per capita (100 000)

Table 1. Descriptive statistics concerning attributes that are used in the hedonic price equation (data observed during 2005–2010)

	Average	Standard dev.	Min	Max
Price	1 871 721	1 192461	200 000	6 850 000
Living area (m^2)	125.0103	33.7113	54	258
Rooms	5.0031	1.2726	1	10
Lot size (m^2)	1157	1215	127	12219
Building period 1 (–1899)	0.0111	0.1147	0	1
Building period 2 (1900–1939)	0.1806	0.3943	0	1
Building period 3 (1940–1959)	0.1664	0.3716	0	1
Building period 4 (1960–1975)	0.3283	0.4664	0	1
Building period 5 (1976–1990)	0.2097	0.4027	0	1
Building period 6 (1991–)	0.0860	0.2753	0	1
Quality index	29.6156	4.1373	18	57
Semi-detached	0.1119	0.3153	0	1
Detached	0.1172	0.3214	0	1
Seafront	0.0042	0.0644	0	1
Sea view	0.0455	0.2085	0	1
Urban	0.4536	0.4978	0	1

32% of the houses were built between 1960 and 1975. Houses at the seafront are rather uncommon, but around 5% of all houses in the dataset are either sea view or seafront houses. Naturally, the descriptive statistics vary from region to region for many of the variables. For instance, the mean and standard deviation of regional prices vary a lot. Furthermore, the lot sizes are usually smaller in in urban areas as compared to those in the rural areas.

The dummy variable "Urban" has been provided to this research by Valueguard Index Sweden AB. A house (transaction) is said to be Urban (has value 1) if the house is located in the conurbation of one of the 100 largest cities in Sweden.

A hedonic price index equation will be estimated for each of the 66 labor markets, and such estimations require a minimum amount of observations in order for the statistical estimation procedures to work with large enough degrees of freedom from a statistical point of view. The average number of observations is 3 500 observations per FA-(sub)region or labor market which corresponds to only 50 observations per month. However there is a large variation between the labor markets: the smallest labor markets have less than 10 observations per month, which makes it impossible to estimate reliable price index series on a monthly basis. As a comparison, Stockholm has more than 500 observations on average per month.

Description of the regions

Based on this sample, we create a table with descriptive statistics on the regions that will be used in the cluster analysis:

– Number of observations;
– Mean price;
– Mean coordinate;
– Average price development;
– Price development pattern.

In Table 2, the regions are described based on all the data in the dataset.

Table 2. Descriptive statistics for each region

Code	Name	Type of region	Population	Observations	Average price	X coord., mean	Y-coord., mean
1a	Stockholm	Big city	1 906 700	34 456	3 103 882	6585123	1626376
1b	Södertälje	Big city	147 044	3 326	2 118 589	6561190	1591590
1c	Uppsala	Big city	298 899	6 668	2 082 493	6646798	1600724
2	Nyköping	Small regional centre	62 143	2 034	1 685 633	6514938	1568459
3a	Eskilstuna	Large regional centre	94 785	1 483	1 597 988	6583076	1538713
3b	Katrineholm	Large regional centre	57 340	1 740	1 061 708	6546189	1529953
4a	Linköping	Large regional centre	163 291	3 896	1 951 823	6469534	1491951
4b	Norrköping	Large regional centre	170 460	2 761	1 619 325	6494894	1521951
4c	Mjölby	Large regional centre	36 264	834	1 104 919	6465376	1460531
4d	Motala	Large regional centre	49 428	962	1 231 990	6490335	1456427
5a	Värnamo	Small regional centre	42 481	700	1 133 092	6339395	1391331
5b	Gislaved	Small regional centre	41 101	1 007	784 678	6360975	1359072
6a	Jönköping	Large regional centre	162 270	3 766	1 682 044	6407746	1403053
6b	Nässjö	Large regional centre	45 823	1 372	922 056	6391463	1442965
7	Vetlanda	Small regional centre	37 226	1 177	749 914	6364538	1449085
8	Tranås	Small regional centre	21 792	666	967 926	6432166	1452530

(Continued)

(Continued)

Code	Name	Type of region	Population	Obser-vations	Average price	X coord., mean	Y-coord., mean
9	Älmhult	Small regional centre	28 059	745	990 424	6264397	1395264
10	Ljungby	Small regional centre	37 030	1 041	1 021 285	6289240	1378074
11	Växjö	Large regional centre	129 783	2 755	1 247 018	6306176	1443344
12a	Kalmar	Large regional centre	93 361	2 843	1 563 330	6283810	1534391
12b	Nybro	Large regional centre	28 888	711	782 314	6290979	1502146
13	Vimmerby	Small region	29 597	909	623 367	6379998	1500528
14	Västervik	Small regional centre	36 356	1 241	1 076 651	6407407	1543372
15	Oskarshamn	Small regional centre	45 195	1 470	921 444	6337867	1533535
16	Gotland	Small regional centre	57 004	1 371	1 630 159	6384632	1657154
17a	Karlskrona	Large regional centre	91 293	2 248	1 300 121	6232091	1483006
17b	Karlshamn	Large regional centre	44 126	1 144	1 017 292	6232953	1436089
18a	Kristianstad	Large regional centre	141 645	4 512	1 177 188	6219285	1388678
18b	Sölvesborg	Large regional centre	29 040	1 186	1 092 259	6216046	1423128
19a	Malmö	Big city	668 074	14 292	2 479 303	6168252	1334240
19b	Ystad	Big city	60 042	1 985	1 640 033	6154138	1387774
19c	Helsingborg	Big city	320 149	8 941	1 852 158	6223924	1317742
20	Halmstad	Large regional centre	164 223	4 864	1 699 047	6295715	1317894
21a	Göteborg	Big city	929 536	21 653	2 558 446	6399954	1278185
21b	Alingsås	Big city	63 128	1 582	1 594 485	6432393	1312081
21c	Stenungsund	Big city	53 947	1 145	2 075 495	6447818	1261194
22	Borås	Large regional centre	134 506	3 513	1 336 042	6402474	1337005
23a	Trollhättan	Large regional centre	106 771	2 015	1 287 018	6476362	1297997
23b	Uddevalla	Large regional centre	92 004	2 293	1 625 087	6479148	1260726
24	Lidköping	Small regional centre	66 875	1 504	1 148 439	6483198	1346708
25a	Skövde	Large regional centre	139 431	3 479	1 084 560	6469925	1385563
25b	Mariestad	Large regional centre	38 452	1 039	943 587	6513273	1393015
26	Strömstad	Small regional centre	23 878	548	1 735 983	6533119	1238508
27	Bengtsfors	Small region	14 685	526	695 645	6544330	1290023
28	Årjäng	Small region	9 952	195	919 329	6594864	1286681

(Continued)

(Continued)

