ERSA Conference August 23 – 27 2005, Amsterdam, The Netherlands

Spatial and multidimensional analysis of the Dutch housing market using the Kohonen Map and GIS

Tom Kauko & Roland Goetgeluk (OTB, Delft University of Technology)

Abstract

In this work the idea is to analyse general spatially identifiable housing market related data on Dutch districts (wijken) with the SOM (self-organizing map, Kohonen Map) and a GIS. The SOM is a neural network technique, which is suitable for data mining. The goal of data mining is to inductively identify structures in complex datasets. However, the SOM can also be applied in a more deductive way, in which clusters are assigned on the basis of *a priori* knowledge of the assumed latent structuring factors given the set of input variables used. One of the authors has earlier carried out purely visual SOM analysis of that data, where patterns formed on a larger 'map' (the output matrix of the SOM) were used as a basis for classification of the Dutch housing market segments on a nationwide level. This way the SOM was used as a method for exploratory data analysis. The outcome of the visual analysis was a classification into five categories: we call these urban, urban periphery, pseudo-rural, traditional, and low-income segments. Now we attempt a more rigorous method of determining the segmentation using a smaller 'map' size, in order to be able to export the SOM-output directly into a GIS-system to analyse it further. Two technical issues interest us: one, the robustness of the results – do the five basic housing market segments found in the earlier analysis prevail when the map size is changed; and two, which classes fit the real situation better and which worse, when using the RMSE for a measure of goodness? While the perspective of the paper is a technical one, we also keep an eye on policy implications and aim at comparing our classifications with the 'actual' ones used in official discourse on urban regeneration.

Paper no. 91
Session Q

Contact address:

Tom Kauko
OTB Research Institute for Housing, Urban and Mobility Studies
Delft University of Technology
P. O. Box 5030
2600 GA Delft
The Netherlands
Email: Kauko@otb.tudelft.nl

Spatial and multidimensional analysis of the Dutch housing market using the Kohonen Map and GIS

1. Introduction

Valid and reliable data together with sound statistical techniques do matter in the process of monitoring and analysing residential patterns and housing market diversity. Normally a set of variables are selected and given meanings, using data-reduction (PCA and factor-analysis) and/or clustering methods. Our contribution aims at a coarse classification of statistical housing market districts in the Netherlands based on district (wijk) level socio-demographic data related to spatial housing market structures. There are many techniques available for this. In this work the idea is to analyse spatially identifiable data with the Kohonen Map (self-organizing map, SOM), and a GIS (Geographical information system), with particular focus on urban restructuring. The emphasis of this paper is on demonstrating the general method, that is to say, the way the SOM and the GIS are combined for this purpose, rather than on the empirical evidence.

The Kohonen Map is a neural network technique that can be interpreted as a close relative to the established *k*-means clustering technique. One of the authors has earlier carried out purely visual SOM analysis of that data, where patterns formed on a larger 'map' (the output matrix of the SOM) were used as a basis for classification of the Dutch housing market segments on a nationwide level (Kauko, 2005). This way the SOM was used as a method for exploratory data analysis. In the mentioned study, the visual analysis of the SOM output resulted in five clusters with permeable boundaries: we call these urban, urban periphery, pseudo-rural, traditional, and low-income segments (see Kauko, 2005). In a follow-up study we now attempt at a more rigorous method of determining the segmentation using the same dataset, but a smaller 'map' size, in order to be able to export the SOM-output directly into a GIS-system to analyse it further. In this setting a number of aspects interest us:

- On the basis of current theory and prior evidence (part of which is anecdotal), we expect three or four basic dimensions: urban/density, economic position, and other composition of socio-demographic (i.e. socio-economic and demographic) factors.

- We generate new knowledge based on *the SOM-clustering* (feature map) of the district-wise aggregated data, together with *the residual mean squared errors (RMSE)* of each observation (i.e. deviation from the model calculations). These spatial relationships will be illustrated with colours for each district in the country.

- If the feature map is too small there is not sufficient resolution; thus, the question arise, whether the *a priori* determined dimensions of the SOM are large enough for resolution to discern all five groups? In such a case the map size has to be enlarged, which is likely to change the multi-dimensional relationships of the input, but not completely distort the structure in terms of the main organising features.

The focus of the results will primarily be on the method and techniques, with a convenient point of reference in previous research applying factor ecological and other quantitative approaches of social area analysis. In a secondary sense, the spatially identifiable SOM-output is related to theory, within the paradigm of social-ecology, and to policy, within the discourse of urban restructuring in the Netherlands. For this last aspect we are interested in to what extent certain assigned neighbourhoods for urban restructuring fit into the SOM-clustering. If

such a fit is achieved, we can in principle also identify other potential neighbourhoods for urban restructuring within these clusters. If such potential cases that are comparable with the actual ones are found, this method can, albeit cautiously, be used in an *ex-ante* cost estimation of restructuring for municipalities and the national government.

The paper is structured in six sections. After this brief introduction, section 2 describes the context of urban restructuring in the Netherlands. In section 3 we present an overview of the methods and techniques of classification. We will also discuss the pros and cons of the SOM-technique. In section 4 we will discuss the data used, the way we applied the SOM and the way we post-processed the results with statistical and spatial tools. Since new techniques can be accepted only if the results – in this case spatial patterns – match existing knowledge, we have applied well-known data of the Central Bureau of Statistics (Kerncijfers Wijken and Buurten, KWB, 1999). In this section we present the results in figures and geographical maps. I section 5 we show to what extent the neighbourhoods selected for urban regeneration fit within our clustering results. In section 6 we discuss the usefulness of the SOM-technique within this problem field.

2. Urban restructuring

The analysis fits into a range of studies that deal with residential mobility, housing choice, housing policy and the dynamics of cities and their neighbourhoods (e.g. Dieleman & Wallet (2003; Goetgeluk & Musterd, 2005). Many of the studies have a direct relationship with the urban restructuring in which physical and social investments are made in order to keep neighbourhoods economically and socially vital. This has resulted in a national program for urban restructuring.

In 1995 fifteen large cities and the national government signed a covenant, which was the basis for *the Big Cities Policy (Grote Steden Beleid, GSB).* The GSB resulted in an inventory of thirty cities (Grote 30 or G30). The GSB aims to improve the economic competitive power of cities and to restrict socio-economic and ethnical divisions within cities. In socio-economic sense urban restructuring should lead to a good mix between bonding and bridging capital between its inhabitants and entrepreneurs. The strategy is based on three pillars: (1) physical improvements (emphasis on urban restructuring), (2) economical improvements (entrepreneurship and labour) and (3) social improvements (education, liveability, safety and care). The elaboration of the first pillar resulted in 1997 in *the National Program for Urban Renewal*. Rather confusing, however, is that its goal and instruments are more or less equal to the GSB targets and pillars (www.kei-centrum.nl).

As expected, urban restructuring or renewal is not an easy task. In 2003 the Ministry of Housing, Spatial Planning and the Environment formulated a *Program for Action Restructuring (Actieprogramma Herstructuering)*. In this program 56 neighbourhoods were assigned. These neighbourhoods are assigned to improve the planning and negotiation processes between the various stakeholders, and also to improve the vacancy chain on the housing market. This last aspect refers to assumption that primary strategic supply of new dwellings (diversity according to tenure, type and so forth) generates housing moves in the existing stock. The increased residential mobility should lead to a different composition of the stock and households, and further to reduce the spatial segmentation, which in its turn should facilitate bonding and bridging capital, investments and so on (see appendix 1).

In this case the budget for urban restructuring is limited, while the number of neighbourhoods that need to be restructured is large. Therefore selections have to be made. Here urban restructuring and renewal is based on a selection of criteria that fit into the three pillars. Especially those neighbourhoods that have multiple negative scores on the indicators of the three pillars are in favour. Neighbourhoods were selected based on statistical analysis, (political) negotiations between various stakeholders, such as the National government, municipalities, housing corporations and so forth. In the selection of neighbourhoods various discussions lead to the criticism from the municipalities that they were omitted from the list. Some of the protests were rewarded, while others were not. Thus, the selection of statistical data matters as well as the methods for ordering the neighbourhoods from being targeted for 'immediate renewal' to 'renewal prospects on the long term.' At the same time the progress made in these neighbourhoods is monitored, partially with spatial and statistical data.

3. Methods and techniques of classification and data

In the recent literature of spatial economics and regional science, we note various research traditions where the housing markets or the residential patterns have been classified across a large set of regions within one and the same country. Especially in the Netherlands this tradition is strongly rooted in the research programmes of the main institutes and government bodies. Dieleman & Wallet (2003) applied Ward's hierarchical cluster algorithm for their analysis of city-suburb differences in income for 24 Dutch regions 1946-94, and noted the following groups:

(1) The three largest metropolitan regions (*declining good areas*): Amsterdam, Rotterdam, and Utrecht; plus five medium sized cities: Arnhem (the only one which is outside Randstad), Dordrecht, Amersfoort, Alkmaar and Leiden). These were wealthy in the 60s (also the central city), until the era of growth centres began in the late 1960s; this prompted an outflow of population from central cities in this group, until the government acted to alter this trend by launching the compact cities policy in the 1980s. Consequently this outflow slowed down.
(2) The Hague, Haarlem and Hilversum (*status quo top areas*) contained already wealthy suburban areas in the 1940s-50s. The average income in these regions is also somewhat higher than in group 1.
(3) The rest of the regions (cities outside Randstad, *improving bad areas*) were in the 1950s-60s surrounded by agricultural municipalities, with a modest average income compared to the national average figure. In the 1970s suburbanisation had already increased the income level of the suburban municipalities above the Dutch average. In the cities the income remained stable in relation to the Dutch average.

Taltavull de La Paz (2003) modelled inter-urban house prices using panel data of 71 Spanish cities (provincial capitals plus others with more than 100,000 inhabitants), and the years 1989-99 (resulting in 780 observations) using 17 variables of socioeconomic and demographic indicators. Using generalised least squares estimation, and model heterogeneity of areas/region based on coefficients, she found some evidence that prices are related to the income, the population and the production structure.

Other contributions worth mentioning are inter-metropolitan hedonic regression modelling within urban economics by Izraeli (1987), and Potepan (1996), respectively from the US; econometric modelling by Tu (2000) from Australia; spatial interpolation and statistics for

regional level analysis of housing market structure and dynamics by Meen (2001) from the UK; and factor ecology based area analysis by Siikanen (1992) from Finland; and Wong (2001) from England.

