Testing for common factors and cross-sectional dependence among individual housing prices

Nicolás Durán and J. Paul Elhorst[1]

Faculty of Economics and Business, University of Groningen
PO Box 800, 9700 AV Groningen, The Netherlands
E-mail: n.duran@rug.nl and j.p.elhorst@rug.nl

8 June 2017

Abstract

This paper extends the cross-sectional dependence (*CD*) test of Pesaran (2004, 2015a) and the exponent of cross-sectional dependence test of Bailey et al. (2016b) to a panel that is unbalanced, both in the time and cross-sectional domain. The modified tests are applied to 163,323 individual housing transactions that took place in the provinces of Groningen, Friesland and Drenthe located in the Netherlands over the period 2003-2014.

K**ey words:** Housing prices, weak and strong cross-sectional dependence

JEL classification: C21, C23, R23

[1] The authors thank Joost Groeneveld (University of Groningen, Netherlands) for research assistance.

1. Introduction

Testing and accounting for cross-sectional dependence when evidence is found in favor of it has become a major research area in the econometrics literature. If a set of cross-sectional observations at a particular point in time are interdependent, treating them as being independent, which is the standard if a linear regression model is estimated by ordinary least squares (OLS), may lead to biased or inefficient parameter estimates. For example, if a house is put for sale on the market and the owner, or a real estate agent representing the owner, uses information of houses with similar characteristics that are for sale or have been sold in the past to set the asking price, known as the sales comparison approach, individual housing prices will no longer be independent of each other. Similarly, if a potential buyer of a house compares quality for money, that is, if he searches for the best possible set of housing characteristics within a particular search area given a particular budget, his bid for one house will depend on the asking price and characteristics of other houses. Finally, if housing prices of all houses go up and down along the business cycle, individual housing prices are not independent either, since they are affected by a third factor.

The first type of cross-sectional dependence is known as (local) spatial dependence and the second type as (global) common factors. Both are also viewed as 'weak' and 'strong' cross-sectional dependence (Chudik and Pesaran, 2015). Two statistics have been developed to test for cross-sectional dependence: the cross-sectional dependence (*CD*) test of Pesaran (2004, 2015a) and the exponent α-test test of Bailey et al. (2016b). Unfortunately, these tests have been developed for a balanced spatial panel only, i.e. for a cross-section of N units over T time periods. For example, the application on housing prices presented in Bailey et al. (2016a) as an empirical illustration of both tests employs aggregated data of 363 MSAs over the period 1975Q1-2010Q4 (T=144). One exception is Pesaran (2015b, Section 29.8; see also Chudik and Pesaran, 2015, section 1.7) who explains how to modify the CD test when having an unbalanced panel due to missing observations in the time domain. However, most studies trying to explain housing prices are based on data which are also unbalanced in the cross-sectional domain. Table 1 provides a simple numerical example of the number of housing transactions in two units and two time periods to illustrate the problem; the number of transaction varies across both space and time.

Table 1: Example of unbalanced panel of housing transaction representative for most studies

# Transactions	Unit 1	Unit 2	Total research area
Period 1	2	3	5
Period 2	1	2	3
Total sample period	3	5	8

In this particular study, we have data on 163,323 housing transactions in the provinces of Groningen, Friesland and Drenthe located in the Netherlands over the period 2003-2014, subdivided over 948 postcode areas. This dataset has been made available by the NVM

(Nederlandse Vereniging van Makelaars en Taxateurs), the largest association of real estate agents in the Netherlands. The variables used in this study are the transaction price, the transaction price per square meter of living space, the number of weeks the house has been on the market, and a variable measuring the physical impact of a series of earthquakes due to gas extraction from the soil in the province of Groningen. The latter variable is provided by the Geo-services office of the University of Groningen. They assembled a dataset with data collected by the KNMI (Koninklijk Nederlands Meteorologisch Instituut), the Dutch meteorology institute, containing the date, geographical location, magnitude, and depth of each earthquake that occurred in the Netherlands since seismic events started being measured in the eighties. Between 1985 and 2015 the Netherlands has been hit by 1100 earthquakes according to this dataset. These earthquakes are all relatively small in magnitude on the scale of Richter (smaller than 4), but when taken together there is empirical evidence that they have affected transaction prices (see Koster and Van Ommeren, 2015 for a study that appeared in English). Although further research is beyond the topic of this study, a major issue in all studies on this topic so far has been to find out whether global common factors and local spatial dependence are relevant extensions to a standard hedonic price model that need to be accounted for. Up to now, this problem has not been systematically analyzed. Most studies select reference areas not affected by earthquakes (Bosker et al., 2016; CBS, 2017).