Code	Name	Type of region	Population	Obser-vations	Average price	X coord., mean	Y-coord., mean
29	Eda	Small region	8 653	233	686 928	6636141	1301428
30a	Karlstad	Large regional centre	198 409	5 367	1 244 361	6598969	1363848
30b	Säffle	Large regional centre	28 329	874	753 634	6557132	1326624
31	Torsby	Small region	12 707	255	757 171	6686443	1343152
32	Hagfors	Small region	12 804	326	539 888	6658302	1379841
33	Filipstad	Small region	10 682	10	791 000	6630662	1409712
34	Örebro	Large regional centre	225 531	5 538	1 404 929	6567742	1462868
35	Hällefors	Small region	7 361	37	470 135	6626438	1430412
36	Karlskoga	Small regional centre	44 094	1 432	764 571	6578392	1424704
37a	Västerås	Large regional centre	181 125	4 631	1 764 276	6616690	1537590
37b	Köping	Large regional centre	46 211	825	1 048 914	6593367	1506258
38	Fagersta	Small region	22 638	528	636 679	6651597	1501533
39	Vansbro	Small region	6 916	49	436 673	6709130	1419465
40	Malung	Small region	10 385	151	643 291	6736740	1382258
41	Mora	Small regional centre	34 430	899	926 234	6776782	1425714
42	Falun/ Borlänge	Large regional centre	150 684	3 832	1 209 143	6717377	1481429
43	Avesta	Small regional centre	37 196	726	714 912	6681235	1516930
44	Ludvika	Small region	41 385	938	715 389	6666995	1466975
45a	Gävle	Large regional centre	108 600	2 705	1 385 628	6731008	1573218
45b	Sandviken	Large regional centre	46 775	1 552	941 863	6720023	1546322
46a	Söderhamn	Small regional centre	25 987	601	692 818	6796800	1565705
46b	Bollnäs	Small regional centre	37 836	1 007	691 610	6806016	1524189
47	Hudiksvall	Small regional centre	46 641	823	897 142	6851668	1564744
48	Ljusdal	Small region	19 133	327	720 813	6853904	1511043
49	Sundsvall	Large regional centre	147 974	3 427	1 091 745	6928147	1582596
50	Kramfors	Small region	19 473	179	482 626	6983506	1606806
51	Sollefteå	Small region	20 538	253	520 375	7008711	1571494
52	Örnsköldsvik	Small regional centre	55 387	1 522	932 072	7022838	1643872
53	Östersund	Small regional centre	116 252	1 579	1 481 488	7014018	1435389
54	Härjedalen	Small region	10 645	59	585 593	6897920	1413239
55	Storuman	Small region	6 304	49	673 014	7253155	1524439

(Continued)

(Continued)

Code	Name	Type of region	Population	Observations	Average price	X coord., mean	Y-coord., mean
56	Lycksele	Small region	15 846	207	678 601	7169545	1637752
57	Dorotea	Small region	2 914	3	736 667	7155357	1501103
58	Vilhelmina	Small region	7 220	20	517 000	7184367	1525345
59	Åsele	Small region	3 180	6	365 833	7116412	1574215
60	Sorsele	Small region	2 733	2	250 000	7278598	1575254
61	Umeå	Large regional centre	143 390	2 682	1 653 478	7087860	1716009
62	Skellefteå	Small regional centre	76 225	1 457	967 047	7191705	1747407
63	Arvidsjaur	Small region	6 665	37	556 486	7282875	1656785
64	Arjeplog	Small region	3 146	11	395 227	7315194	1597937
65	Luleå	Large regional centre	167 470	4 761	1 005 182	7291709	1779966
66	Överkalix	Small region	3 715	26	323 173	7376047	1815155
67	Övertorneå	Small region	4 972	23	476 891	7384784	1851035
68	Haparanda	Small region	10 173	88	765 210	7328661	1874003
69	Pajala	Small region	6 429	3	465 000	7477218	1826688
70	Jokkmokk	Small region	5 305	2	497 500	7361828	1701574
71	Gällivare	Small region	18 703	200	899 954	7458689	1712038
72	Kiruna	Small regional centre	23 099	266	943 579	7535039	1689671

Figure 9 that shows all the regions can be found in the Appendix.

Removal of measurement errors

We have removed extreme values from the dataset, or rather data that probably is incorrect. Much of the data is originally entered manually from real estate agents, and there are some errors in the dataset. For instance, when there are no neighbors within 10 000 meters, the coordinates might be wrong.

We have used the following criteria to remove extreme values and incorrect data:

- Coordinates that indicate sales more than 10 000 meters from the closest neighbor are removed.
- Houses built before 1850.
- Houses with a lot area bigger than 7000.
- Houses with more than 10 rooms.
- Observations were the total building area (including garage etc.) is more than 200 meters larger than the living area.
- Houses with extreme values on price or price per square meter in their respective municipality.

Hedonic price indexes for each labor market region (step 1)

In the first step hedonic price equations are estimated for the 66 labor market regions (FA-(sub)regions). A temporal aggregation is carried out as we are only estimating a yearly price index. In Table 3, three results are presented – three very different labor markets. The first is Stockholm, the capital of Sweden, the second is a medium sized city and the third is a small labor market with very few observations. Based on the number of transactions a monthly index can be estimated for Stockholm (on average 500 observations per month), a quarterly index for Västerås (on average 60 observations per month) and yearly index for Mora (on average 13 observations per month).

The results from these regressions vary a little between the replications because only 90% of the data is used. In the Appendix, the index values can be found for all the regions based on all observations.

Table 3. Regression results (Labor markets of Stockholm, Västerås, and Mora)

	Stockholm		Västerås		Mora	
	coefficients	t-value	coefficients	t-value	coefficients	t-value
Living area	0.3906	85.9	0.4560	25.9	0.3476	6.4
Room 2	0.0770	1.6	–0.1475	–1.8	0.1734	1.2
Room 3	0.1414	3.0	–0.1105	–1.4	0.3334	2.4
Room 4	0.1878	5.9	–0.0501	–0.6	0.3671	2.6
Room 5	0.2049	4.3	–0.0329	–0.4	0.3671	3.0
Room 6	0.2198	4.6	–0.0087	–0.1	0.4669	3.0
Room 7	0.2294	4.8	0.0010	0.0	0.5109	3.3
Room 8	0.2377	5.0	0.0458	0.5	0.4315	2.5
Room 9	0.2500	5.2	–0.0243	–0.3	0.3167	1.6
Room 10	0.2790	5.7	0.0341	0.4	0.7106	3.4
Quality index	0.0132	10.9	0.0585	11.7	0.1427	10.4
Quality index sq.	–0.0001	–7.9	–0.0008	–10.1	–0.0021	–9.3
Sea front	0.3829	23.6	0.6049	14.5	0.5272	4.8
Sea view	0.0917	19.1	0.0362	2.1	0.1219	3.2
Semi-detached	–0.0399	–9.8	–0.0449	–3.8	–0.0765	–1.5
Detached	–0.0634	–12.4	–0.0301	–1.9	–0.1433	–1.4
Building period 1	0.0711	4.2	0.0451	1.2	0.1806	1.2
Building period 2	–0.0117	–1.5	–0.1432	–5.2	–0.0204	–0.3
Building period 3	–0.0755	–9.3	–0.1442	–5.1	–0.0984	–1.2
Building period 4	–0.0798	–10.0	–0.1286	–4.6	0.0140	0.2
Building period 5	–0.0278	0.6	–0.0229	–0.8	0.1145	1.4
Building period 6	0.0061	0.6	0.0034	0.1	0.4257	3.0
Urban	–0.0193	–4.1	–0.0293	–1.5	0.1120	3.1
2006	0.1050	41.5	0.0964	10.4	0.1062	3.1
2007	0.2288	92.0	0.1577	17.2	0.2582	7.5
2008	0.2239	85.3	0.2050	21.8	0.2873	8.2
2009	0.2495	96.1	0.2063	22.2	0.3362	9.1
2010	0.3251	124.6	0.2289	24.2	0.3362	9.1
R^2	0.905		0.880		0.732	
No of obs.	36 158		4 633		912	
No. of obs./month	502		64		13	

Note: Statistics concerning value-area, lot size, coordinates and age are not shown in the table.

The overall goodness-of-fit is good. In the labor market of Stockholm, the estimated model can explain more than 90% of the variation in price. R-square in Västerås is more or less in the same magnitude, but it is lower in the smallest labor market Mora. A possible reason for why the high goodness-of-fit figures obtained, is that we use a sort of weighted least square (WLS) in order to down-weight outliers.