The Kohonen Map fits well here too. According to the basic idea of the SOM-algorithm an *n*-dimensional input data matrix, where *n* is the number of variables, is compressed to an output, where the array of nodes comprises two dimensions and numerical values of each node (feature map). Each *layer* in the feature map represents one variable. The array of nodes may form *patterns and clusters*, and for every node there are *'typical values'* describing the variation of the data set. Also the empty nodes (i.e. the nodes which do not 'win' any observations and remain without label) obtain an estimate of their 'typical values'. (Kohonen, 1995)

The operation and processing of the SOM is briefly explained in the beginning of next section. We can note that the technique itself is easy to use, as it like other neural network techniques 'eats' all kinds of data, and even allows for missing entries in the input matrix. On the other hand, it is really when the output-matrix is obtained that the analysis begin in earnest. Then the task is to make sense of the resulting patterns, cluster and estimates, and try to relate them to relevant external knowledge – either to empirical context or more general theoretic considerations. We need therefore expectations about the phenomenon under study, and for that, we always need to be acquainted with the context.

For prior work, the most relevant is probably the geo-demographic classification work by Openshaw et al. (1994), which uses the SOM on British census data. Most recently, Hatzichristos (2004) carried out a demographic classification of Athens, Greece, using the SOM algorithm in combination with Fuzzy logic, much following the pioneering work by Openshaw and others in the 1990s. The starting point for this study is in prior studies on clustering the housing market areas/regions of Finland by Kauko (1997, 2002, and of The Netherlands and Hungary, respectively by Kauko (2005).

The choice for conducting the analysis involves a trade-off between the pros and cons of the SOM and the more traditional techniques, when the goal of the computing is a match between the expected clusters and those determined by the clustering routine. Which method produces the most valid classification or estimation, given the available resources. The Kohonen Map is a more pragmatic and holistic method than the more traditional methodologies; however, it has been criticised for being too much of a black box approach (see Kauko, 2002, for a profound evaluation of the method).

4. The analysis and display of results

4.1. The research strategy and the data

To reiterate our research aims, first we use the SOM for the classification of the Dutch housing market on the basis of general spatially identifiable housing market related data on Dutch districts. The input variables are listed in Table 1 (for the SOM processing each field range was transformed into the interval 0-1). We define as small map dimensions x and y as possible in order to retain tractability and a deterministic clustering element – defining a larger map increase the resolution and makes the analysis more fuzzy. We also need to define the learning process in terms of running time, neighbourhood function, learning rate and

radius of the area to be affected for each new input that contributes to the organization process. This is a simulation approach, where the map learns by trial-and-error until the organization has reached a sufficient degree of stability. In this process, the dimensions of the map and the other network parameters are compulsory and more or less *ad hoc* choices to make by the analyst in this application, which is based on the SOM_PAK software (See Kohonen et al., 1996). Second, we illustrate the SOM output, the feature map, as a geographical map. This is possible by exporting the SOM output to a GIS. In other words, the topological context of the SOM will be transformed into the 'real' context of geographical relations, as defined by the KWB district data-set of 2382 observations. The following questions arise:

- Where do we find high/low values (as estimated by the SOM) on the selected twenty dimensions related to urban density and socio-demographic indicators?

- Which 'wijken' fit particularly well/poorly into this model in terms of RMSE?

Table 1: The variables used

Aggregated/mean values for each 'wijk'	*1999-data*
(1) Addresses per neighbourhood (density-proxy)	1 – 11856
(2) Extent of urbanisation	Classification 1: highly urban; 5: least urban
(3) Population density	0 – 31001 inhabitants per sq.km
(4) Percentage of 0-14 years old children	1 – 65
(5) Percentage of 15-24 years old	2 – 96
(6) Percentage of 25-44 years old	3 – 76
(7) Percentage of 45-64 years old	1 – 71
(8) Percentage of 65+ years old	1 – 98
(9) Percentage of non-westerners (1th and 2nd generation immigrants)	0 – 89
(10) Percentage of 1-person households	3 – 99
(11) Number of families	0 – 141280
(12) Percentage of families with kids	7 – 93
(13) Average family size	2.1 - 4.6
(14) Percentage of low income takers	12 – 95
(15) Percentage of high income takers	4 – 74
(16) Percentage of 15-65 years old with unemployment benefit as the primary source of income	0 – 95
(17) Assessed market value of dwelling (total price, 1000 NLG)	43 – 1170
(18) Percentage of industrial enterprises (including construction)	0 – 63 %
(19) Percentage of commercial enterprises	24 – 99 %
(20) Percentage of non-commercial enterprises	0 – 61 %

4.2. The small feature map

We selected a 3 by 2 map of wijk-level information as a starting point. The data used for the SOM analysis is the same as in the study reported in Kauko (2005). However, instead of a matrix of 192 nodes the data was now transformed onto a matrix of six nodes. Then these six nodes were labelled based on the original observations. According to the principle of calibration the feature map, each neuron is named after the most typical district captured by

that neuron, or in case of no 'hits' for that neuron, it remains without a label. This way, we used the original observations as labels to calibrate the output with, when the corresponding nodes (i.e. output of the original analysis) where used as input for the GIS analysis. The second feature of this method was to use the RMSE of the analysis as a measure of fit between each node and its corresponding input.

We tried to use the same interpretation as in Kauko (2005), where the Netherlands housing context was segmented into five different area types: truly urban; urban periphery; pseudo-rural; traditional rural (possibly agricultural); and low-income (possibly disadvantaged). This kind of correspondence we did not however achieve, which is not surprising as the number of nodes were reduced to six from 192. Thus compared to the original analysis the cluster structure was different, which meant that the meaning of the clusters became less clear. The SOM had arrived at the following classification of nodes (note the hexagonal shape of the SOM array):

(0,0) 1400: urban; bad fit	(1,0) 300: low-income; bad fit	(2,0) 700: traditional; good fit
(0,1) 1000: pseudo-rural; good fit	(1,1) 500: no typical feature in the context of the prior analysis; bad fit	(2,1) 702: no typical feature in the context of the prior analysis; good fit

Here the first figure(s) in brackets denote the x- and y- coordinates of the node in the 3 by 2 map; the second figure is the label of the typical district, identified in terms of district code; the remaining description characterise the typical features of that cluster, and the fit between the actual values and computed typical values for each of the six categories (define via the RMSE statistic), in terms of the 20 dimensional space defined by the input variables.

We can see that the map is organised based on two basic features: the first is the urban/rural and density balance, which is also associated with many of the socio-economic and demographic features, and the shares between commercial, non-commercial and industrial enterprises; the second is the input variable 14 – the share of poor people is low elsewhere, and high in the cluster (1,0).

We could also find that some of the nodes have a valid interpretation with the real situation, whereas others proved to have more problematic, complex and confusing interpretations. Only two nodes seemed clear to us: (0,0) is truly urban including a high share of elderly people, non-western immigrants, and single person households. Node (1,0) in turn contains the 'low-income' category of moderately urban areas, thus here is a rationale for policy suggestions – such as urban renewal. The remaining four nodes seem trickier to name: (2,0), (1,1) and (2,1) are rural, sparse, family-areas, with high income and property values. They are qualitatively similar, only the magnitudes of the 20 variables differ: of these three nodes, the corner node (2,1) is the one with highest typical values for relevant determinants. The SOM actually often places the observations with the highest and lowest magnitudes for the variables in the corners. Node (0,1) is completely average in every possible way and could be interpreted as 'pseudo-rural' areas.

An alternative way of presenting the map is displayed in Figure 1. By using grey-scales the heterogeneity and homogeneity of the structure may be illustrated for various parts of the 'map', when distances between the vectors of the nodes are interpreted as a degree of similarity in terms of the measured dimensions. Thus, the further the neighbouring nodes are from each other (i.e. the lighter the colour), the greater the dissimilarity between them, and the closer the neighbouring nodes are to each other (i.e. the darker the colour), the greater the similarity between them. This is shown in Figure 1. The dark map units (labels have white text) in the upper middle row, lower left corner and lower middle row are particularly similar to each others, and the two corner neurons on the right side of the map particularly dissimilar to each other, and to the units in the centre of the map. This implies a considerable homogeneity in the centre and the lower left of the map, and a heterogeneity on the right side of the map where the least urban (and also best fitting) cases are situated.

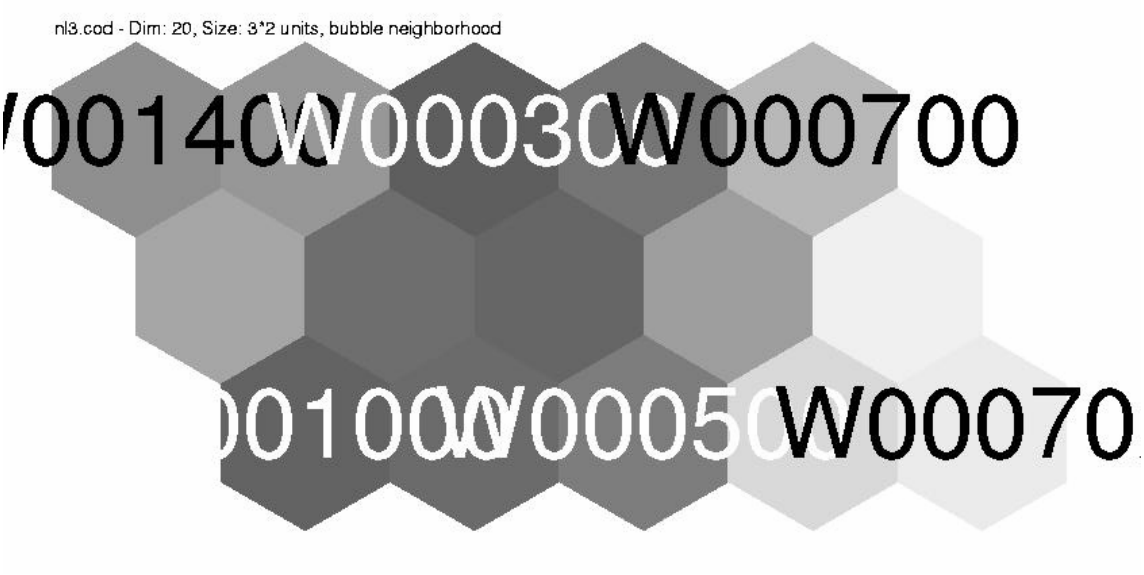


Figure 1: Distances between the reference vectors of the neighbouring map units (the lighter the colour, the more dissimilar two neighbouring map units are from each other)

From a purely model fit perspective the problem cases are the nodes (0,0), (1,0) and (1,1). For all these the RMSE was large, which ideally could be due to these cases being outliers. However, this was not the case; when we placed these clusters on the (geographical) map we realised that they were all relatively normal residential segments in the Dutch context. For example, 14 out of 15 districts in the municipality of Amsterdam were correctly classified as (0,0), apart from Westerpoort in the western part of the city, which is classified differently: after the node (1,1). These classifications are illustrated in one of the maps displayed in Figure 2 (we come back to this figure later).