Just as the illustration in Table 1, our data set of individual housing transactions is anything but balanced. The majority of houses has been sold only once during the observation period, as a result of which it is not possible to treat individual houses as units. Just as Bailey et al. (2016a), we can aggregate the data to a smaller sample of N geographical units and T time periods based on the location of the houses and the transaction dates, and then calculate the statistics based on these N times T observations, but this might lead to a considerable aggregation bias. Figure 1 shows the average number of transactions per year in descending order for the zip code areas in the sample and so indicates the amount of information that is lost when using aggregated data. The aim of this paper is modify the expressions of the two test statistics such that they can also be calculated based on an unbalanced panel of individual data observations. For this purpose, the paper is set up as follows. Section 2 provides detailed mathematical descriptions and background explanations of the original CD and exponent α-tests. Section 3 presents the modifications that are proposed in this paper. Section 4 presents and discusses results, and Section 5 concludes.

<< Insert Figure 1 here >>

2. Cross-sectional dependence tests for balanced data

Suppose a balanced spatial panel of N cross-sectional units over T time periods for a particular variable x_{it} (i=1,…,N; t=1,…,T). The Pesaran (2015a, eq.10) *CD* test is then defined as

$$CD = \sqrt{2\,T/N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{ij}, \quad (1)$$

where $\hat{\rho}_{ij}$ denotes one of the N times N-1 mutual correlation coefficients between the time-series of each pair of units i to j, and T is the number of observations on each unit. The correlation coefficients are obtained by

$$\hat{\rho}_{ij} = \frac{\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{jt}-\bar{x}_j)}{\sqrt{\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2}\sqrt{\sum_{t=1}^{T}(x_{jt}-\bar{x}_j)^2}}, \text{ where } \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T}x_{it}. \quad (2)$$

The expression $\sqrt{2T/N(N-1)}$ is taken up before the two summation signs in (1) since each correlation coefficient has the same weight. There are $N(N\text{-}1)$ mutual correlation coefficients, which explains the division by $N(N\text{-}1)$, calculated over T observations, which explains the multiplication by T. The number 2 is added since the correlation matrix is symmetric. Consequently, it is sufficient to calculate the CD statistic over the upper triangular elements of the correlation matrix only and to multiply the outcome by 2 so as to also represent the impact of the lower triangular elements. Importantly, the CD test does not require any (arbitrary) specification of a spatial weight matrix describing the spatial arrangement of the cross-sectional units in the sample, as is standard in the spatial econometrics literature.

The null hypothesis of the CD-test is weak cross-sectional or local spatial dependence, while the alternative hypothesis reflects strong cross-sectional dependence or the presence of common factors. Weak cross-sectional dependence implies that housing prices are related to each other but that the strength of this relationship falls with distance and goes to zero if the distance separating two units becomes sufficiently large. By contrast, strong cross-sectional dependence implies that housing prices remain related to each other also when the distance separating two units goes to infinity (Chudik and Pesaran, 2015; Elhorst et al., 2017). The CD statistic is a two-sided test statistic whose limiting distribution converges to the standard normal distribution $N(0,1)$, as N and T go to infinity. This implies that the critical values of this two-sided test are -1.96 and 1.96 at the 5% significance level. If the test statistic takes a value outside the interval (-1.96,+1.96), thereby rejecting the existence of weak in favor of strong cross-sectional dependence, another question is whether the degree of strong cross-sectional dependence can be determined. For this purpose, Bailey et al. (2016b) developed the exponent α-test.

The mathematical form of the α-test consists of three right-hand side components that need to be computed

$$\alpha = 1 + \frac{1}{2}\frac{\ln\sigma_{\bar{x}}^2}{\ln(N)} - \frac{1}{2}\frac{c_N}{(N\ln N)\sigma_{\bar{x}}^2} - \frac{1}{2}\frac{\ln u_{\bar{v}}^2}{\ln(N)} \quad (2)$$

The first component is the dominating term, the second and third components are bias correction terms. These three components are added to the constant 1. Prior to any

calculations, the data need to be standardized for each single unit in the sample, to get $x_{it} \equiv (x_{it} - \bar{x}_i)/\frac{1}{N}\sum_{i=1}^{N}(x_{it}-\bar{x}_i)^2$. It is to be noted that standardization is not required for the CD test since the pairwise correlation coefficients do not change when the data are standardized.