In the first step, we have estimated 66 different hedonic price equations with OLS, one for each labor markets. The Moran's I statistic is on average equal to 26.498 with a standard deviation of 13.233. That is, the result indi-

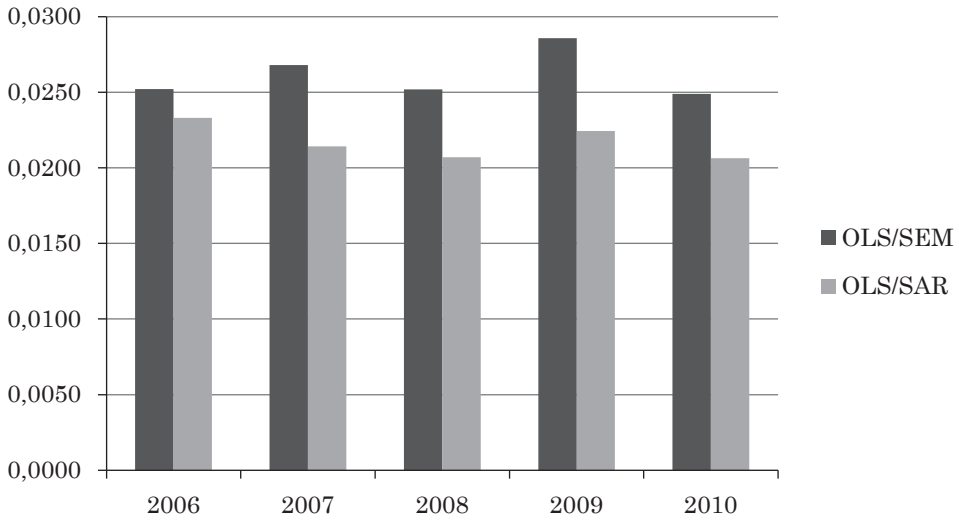

Figure 3. The average difference (absolute value) for each of the 66 labor markets between OLS estimates and SAR and SEM, respectively

cates that spatial dependency is present; we reject the hypothesis of no spatial correlation. However, the question is whether the hedonic house price indexes in each labor market are affected or not. In order to test for bias in the coefficients for the time dummies we have estimated a SEM and a SAR model for each labor market. We have limited the sample size to be maximum 900 observations in each labor market.

In the Figure 3, the average difference (absolute value) between the aggregated house price index and the two spatial regression indexes are displayed. Although there is a presence of spatial dependency in our hedonic house prices models; this seems not to spill over to the price indexes. In fact, the differences between the house price index estimated by OLS, the SEM and the SAR seem not to be significant. The difference is on average around 0.026 between OLS and SEM and 0.022 between OLS and SAR, which is low compared to the coefficient concerning the time dummies.

We have tested whether the differences are statistically significant or not with a Hausman-test. The null hypothesis is that the difference in estimates is equal to zero. The average t-value (absolute value) is estimated to be equal to 0.83 with standard deviation equal to 0.70. That is, on average we cannot reject the null hypothesis of equality in any of the coefficients concerning the time effects and, accordingly, our OLS-estimates can be used in the cluster analysis.

Cluster analysis (step 2)

The results of the cluster analysis are presented in the Table 4 and the Figure 3. Table 4 presents the average number of clusters, and its standard deviation as well as average number of observations for each cluster method is presented. The figures depict the identified clusters on maps.

As can be noted in the Table 4, the average number of clusters for each method varies from 9 to 12 and thereby the average number of observations per cluster.

An interesting result is that we obtain very large differences between the replications of the clustering method. This also occurs even if the data is almost the same. The number of clusters varies a lot, as can be seen in the standard deviation of the number of clusters in Table 4.

Table 4. Results of cluster analysis, methods C1-C6

Cluster method	Average number of clusters	Standard dev. in number of clusters	Average number of observations per cluster (10% out of sample)
C1 (Price development)	9.6	1.9	2 077
C2 (Mean price)	9.5	1.9	2 097
C3 (Price dev. pattern)	9.3	1.8	2 136
C4 (Coordinates)	12.0	1.6	1 656
C5 (Coord. and price dev.)	10.7	2.0	1 843
C6 (All variables)	10.6	1.9	1 867
C7 (Random)	11.0	1.6	1 805
Mean	10.4	1.8	1 926

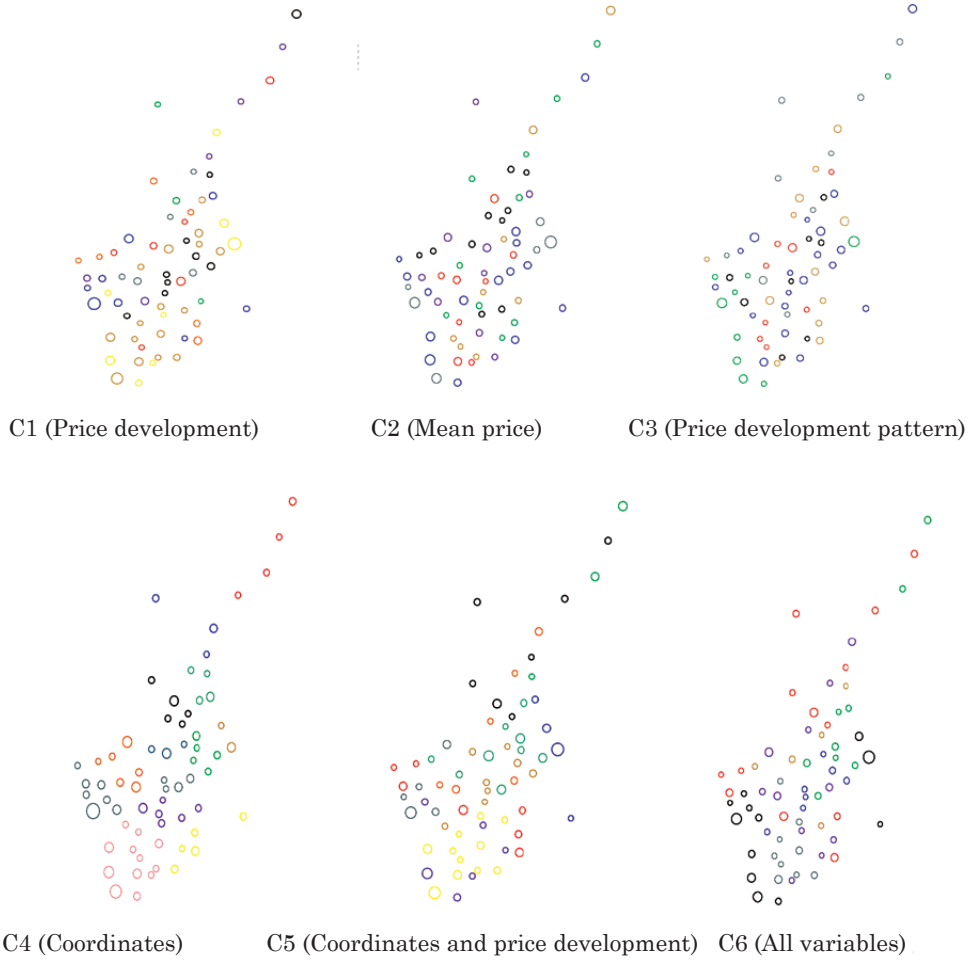

Figure 4. Maps created from clustering methods C1-C6 (cluster variables within parenthesis)

In the diagrams in Figure 4, we can see examples of maps that are created with the different methods. Each cluster is represented with a color. One can see clearly that the method C4, that uses coordinates only, creates clusters based on geographical proximity. It is interesting to see that other clustering commands also creates some clusters with nearby regions, the price development pattern (C3) for instance.