We suspected another reason for the occurrence of nodes with remarkably poor fit, but which represent normal cases. Namely, that the resolution is insufficient for the less well represented cases to show up and thus we need a larger feature map. Would a larger map reveal new features that would add to the detail, while keeping the rough structure constant? Thus the next logical step was to create a larger map.

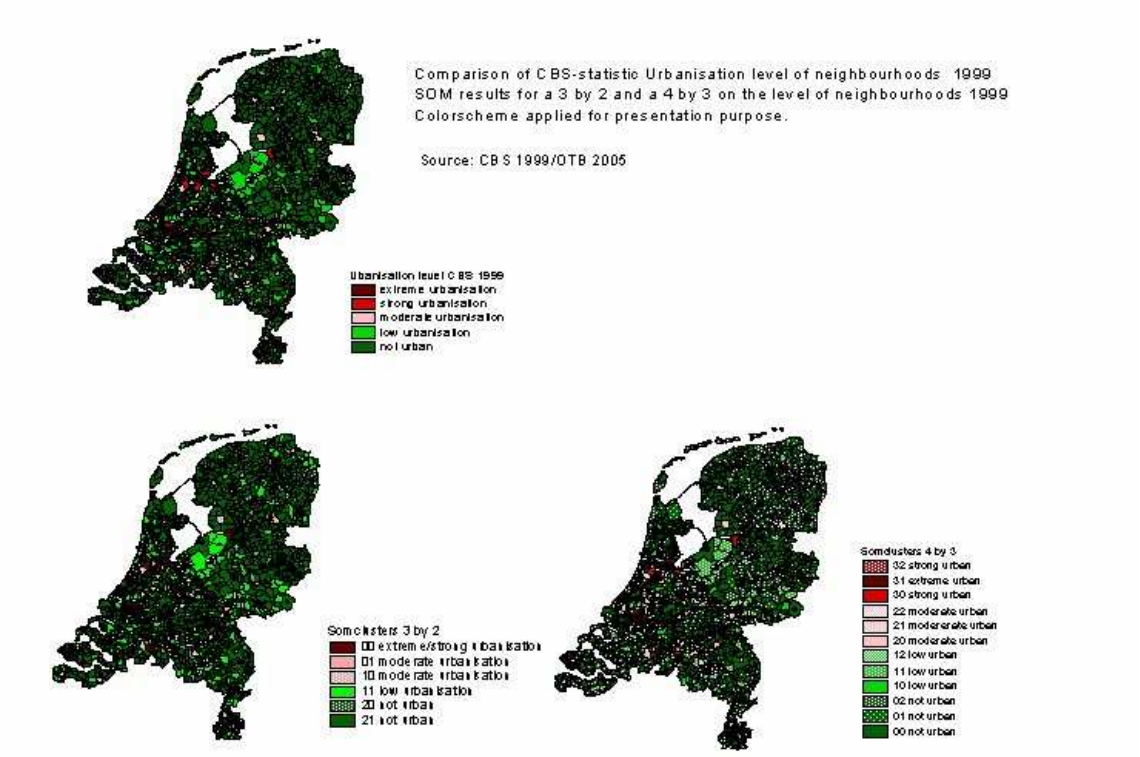


Figure 2: The classifications of the residential districts based on the official assignments (top), and the SOM for 3 by 2 and 4 by 3 maps respectively (bottom)

To sum up the analyses so far, we have now two kinds of outcome from the SOM analysis of district. One pertains to the six clusters that have formed on the map surface based on the input dimensions; the other pertains to the model fit. To illustrate with an example, the variation in RMSE-scores is geo-coded and visualized for one selected dimension (the urban level is selected because of the policy discussion in section 2) and cluster (the most urban node) in Figure 3. (See appendix 2 for a documentation of statistical box-plot analyses of the SOM outcome.) RMSE is the standard error, which here is used to determine the fit between the vectors of each node and observation. In other words, RMSE indicates the fit for each of the clusters/classes generated by the SOM: the larger the RMSE, the more the observation (wijk) deviates from the model estimation. This way, we obtain information about whether the area is well captured by the model (smaller RMSE) or worse captured by the model (larger RMSE); the latter case may indicate an outlier observation. However, if we know based on expert knowledge of the context that the district-label does not represent an outlier, the likely reason for a large RMSE score is that the resolution of the map is not sufficient to discern this particular, rather heterogeneous, but by no means rare, category of district.

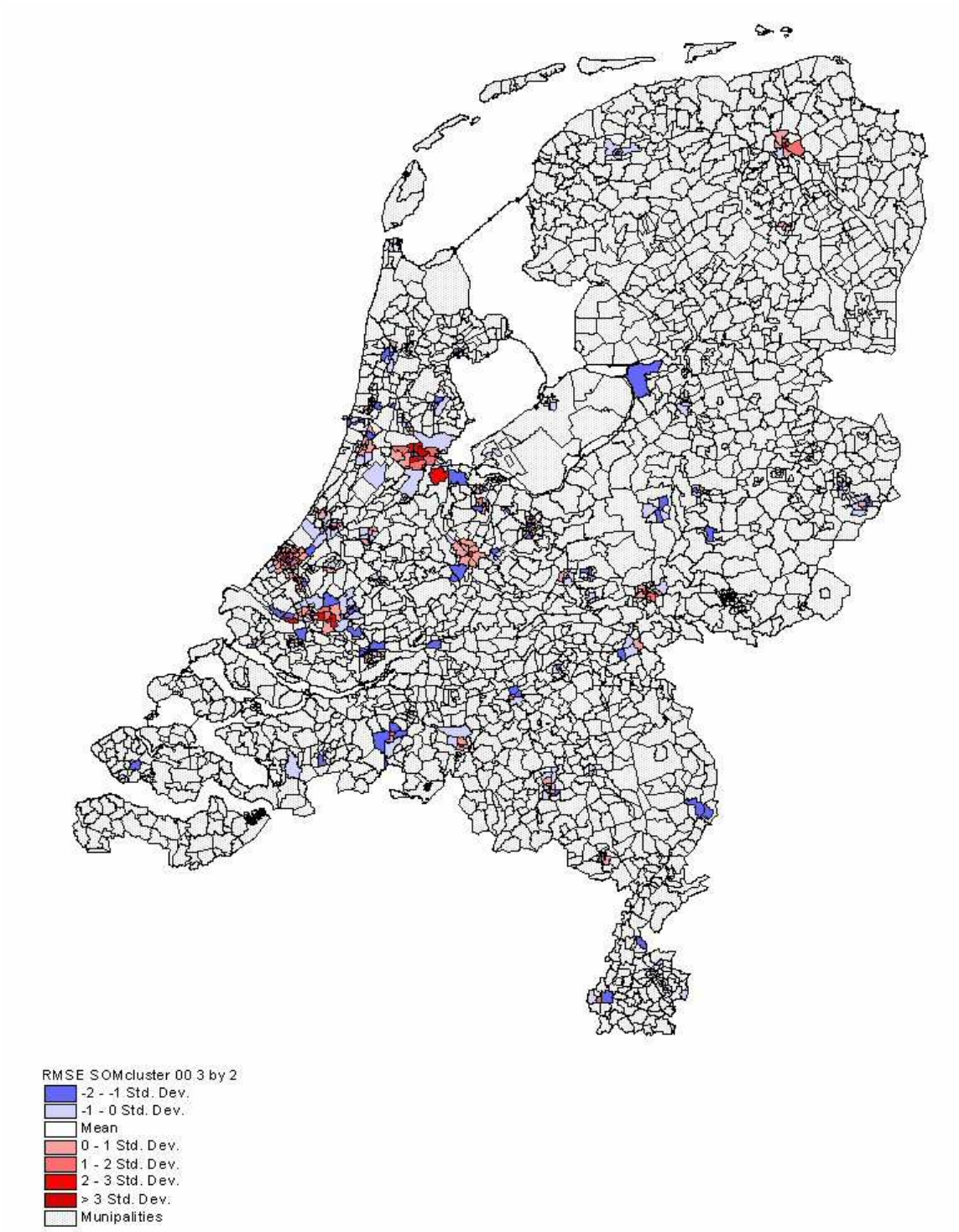


Figure 3: The urban level of districts compared to the mean value of RMSE for cluster (0,0) in the 3 by 2 map: more blue => standard deviation negative => more rural (i.e. large values for this dimension); more red => standard deviation positive => more urban (i.e. small values)

4.2. The larger feature map

The 4 by 3map classification of nodes resulted in a much similar pattern as above. The x- and y-coordinates of the node (in brackets), the typical ‘wijk’- code’, the typical characteristics of the category of observations, and the model fit are given below. In this case the characteristics are more detailed than above, as the enlargement of the map inevitably has increased the resolution, which was the reason for doing this adjustment. (Note that ‘no property x’ does not mean literally that this characteristic is missing, but that this is a dominating feature for this category of ‘wijken’.)

