Modelling regional house prices

Bram van Dijk , Philip Hans Franses , Richard Paap & Dick van Dijk

Modelling regional house prices

Bram van Dijk, Philip Hans Franses, Richard Paap and
Dick van Dijk*

*Econometric Institute, Erasmus University Rotterdam, PO Box 1738,
3000 DR Rotterdam, The Netherlands*

We develop a panel model for regional house prices, for which both the
cross-section and the time series dimension is large. The model allows for
stochastic trends, cointegration, cross-equation correlations and, most
importantly, latent-class clustering of regions. Class membership is fully
data-driven and based on the average growth rates of house prices, and the
relationship of house prices with economic growth. We apply the model to
quarterly data for the Netherlands. The results suggest that there is
convincing evidence for the existence of two distinct clusters of regions
with pronounced differences in house price dynamics.

I. Introduction

Real estate prices in many countries have experienced
a dramatic boom in recent years (International
Monetary Fund (IMF), 2004). At the same time, the
extent of the price increase appears to vary substan-
tially across different regions within a given country.
In the Netherlands, for example, it is commonly
believed that the house prices in *Amsterdam* and the
densely populated western part of the country have
increased far more than prices in the smaller cities and
rural areas in the east. As house prices are typically
available per region or city, we may analyse these data
at such a disaggregate level, to examine whether
indeed regions or cities behave differently, perhaps
in terms of trends but also in terms of response to
outside economic shocks. In this article we develop a
time series model that suits this purpose.

Most regional house prices have the following
properties. First, they tend to display a trend, and
historical price patterns suggest that this trend
probably is not deterministic but stochastic. In
particular, house prices show 'bubble'-type behav-
iour, where prolonged periods of steady increases of
the price level suddenly end with a sharp drop
followed by a period of low price levels, suggesting

that trends are unlikely to be deterministic. Second,
for different regions within a country these stochastic
trends should somehow be linked. It is not plausible
that prices in different regions would diverge indefi-
nitely or that certain regions would not respond to
common macroeconomic shocks. So, a model for
regional house prices should allow for some form of
common trends. Third, it can be expected that
adjacent regions show similar price patterns,
although this may also be the case for regions far
apart geographically but with similar economic or
demographic characteristics. Hence, a suitable model
should allow for similarities in the dynamic behaviour
of house prices across regions. An intuitively appeal-
ing possibility is to consider a model that allows for
groups or clusters of regions, where house price
dynamics in regions within a given cluster are the
same, while they are different across clusters.
Preferably, such a model should not require *ex ante*
or exogenous assignment of regions to specific
clusters. In fact, it would be best if the data
themselves were allowed to indicate if clusters exist
and if so, which regions belong to which cluster.

In this article we extend the latent-class panel time
series model introduced by Paap *et al.* (2005) to cap-
ture these different properties of regional house prices.

*Corresponding author. E-mail: djvandijk@ese.eur.nl

The key feature of this model is that the clustering of regions is purely data-driven, where cluster membership is based on characteristics corresponding to two specific research questions we want to address. The first question is whether prices in all regions have the same average growth rate. Note that a common trend specification across the regions entails that their growth rates must be somehow compatible, but it still leaves open the possibility that house prices in some regions grow faster than in others. The second question we consider is the way the house prices in each region react to changes in the overall economic situation, which we measure by Gross Domestic Product (GDP). We examine both the size of the effect from GDP on the house prices and the speed at which regions react to changes in GDP.

We apply our model to house price data for the Netherlands, comprising 76 regions for which we have quarterly data for the period 1985Q1 to 2005Q4. We find that the 76 regions can be grouped into two clusters. The first cluster consists mainly of regions in the east of the country. These are mainly rural areas that are close to the larger cities, especially close to the *Randstad* (consisting of *Utrecht*, *Amsterdam*, *Den Haag*, *Rotterdam* and other cities in the area). This cluster reacts both stronger and faster to changes in GDP. The average growth rate does not vary over the regions.

There are not many studies that describe regional house prices. Cameron *et al.* (2006) build a model from inverse demand equations. They have, however, only a limited number (nine) of regions, and their model would not work in our situation where we have many more (76) regions, as we will describe below. Malpezzi (1999) constructs an error correction model for regional house prices. The parameters of this model are, however, not allowed to vary across regions. Holly *et al.* (2008) model US house prices at the state level. Their model is 'fully heterogenous' in the sense that it has different parameters for each region. In this article we cover the middle ground, that is, the model parameters are allowed to vary across *groups* of regions but not across each region individually.

Before we propose our latent-class model, we first provide some details on the house price data in Section II. We consider two decades of quarterly house prices on 76 regions in the Netherlands. We discuss their trending behaviour by performing panel unit root tests and we also show that the growth rates in different regions show strong cross-correlations. Using Multi-Dimensional Scaling (MDS) techniques we get a first impression if and how these 76 regions could get clustered. Then, in Section III, we put forward our model specification, highlighting the

underlying data-driven clustering mechanism. In addition, we describe the method used for parameter estimation. In Section IV, we present our estimation results, and interpret the various outcomes. Next, we take a look at impulse-response functions of the house prices with respect to a shock in GDP and in the interest rate. In Section V, we conclude and we outline topics for further research.

II. Data

The Dutch real estate agent association (NVM) publishes quarterly data on house prices for $N = 76$ regions in the Netherlands. Our dataset covers the sample period 1985Q1 to 2005Q4 ($T = 84$ quarters). Hence, we have a panel database where both the cross-section dimension N and the time dimension T are fairly large.

The way the country is divided into 76 regions is determined by the NVM. Macroeconomic data, such as output and inflation, are not available for this particular specification of regions. Other (macro) variables that we use in our model are therefore measured at the country level. In particular, this concerns the interest rate (obtained from the Dutch Central Bank) and quarterly real GDP (from Statistics Netherlands). The GDP series is available until 2005Q2. We obtain real house prices by deflating with the Consumer Price Index (CPI) (from Statistics Netherlands). In addition, we seasonally adjust the real GDP series using the Census X-12 algorithm (available in EViews 5.1). We denote the real house price in region i at time t as $p_{i,t}$, and real GDP as y_t.