The term $\sigma_{\bar{x}}^2$ in the first component is defined as $\sigma_{\bar{x}}^2 = \frac{1}{T}\sum_{t=1}^{T}(\bar{x}_t - \bar{x})^2$, where $\bar{x} = \frac{1}{T}\sum_{t=1}^{T}\bar{x}_t$. These expressions state that, firstly, the cross-sectional mean ($\bar{x}_t$) needs to be determined in each time period, secondly, the overall mean $\bar{x}$ over these *T* cross-sectional means and, finally, the standard deviation $\sigma_{\bar{x}}^2$ of this overall mean. Due to the standardization of the data $\sigma_{\bar{x}}^2 < 1$, as a result of which $\ln \sigma_{\bar{x}}^2 < 0$ and $1 + \frac{1\ln \sigma_{\bar{x}}^2}{2\ln(N)} < 1$.

The term c_N in the second component is a small sample bias-correction term that is obtained successively (i) by running separate regressions of x_{it} on $\bar{x}_t$ with coefficient δ_i for each unit *i* in the sample, each based on *T* observations, (ii) by estimating the standard deviation σ_i^2 of each of these regressions, yielding $\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}(x_{it}-\hat{\delta}_i\bar{x}_t)^2$, and (iii) by computing the average of these estimated standard deviations over all units in the sample, $c_N = \frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2$ (Bailey et al., 2016b, equation 12). An alternative approach to estimate the standard deviation σ_i^2 is based on using a set of principal components (Bailey et al., 2016b, equations 30-31).

The term u_v^2 in the third component is also a small sample bias-correction term (Bailey et al., 2016b, four-step procedure in section 3.1). It is obtained by running separate regressions of x_{it} on a constant and $\bar{x}_t$ with coefficients γ_{i0} and γ_{i1} for each unit *i* in the sample. Next, the average value of x_{it} is computed over all units *i* in the sample at time *t* for which γ_{i1} is significant, to get $\bar{x}_t^0$ (*t*=1,…,*T*). Finally, u_v^2 is determined by $u_v^2 = \frac{1}{T}\sum_{t=1}^{T}(\bar{x}_t^0 - \frac{1}{T}\sum_{t=1}^{T}\bar{x}_t^0)^2$. To determine whether the γ_{i1} parameter estimates are significant a procedure is used developed by Holm (1979). First, their t-values are ordered in descending order and then the *i*-th critical value is determined by $c_i = \Phi^{-1}(1 - \frac{0.05}{2N})$ for *i*=1,...,*N*, where 0.05 reflects the standard significance level of 5% and Φ^{-1} is the inverse of the standard normal distribution. The Holm procedure has the effect that the critical values decrease from approximately 4 to 0.

The exponent α-test can take values on the interval (0,1]: $\alpha \leq \frac{1}{2}$ points to weak cross-sectional dependence and corresponds to values of the CD test statistic within the interval (-1.96,+1.96); $\alpha = 1$ points to the strongest form of cross-sectional dependence of no distance decay effect at all, while values in between indicate moderate ($\frac{1}{2} < \alpha \leq \frac{3}{4}$) to strong ($\frac{3}{4} < \alpha <$ 1) cross-sectional dependence. According to Elhorst et al. (2017), following Lee (2002, 2004), $\alpha = \frac{3}{4}$ represents a turning point. Values of α below or above this turning point have implications for the method that should be used to estimate the hedonic model explaining housing prices, i.e., whether or not the housing prices of competing houses used in the sales comparison approach should be treated as endogenous or may be treated as weakly exogenous.

Importantly, evidence in favor of weak cross-sectional dependence (i.e., the null hypothesis of weak cross-sectional dependence based on the CD test is not rejected) excludes

strong cross-sectional dependence, but not vice versa. If evidence is found in favor of strong cross-sectional dependence and is subsequently accounted for, the residuals of x_{it} modified for the contribution of strong cross-sectional dependence in the form of global common factors might still be due to weak cross-sectional dependence. This can be tested by running the CD test and the α-test on these residuals. Halleck Vega and Elhorst (2016) provide an application to regional unemployment rates in the Netherlands. They find empirical evidence in favor of both local spatial dependence and global common factors, and demonstrate that both should be accounted for within one simultaneous framework to get unbiased results.