(0,0) 901: sparse, rural, large families, kids, no elderly people, no foreigners, no singles, no unemployed, small population, expensive, industry, no commercial/non-commercial enterprises; good fit	(1,0) 11413: sparse, rural, no elderly people, no foreigners, no unemployed, expensive; average fit	(2,0) 1409: no elderly people, no low income takers; average fit	(3,0) 3400: dense, urban, foreigners, 15-24 and 25-44 years old, but not 45-64 years old, large population, quite cheap; average fit	
(0,1) 900: sparse, rural, no foreigners, small population, expensive; good fit	(1,1) 500: expensive; good fit	(2,1) 1801: average in all dimensions; average fit	(3,1) 1400: dense, urban, foreigners, singles, families, low income takers, unemployed, no high income takers, large population, cheap, commercial enterprises, no industry; bad fit	
(0,2) 700: sparse, rural, 45-64 years old, but not 15-24 or 25-44 years old, no foreigners, small population, very expensive; good fit	(1,2) 20213: sparse, rural, no 15-24 or 25-44 years old, no foreigners, expensive; good fit	(2,2) 300: no 25-44 years old, high income takers; average fit	(3,2) 1407: dense, urban, foreigners, singles, elderly people, low income takers, unemployed, small families, no kids, large population, quite cheap, commercial/non-commercial enterprises, no industry; average fit	

We see how three basic features matter for the organization of the map: (1) density and urban character together with share of foreigners, singles, number of families, and market value organize the data along the longer axis; in particular, the urban dimension was a key feature in the organisation of the map along the longer (x-) axis; (2) the indicators of the shares of each age-group, the shares of families with kids, family size and unemployment now seem to organize the data diagonally, based on each diagonal; (3) what is the most interesting difference is that this resolution allows the occurrence of larger clusters than just one node: both the share of low-income takers, and the share of high-income takers partition the map into three clusters each, although the clustering does not overlap. Furthermore, compared to the original structure formed in the map 3 by 2, new features have occurred and also new clusters. We can also see that four of the six labels of the smaller map are retained (but two of them: 702 and 1000, have disappeared). One last observation is that this map (too) is well organized: all 12 nodes have 'won' observations.

Similarly as with the smaller map, figure 4 illustrates the heterogeneity for various parts of the data structure. We see at a glance that the more homogeneous part of the structure is the middle cluster(s) of the map: the nodes (2,0), (1,1), (2,1) and (2,2): for this block the neighbouring nodes are more similar to each other in terms of measured dimensions than elsewhere on the map surface. The most heterogeneous units in turn are the ones on the left side of the map: the nodes (0,0), (0,1) and (0,2), and the urban nodes (3,0), (3,1) and (3,2) on the right side of the map are fairly heterogeneous too. As with the smaller map, the most heterogeneous part of the map is here too the clusters that are the least urban (and also here the one with the best fit in RMSE terms – see below).

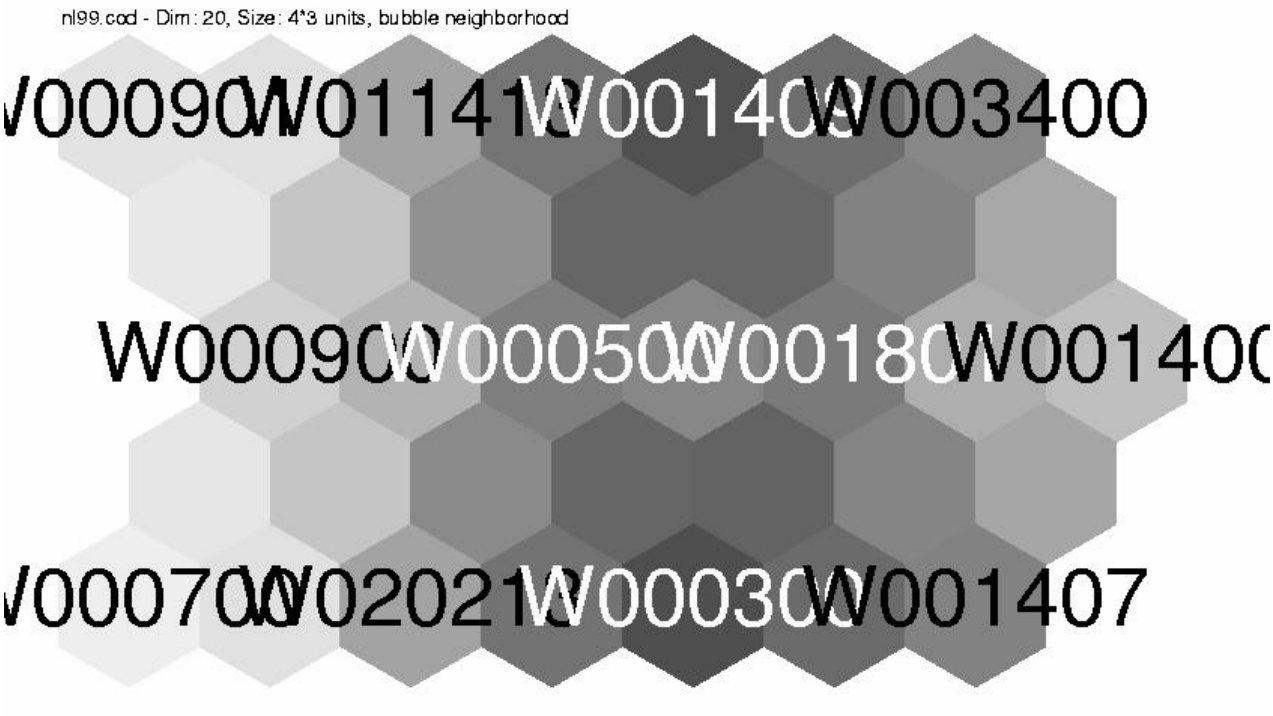


Figure 4: Distances between the reference vectors of the neighbouring map units (the lighter the colour, the more dissimilar two neighbouring map units are from each other)

In the Figure 2 above the SOM output is displayed as GIS classifications, which enables convenient comparison with the corresponding classifications resulting from the smaller map, and from the official discourse. Furthermore, what was discovered in the 3 by 2 map for Amsterdam applies here too: 14 of the 15 Amsterdam districts are also here classified as the most urban district type, and the remaining district Westerpoort in the western part of the city differently as a less urban area.

For the model fit aspect of the SOM output, clearly the urban neuron in the middle of the right side of the map has the worst fit; in this sense the result is the same as with the 3 by 2 map: the urban neurons fit worse than the more rural ones. When this aspect is investigated further based on the box-plot analyses documented in appendix 2, we note that the twelve neuron clustering fits better with the variable urban level than the six neuron clustering which is nevertheless adequate too. This variation in RMSE-scores is illustrated as a geographical map in figure 5 for the 4 by 3 map (cf. figure 3 for the 3 by 2 map).

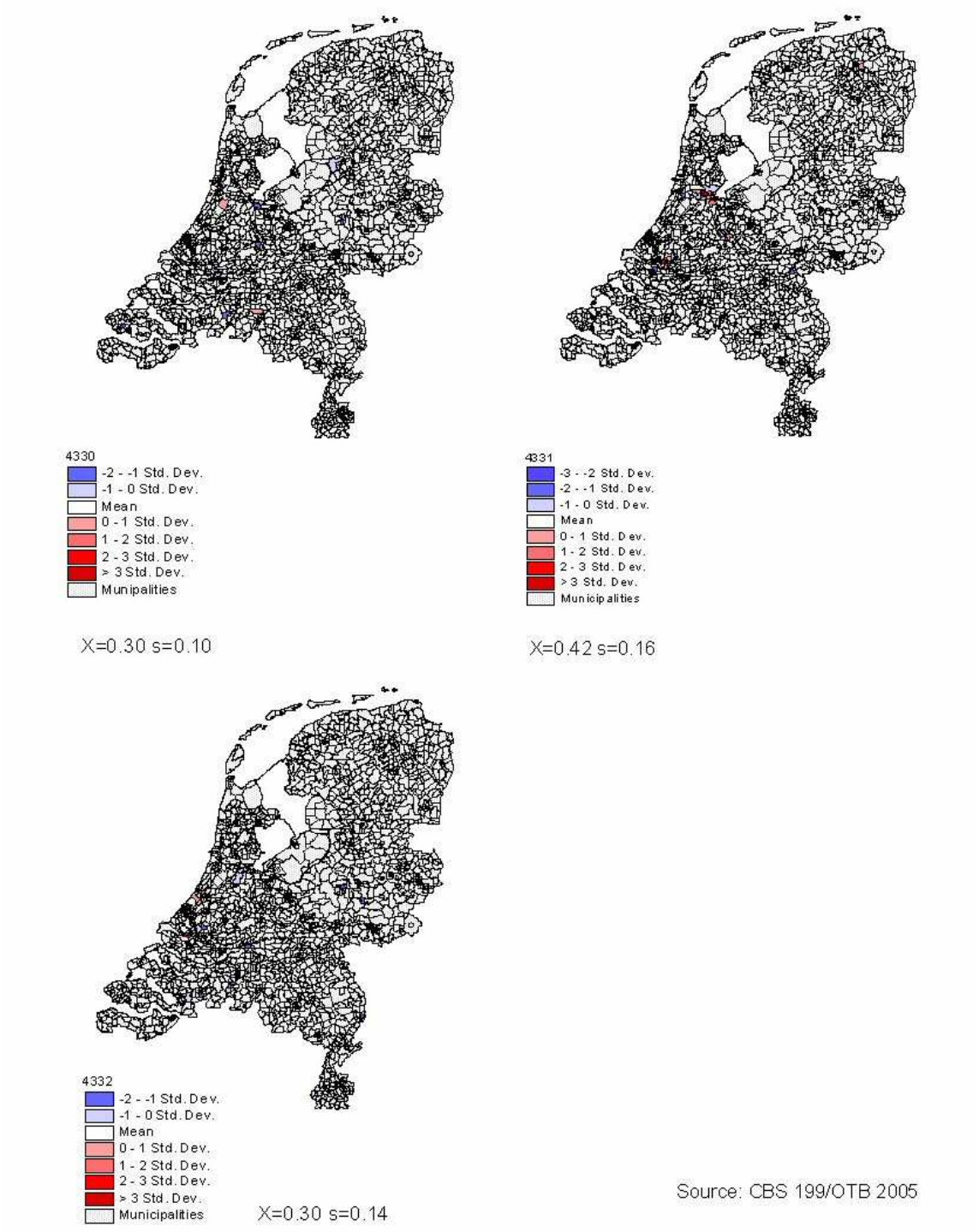


Figure 5: The urban level of districts compared to the mean value of RMSE for the urban clusters (3,0), (3,1) and (3,2) in the 4by3 map (thus the maps are referred to as 4330, 4331 and 4332): more blue => standard deviation negative => more rural (i.e. large values for this dimension); more red => standard deviation positive => more urban (i.e. small values)

5. The SOM classification in relation to urban regeneration categories

To test the predictive value of our SOM-analysis, we have confronted the typology of all neighbourhoods according to the 4 by 3 SOM analysis (henceforth SOM43) with actual policy typology in urban restructuring (Arnoldus et al. 2005). We will not discuss the pros and the cons of the selection by policymakers, but we just accept them since they are applied. Since the number of neighbourhoods in this list made by policymakers of the Ministry of Housing, Spatial Planning and the Environment (VROM) and the ministry of the Interior and Kingdom Relations (BZK) have a bias towards the larger cities and municipalities (G30 of Big 30) we expect two results. First, we expect that the most urban SOM clusters, most notably the nodes (3,0), (3,1) and (3,2) and to a lesser extent (2,0), (2,1) and (2,2), will be stronger related to this list. Second, we expect that the description of these nodes reveal 'problems'.