Figure 1 shows time series of $\log(p_{i,t})$ for three specific regions: *Noordwest–Friesland*, which usually is the least expensive region, *Bunnik/Zeist*, which usually is the most expensive region, and *Amsterdam*, which is in between. On top we also plot $\log(y_t)$ (scaled to limit the size of the vertical axis in the graph). Comparing the graphs in Fig. 1 suggests that real house prices increase slightly faster than real GDP. Prices in *Bunnik/Zeist* and *Amsterdam* show substantial variations in the trend growth rate over time, with alternating periods of steep price increases and of stable or falling prices. Especially, the 'hump' in the prices around 2000 stands out clearly. This suggests that the trend in the house prices is stochastic rather than deterministic. Furthermore, as the trending behaviour of the different price series seems quite similar regional house prices may well be cointegrated.
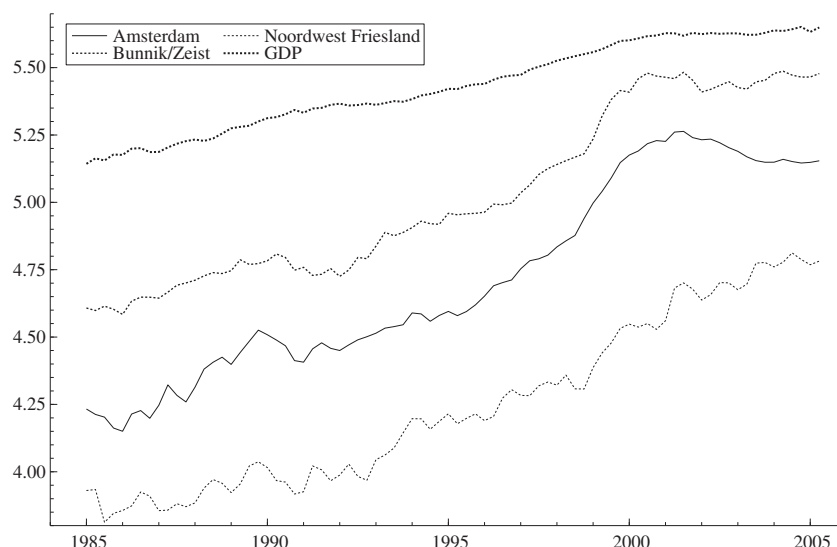
Fig. 1. Log house prices for three distinct regions and log GDP

Unit roots and cointegration

To test whether these visual impressions from Fig. 1 can be given more formal statistical support, we perform panel unit root tests on the regional house prices. Two of the most popular tests in the literature are those from Levin, Lin and Chu (LLC) (2002) and Im, Pesaran and Shin (IPS) (2003), see also Breitung and Pesaran (2008). These tests have as null hypothesis the presence of a unit root in all the series in the panel. The alternative hypotheses are different however. LLC (2002) assume that the house price dynamics are the same for each region, and therefore the alternative hypothesis is that all regional house prices are stationary. IPS (2003), however, have as alternative hypothesis that at least one regional house price is stationary. Both these tests assume that there is no cross-correlation between different series in the panel. In fact, they are not consistent if such a dependence is present, which is quite likely in our case. Alternative tests that do allow for Cross-section Dependence (CD) are available, like the one in Moon and Perron (2004), but these usually rely on asymptotics that require T to be much larger then N, while in our case they are about equal.

To meet our data characteristics, we therefore employ the Cross-sectionally augmented IPS (CIPS) test, recently developed in Pesaran (2007). This allows for cross-sectional dependence, and is also valid when N is larger than T. The idea of the CIPS test is to add the cross-section averages of the lagged levels and first differences to the familiar Augmented Dickey–Fuller (ADF) regression equation. If it can be assumed that the cross-correlations are caused by a common factor, then this common factor must also be present in the cross-section averages. Adding these to the ADF equations should then get rid of the common factor in the residuals and thus correct for the presence of cross-correlations.

As the CIPS test is known to have reduced power relative to the IPS and LLC tests in case cross-correlation is not present, we test whether we really should use the CIPS test instead of these simpler tests. For this purpose, we use the CD test of Pesaran (2004) and the adjusted Lagrange Multiplier (LM_{adj}) test of Pesaran *et al.* (2008). These tests both use the cross-correlations between the residuals of the individual ADF regressions for the different regions. The CD test takes a simple sum which is scaled such that it has a standard normal distribution under the null hypothesis of no cross-sectional dependence. Therefore, the CD test has little power in case there are both positive and negative correlations such that the average is close to zero. The LM_{adj} test, however, is also valid in this case as it employs the squares of the cross correlations in the construction of the test statistic. However, the LM_{adj} test is less robust against nonnormally distributed error terms and exhibits size distortions, especially when N is much larger than T.

Table 1 gives the result of these tests for the panel of quarterly growth rates in house prices $\Delta\log(p_{i,t})$, where Δ denotes the first-difference filter, and of $\log(p_{i,t}) - \log(p_{34,t})$, which is the difference of each series with the log house prices in *Amsterdam* (region 34, see Appendix). The number of lagged (first) differences is allowed to vary across each Covariate ADF (CADF) equation and is determined by minimizing Bayesian Information Criterion (BIC).

Table 1. Results of the CD test, the LM$_{adj}$ test and three different tests for a unit root for two series

Test series	$\Delta[\log(p_{i,t})]$	$\log(p_{i,t}) - \log(p_{34,t})$
CD[a]	**92.0**	**144.2**
LM$_{adj}$[a]	**60.4**	**175.1**
LLC[a]	**−61.2**	2.0
IPS[a]	**−55.9**	1.9
CIPS[b]	**−8.9**	**−3.5**

Notes: Boldface numbers indicate rejection of the null hypothesis.
[a]Test statistic is asymptotically distributed as normal.
[b]Tables with critical values for various values for N and T are given by Pesaran (2007), in the presence of an intercept and a trend in the CADF equations and for $N = T = 70$ the critical value at the 95% level is -2.58, for $N = T = 100$ it is -2.56.

Adding a lagged variable means losing one observation, therefore we actually minimize BIC/T, see Cameron and Trivedi (2005, p. 279) or the definition of BIC given in Franses and Paap (2001). Each CADF regression equation contains an intercept and a trend.

From the second column of Table 1, we see that for the first difference of the log house prices there is substantial cross-sectional dependence, according to both the CD and LM$_{adj}$ tests. Next, we see that all three unit root tests reject the presence of a unit root in these growth rate series. Results for the difference between the log price in a region and the log price in *Amsterdam* (region 34) appear in the third column of Table 1. The reason for examining the log price differences with respect to *Amsterdam* is that finding these to be stationary, we can conclude that the house prices in each region are cointegrated. Again, the CD and LM$_{adj}$ tests indicate that there is substantial cross-sectional dependence. Next, the LLC and IPS unit root tests do not reject the presence of a unit root, but the CIPS test does. Since the LLC and IPS tests are not valid in case of cross-sectional dependence, we rely on the CIPS test and conclude that the log house prices in each region are cointegrated. Note that the $(1, -1)$ cointegration relationships suggested by the results in Table 1 are quite plausible. It means that the difference between the log of house prices, or, equivalently the ratio of house prices, in each region is a stationary process. This constrains the long-term growth of house prices in each region to be about the same.

Clusters

Before we turn to our conditional clustering analysis using latent class techniques, we consider unconditional clustering based on the correlations of the house price growth rates or of the residuals of the ADF regressions used above. For this purpose, we use MDS, which results in the graphs shows in Figs 2 and 3.

Although the graphs in these figures are rather different, they basically lead to the same conclusion that there are no apparent clusters. Hence, dividing the regions into different groups based only on the cross-correlations of the regional house prices is not a meaningful possibility. Apparently, we need a more sophisticated clustering method, perhaps based on latent classes, as we will propose in the next section.

III. The Model

In this section, we put forward the specification of the latent-class panel time series model for describing the regional house prices. We first discuss the characteristics of the model, and then we outline the parameter estimation procedure.