Comparison and evaluation of regional prices indexes (step 3)

In stage three, a hedonic price equation for each cluster is estimated (all the estimates are available upon request). In the Table 5 the root mean squared errors (RMSE) have been estimated. MSE is here defined as:

$$RMSE = \sqrt{\sum_{j=1}^{4}\left[\frac{\sum_{i=1}^{n} e_{ij}^2}{n_j}\right]} \quad i = 1, \ldots, n, j = 1\text{–}4. \quad (2)$$

For the calculation of RMSE we use the 10% out-of-sample data for each of the 100 replications. The price is used on logarithmic scale in the regression, and the RMSE is also measured on this scale.

Results

Table 5 displays the results of the different clustering methods, including C7 (random). Surprisingly, there seems to be no differences at all between the different clustering methods.

We know that the problem with thin markets is bigger in small regions. Maybe we can observe some differences in different types of regions? We have therefore divided the regions into three groups: (1) regions with the least number of sales, totally 10% of the transactions in the dataset or the 25 smallest regions, (2) medium sized regions and the three largest regions, with one third of all the transactions in the dataset. In Table 6 are we comparing the differences in RMSE between the cluster methods and the type of the region.

Even if there are some differences, they are very small and not statistically different from each within each type of region. However, results suggest that the average prediction error is largest in small regions and smallest in large regions. Moreover, we have also noted that there are rather substantial differences between the different replications. The number of clusters varies a lot – even if the data used are almost the same. The average result seem to be almost exactly the same for all the clustering methods, but maybe some methods produce better clusters in some of the replications?

Figure 5 shows the differences between the replications. For each clustering method, the resulting RMSE has been sorted from smallest to largest. The figure shows these sorted results for each method.

The differences are not very big, but the results are clear. The clustering with a random variable seems to be the worst method, even if it is rather stable. Coordinates seem to be a good method, but sometimes we get bad results. Price and price development seem also to be good variables to use. A clear conclusion is that geography is important, but using geographical location only might lead to a bad result. In the Figure 6, we see similar results for the smallest regions (based on number of transactions).

Table 5. Comparison of different clustering methods

Cluster method	Average number of clusters	Standard dev. in nr. of clusters	Average no of observations per cluster (10% out of sample)	Average standard error (out of sample)
C1 (Price development)	9.6	1.9	2 077	0.284
C2 (Mean price)	9.5	1.9	2 097	0.283
C3 (Price dev. pattern)	9.3	1.8	2 136	0.284
C4 (Coordinates)	12.0	1.6	1 656	0.283
C5 (Coord. and price dev.)	10.7	2.0	1 843	0.283
C6 (All variables)	10.6	1.9	1 867	0.283
C7 (Random)	11.0	1.6	1 805	0.285
Mean	10.4	1.8	1 926	0.283

Table 6. Comparison of different clustering methods

Cluster method	Average out-of-sample error				
	Average nr of clusters	All regions	Smallest regions	Medium regions	Largest regions
C1 (Price development)	9.6	0.284	0.339	0.311	0.210
C2 (Mean price)	9.5	0.283	0.337	0.310	0.210
C3 (Price dev. pattern)	9.3	0.284	0.339	0.311	0.210
C4 (Coordinates)	12.0	0.283	0.343	0.309	0.211
C5 (Coord. and price dev.)	10.8	0.283	0.341	0.309	0.210
C6 (All variables)	10.6	0.283	0.338	0.310	0.210
C7 (Random)	11.0	0.285	0.343	0.312	0.212
Mean	10.4	0.284	0.340	0.310	0.211

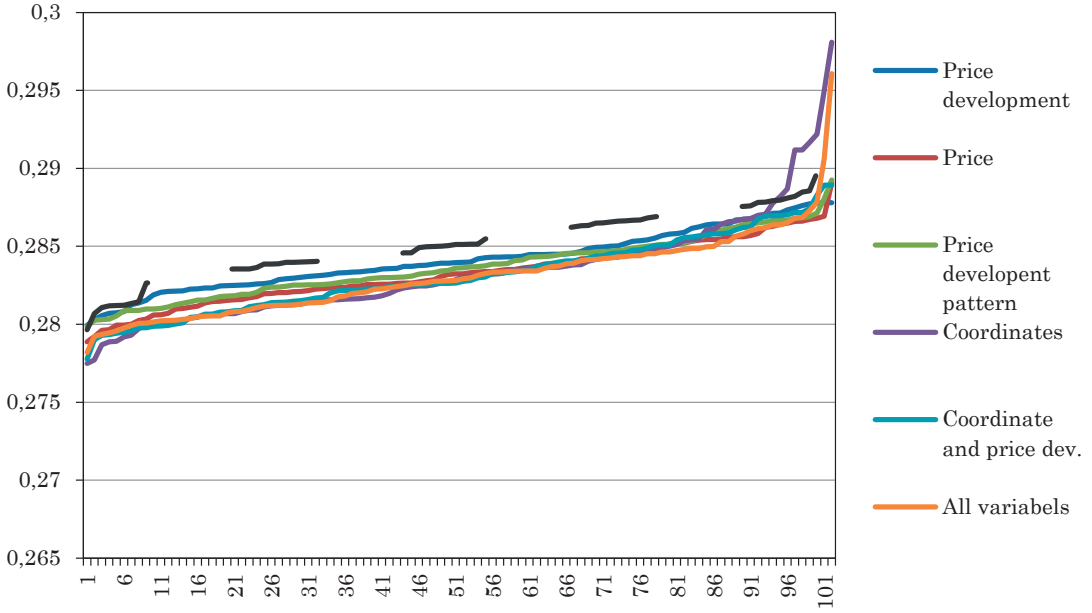

Figure 5. Variation of RMSE between the replications for the clustering methods

We notice the same result – we get some really bad results from using coordinates only. We also note that the differences between the replications are bigger. The average RMSE vary from 0.32 to 0.44.

An alternative out-of-sample

We conduct also an alternative out-of-sample test by removing 30% of the observations from the last year only (2010).

The RMSE is on average somewhat bigger as compared to the original test (see Table 7). The differences are also somewhat larger between the clustering methods. For the smallest regions, the clustering method that uses coordinates (C4); seem to be the worst method, even worse than using a random variable. In order to understand the differences better, we illustrate the RMSE from each replication in Figure 7.

Table 7. Comparison of different clustering methods

Cluster method	Average out-of-sample error				
	Average nr of clusters	All regions	Smallest regions	Medium regions	Largest regions
C1 (Price development)	9,5	0,296	0,357	0,325	0,214
C2 (Mean price)	9,7	0,296	0,355	0,325	0,214
C3 (Price dev. pattern)	9,4	0,296	0,358	0,325	0,214
C4 (Coordinates)	12,0	0,298	0,371	0,324	0,215
C5 (Coord. and price dev.)	10,5	0,296	0,357	0,324	0,214
C6 (All variables)	11,1	0,295	0,358	0,323	0,214
C7 (Random)	10,6	0,299	0,366	0,326	0,216
Mean	10,4	0,296	0,360	0,325	0,214