Gauss code to calculate the CD and the α-tests are made available in an online appendix to Bailey et al.'s (2016) paper. As part of this paper, these routines have been reprogrammed in Matlab, which will be made available.

3. Towards a CD-test for unbalanced data

Time unbalances

Pesaran (2015b, Section 29.8; see also Chudik and Pesaran, 2015, section 1.7) explains how to modify the CD test when having an unbalanced panel due to missing observations in the time domain

$$CD = \sqrt{2/[N(N-1)]}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sqrt{T_{ij}}\hat{\rho}_{ij}. \quad (3)$$

The square root of T to the left of the two summation signs in (1) is moved to the right of them in (3), since the number of observations on which the correlation coefficients are based is different for every pair of units when having an unbalanced panel; let T_i and T_j ($T_i, T_j \leq T$) denote the number of observations available for units i and j, then T_{ij} ($T_{ij} = T_i \cap T_j$) represents the number of observations each pair has in common. The calculation of the mutual correlation coefficients also needs to be changed. The simplest approach only determines these coefficients over the number of observations each pair of units has in common, yielding

$$\hat{\rho}_{ij} = \frac{\sum_{t\epsilon T_{ij}}(x_{it}-\bar{x}_i)(x_{jt}-\bar{x}_j)}{\sqrt{\sum_{t\epsilon T_{ij}}(x_{it}-\bar{x}_i)^2}\sqrt{\sum_{t\epsilon T_{ij}}\left(x_{jt}-\bar{x}_j\right)^2}}, \text{ where } \bar{x}_i = \frac{1}{T_{ij}}\sum_{t\in T_{ij}} x_{it}. \quad (4)$$

However, to utilize the data in a more efficient way this expression is better extended to

$$\hat{\rho}_{ij} = \frac{\frac{1}{T_{ij}}\sum_{t\epsilon T_{ij}}(x_{it}-\bar{x}_i)(x_{jt}-\bar{x}_j)}{\sqrt{\frac{1}{T_i}\sum_{t\epsilon T_i}(x_{it}-\bar{x}_i)^2}\sqrt{\frac{1}{T_j}\sum_{t\epsilon T_j}\left(x_{jt}-\bar{x}_j\right)^2}}, \text{ where } \bar{x}_i = \frac{1}{T_i}\sum_{t\in T_i} x_{it}. \quad (5)$$

From this expression it can be seen that all observations for both unit i and unit j are used to compute respectively the mean and the standard deviation of x_i and x_j, while the covariance between these two units is based on the observations they have in common. In contrast to equation (4), the restriction $T_i = T_j = T_{ij}$ will not hold for all i and j in an unbalanced panel,

as a result of which the ratios $1/T_{ij}$, $1/T_i$ and $1/T_j$ do not drop out of equation (5). From equation (3) it can further be seen that if one pair of units has more observations in common than another pair, the former gets a higher weight in the determination of the CD statistic. Although Pesaran (2015b, section 29.8) discusses the option to calculate cross-sectional averages over all available observations in order to "utilize data in a more efficient way" (p.793) and to switch from equation (1) to equation (3), he does not extend the determination of the correlation coefficients in equation (4) to that in equation (5), which represents a further step forward.

Cross-sectional unbalances

To be able to modify the determination of $\hat{\rho}_{ij}$ and its weighting scheme within the CD statistic when having also different numbers of observation in the cross-sectional domain, we first need to change our point of departure of having a balanced spatial panel of N cross-sectional units over T time periods for a particular variable x_{it} (i=1,…,N; t=1,…,T). Instead of a single observation, we now have a series of N_{it} observations for each geographical unit i at time t. To avoid confusion, we denote this series of observations by the vector z_{it}, while individual observations are indexed by h, i.e., z_{hit} reflects an observation of house h in unit i at time t. N now represents the total number of geographical units.