Representation

Our starting point is the latent-class panel time series model developed by Paap *et al.* (2005). The crucial idea behind this model is that the individual time series may be grouped into a limited number of clusters. Within each cluster, a linear model is assumed to describe the dynamic behaviour of the time series. The clusters are defined such that the model parameters are the same for all time series within a cluster, but they are different across clusters. Hence, this model covers the middle ground between a pooled regression model, where the model parameters are constrained to be the same for all regions, and a 'fully heterogenous' model, where the parameters are allowed to be different for each individual region. Whereas a pooled regression model may be too restrictive, a fully heterogenous model may be too flexible and ignores the possible similarities between regions. Finally, the key feature of the model of Paap *et al.* (2005) is that the number of clusters in the model as well as the allocation of the individual time series to different clusters is purely data-based. This avoids *ex ante*, and necessarily subjective, grouping of regions according to geographical location or economic or demographic characteristics, for example.

In our model for quarterly growth rates of house prices, we allow for more flexibility than was done in Paap *et al.* (2005). As mentioned, there are two research questions we want to answer with our model and each question corresponds to different
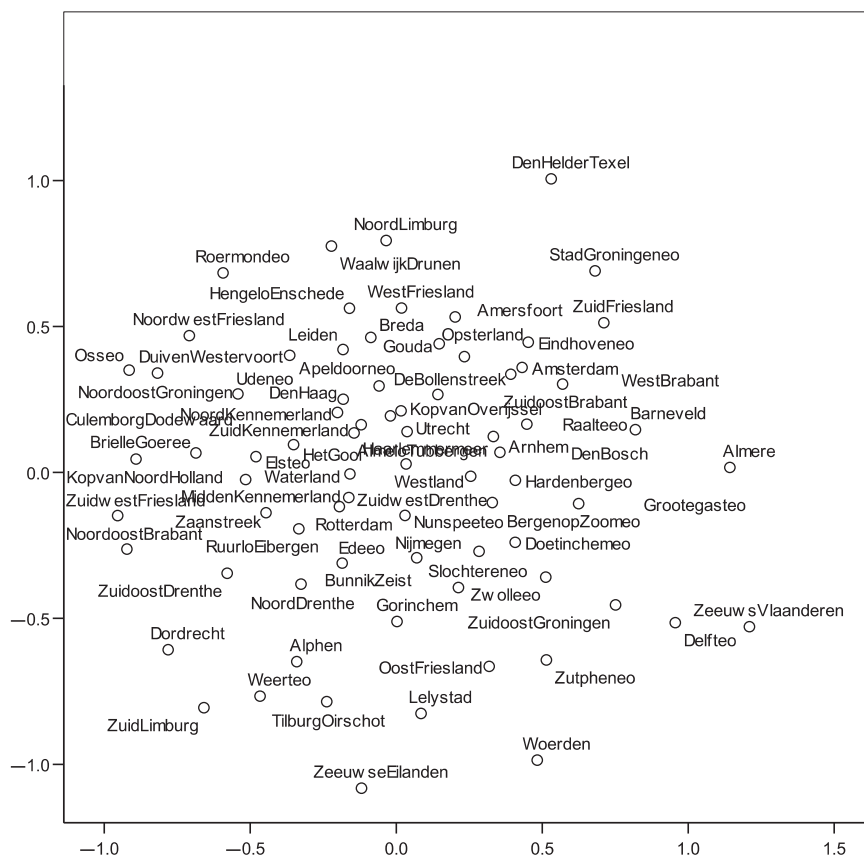
Fig. 2. MDS plot of the regions, based on the correlations of the first difference of the log house prices over the period 1985Q1 to 2005Q4

parameters that can vary across the latent classes. The first is whether the mean growth rates of house prices are the same across all regions. We therefore allow the clusters to have a different average growth rate by allowing for a class-specific intercept. To facilitate interpretation, we demean all other variables in the model such that the intercept is equal to the average growth rate of the house prices in the regions in a cluster.

The second question we wish to answer with our model is whether the house prices in regions follow the trend in real GDP. We add an error correction variable linking regional real house prices and real GDP, where the long-run parameter should be estimated. This long-run parameter determines the size of the effect of GDP on the house prices. The adjustment parameter indicates how fast the house prices in a region react to changes in GDP.

Based on the above discussion, we propose the following latent-class panel time series model for regional house prices in the Netherlands:

$$\Delta \log(p_{i,t}) = \beta_{0,k_i} + \beta_{1,k_i}[\log(p_{i,t-1}) + \gamma_{k_i} \log(y_{t-1})] + \eta_{i,t} \tag{1}$$

The β and γ parameters are class-specific parameters, where the subscript $k_i = 1, \ldots, K$ denotes the latent class which region i belongs to with K being the number of latent classes. We denote the probability that a region belongs to latent class k, or the mixing proportions, as π_k. Naturally, it must hold that $0 < \pi_k < 1$ and that $\sum_{k=1}^{K} \pi_k = 1$.

As the house prices of each regions are cointegrated with GDP, they are also cointegrated amongst themselves. This can easily be seen in the following way. Both $p_{i,t} - \gamma_{k_i} y_t$ and $p_{j,t} - \gamma_{k_j} y_t$ are stationary series. Now, consider the following expression:

$$(p_{i,t} - y_{k_i} y_t) - \delta(p_{j,t} - \gamma_{k_j} y_t)$$
$$= (p_{i,t} - \delta p_{j,t}) - (\gamma_{k_i} - \delta \gamma_{k_j}) y_t \tag{2}$$

The Left-Hand Side (LHS) of (2) is stationary, therefore the Right-Hand Side (RHS) is also a stationary series. For $\delta = \gamma_{k_i}/\gamma_{k_j}$ the second term on the RHS of (2) will disappear, therefore regions i and j must have a $(1, -\delta)$ cointegration relationship. Two regions in the same cluster will therefore have a $(1, -1)$ cointegration relationship, because they share the same γ parameter. As we have seen in the
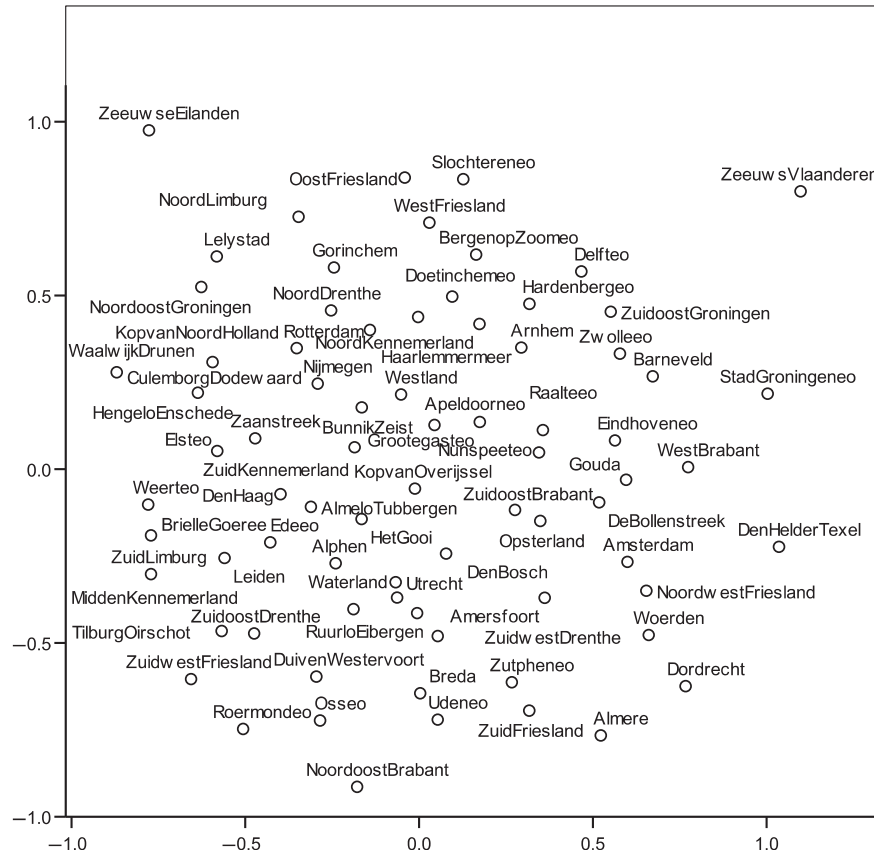
Fig. 3. **MDS plot of the regions, based on the correlations of the residuals of the ADF regressions for the log house prices over the period 1985Q1 to 2005Q4**

Section 'Unit roots and cointegration' there is support for exactly this relationship.