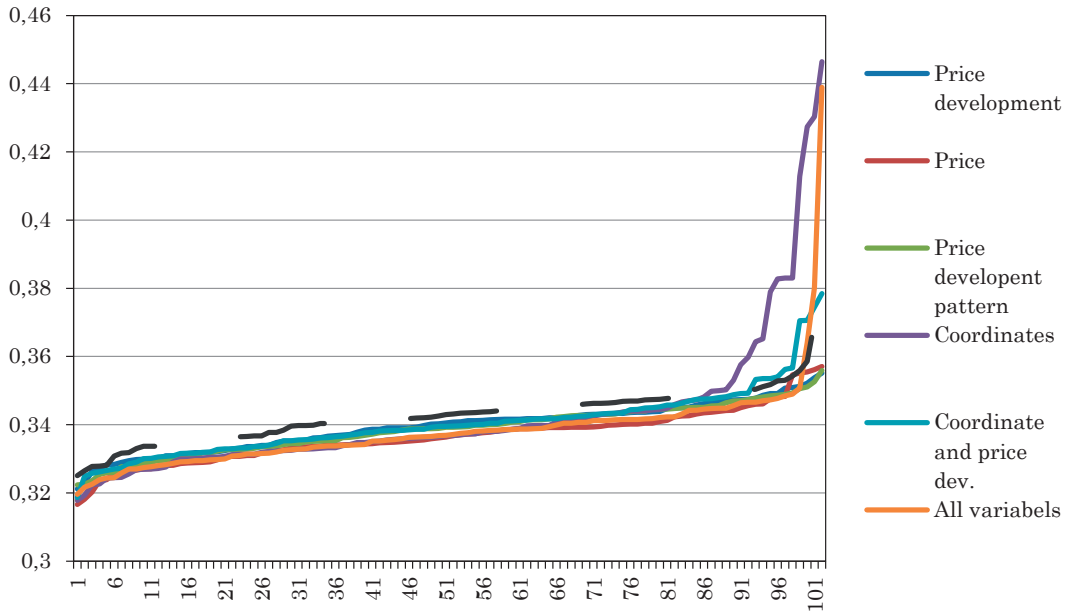

Figure 6. Variation of RMSE between the replications for the clustering methods in the smallest regions

We have some extreme differences in some of the replications. We sometimes get a very high RMSE with many of the clustering methods. We can clearly see that clustering using coordinates only will result in the highest RMSE in all these extreme replications. In Figure 8, we show the same results after sorting after RMSE individually for each clustering method.

5. CONCLUSIONS

Our goal is to create clusters which have large enough number of observations needed for constructing reliable indices based on hedonic regression analysis. However, it may require several clusters to reproduce the different characteristics among the many housing markets. But segmenting the market in too many small clusters creates problems with thin markets.

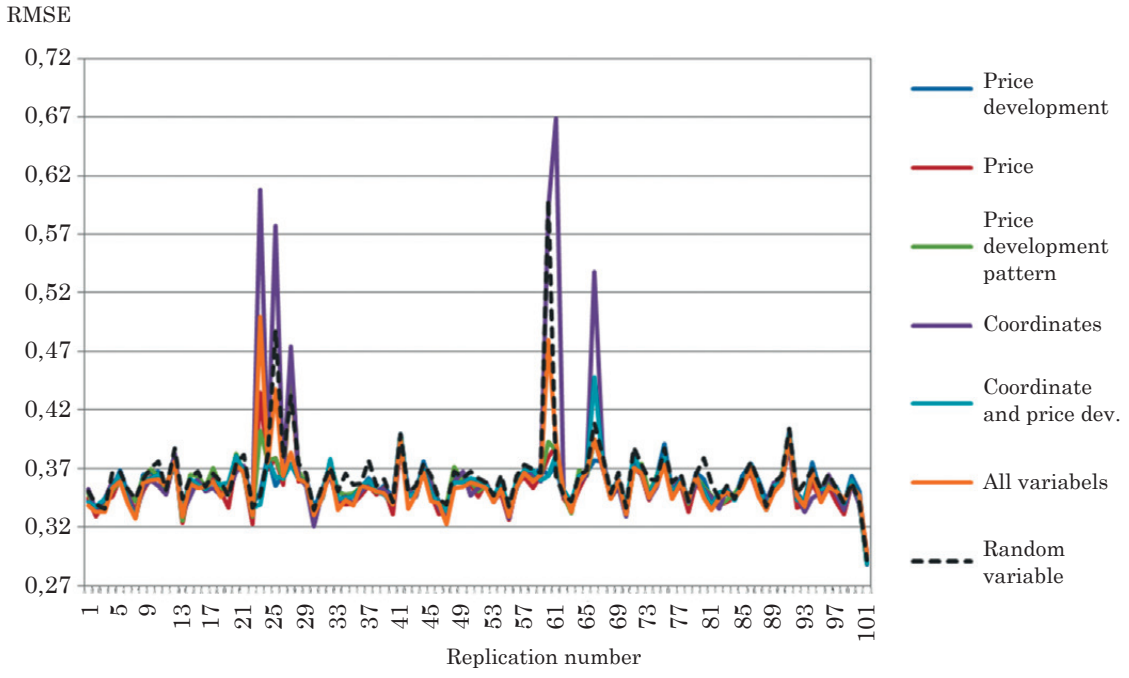

Figure 7. Variation of RMSE between the replications for the clustering methods in the smallest – not sorted

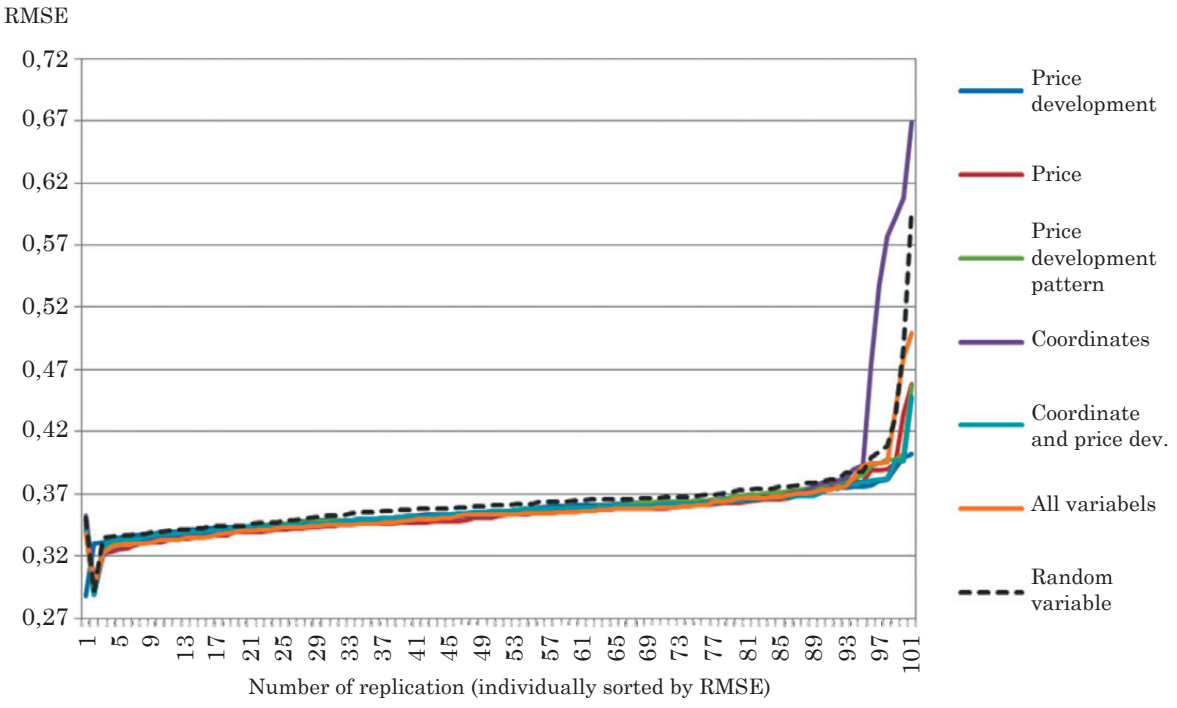

Figure 8. Resembles Figure 6, but in the extreme replications the RMSE is higher

Thus, there is a trade-off between identifying large enough clusters and to identify as many clusters as needed to reflect the heterogeneity among the housing submarkets.