The spatial data is actually collected on the level of postcodes (zip-codes). By means of fairly simple GIS-overlay all neighbourhoods are assigned 'urban restructuring' status even if they have a very limited score for that criterion on a postcode level. Thus, we will have a slight overestimation of neighbourhoods compared to the real situation. Of the total number of neighbourhoods (2400) approximately 220 are assigned in this way. These scores are linked to the SOM results in the GIS and the statistical package SPSS. We formulate two tests to analyse the fit between real and estimated area classifications.

The first test compares the prior dimensions of the official policy-list and the relevant dimension(s) of SOM43. In SPSS we have applied a multidimensional scaling method known as HOMALS (see appendix 3 for an explanation). *A priori* we wanted two dimensions since the policy-list defines a binary of nature (urban restructuring versus not urban restructuring). The HOMALS analysis leads to factor scores of each of 12 SOM43 clusters per dimension. Figure 6 shows the results. Here it is sufficient to say that the more the score deviates from zero, the more a cluster deviates from a 'mean' per dimension. By combining the orthogonal dimension HOMALS shows the match between the original variables of SOM43 and the binary list of the policymakers (green in which ISV refers to urban restructuring). Table 2 shows the corresponding scores for both dimensions.

The HOMALS shows that only one dimension really matters (ANOVA on the factor scores on the binary segmentation confirms that): highly urban versus the rest as table 2 shows. There is one striking deviation: cluster (2,2) is the grey zone between countryside and influence-cities nearby. These are often pleasant neighbourhoods with wealthier and older people. It refers to mostly an urban level 3 (moderate) but also to an urban level 4 (nearly countryside), which is in contrast to (2,0) and (2,1) that score full on urban level 3. Thus, (2,2) is for some aspects indeed belonging to the less urban six clusters on the left side of the map. Therefore the HOMALS (2,2) score on dimension 1 scores negative. Below we compare this cluster with the clusters (2,0) and (2,1) with respect to some of the original input variables:

- Population density in (2,2) is lower (1701) than in (2,0) and (2,1)
- Address density in (2,2) is therefore lower (1133) than in (2,0) and (2,1)
- Income in (2,2) is higher (24000) than in (2,0) and (2,1)
- Percentage high income is therefore higher in (2,2) (25,5%) than in (2,0) and (2,1) as well
- Percent age 00-14 is lower (15) in (2,2) than in (2,0) and (2,1)
- Percentage of 45+ years old is higher in (2,2) than in (2,0) and (2,1)
- Percentage of non-western immigrants (4%) in (2,2) is lower than in (2,0) and (2,1)

- Percentage non-workers (see age structure) in (2,2) is in-between (17%) (2,0) (12) and (2,1) (21)
- Percentage of pensioners is most in (2,2)
- Percentage industry is lower in (2,2) (11,5%) than in (2,0) and (2,1).

Using box-plots for each cluster of SOM43 it could be checked why (2,2) has a negative score. It turned out that this node not only deviates on income, but surely also on age and so forth. The result also makes sense. The areas characterised as (2,2) in general have older and more affluent residents, and these are also less 'working areas' than (2,0) and (2,1). These are neighbourhoods in small cities in the countryside along infrastructure or the outskirt of urban conurbations (for example, De Bilt nearby Utrecht). More precisely, the cluster (3,1) (comprising districts of the four largest Dutch cities: Amsterdam, Rotterdam, The Hague, and Utrecht) nearly coincides with the classification 'Urban restructuring' and scores extremely highly, as expected. The same is more or less true for cluster (3,2). The scores for (3,0), (2,0) and (2,1) are similar. Thus, urban restructuring is indeed a biased towards the larger and largest cities. The figure 6 might reveal that the result is less promising since, except for (3,1), nearly all other clusters are located further away from the ISV node. However, the HOMALS and the ANOVA show that the second dimension has no real meaning at all (see table 2).

We may conclude that the predictive power of the SOM43 clustering is good. However, the results show that a simpler SOM structure – that is a smaller map – might be applied as well based on the urban level. However, it would be incorrect to conclude that only the four largest cities have real problems. The statistical analysis thus suggests that urban restructuring has an urban bias, which raises the concerns about the validity of the policy-makers list (app. 1).

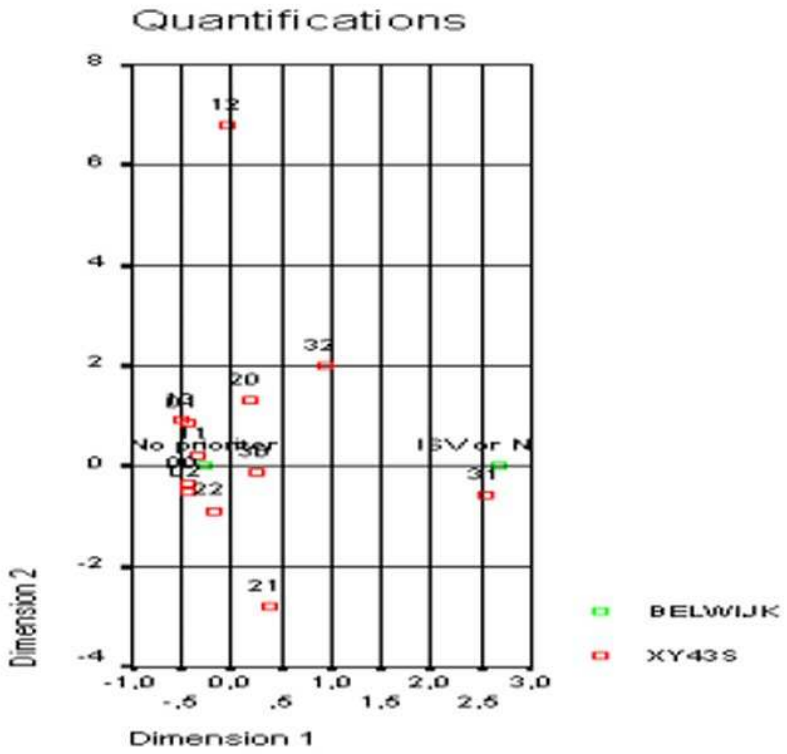


Figure 6: HOMALS with respect to urban restructuring cases (ISV)

Table 2: HOMALS scores for the dimensions 1 and 2

	Marginal Frequency	Category Quantifications	
		Dimension	
		1	2
1 00	530	-,458	-,341
2 01	219	-,447	,855
3 02	500	-,435	-,506
4 10	56	-,496	,920
5 11	287	-,350	,221
6 12	12	-,037	6,783
7 20	138	,168	1,319
8 21	61	,369	-2,789
9 22	114	-,193	-,888
10 30	131	,249	-,132
11 31	198	2,548	-,582
12 32	136	,927	

5.3. Second test

The second test compares the now 'crisp' classification based on the first test (i.e. official policy + HOMALS scores) with the 'fuzzy' outcome of the SOM43 clustering. In table 3 we describe the 12 clusters in relation to policy and SOM outcome; more precisely, we characterise each cluster with respect to the urban regeneration criteria.

Table 3: Synthesis of policy, HOMALS scores and SOM outcome in order to identify urban regeneration targets (Y=yes; N=no)

SOM43 clusters	Official policy and HOMALS scores for the dimensions 1 and 2	SOM clustering outcome (see sub-section 4.2.)
1 (0,0)	N	N
2 (0,1)	N	N
3 (0,2)	N	N
4 (1,0)	N	N
5 (1,1)	N	N
6 (1,2)	N	N
7 (2,0)	Y	N
8 (2,1)	Y	N
9 (2,2)	N	N
10 (3,0)	Y	Y
11 (3,1)	Y	Y (however, note the poor fit of this most urban cluster in the SOM analyses)
12 (3,2)	Y	Y

We can see that the overlap between official and SOM classified– potential and actual – urban restructuring areas is relatively good. However, for the nodes (2,0) and (2,1) the assignment cannot be sustained based on the SOM analysis alone as these clusters do not unambiguously point to particularly 'urban' or 'low-income' dimensions. In the last two tests we assumed

for practical reasons that the list of the policy-makers is validly made. Based on the first test, this might be assumed as well. However, does the second test falsify this? Is it possible that the SOM43 results show that the list is unjustly biased? As with Kauko (2005), also these analyses show that the mixture of districts apply across the whole country and not just for the larger cities, as assumed by the policymakers.

6. Summary and conclusions

We can now ask a variety of questions with regard to the validity and reliability of the analysis. What can we say about the match between the classification and the reality? The prior analysis by Kauko (2005) suggested that the districts, that are of same SOM-determined class, would be scattered all over the Netherlands. This finding pertains to the absolute characteristics of districts and is hardly subject to any debate. However, the argument concerns the relative context of the classification – does a neighbourhood in a large city and one in a small place have different positions in the reality even if they belong to the same classes in the SOM output? The classification in Kauko (2005) did in most, but not all of the cases separate between districts that were of clearly different characteristic, and/or located far away from each other. For example, one of the 15 Amsterdam districts was labelled after a municipality on the Frisian coast, and in some cases the core district of very small municipalities were labelled after core districts in middle-sized towns; such (mis)classification should not have happened given *a prior* information about official categorisations.