Even though model (1) includes $\log(y_{t-1})$, which is the same for all regions, there may still be some cross-section correlation among the house prices that is not captured. Therefore, following Holly *et al.* (2008), we allow the error term $\eta_{i,t}$ in (1) to be correlated across regions, but assume that this correlation is due to dependence on certain common factors. To be precise, we consider the specification

$$\eta_{i,t} = \alpha_{1,i}\Delta\log(y_{t-1}) + \alpha_{2,i}I_{t-1} + \alpha_{3,i}\Delta\log(p_{t-1}) + \varepsilon_{i,t} \tag{3}$$

where I_{t-1} denotes the interest rate at time $t-1$, p_{t-1} denotes the average house price in the Netherlands at time $t-1$ and where $\alpha_{l,i}$ for $l=1,2,3$ are region-specific parameters. The residuals $\varepsilon_{i,t}$ are now assumed to be independently normally distributed with a region-specific variance σ_i^2.

In the application below, we demean all variables in (1) and (3) and hence the intercepts β_{0,k_i} in (1) are equal to the average growth rates of the house prices in the latent classes k_i for $k_i = 1, \ldots, K$.

Estimation

The parameters in our model (1) with (3) can be estimated as outlined in Paap *et al.* (2005), using the Expectation–Maximization (EM) algorithm of Dempster *et al.* (1977). This makes use of the full data log-likelihood function, that is, the joint density of the house prices and the latent classes k_i, which we specify in detail below. The EM algorithm is an iterative maximization algorithm, which alternates between two steps until convergence occurs. In the first step (E-step) we compute the expected value of the full data log-likelihood function with respect to the latent classes k_i, $i = 1, \ldots, N$, given the house prices and the current values of the model parameters. In the second step (M-step), we maximize the expected value of the full data log-likelihood function with respect to the model parameters. As the model given the class memberships can be written as a standard linear regression, the M-step amounts to a series of (weighted) regressions. As the EM algorithm maximizes the log-likelihood function, the resulting estimates of the model parameters are equal to the Maximum Likelihood (ML) estimates. We can

therefore compute SEs of the estimates using the second derivative of the log-likelihood function.

Note that due to the presence of the term $\beta_{1,k_i}[\log(p_{i,t-1}) + \gamma_{k_i}\log(y_{t-1})]$ the model in (1) is actually nonlinear in the parameters. To deal with this issue, we follow Boswijk (1994) and rewrite the model as

$$\Delta\log(p_{i,t}) = \beta_{0,k_i} + \beta_{1,k_i}\log(p_{i,t-1}) + \beta_{2,k_i}\log(y_{t-1}) + \eta_{i,t} \tag{4}$$

where $\beta_{2,k_i} = \beta_{1,k_i}\gamma_{k_i}$. Note that (4) is linear in the parameters, which facilitates estimation. The ML estimate $\hat{\gamma}_{k_i}$ can be obtained from the ML estimates of β_{1,k_i} and β_{2,k_i} as $\hat{\beta}_{2,k_i}/\hat{\beta}_{1,k_i}$.

The full data likelihood function, that is, the joint density of $\mathcal{P} = \{\{\Delta\log p_{i,t}\}_{t=1}^T\}_{i=1}^N$ and $\mathcal{K} = \{k_i\}_{i=1}^N$ is given by

$$l(\mathcal{P}, \mathcal{K}; \theta) = \prod_{i=1}^N \left(\prod_{k=1}^K \left(\pi_k \prod_{t=1}^T \frac{1}{\sigma_i}\phi(\varepsilon_{i,t}^k/\sigma_i) \right)^{I[k_i=k]} \right) \tag{5}$$

where $\phi(\cdot)$ denotes the probability density function of a standard normal random variable and θ is a vector containing all model parameters. The error term at time t for region I belonging to cluster k is defined as

$$\varepsilon_{i,t}^k = \Delta\log p_{i,t} - x_{i,t}'\beta_k - w_t'\alpha_i \tag{6}$$

where $x_{i,t}$ is the (3×1) vector with the regressors appearing in (4) and β_k contains the corresponding parameters for cluster k. Similarly, w_t is the (3×1) vector with common factors in the specification for $\eta_{i,t}$ in (3), and $\alpha_i = (\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i})'$ containing the parameters for region i.

The expectation of the full data log-likelihood function with respect to $\mathcal{K}|\mathcal{P}, \theta$ [E-step] is given by

$$\mathcal{L}(\mathcal{P}; \theta) = \sum_{i=1}^N \left(\sum_{k=1}^K \hat{\pi}_{i,k}\left(\ln\pi_k + \sum_{t=1}^T \right. \right.$$
$$\left. \left. -\frac{1}{2}\ln\sigma_i^2 - \frac{1}{2}\ln 2\pi - \frac{(\varepsilon_{i,t}^k)^2}{2\sigma_i^2} \right) \right) \tag{7}$$

where $\hat{\pi}_{i,k}$ denotes the conditional probability that region i belongs to class k. This is equal to

$$\hat{\pi}_{i,k} = \frac{\pi_k \prod_{t=1}^T \frac{1}{\sigma_i}\phi(\varepsilon_{i,t}^k/\sigma_i)}{\sum_{l=1}^K \pi_l \prod_{t=1}^{T_i} \frac{1}{\sigma_i}\phi(\varepsilon_{i,t}^l/\sigma_i)} \tag{8}$$

In the M-step, we need to maximize (7) with respect to the parameters β_k, π_k, $k=1,\ldots,K$ and α_i, σ_i^2 for $i=1,\ldots,N$. We perform this maximization step sequentially. First, we optimize over β_k keeping the other parameters fixed. This can be done by a simple

weighted regression of $\Delta\log(p_{i,t}) - w_t'a_i$ on $x_{i,t}$ with weight given by $\sqrt{\hat{\pi}_{i,k}}/\sigma_i$. Clearly, we want regions with a larger probability of belonging to class k to have a larger weight in estimating β_k. At the same time, regions with a larger SD of the error term should get a smaller weight, as their house prices contain relatively more noise and less information about β_k. Each β_k, $k=1,\ldots,K$ is estimated in a separate weighted regression.