First of all, we have found that the methods we have used for cluster analysis produce very different results, based on very small variations in the used dataset. It would of course have been more satisfying to find a stable method that would always produce good results.

The Swedish housing market is homogeneous in many aspects. The price development between different regions is closely correlated with each other. Furthermore, most people value the same attributes in their homes. Because of this, different regions might work very well in the same regression model.

We have found that geographical proximity is a good variable for clustering regions, but is should not be used alone. The results from this study suggest that regions should be clustered together based on regional price levels and/or price development as clustering variables. If only geographical proximity is used as clustering variable, our computations show that there is a high risk that we end up with some clusters having large standard errors in the regression models, which in turn might result in inaccurate indexes.

The differences are biggest in small regions. Large regions are often the center of their clusters if only geographical proximity is used, while small regions often are clustered with the closest large region. This means that the parameter-estimates in the regression model will not adapt to the small region. If many smaller regions are clustered together, maybe because price level is used in the clustering, the regression model is more likely to adjust to the small regions.

The results are not so clear. There is not one method that would produce stable results with big improvements in standard errors. Even if we could not find such a method, we have made some conclusions and we also find the method interesting. It might be used to test other models for the purpose of clustering regions. In this paper, we have only used information from within the dataset of sold houses. There are many other sources of information that can be used to create other variables to describe the regions such as population density, income level, and unemployment rates and so on.

It could also be interesting in future research to analyze the best and the worst clusters in order to explain the differences. Moreover, we could also test other regions as smallest unit for clustering. In this paper we use functional labor market regions, where each region consists of a number of municipalities. We could do the same study with smaller areas, maybe concentration the study to a part of Sweden, Finally, the same analysis could also be used with longer time series based on another dataset – perhaps price development will be more important for clustering with a longer period. We have worked with a period of six years.

ACKNOWLEDGEMENTS

We gratefully acknowledge the financial support from Nasdaq OMX Nordic Foundation that made this research possible. We also acknowledge the help from Valueguard AB with data and comments during the project. We do also appreciate the helpful comments from two anonymous referees.

REFERENCES

Anselin, L. (1988) *Spatial econometrics: method and models*. Dordrecht: Kluwer Academic Publishers.

Berndt, E. R. and Rappaport, N. J. (2001) Price and quality of desktop and mobile personal computers: a quarter-century historical overview, *American Economic Review*, 91(2), pp. 268–273. http://dx.doi.org/10.1257/aer.91.2.268

Ceccato, V. and Wilhelmsson, M. (2011) The impact of crime on apartment prices: evidence from Stockholm, Sweden, *Geografiska Annaler Series B-Human Geography*, 93(1), pp. 81–103. http://dx.doi.org/10.1111/j.1468-0467.2011.00362.x

Costello, G., Goh, Y. M. and Schwann, G. M. (2009) The accuracy and robustness of real estate price index methods. In: *16th Annual European Real Estate So-*

ciety (ERES) Conference, 24–27 June, 2009, Stockholm, Sweden.

Diewert, W. E., Heravi, S. and Silver, M. (2009) Hedonic Imputation versus time dummy hedonic indexes. In: Diewert, W. E., Greenlees, J. S. and Hulten, C. R. (Eds.). *Price index concepts and measurement*. University of Chicago Press.

Englund, P., Quigley, J. M. and Redfearn, C. (1999) The choice of methodology for computing housing price indexes: comparison of temporal aggregation and sample definition, *Journal of Real Estate Finance and Economics*, 19(2), pp. 91–112. http://dx.doi.org/10.1023/A:1007846404582

Francke, M. K. (2010) Repeat sales index for thin markets, *Journal of Real Estate Finance and Economics*, 41(1), pp. 24–52. http://dx.doi.org/10.1007/s11146-009-9203-1

Francke, M. K. and Vos, G. A. (2004) The hierarchical trend model for property valuation and local price indices, *Journal of Real Estate Finance and Economics*, 28(2-3), pp. 179–208. http://dx.doi.org/10.1023/B:REAL.0000011153.04496.42

Galster, G., Tatian, P. and Pettit, K. (2004) Supportive housing and neighborhood property value externalities, *Land Economics*, 80(1), pp. 33–54. http://dx.doi.org/doi:10.3368/le.80.1.33

Goh, Y. M., Costello, G. and Schwann, G. (2012) Accuracy and robustness of house price index methods, *Housing Studies*, 27(5), pp. 643–666. http://dx.doi.org/10.1080/02673037.2012.697551

Gouriéroux, C. and Laferrère, A. (2009) Managing hedonic housing price indexes: The French experience, *Journal of Housing Economics*, 18(3), pp. 206–213. http://dx.doi.org/10.1016/j.jhe.2009.07.012

Geltner, D. (1997) Bias and precision of estimates of housing investment risk based on repeat-sales indices: a simulation analysis, *Journal of Real Estate Finance and Economics*, 14(1–2), pp. 155–171. http://dx.doi.org/10.1023/A:1007732320832

McMillen, D. P. (2003) Neighborhood house price indexes in Chicago: a Fourier repeat sales approach, *Journal of Economic Geography*, 3(1), pp. 57–73. http://dx.doi.org/10.1093/jeg/3.1.57

Mooi, E. and M. Sarstedt, A. (2011) *Concise guide to market research*. Berlin Heidelberg: Springer-Verlag.

Pakes, A. (2003) A reconsideration of hedonic price indexes with an application to PCs, *American Economic Review*, 93(5), pp. 1578–1596. http://dx.doi.org/10.1257/000282803322655455

Schwann, G. M. (1998) A real estate price index for thin markets, *Journal of Real Estate Finance and Economics*, 16(3), pp. 269–287. http://dx.doi.org/10.1023/A:1007719513787

Song, H.-S. and Wilhelmsson, M. (2010) Improved price index for condominiums, *Journal of Property Research*, 27(1), pp. 39–60. http://dx.doi.org/10.1080/09599916.2010.500394

Wilhelmsson, M. (2002) Spatial models in real estate economics, *Housing, Theory and Society*, 19(2), pp. 92–101. http://dx.doi.org/10.1080/140360902760385646

Wilhelmsson, M. (2009) Construction and updating of property price index series: the case of segmented markets in Stockholm, *Property Management*, 27(2), pp. 119–137. http://dx.doi.org/10.1108/02637470910946426

APPENDIX

Difference between different clustering commands. Results from 100 replications for each clustering, in total 400 replications

Based on an early analysis, where we using the sales prices for calculating RMSE instead of the logarithmic prices, the results differ a little.