The CD statistics modified for the number of observations in each unit and each time period then takes the form

$$CD = \sum_{t=1}^{T} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sqrt{\frac{2N_{it}N_{jt}}{\sum_{k=1}^{N}\sum_{l=1}^{N} N_{kt}N_{lt} - \sum_{p=1}^{N} N_{pt}^2}}\, \hat{\rho}_{ij} \,. \qquad (6)$$

In this modification not only $\sqrt{T_{ij}}$, but also 1/[N(N-1)] is moved to the right of the summation signs. The product term $N_{it}N_{jt}$ in the numerator of this ratio, which replaces T_{ij}, denotes how many observations in both unit i and unit j in period t are used to determine the correlation coefficient $\hat{\rho}_{ij}$. If both values N_{it} and N_{jt} are large (small), so will be this product term. If either N_{it} or N_{jt} is large and the other is small, the product term may still be limited. For example, it is better to have 2 observations in both units (product is 4) than to have 3 observations in one unit and 1 in the other (product is 3). If no observations are available for a particular unit in a particular time period, the product term will be zero. This is in line with the modification made in equation (3) when no observations are available for a particular unit in the time domain. The term $\sum_{k=1}^{N}\sum_{l=1}^{N} N_{kt}N_{lt} - \sum_{p=1}^{N} N_{pt}^2$ in the denominator of the above-mentioned ratio denotes the total sum of these product terms, where the contribution of product terms with respect to the own units themselves is subtracted, since its correlation coefficient is also excluded from the summation. Finally, since the number of observations in two units that are related to each other is different for every time period and $N_{it}N_{jt}$ replaces T_{ij}, we cannot multiply $\hat{\rho}_{ij}$ by a fixed number T_{ij} representative for all time periods, as in (3); instead we repeat this calculation for every time period. This explains the addition of the third summation sign with index t. If the number of observations in a particular time period is

greater (smaller) than in another time period, this time period also get a larger weight in (6), just as in equation (3).

The calculation of the mutual correlation coefficients may also be further extended. The simplest extension is to approach the correlation coefficients by either equation (3) or (4), thereby, replacing z_{it} by $\bar{z}_{it}$ for every *i* and *t*, which denotes the average over all individual housing prices in a particular unit at a particular time period. This approach is tempting since it is not immediately clear whether the correlation coefficient between two unequal series of individual housing observations exists and can be determined. This question has been posed on internet several times, but an adequate answer has not been provided.[2] In addition, this principle to work with cross-sectional averages within each unit in each time period ($\bar{z}_{it}$) is also employed in Bailey et al. (2016a). The downside of this approach is that substantial information gets lost about the variation of housing prices around the mean within each cross-sectional unit at each point in time, as earlier illustrated in Figure 1. Starting with 14.35 observations within a particular unit at one moment in time, the average in our sample, the only information that will then be utilized is the mean and standard deviation of this set of observations. Similarly, the number of observations reported in Table 1 would reduce from 8 to 4 (*N*=2, *T*=2). To utilize all available information much more efficiently, the correlation coefficients can be extended as follows

$$\hat{\rho}_{ij} = \frac{s_{z_i z_j}}{\sqrt{s_{z_i}}\sqrt{s_{z_j}}}, \tag{7}$$

where

$$\bar{z}_i = \frac{1}{\sum_{t\epsilon T_i} N_{it}} \sum_{t\epsilon T_i} \sum_{h=1}^{N_{it}} z_{hit}, \tag{7a}$$

$$s_{z_i} = \frac{1}{\sum_{t\epsilon T_i} N_{it}} \sum_{t\epsilon T_i} \sum_{h=1}^{N_{it}} (z_{hit} - \bar{z}_i)^2, \tag{7b}$$

$$s_{z_i z_j} = \frac{1}{\sum_{t\epsilon T_{ij}} N_{it} N_{jt}} \begin{pmatrix} z_{i1} \otimes \iota_{N_{j1}} - \bar{z}_i \otimes \iota_{N_{i1} N_{j1}} \\ \vdots \\ z_{i1} \otimes \iota_{N_{jT}} - \bar{z}_i \otimes \iota_{N_{iT} N_{jT}} \end{pmatrix}' \begin{pmatrix} \iota_{N_{i1}} \otimes z_{j1} - \bar{z}_i \otimes \iota_{N_{i1} N_{j1}} \\ \vdots \\ \iota_{N_{iT}} \otimes z_{jT} - \bar{z}_i \otimes \iota_{N_{iT} N_{jT}} \end{pmatrix}, \tag{7c}$$