Second, we optimize the log-likelihood function over α_i for $i=1,\ldots,N$. We do this by regressing $\sum_{k=1}^K \hat{\pi}_{i,k}[\Delta\log(p_{i,t}) - x_{i,t}\beta_k]$ on w_t. The dependent variable in this regression is the conditional expectation of $\eta_{i,t}$. We perform these regressions for each region separately.

Next, the new estimate of σ_i^2 is given by

$$\sigma_i^2 = \frac{1}{T}\sum_{t=1}^T \sum_{k=1}^K \hat{\pi}_{i,k}(\varepsilon_{i,t}^k)^2 \tag{9}$$

for $i=1,\ldots,N$. Finally, the mixing proportions are updated by averaging the conditional class membership probabilities, that is,

$$\pi_k = \frac{1}{N}\sum_{i=1}^N \hat{\pi}_{i,k} \tag{10}$$

for $k=1,\ldots,K$.

As we maximize over the parameters sequentially in the M-step, we do not reach the optimum of the expected full data log-likelihood function (7) in each iteration of the EM-algorithm. We can repeat the individual update steps until convergence, but this is not necessary. Indeed, Meng and Rubin (1993) have shown that an increase in the full-data log-likelihood function in the M-step is sufficient for the EM algorithm to converge to the maximum of the log-likelihood function.

Determining the appropriate number of latent classes is not straightforward. We cannot use a standard statistical test, due to the Davies (1977) problem of unidentified nuisance parameters under the null hypothesis. The usual approach is using a criterion function balancing the fit and the complexity of the model, where the model fit is measured by the value of the log-likelihood function while the number of model parameters provides a measure of complexity. The most well-known criteria are the Akaike information Criterion (AIC) and the BIC. Bozdogan (1994) suggests that the AIC should have a penalty factor of three instead of two in the case of mixture models. Indeed, Andrews and Currim (2003) show that this AIC-3 criterion outperforms other criteria. Bozdogan (1987) modifies the AIC into the so-called Consistent AIC (CAIC), which is almost

Table 2. Results for the fully heterogenous model

Parameter	Mean	Minimum	Maximum
β_0	0.011	0.006	0.013
β_1	−0.363	−0.692	0.125
γ	−0.591	−7.601	2.127
α_1	−0.271	−1.851	0.703
α_2	−0.004	−0.013	0.004
α_3	0.277	−0.340	0.739

equal to BIC. He shows that when the sample size is large the CAIC and BIC criteria perform better than AIC. We will consider all four criteria below.

IV. Empirical Results

In this section we discuss the results of applying our model to the regional house price data for the Netherlands described in Section II. The effective sample period ranges from 1985Q3 (because we have $\Delta\log(p_{t-1}) = \log(p_{t-1}) - \log(p_{t-2})$ in our model) to 2005Q2 (because we only have real GDP data until 2005Q2), giving $T = 80$ data points in the time series dimension. To obtain a first impression of the extent of similarities across regions, we start by estimating a fully heterogenous model allowing for different parameters for each region. Next, we provide estimation results for the model with a limited number of latent classes. Finally, we consider impulse-response functions for two interesting scenarios to provide further interpretation of the model.

A fully heterogenous model

We first estimate the parameters in a fully heterogenous model, that is, we estimate the model in (1) with (3) allowing for different parameters for each individual region. This essentially is a model with $K = 76$ latent classes, in which case each region forms a separate class.

The mean, minimum and maximum of each parameter of the 76 regions can be found in Table 2. Figure 4 displays the histograms for the 76 estimated values for each of the parameters β_j, $j = 0$, 1, and γ in (1). The top panel shows the intercepts, β_0, which equal the quarterly growth rates. These are all positive, reflecting the upward trend in the house prices, and range between 0.6% and 1.3% per quarter. The middle panel of Fig. 4 shows the results for the adjustment parameter for the cointegrating

term with GDP. We find some positive values, which is not as expected, as these imply divergence between GDP and the house prices in that region. Finally, the histogram in the bottom panel shows the parameter γ in the cointegration relationship with GDP, which we except to be negative as we expect the house prices and GDP to move in the same direction. Table 2 also shows the results for the α parameters from (3). Again we find that they show some counterintuitive signs and a relatively large spread.

We can see from these results that some form of aggregation may be useful, as we now get a wide variety of parameter estimates, with sometimes quite implausible results. At the same time, this variety also suggests that we should perhaps better not restrict the parameters to be the same across all regions. Hence, it may be optimal to allow for a limited number of different clusters.

A model with latent classes

A major issue for successful application of the latent-class panel time series model is of course determining the appropriate number of latent classes. As discussed in the section 'Estimation', we consider four different information criteria for this purpose. Table 3 shows the values of these criteria for models with one to four and 76 classes. For all criteria, we see that going from a homogenous model (with a single class) to two classes amounts to a relatively large improvement in the balance of model fit and complexity. After this, adding more classes does not improve any of the criteria. We therefore focus on the model with two latent classes.

The estimation results for the model with two latent classes are given in Table 4. Additionally, Table 5 gives the results for a series of Wald tests which we use to examine whether the parameters for the different classes are significantly different from each other. The estimation results show that the regions in the two latent classes do indeed differ from each other in several important respects. First, the estimated intercepts show that the average growth rate in class 1 is slightly higher than in class 2.[1] This difference is not significant though, as can be seen from the second row of Table 5. The average growth rate in class 1 is equal to 1.2% per quarter, or 4.8% annually, while the house prices in class 2 grow with 1.1% per quarter, or 4.4% annually.

Second, examining the cointegrating relationship with GDP, we find that class 1 has a significantly larger adjustment parameter. Thus, the house prices