Table 8. Results from 100 replications of each clustering, in total 400 replications

Cluster command	Similiarity measure	Varibles used	Average nr of clusters	Average out-of-sample error			
				All regions	Smallest regions	Medium regions	Largest regions
Kmeans	Euclidean	Price development	10,3	0,293	0,368	0,319	0,224
Kmeans	Euclidean	Mean price	9,4	0,296	0,360	0,323	0,224
Kmeans	Euclidean	Price dev. pattern	9,6	0,294	0,376	0,318	0,223
Kmeans	Euclidean	Coordinates	11,6	0,291	0,363	0,316	0,226
Kmeans	Euclidean	Coordinates and development	11,2	0,291	0,366	0,316	0,225
Kmeans	Euclidean	All	11,3	0,293	0,360	0,318	0,224
Kmeans	Canberra	Price development	10,1	0,293	0,369	0,318	0,225
Kmeans	Canberra	Mean price	9,7	0,296	0,359	0,322	0,224
Kmeans	Canberra	Price dev. pattern	9,6	0,293	0,376	0,317	0,223
Kmeans	Canberra	Coordinates	11,6	0,291	0,362	0,315	0,226
Kmeans	Canberra	Coordinates and development	11,3	0,291	0,366	0,315	0,226
Kmeans	Canberra	All	11,1	0,293	0,362	0,317	0,224
Kmedians	Euclidean	Price development	10,4	0,293	0,371	0,318	0,225
Kmedians	Euclidean	Mean price	9,8	0,296	0,362	0,322	0,224
Kmedians	Euclidean	Price dev. pattern	9,5	0,294	0,376	0,318	0,224
Kmedians	Euclidean	Coordinates	11,9	0,291	0,365	0,315	0,226
Kmedians	Euclidean	Coordinates and development	11,2	0,292	0,369	0,316	0,226
Kmedians	Euclidean	All	11,5	0,293	0,363	0,317	0,224
Kmedians	Canberra	Price development	10,1	0,293	0,368	0,319	0,224
Kmedians	Canberra	Mean price	9,6	0,296	0,359	0,322	0,223
Kmedians	Canberra	Price dev. pattern	9,6	0,293	0,374	0,317	0,223
Kmedians	Canberra	Coordinates	11,6	0,291	0,361	0,316	0,226
Kmedians	Canberra	Coordinates and development	11,1	0,290	0,365	0,315	0,225
Kmedians	Canberra	All	11,4	0,292	0,359	0,317	0,224
All	All	Price development	10,3	0,293	0,369	0,318	0,225
All	All	Mean price	9,6	0,296	0,360	0,322	0,224
All	All	Price dev. pattern	9,6	0,293	0,376	0,317	0,223
All	All	Coordinates	11,7	0,291	0,363	0,315	0,226
All	All	Coordinates and development	11,2	0,291	0,366	0,315	0,226
All	All	All	11,3	0,293	0,361	0,317	0,224

Table 9. Descriptive statistics concerning all labor markets (see Figure 9)

Code	Region	Observ.	Mean price	Excluded
1a	Stockholm	34 456	3 103 882	
1b	Södertälje	3 326	2 118 589	
1c	Uppsala	6 668	2 082 493	
2	Nyköping	2 034	1 685 633	
3a	Eskilstuna	1 483	1 597 988	
3b	Katrineholm	1 740	1 061 708	
4a	Linköping	3 896	1 951 823	
4b	Norrköping	2 761	1 619 325	
4c	Mjölby	834	1 104 919	
4d	Motala	962	1 231 990	
5a	Värnamo	700	1 133 092	
5b	Gislaved	1 007	784 678	
6a	Jönköping	3 766	1 682 044	
6b	Nässjö	1 372	922 056	
7	Vetlanda	1 177	749 914	
8	Tranås	666	967 926	
9	Älmhult	745	990 424	
10	Ljungby	1 041	1 021 285	
11	Växjö	2 755	1 247 018	
12a	Kalmar	2 843	1 563 330	
12b	Nybro	711	782 314	
13	Vimmerby	909	623 367	
14	Västervik	1 241	1 076 651	
15	Oskarshamn	1 470	921 444	
16	Gotland	1 371	1 630 159	
17a	Karlskrona	2 248	1 300 121	
17b	Karlshamn	1 144	1 017 292	
18a	Kristianstad	4 512	1 177 188	
18b	Sölvesborg	1 186	1 092 259	
19a	Malmö	14 292	2 479 303	
19b	Ystad	1 985	1 640 033	
19c	Helsingborg	8 941	1 852 158	
20	Halmstad	4 864	1 699 047	
21a	Göteborg	21 653	2 558 446	
21b	Alingsås	1 582	1 594 485	
21c	Stenungsund	1 145	2 075 495	
22	Borås	3 513	1 336 042	
23a	Trollhättan	2 015	1 287 018	
23b	Uddevalla	2 293	1 625 087	
24	Lidköping	1 504	1 148 439	
25a	Skövde	3 479	1 084 560	
25b	Mariestad	1 039	943 587	
26	Strömstad	548	1 735 983	
27	Bengtsfors	526	695 645	
28	Årjäng	195	919 329	Y
29	Eda	233	686 928	Y
30a	Karlstad	5 367	1 244 361	
30b	Säffle	874	753 634	
31	Torsby	255	757 171	Y
32	Hagfors	326	539 888	Y
33	Filipstad	10	791 000	Y
34	Örebro	5 538	1 404 929	
35	Hällefors	37	470 135	Y
36	Karlskoga	1 432	764 571	
37a	Västerås	4 631	1 764 276	
37b	Köping	825	1 048 914	
38	Fagersta	528	636 679	
39	Vansbro	49	436 673	Y
40	Malung	151	643 291	Y
41	Mora	899	926 234	
42	Falun/ Borlänge	3 832	1 209 143	
43	Avesta	726	714 912	
44	Ludvika	938	715 389	
45a	Gävle	2 705	1 385 628	
45b	Sandviken	1 552	941 863	
46a	Söderhamn	601	692 818	
46b	Bollnäs	1 007	691 610	
47	Hudiksvall	823	897 142	
48	Ljusdal	327	720 813	Y
49	Sundsvall	3 427	1 091 745	
50	Kramfors	179	482 626	Y
51	Sollefteå	253	520 375	Y
52	Örnsköldsvik	1 522	932 072	
53	Östersund	1 579	1 481 488	
54	Härjedalen	59	585 593	Y
55	Storuman	49	673 014	Y
56	Lycksele	207	678 601	Y
57	Dorotea	3	736 667	Y
58	Vilhelmina	20	517 000	Y
59	Åsele	6	365 833	Y
60	Sorsele	2	250 000	Y
61	Umeå	2 682	1 653 478	
62	Skellefteå	1 457	967 047	
63	Arvidsjaur	37	556 486	Y
64	Arjeplog	11	395 227	Y
65	Luleå	4 761	1 005 182	
66	Överkalix	26	323 173	Y
67	Övertorneå	23	476 891	Y
68	Haparanda	88	765 210	Y
69	Pajala	3	465 000	Y
70	Jokkmokk	2	497 500	Y
71	Gällivare	200	899 954	Y
72	Kiruna	266	943 579	Y
	Total	209 126	1 859 967	