Also after this follow-up analysis the answer is that, in general, the SOM classifications are logical but obviously exceptions may be found all over the Netherlands, and then the question is whether the SOM method is more or less valid than official classifications. The fundamental question arising from the analyses is whether the official policy-list is sound. Since the fit is fairly good, we may conclude that neighbourhoods belonging to the five clusters on the right half of the 4 by 3 map, or at least to the three clusters most on the right side of this map, and which are not assigned as problem neighbourhoods, may have a fairly high probability of having problems. However, here is a word of warning when we look at potential new cases for urban restructuring: such urban and not-affluent districts seem to involve a considerable heterogeneity both across neighbouring clusters (i.e. nodes in the SOM analysis) as within them with respect to our selected 20 input characteristics. Therefore, we cannot be sure that, if we find a ‘non-assigned’ district grouped together with an assigned case of urban restructuring, that this area too is a suitable target. It may be likely, however.

While a reasonably good ‘match’ between actual clusters and the ones generated by the SOM was obtained, the results are not robust to a change in SOM map dimensions. The conclusion from Kauko (2005) was that prosperous cities and areas should be separated from low-income areas. However, outlier cases should not be treated within the same general framework. And outlier cases are possible to ascertain only when a sufficiently large map. In this analysis we were able to reliably identify the particularly urban cluster with outlier characteristics (3,0) and (3,2) only after the enlargement of the map from 6 to 12 nodes.

We can also ask questions about the feasibility of the proposed methodological innovation. Only if the match is adequate (in section 5, for example the validity was determined based on the overlap with the 56 municipalities for urban regeneration) and the use of SOM is easy (see section 3) can this method be applied for selection and monitoring purposes. Since selection is

indeed important in policy-making, this 'trick' is to combine the technical and the policy-relevancy. (Let the policy ends justify the technical means, so to speak.)

We can conclude that this method suits well for monitoring purposes and applied research. However, whether it can be considered better than other methods depends on evaluation criteria. The method of combining the SOM and GIS proved to be informative although not entirely straightforward. The downside is that, like in simulation approaches, in general the result is not fully repeatable in quantitative terms unless we have run the SOM until it has reached a global optimum, and this we cannot know about. Moreover, to interpret the map outcome may be time-consuming, if the result does not appear to 'fit'. Nevertheless, if a fuzzy and pragmatic analysis is preferred to a more precise one based on isolating influences, and if we want to go beyond number-crunching and utilise a more qualitative approach based on visualizing patterns, the SOM offers tremendous possibilities for the analyst of spatial phenomena of high social-economic relevance.

References

Arnoldus, M, F. Wassenberg & H. Kruythoff (2005) Behoefteraming Stedelijke Vernieuwing 2010-2019, ramingsmethode. Onderzoeksinstituut OTB TU Delft. In Opdracht voor Ministerie van VROM [Estimation Demand Urban Renewal 2010-2019].

Bootsma, H. (1998) *The Myth of Reurbanization. Location Dynamics of Households in the Netherlands* (NETHUR D Publications. Amsterdam).

Dieleman, F. M. & C. Wallet (2003) Income differences between central cities and suburbs in Dutch urban regions, *Tijdschrift voor Economische en Sociale Geografie* 94(2) 265 – 275.

Goetgeluk, R. & S. Musterd (2005) Residential Mobility and Urban Change. In: R. Goetgeluk & S. Musterd (eds.), Residential Mobility and Urban Change. *Open House International*, Vol. 29, June/July, 2005 (in press).

Hatzichristos, T. (2004) Delineation of demographic regions with GIS and computational intelligence, *Environment and Planning B* **31** 39 - 49.

Izraeli, O. (1987) The Effect of Environmental Attributes on Earnings and Housing Values across SMSAs, *Journal of Urban Economics* **22** 361 – 376.

Kauko, T. (1997) Asuinalueiden attraktiivisuuserojen havainnollistaminen neuroverkon avulla (Visualizing differences in attractiveness between localities with a neural network, in Finnish), *Maanmittaus*, 1-2/72, 37-70.

Kauko, T, (2002) Modelling locational determinants of house prices: neural network and value tree approaches (Doctoral dissertation), Utrecht, (Internet: www.library.uu.nl/decollectie/proefschriften/11688main.html).

Kauko, T, (2005) Using the self-organizing map to identify regularities across country-specific housing market contexts, *Environment and Planning B*, 32(1), January, 89-110.

Kohonen T, (1995) Self-Organizing Maps (Springer Series in Information sciences, Springer-Verlag, Germany).

Kohonen, T., Hynninen, J., Kangas, J. & Laaksonen, J. (1996) SOM_PAK: The Self-Organizing Map Program Package. Helsinki University of Technology, Faculty of Information Technology, Laboratory of Computer and Information Science. Report A31.

Meen, G, (2001) Modelling Spatial Housing Markets: Theory, Analysis and Policy (Advances in Urban and Regional Economics, V.2, Kluwer, USA),

Openshaw, S, Blake, M & Wymer, C, (1994) Using neurocomputing methods to classify Britain's residential areas", Working paper 94/17, School of Geography, University of Leeds.

Potepan M J, (1996) Explaining Intermetropolitan Variation in Housing Prices, Rents and Land Prices, *Real Estate Economics*, **24**(2) 219 – 245.
Siikanen A. (1992) Asuntojen kysyntä, tarjonta ja alueellinen erilaistuneisuus (in Finnish). Asuntohallitus, tutkimus- ja suunnitteluosasto, asuntotutkimuksia 4:1992. Helsinki.
Taltavull de La Paz, P. (2003) Determinants of housing prices in Spanish cities, *Journal of Property Investment and Finance*, 21(2), 109 – 135.
Tu, Y. (2000) Segmentation of Australian housing markets: 1989-98, *Journal of Property Research* **17**(4) 311 – 327.
Wong, C. (2001) The Relationship Between Quality of Life and Local Economic Development: An Empirical Study of Local Authority Areas in England, *Cities* **18**(1) 25 – 32.

Internet:

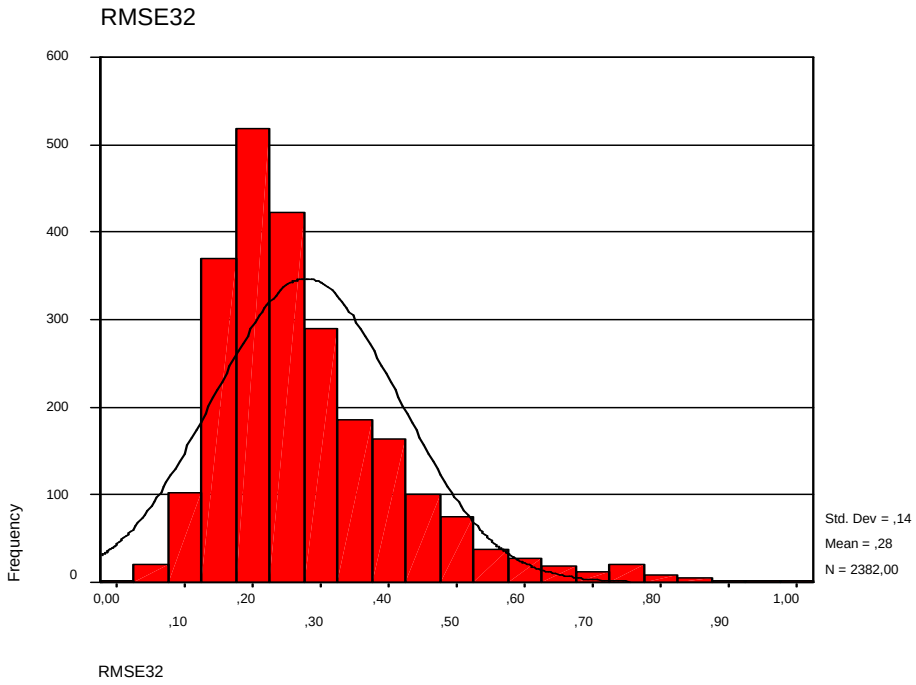
http://www.kei-centrum.nl/ KEI Stedelijke Vernieuwing (Urban renewal in the Netherlands).

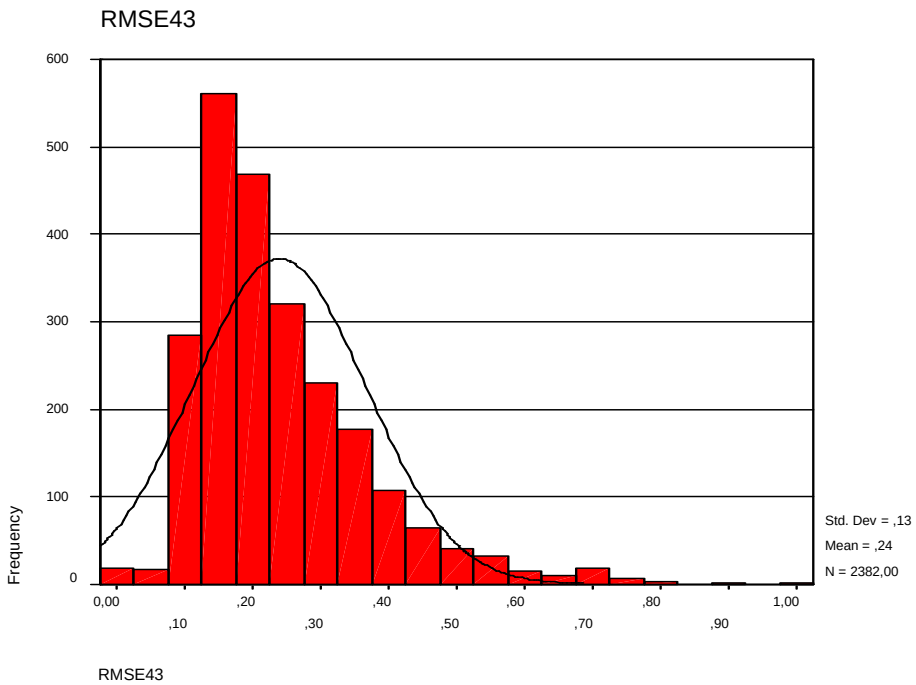
Appendix 1 56 Municipalities with selected urban regeneration areas