[1] Recall that we demeaned all other variables the model, so the intercepts represent the average growth rates.
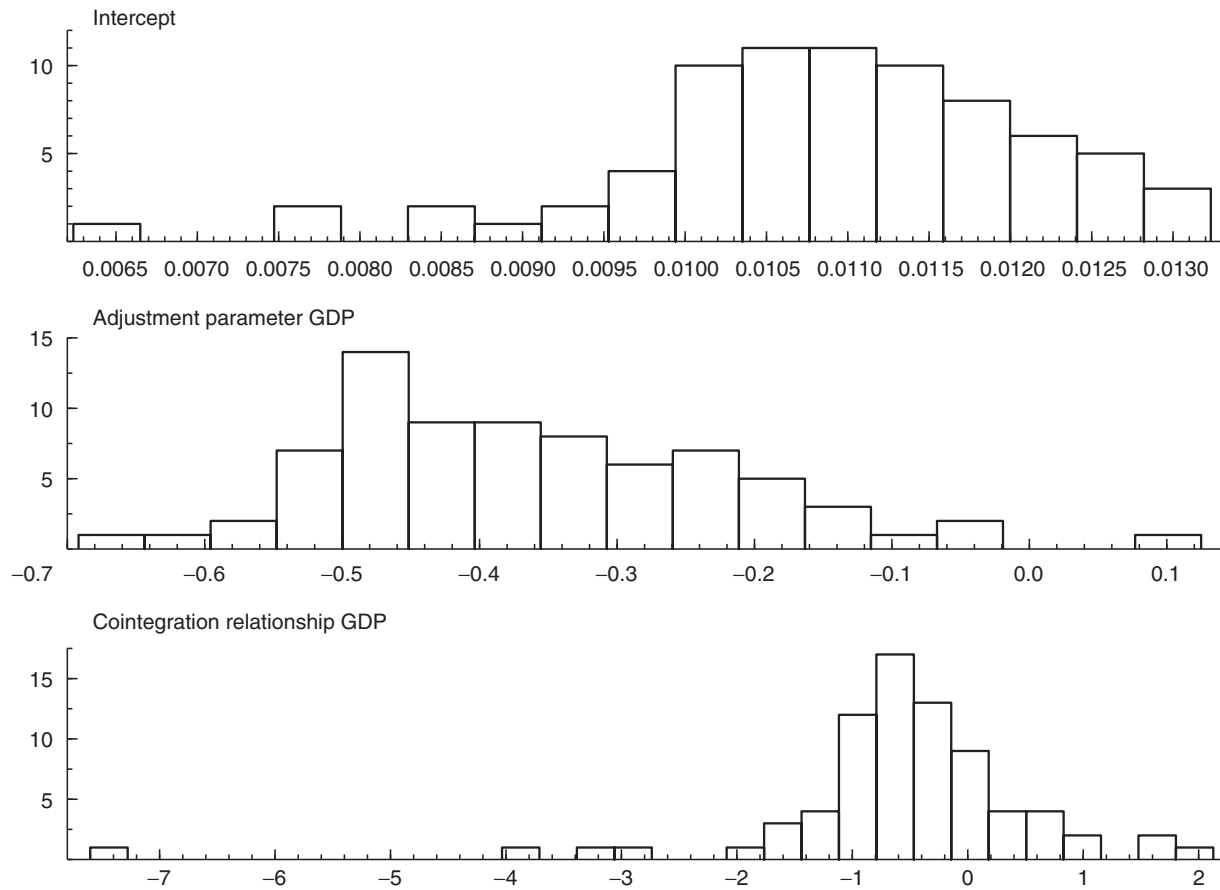
Fig. 4. Histograms of the estimated values of the parameters β_j, $j = 0, 1$ and γ in (1) in the fully heterogeneous model with 76 classes

Table 3. Criteria values for different numbers of latent classes

Criterion\K	1	2	3	4	76
AIC	3.937	**−4.059**	−4.058	−4.057	−3.962
AIC-3	3.887	**−4.008**	−4.006	−4.004	−3.862
BIC	3.598	**−3.716**	−3.710	−3.704	−3.292
CAIC	3.584	**−3.665**	−3.658	−3.652	−3.192

Note: Boldface numbers indicate the optimum.

in regions belonging to cluster 1 react faster to changes in GDP than the house prices in class 2.

Finally, the cointegrating relationship between house prices and GDP itself is also significantly different across the classes. For class 1, it is $(1, -1.89)$, meaning that in the long run the house prices in the regions in this cluster grow almost twice as fast as GDP. In class 2 the cointegration relationship is $(1, -1.68)$. These long-term relationships may not be very plausible, however, as we could already see from Fig. 1, they are a good description of the development of house prices and GDP in the sample period.

Table 4. Estimation results for $K = 2$ latent classes

Class	Estimate	SE
Intercept $\beta_{0,k}$		
1	0.012	0.001
2	0.011	0.000
Adjustment parameter GDP $\beta_{1,k}$		
1	−0.178	0.019
2	−0.131	0.007
Cointegration relationship GDP γ_k		
1	−1.888	0.083
2	−1.684	0.035
Mixing proportions π_k		
1	0.202	0.159
2	0.798	–

Table 5. Wald tests for equality of the parameters across the two classes in (1)

Restriction	Wald statistic	p-value
$\beta_{0,1} = \beta_{0,2}$	0.61	0.41
$\beta_{1,1} = \beta_{1,2}$	7.10	0.01
$\gamma_1 = \gamma_2$	6.90	0.01

Table 6. Cointegration relationships between the regions from clusters *i* and *j*

$i\backslash j$	1	2
1	$(1, -1)$	$(1, -1.12)$
2	$(1, -0.89)$	$(1, -1)$

As shown in Section III, the cointegration relationship of each region with GDP entails that the regions are also cointegrated among themselves. The long-term parameter is only influenced by the γ parameters of the two regions involved, and thus only depends on the class membership of the two regions. Table 6 shows these cointegration relationships between the house prices of regions from any of the two clusters. First, we see that two regions that belong to the same cluster are $(1, -1)$ cointegrated. This is actually very intuitive, as they follow the same trend relative to the trend of GDP, they must follow the exact same trend themselves. Next, we find that a region from cluster 1 is $(1, -1.12)$ cointegrated with a region from cluster 2. This corresponds with the slightly higher growth rate in class 1.

The parameters in (3) are region-specific, and full estimation results are not reported to save space. Only 11% of the $\alpha_{1,i}$ parameters is significant, suggesting that the impact of GDP on the house prices is mostly captured by the cointegration term. Moreover, only 22% has the expected positive sign. The $\alpha_{2,i}$ parameters are mostly negative, and only one region has an (insignificant) positive value. Furthermore, for 63% of the regions the $\alpha_{2,i}$ parameter is significant at the 5% level, indicating that the interest rate indeed influences the house prices in the expected direction. The $\alpha_{3,i}$ parameters, relating the growth of the house price in a region to growth of the average house price in the Netherlands in the previous quarter, is positive for 88% of the regions, but only significant for 42% of these regions.

The latent classes. The parameter estimation results obviously become more interesting if we know which regions belong to each of the two classes. Therefore, we compute the conditional class membership probabilities using (8). The resulting classification of the regions is shown in Fig. 5. Regions are coloured based on $\hat{\pi}_{i,1}$, the probability of belonging to class 1. Regions are coloured in four shades of grey. For the regions that are coloured in the lightest shade it holds that $\hat{\pi}_{i,1} \leq 0.2$. For regions coloured in subsequently darker shades of grey it holds that $0.2 < \hat{\pi}_{i,1} \leq 0.4$, $0.4 < \hat{\pi}_{i,1} \leq 0.6$ or $0.6 < \hat{\pi}_{i,1} \leq 0.8$. There were no regions with $\hat{\pi}_{i,1} > 0.8$. It can be seen that most

regions are either very dark or very light, suggesting that the classification is very clear for most regions. In fact, the average value of $\max(\hat{\pi}_{i,1}, \hat{\pi}_{i,2})$ is equal to 0.83.

We find that class 1 contains mainly rural regions surrounding the big cities in the Netherlands. The regions in this class mainly cover parts of *Noord-Brabant* and the *Veluwe*. Even though the East belongs almost completely to class 1, the larger cities of the East, like *Zwolle*, *Almelo*, *Hengelo*, *Enschede* and *Arnhem* are part of class 2.

Class 2 contains different types of regions. First, it contains many large cities in different parts of the country, like *Breda* and *Groningen*, as well as almost all of the regions in the *Randstad*, the densely populated western part of the country. At the same time some rural regions, like *Zeeland, Zuid-Limburg* and regions in the North belong to this class with high probability. Note that these rural regions are not as close to the *Randstad* as most of those in class 1.