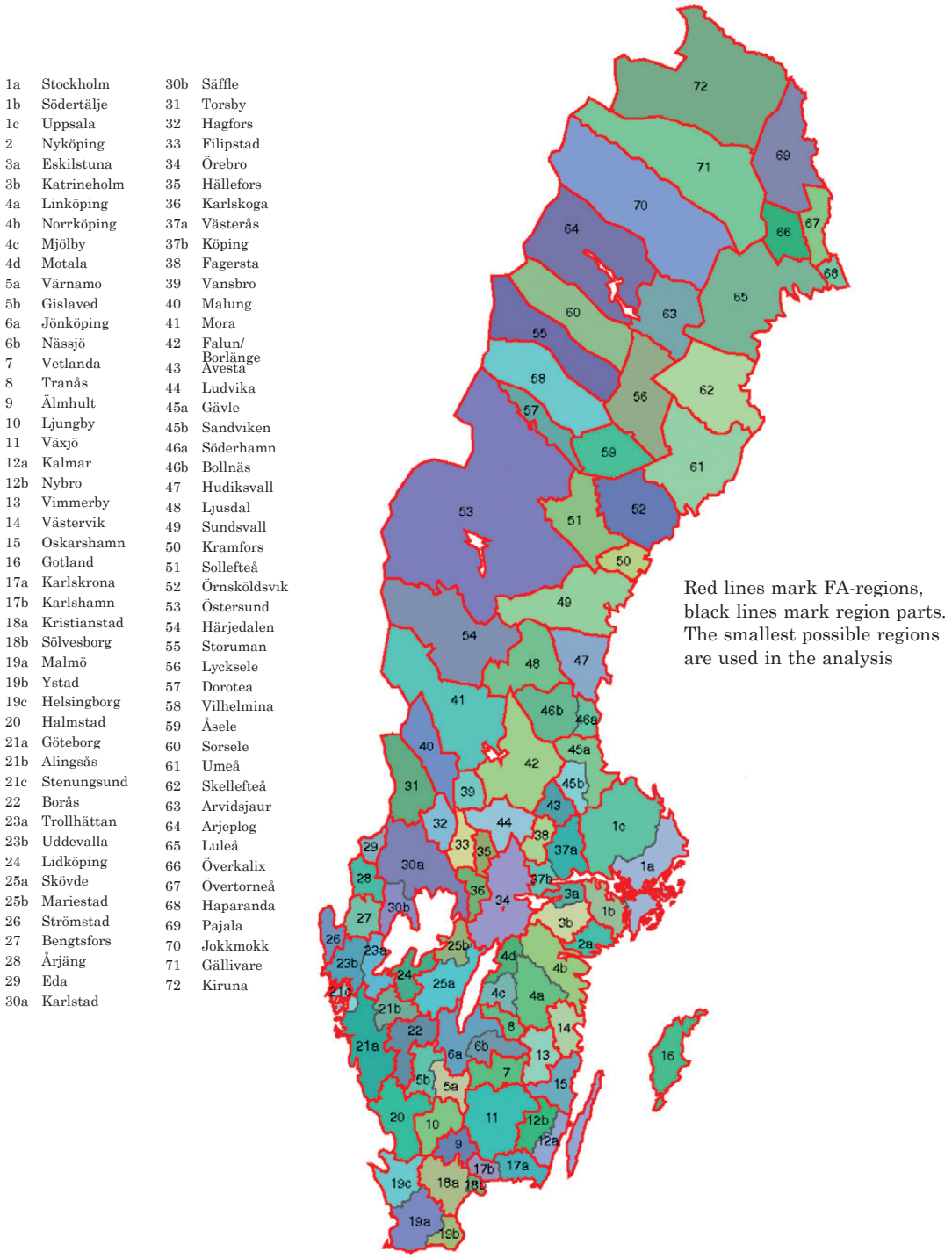

Figure 9. Map of FA-regions and region parts
Note: The mean price is calculated from all observations 2005 to 2010
Source: Swedish Agency for Economic and Regional Growth http://www.tillvaxtverket.se/ http://www.tillvaxtverket.se/download/18.21099e4211fdba8c87b800035198/FA-regioner_med_delregioner2.pdf

Table 10. Yearly price index series for each region

Code	Name	2005	2006	2007	2008	2009	2010
	Stockholm	100	111	126	125	128	138
1b	Södertälje	100	107	117	119	121	127
1c	Uppsala	100	107	116	119	125	135
2	Nyköping	100	106	114	115	119	120
3a	Eskilstuna	100	114	125	127	129	130
3b	Katrineholm	100	108	115	117	116	118
4a	Linköping	100	107	112	113	120	124
4b	Norrköping	100	107	121	124	130	136
4c	Mjölby	100	110	122	121	119	120
4d	Motala	100	113	118	120	119	125
5a	Värnamo	100	109	119	124	129	123
5b	Gislaved	100	108	117	117	116	122
6a	Jönköping	100	108	118	124	131	142
6b	Nässjö	100	109	116	118	123	127
7	Vetlanda	100	110	124	130	132	133
8	Tranås	100	112	114	114	124	120
9	Älmhult	100	110	122	128	122	121
10	Ljungby	100	105	124	126	128	122
11	Växjö	100	111	120	122	123	129
12a	Kalmar	100	112	124	130	138	144
12b	Nybro	100	115	130	134	136	131
13	Vimmerby	100	97	112	120	120	120
14	Västervik	100	112	124	133	136	151
15	Oskarshamn	100	103	113	125	131	139
16	Gotland	100	110	115	118	122	134
17a	Karlskrona	100	108	115	120	123	127
17b	Karlshamn	100	110	124	127	123	127
18a	Kristianstad	100	108	117	118	123	127
18b	Sölvesborg	100	110	119	125	127	136
19a	Malmö	100	112	123	119	123	130
19b	Ystad	100	113	125	124	128	139
19c	Helsingborg	100	113	126	128	130	136
20	Halmstad	100	112	125	123	125	130
21a	Göteborg	100	110	120	120	124	134
21b	Alingsås	100	108	126	124	126	139
21c	Stenungsund	100	114	118	125	122	135
22	Borås	100	110	123	125	125	132
23a	Trollhättan	100	111	125	127	127	131
23b	Uddevalla	100	117	130	132	136	144
24	Lidköping	100	110	124	121	132	137
25a	Skövde	100	112	130	133	132	138
25b	Mariestad	100	103	113	123	124	123
26	Strömstad	100	105	119	129	132	140
27	Bengtsfors	100	104	126	132	133	144
28	Årjäng	–	–	–	–	–	–
29	Eda	–	–	–	–	–	–
30a	Karlstad	100	108	118	124	126	131
30b	Säffle	100	109	115	114	119	126
31	Torsby	100	114	131	134	125	131
32	Hagfors	100	100	103	106	107	107
33	Filipstad	–	–	–	–	–	–
34	Örebro	100	110	118	124	124	125
35	Hällefors	–	–	–	–	–	–
36	Karlskoga	100	107	117	120	122	120
37a	Västerås	100	110	117	123	123	126
37b	Köping	100	113	124	122	124	121
38	Fagersta	100	109	121	119	131	121
39	Vansbro	–	–	–	–	–	–
40	Malung	–	–	–	–	–	–
41	Mora	100	111	129	133	139	140
42	Falun/ Borlänge	100	113	124	131	138	149
43	Avesta	100	114	127	142	140	140
44	Ludvika	100	108	115	122	133	136
45a	Gävle	100	108	119	123	126	133
45b	Sandviken	100	112	123	128	127	129
46a	Söderhamn	100	105	116	123	115	116
46b	Bollnäs	100	111	119	126	133	135
47	Hudiksvall	100	116	126	139	136	142
48	Ljusdal	100	94	121	113	133	129
49	Sundsvall	100	110	119	124	129	135
50	Kramfors	100	102	100	129	107	118
51	Sollefteå	100	113	120	115	121	120
52	Örnsköldsvik	100	119	135	138	146	138
53	Östersund	100	119	133	140	142	153
54	Härjedalen	–	–	–	–	–	–
55	Storuman	–	–	–	–	–	–
56	Lycksele	–	–	–	–	–	–
57	Dorotea	–	–	–	–	–	–
58	Vilhelmina	–	–	–	–	–	–
59	Åsele	–	–	–	–	–	–
60	Sorsele	–	–	–	–	–	–
61	Umeå	100	110	120	118	121	127
62	Skellefteå	100	110	131	134	133	140
63	Arvidsjaur	–	–	–	–	–	–
64	Arjeplog	–	–	–	–	–	–
65	Luleå	100	106	114	117	115	122
66	Överkalix	–	–	–	–	–	–
67	Övertorneå	–	–	–	–	–	–
68	Haparanda	–	–	–	–	–	–
69	Pajala	–	–	–	–	–	–
70	Jokkmokk	–	–	–	–	–	–
71	Gällivare	–	–	–	–	–	–
72	Kiruna	100	111	134	126	131	149

"–" means that the number of observations is not enough for calculating an hedonic index. These regions are not included in the clustering analysis.