Municipality	*Neighbourhood*	*Municipality*	*Neighbourhood*
Alkmaar	Overdie/Schermereiland*	Heerlen	Heerlen Stad Oost
Almelo	Almelo Zuidwest* (Ossenkoppelerhoek/Kerkelanden)		Grasbroek/Musschemig/Schandelen*
Amersfoort	De Kruiskamp/Koppel	Helmond	Binnenstad*
Amsterdam	Westelijke Tuinsteden*	Hengelo (Ov)	Berflo Es
	ZuidOost*	Leeuwarden	Achter de Hove - Vegelin
	Noord (De Banne*/Nieuwendam Noord)		Vrijheidswijk*
Arnhem	Presikhaaf	Leiden	Leiden Noord* (Groenoord/Noorderkwartier/De Kooi)
	Malburgen*		Leiden Zuid-West
Breda	Breda Noord-Oost* (Hoge Vlucht/Doornbos-Linie)	Lelystad	Zuiderzee/Atol*
	De Heuvel*	Maastricht	Maastricht Noordwest (Malberg*-Bospoort)
Den Bosch	Boschveld*	Nijmegen	Willemskwartier*
	Barten/Eikendonk/Hofstad	Rotterdam	Zuidelijke Tuinsteden
Den Haag	Den Haag Zuidwest*		Oud Zuid (Katendrecht*, Afrikaanderbuurt, Tarwewijk*, Bloemhof)
	Transvaal*		Crooswijk Noord*
	Duindorp		Rotterdam West
	Laakkwartier/Spoorwijk		Hoogvliet*
	Rustenburg/ Oostbroek	Schiedam	Nieuwland/Groenoord*
Deventer	Rivierenwijk *	Tilburg	Oud Zuid
	Keizerslanden*		Nieuw Noord*
Dordrecht	Dordrecht West * (Oud Krispijn*/Nieuw Krispijn/Wielwijk/Crabbehof)	Utrecht	Overvecht Zuid
Eindhoven	Woensel Zuid (Hemelrijken)		Kanaleneiland Noord*
	Tongelre* (Doornakkers en gedeeltelijk Lakerlopen)		Hoograven
Emmen	Emmen Revisited*(Angelslo, Bargeres, Emmerhout)		Zuilen/Ondiep
Enschede	Wesselerbrink*	Venlo	Q4*
	De Velve Lindenhof	Zaanstad	Zaandam Zuidoost*
Groningen	Vinkhuizen*	Zwolle	Holtenbroek*
	Lewenborg		
Haarlem	Delftwijk*		
	Europawijk Zuid		

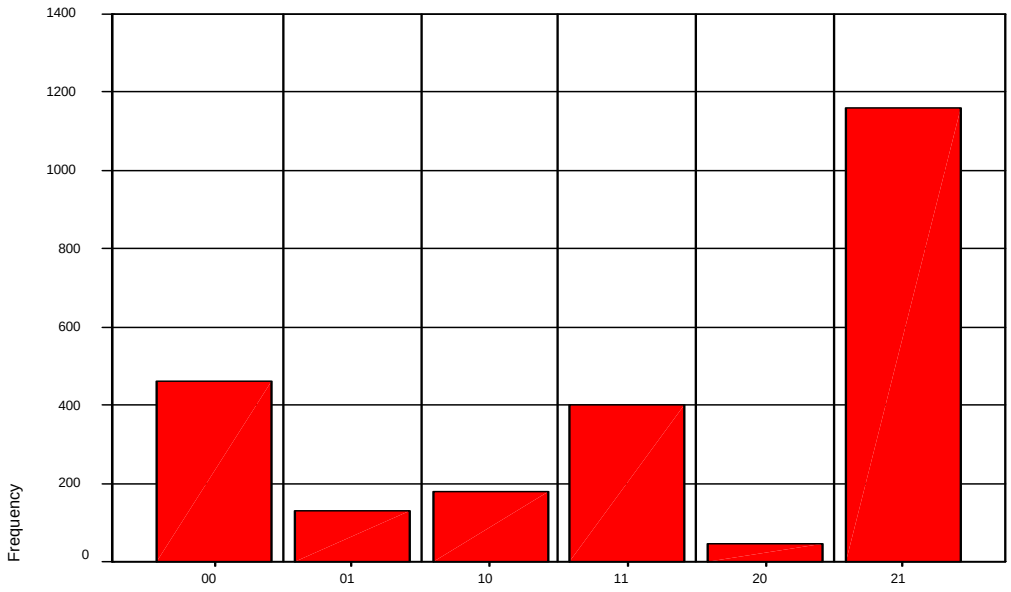
Appendix 2 The Box-plot of statistics for the SOM output

Boxplots: Summaries for Groups Example (SPPS HELP).
Simple Boxplot: Summaries for Groups of Cases.
A single numeric variable is summarized within categories of another variable. Each box shows the median, quartiles, and extreme values within a category.
Minimum specifications.
A numeric summary variable.
A Category Axis variable.