A possible explanation for our results is the increased number of commuters that live in the regions belonging to class 1 and who work in the large western cities. If the number of commuters increases, it is likely that they move to regions in cluster 1, as these are still at travelling distance from the *Randstad*. This development has two consequences for the regions in class 1. First, the average income in these regions is likely to increase, as the individuals who move away from the cities are relatively wealthy. The second consequence is an increase of housing quality in these regions, as wealthier people leaving the cities will increase the demand for more luxurious houses.

These potential structural changes within the regions of cluster 1 are consistent with all of our findings. First, the increase in housing quality will result in a larger increase in the average house prices in class 1 as compared to class 2. Our second finding is that house prices in these regions react faster to changes in GDP. This may be caused by the fact that the increase of their income may influence the decision of these individuals to move and start commuting. Our last and most striking finding is that the house prices in class 1 increase almost twice as fast as GDP. Note, however, that the increase is not corrected for higher housing quality.

Impulse-response functions

To give further interpretation to our estimation results we compute impulse-response functions for two interesting scenarios, each occurring in the second quarter of 2005. In the first scenario real GDP receives a shock of 1%. In the second scenario
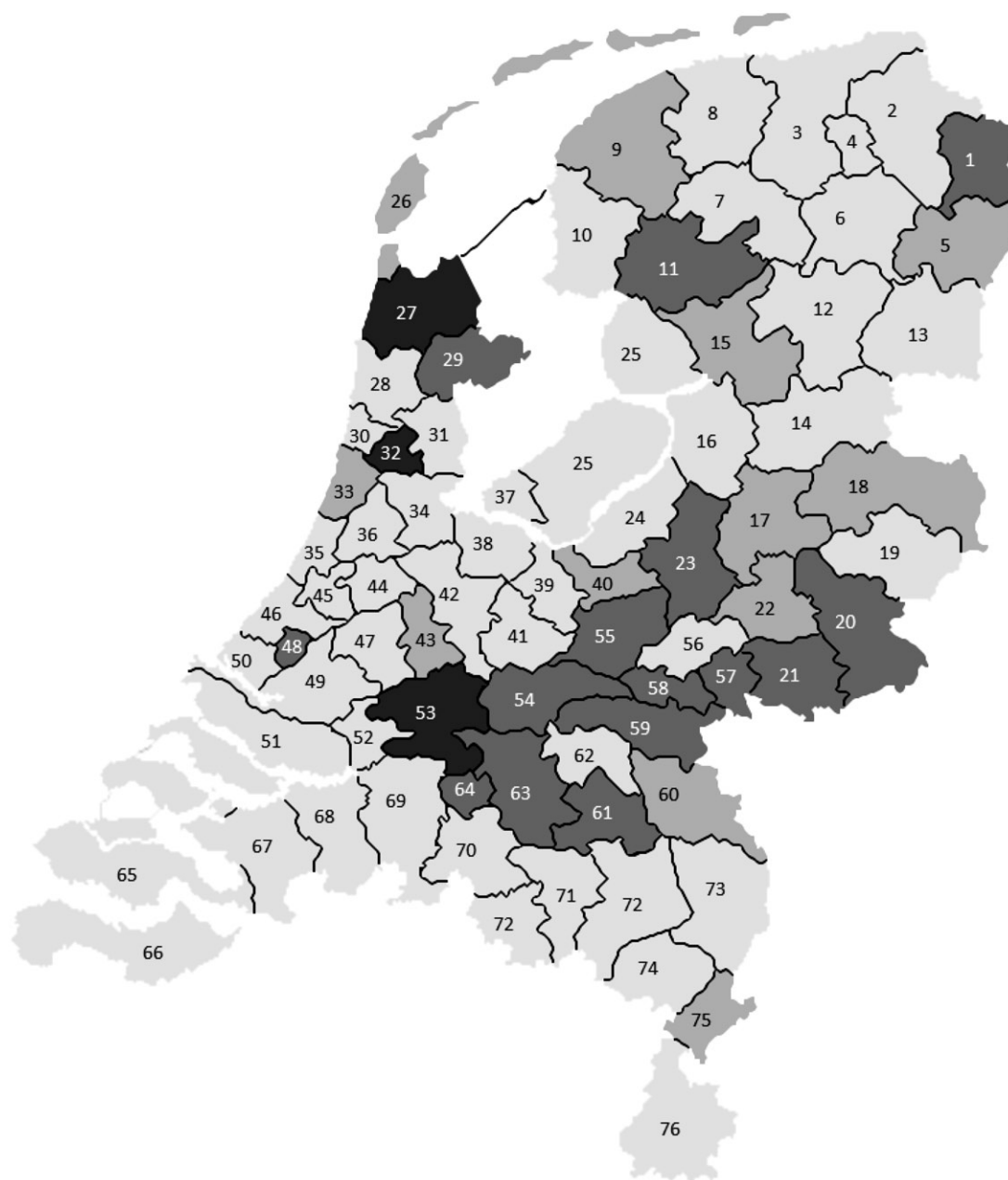
Fig. 5. Clustering of regions. Regions with a high probability belonging to class 1 are coloured dark, regions with a low probability belonging to class 1 are coloured lighter. The numbers inside the regions correspond to the ones in Appendix

real GDP stays the same, but the interest rate receives a shock of 1% point. We forecast the house prices for each of the scenarios and compare with a no-change scenario, for the subsequent 3-year period from 2005Q3 to 2008Q2.

In order to compute the impulse responses up to 12 quarters ahead, we also need forecasts for GDP and the interest rate, as these variables also affect house prices, see (1). Here we assume that the interest rate stays the same during the forecast period. In scenario 3, the interest rate is higher, but still assumed

to be constant over the whole forecast period. To obtain forecasts for GDP we construct a simple AR(q) model with intercept for $\Delta \log y_t$. We choose q based on out-of-sample forecasting performance, where we use the last 3 years as a hold-out sample. It turns out that $q = 8$ gives the best performance.

Figure 6 shows the impulse-response functions of the log house prices with respect to the log of GDP. The y-axis gives the relative change in house prices between the two scenarios, that is, a value of 0.01 means that the house price is 1% higher than the

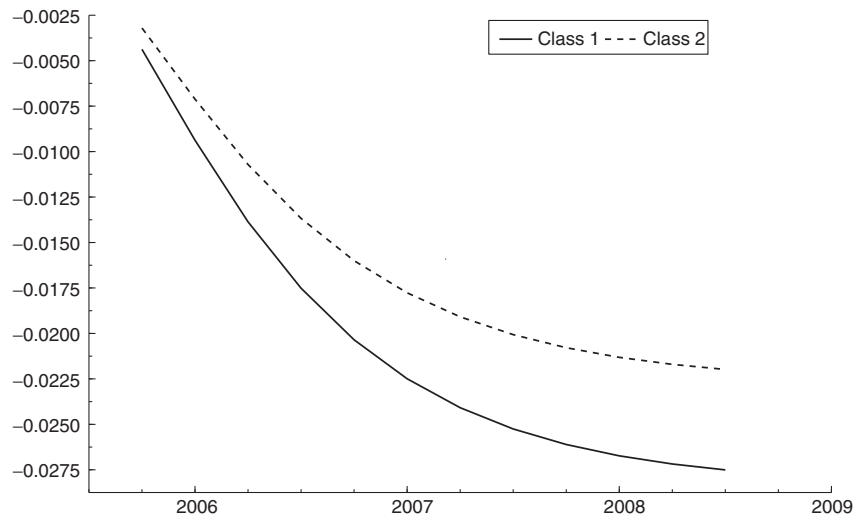**Fig. 6.** Impulse-response function of $\log(p_{i,t})$ with respect to $\log(y_t)$ for three regions

Fig. 7. Impulse-response function of $\log(p_{i,t})$ with respect to $\log(I_t)$ for three regions

reference forecast. We calculate the impulse-response functions for each of the 76 regions. We then aggregate these to average responses in the two clusters.

We find that the effect of an increase in GDP is initially negative in both clusters, which is caused by the many negative $\alpha_{1,i}$ parameter in both clusters. However, this negative effect lasts only one quarter, and after that the house prices are higher compared to the reference forecasts. As expected, we find that the house prices in cluster 1 react both faster and more to the change in GDP.

In the second scenario, the interest rate receives a shock, and increases from 2.06% to 3.06% (Figure 7). We find that the house prices are falling. After 3 years

the house prices are about 2% lower in cluster 2 and almost 3% lower in cluster 1, as compared to the reference forecasts.

V. Conclusions

In this article, we developed a latent-class panel time series model for describing several key characteristics of regional house prices in the Netherlands between 1985 and 2005. An important feature of the model is that we cluster the regions in separate classes, where the price dynamics of house prices in regions within the same class are similar, while they are different

across the classes. For the 76 regions in the Netherlands we find that two classes are sufficient. The first class contains mainly rural regions close to large cities. The second class contains both the larger cities and some more remote rural regions. The house prices in regions in the first class are characterized by slightly higher average growth rates, and stronger and faster reactions to changes in GDP. These findings may be caused by the increased number of commuters. Indeed, the number of people working in the larger cities, but living in the regions of class 1, has increased substantially during our sample period.

Our model allows for the analysis of rather detailed data. To fully exploit its properties, one would want to analyse even further disaggregated data. The collection of such more detailed series is left to further research. Another issue for further research is to make the class probabilities dependent on certain explanatory variables.

References

Andrews, R. L. and Currim, I. S. (2003) A comparison of segment retention criteria for finite mixture logit models, *Journal of Marketing Research*, **40**, 235–43.

Boswijk, H. P. (1994) Testing for an unstable root in conditional and structural error correction models, *Journal of Econometrics*, **63**, 37–60.

Bozdogan, H. (1987) Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions, *Psychometrica*, **52**, 345–70.

Bozdogan, H. (1994) Mixture-model cluster analysis using model selection criteria and a new information measure of complexity, in *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach* (Ed.) H. Bozdogan, Kluwer, Boston.

Breitung, J. and Pesaran, M. H. (2008) Unit roots and cointegration in panels, in *The Econometrics of Panel Data*, 3rd edn (Eds) L. Matyas and P. Sevestre, Kluwer, Boston, pp. 115–85.

Cameron, G., Muellbauer, J. and Murphy, A. (2006) Was there a British house price bubble? Evidence from a regional panel, Economics Series Working Papers No. 276, Department of Economics, University of Oxford.

Cameron, A. C. and Trivedi, P. K. (2005) *Microeconometrics*, Cambridge University Press, New York.

Davies, R. B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika*, **64**, 247–54.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Franses, P. H. and Paap, R. (2001) *Quantitative Models in Marketing Research*, Cambridge University Press, Cambridge.

Holly, S., Pesaran, M. H. and Yamagata, T. (2008) A spatio-temporal model of house prices in the US, *Journal of Econometrics*, Cambridge Working Papers in Economics No. 654, University of Cambridge and CESInfo Working Paper Series No. 1826, forthcoming.

Im, K. S., Pesaran, M. H. and Shin, Y. (IPS) (2003) Testing for unit roots in heterogenous panels, *Journal of Econometrics*, **115**, 53–74.

International Monetary Fund (IMF) (2004) The global house price boom, *World Economic Outlook*, 1491–517.

Levin, A., Lin, C. F. and Chu, C. J. (LLC) (2002) Unit root tests in panel data: asymptotic and finite-sample asymptotics, *Journal of Econometrics*, **108**, 1–24.

Malpezzi, S. (1999) A simple error correction model of house prices, *Journal of Housing Economics*, **8**, 27–62.

Meng, X.-L. and Rubin, D. B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika*, **80**, 267–78.

Moon, H. R. and Perron, B. (2004) Testing for a unit root in panels with dynamic factors, *Journal of Econometrics*, **122**, 81–126.

Paap, R., Franses, P. H. and van Dijk, D. J. C. (2005) Does Africa grow slower than Asia, Latin America and the Middle East? Evidence from a new data-based classification method, *Journal of Development Economics*, **77**, 553–70.

Pesaran, M. H. (2004) General diagnostic tests for cross section dependence in panels, Cambridge Working Papers in Economics No. 435, University of Cambridge and CESInfo Working Paper Series No. 1229.

Pesaran, M. H. (2007) A simple panel unit root test in the presence of cross-section dependence, *Journal of Applied Econometrics*, **22**, 265–312.

Pesaran, M. H., Ullah, A. and Yamagata, T. (2008) A bias-adjusted LM test of error cross section independence, *The Econometrics Journal*, **11**, 105–27.

Appendix

Regions by number

1 Noordoost-Groningen	27 Kop v. Noord-Holland	52 Dordrecht
2 Slochteren +s	28 Noord-Kennemerland	53 Gorinchem
3 Grootegast +s	29 West-Friesland	54 Culemborg/Dodewaard
4 Stad Groningen +s	30 Midden-Kennemerland	55 Ede +s
5 Zuidoost-Groningen	31 Waterland	56 Arnhem
6 Noord-Drenthe	32 Zaanstreek	57 Duiven/Westervoort
7 Opsterland	33 Zuid-Kennemerland	58 Elst +s
8 Oost-Friesland	34 Amsterdam	59 Nijmegen
9 Noordwest-Friesland	35 De Bollenstreek	60 Noordoost-Brabant
10 Zuidwest-Friesland	36 Haarlemmermeer	61 Uden +s
11 Zuid-Friesland	37 Almere	62 Oss +s
12 Zuidwest-Drenthe	38 Het Gooi	63 Den Bosch
13 Zuidoost-Drenthe	39 Amersfoort	64 Waalwijk/Drunen
14 Hardenberg +s	40 Barneveld	65 Zeeuwse Eilanden
15 Kop van Overijssel	41 Bunnik/Zeist	66 Zeeuws-Vlaanderen
16 Zwolle +s	42 Utrecht	67 Bergen op Zoom +s
17 Raalte + s	43 Woerden	68 West-Brabant
18 Almelo Tubbergen	44 Alphen	69 Breda
19 Hengelo Enschede	45 Leiden	70 Tilburg/Oirschot
20 Ruurlo Eibergen	46 Den Haag	71 Eindhoven +s
21 Doetinchem +s	47 Gouda	72 Zuidoost-Brabant
22 Zutphen +s	48 Delft +s	73 Noord-Limburg
23 Apeldoorn +s	49 Rotterdam	74 Weert +s
24 Nunspeet +s	50 Westland	75 Roermond +s
25 Lelystad	51 Brielle/Goeree	76 Zuid-Limburg
26 Den Helder/Texel		

Note: +s means including surrounding area.