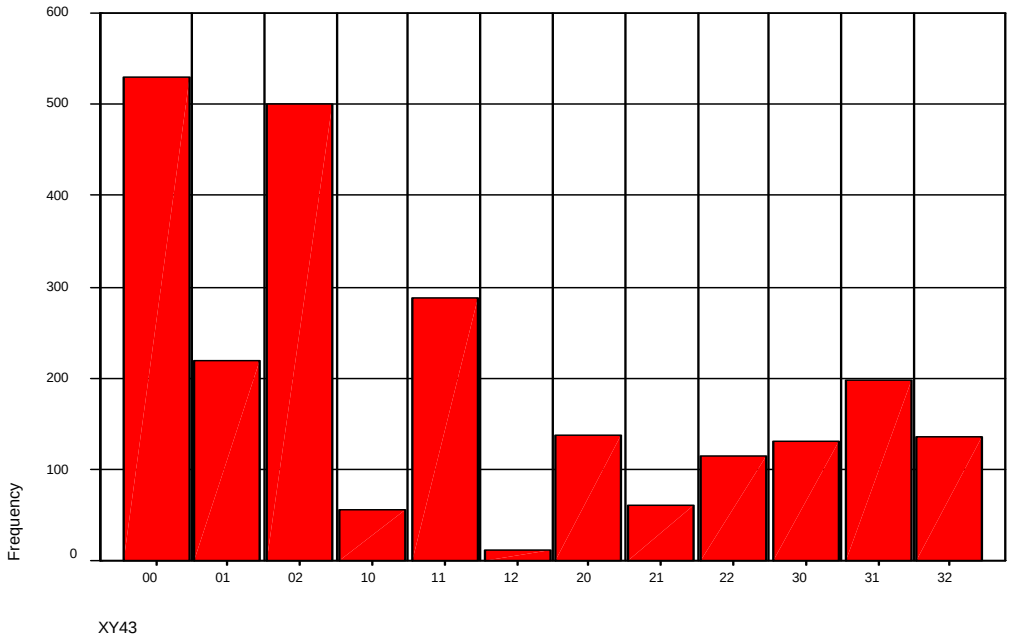


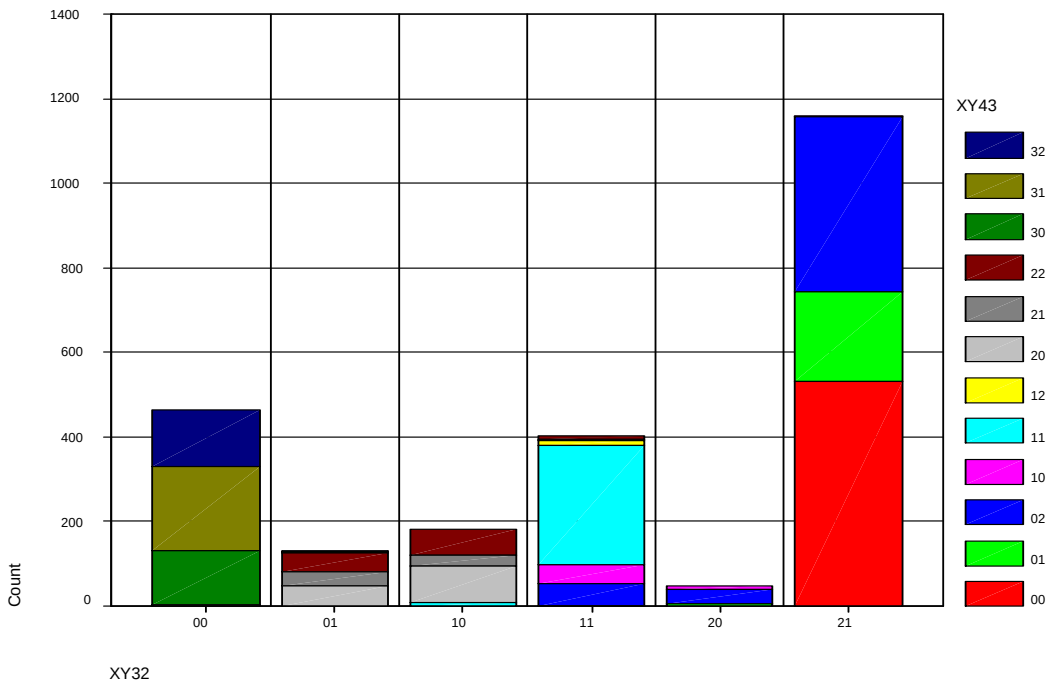


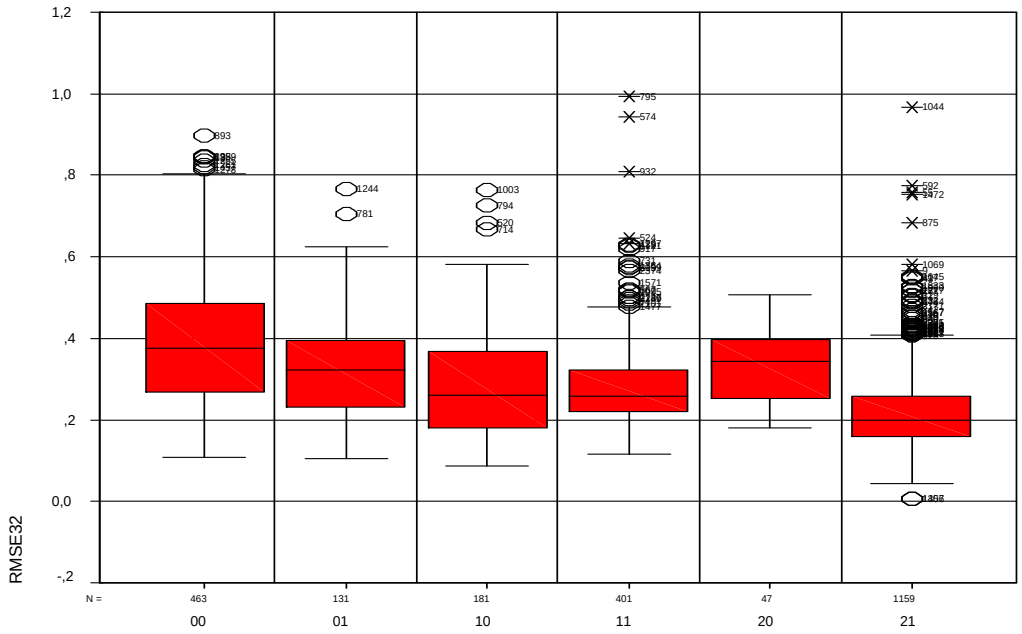
XY32


1400
1200
1000
800
600
400
200
0
Frequency
00
01
10
11
20
21
XY32

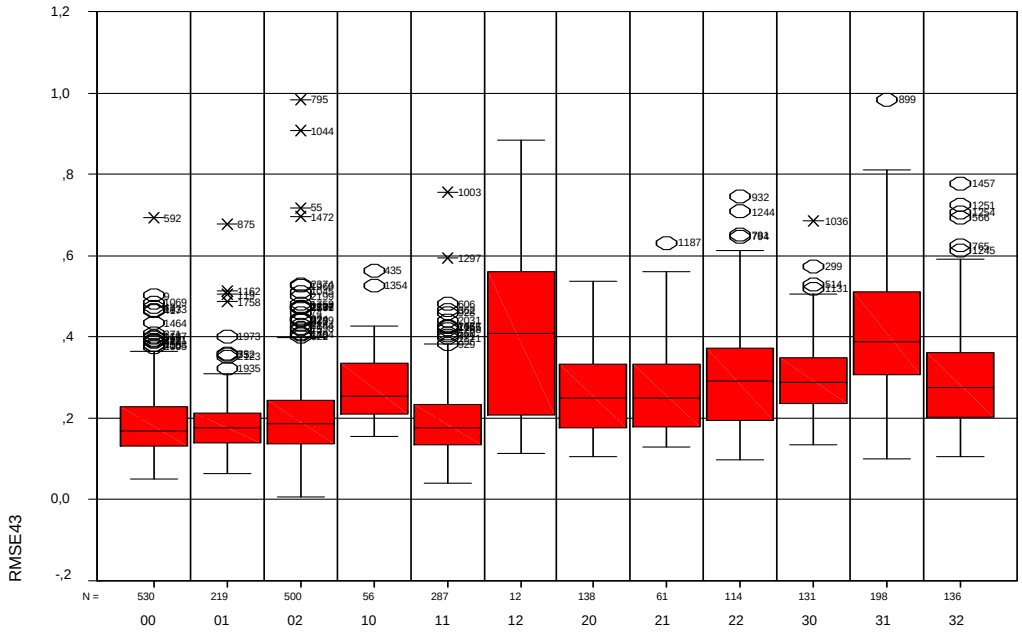
XY43


600
500
400
300
200
100
0
Frequency
00
01
02
10
11
12
20
21
22
30
31
32
XY43

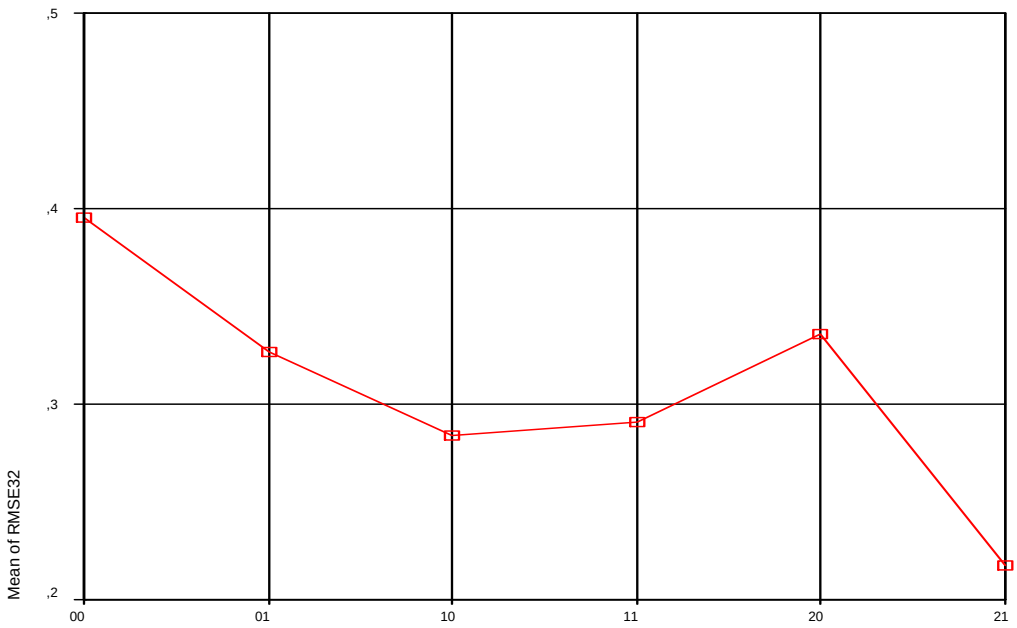

1400
1200
1000
800
600
400
200
0
Count
00
01
10
11
20
21
XY32
XY43
32
31
30
22
21
20
12
11
10
02
01
00


RMSE32
1,2
1,0
,8
,6
,4
,2
0,0
-,2
N =
463
131
181
401
47
1159
00
01
10
11
20
21

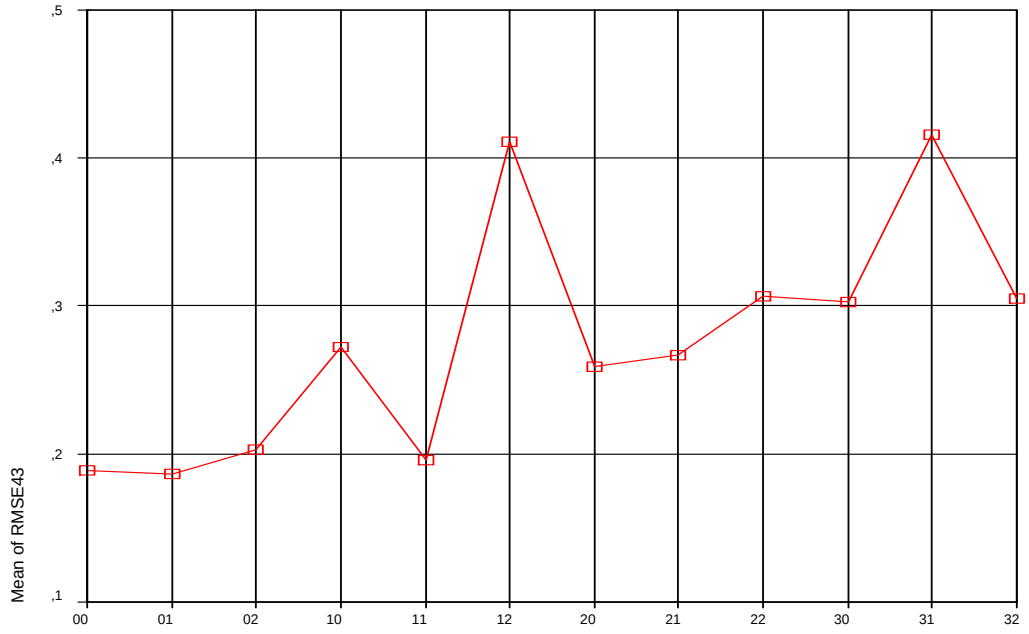
XY32


RMSE43
1,2
1,0
,8
,6
,4
,2
0,0
-,2
N =
530
219
500
56
287
12
138
61
114
131
198
136
00
01
02
10
11
12
20
21
22
30
31
32

XY43


,5
,4
,3
,2
Mean of RMSE32
00
01
10
11
20
21

XY32S


,5
,4
,3
,2
,1
Mean of RMSE43
00
01
02
10
11
12
20
21
22
30
31
32

XY43S

Appendix 3 Homogeneity Analysis (SPPS HELP)

Homogeneity analysis quantifies nominal (categorical) data by assigning numerical values to the cases (objects) and categories. Homogeneity analysis is also known by the acronym HOMALS, for homogeneity analysis by means of alternating least squares. The goal of HOMALS is to describe the relationships between two or more nominal variables in a low-dimensional space containing the variable categories as well as the objects in those categories. Objects within the same category are plotted close to each other, whereas objects in different categories are plotted far apart. Each object is as close as possible to the category points for categories that contain that object.

Homogeneity analysis is similar to correspondence analysis but is not limited to two variables. As a result, homogeneity analysis is also known in the literature as multiple correspondence analysis. Homogeneity analysis can also be viewed as a principal components analysis of nominal data. Homogeneity analysis is preferred over standard principal components analysis when linear relationships between the variables may not hold or when variables are measured at a nominal level. Moreover, output interpretation is more straightforward in HOMALS than in other categorical techniques, such as cross-tabulation tables and log-linear modelling. Because variable categories are quantified, techniques that require numerical data can be applied to the quantifications in subsequent analyses. For example, homogeneity analysis could be used to graphically display the relationship between job category, region of residence, and gender. One might find that region of residence and gender discriminate between people, but that job category does not.

The results of the analysis is displayed by statistics and plots concerning frequencies, eigenvalues, iteration history, object scores, category quantifications, discrimination measures, object scores plots, category quantifications plots, and discrimination measures plots.

The results of the analysis show that only one dimensions/factor fits the data pretty well and be characterised as urban vs countryside with an exception for (2,2): the score of (2,2) is negative for groups on the right side of the map (i.e. the six nodes with x-values 2 and 3), which are labelled ‘Urban target district’.