where the symbol $\otimes$ represents the Kronecker product between two vectors, and ι_p represents a vector of ones of length *p*. The expressions (7a) and (7b) measure the mean and standard deviation over all available observations for a particular unit *i*. Expression (7c) determines the covariance between two units *i* and *j* over all their observations. Take the example provided in Table 1. The first right-hand side term in (7c) based on the numbers provided in this table is 1/(2*3+2*1)=1/8. In the first period there are 2 observations on unit 1 and 3 observations on unit 2. The Kronecker products within the two right-hand side vectors of (7c) establish a comparison of all price combinations between these two units, which sum up to a total of six; the first housing price in unit 1 is compared first with the three housing prices in unit 2, and then the second housing price in unit 1 is compared with the three housing prices in unit 2.[3]

[2] See the results of a search process using google based on the search terms: correlation and unequal.
[3] Note that the order in which the data are provided does not matter when using this setup.

Similarly, there will be 2*1 price comparisons in the second time period. Both right-hand side vectors are thus of length 8. If there are no data available for a particular unit in a particular time period, the corresponding elements in the two right-hand side vectors drop out (which has been left aside mathematically to simplify notation). The idea to compare all price combinations between two units in a particular time period is the main step forward of the proposed modification, since only then each individual observation will be used in the calculation.

4. Towards an α-test for unbalanced data

When having an unbalanced panel both in the cross-sectional and the time domain, the term $\sigma_{\bar{x}}^2$ on the right hand side of the α-test statistic in (2) can readily be replaced by

$$\sigma_{\bar{z}}^2 = \frac{1}{T}\sum_{t=1}^{T}(\bar{z}_t - \bar{z})^2, \tag{8}$$

where

$$\bar{z}_t = \frac{1}{\sum_{i=1}^{N} N_{it}}\sum_{i=1}^{N}\sum_{h=1}^{N_{it}} z_{hit}, \tag{8a}$$

$$\bar{z} = \frac{1}{T}\sum_{t=1}^{T}\bar{z}_t. \tag{8b}$$

Similarly, separate regressions can be run of z_{hit} first on $\bar{z}_t$ and then on a constant and $\bar{z}_t$ for each unit i. Due to unbalances, the number of observations on which these regressions are based are unit-specific, namely $\sum_{t=1}^{T} N_{it}$, rather than a fixed number of T observations. The first of these two regressions is used to determine

$$c_N = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{t=1}^{T} N_{it}}{\sum_{t=1}^{T}\sum_{j=1}^{N} N_{jt}}\hat{\sigma}_i^2 \tag{9}$$

where

$$\hat{\sigma}_i^2 = \frac{1}{\sum_{t=1}^{T} N_{it}}\sum_{t=1}^{T}\sum_{h=1}^{N_{it}}(z_{hit} - \hat{\delta}_i\bar{z}_t)^2 \tag{9a}$$

Note that the estimates of the standard deviations of the regressions are weighted with the total number of observations that is available for each unit. The second of these two regressions is used to determine which parameters γ_{i1} are significant according to Holm's procedure. Finally, u_v^2 is computed by

$$u_v^2 = \frac{1}{T}\sum_{t=1}^{T}(\bar{z}_t^0 - \frac{1}{T}\sum_{t=1}^{T}\bar{z}_t^0)^2 \tag{10}$$

where $\bar{z}_t^0$ is determined in a similar fashion as in (8a), though only over those observations for which γ_{i1} is found to be significant.

5. Results

The numerical results of this study are recorded in Table 1. Column (1) reports the four variables and column (2) the statistics that are considered, among which the outcomes of the modified CD and α-tests. The statistics in column (3) are based on imbalances in the time domain only and reflect the method applied in Bailey et al. (2016a). First, we calculated $\bar{z}_{it}$ for every *i* and *t* for which data are available. Figure 2 shows the imbalances in the time domain when following this approach. For 511 of the 948 zip codes the average transaction price is available for every time period. For 85 zip codes 1 aggregated observation is missing, for 65 zip codes 2 are missing, which goes on to 34 zip codes for which 11 aggregated observations are missing. The latter represent zip code areas that hardly contain any houses, for example, because it concerns industrial sites, nature areas, or districts that have been into taken into housing production only recently. The CD statistic is calculated using (3), the correlation coefficients using (5), and the α-test using (2). To verify whether the α-test is sensitive to zero averages, which occur if there are no observations available for a particular zip code in a particular period, we also calculated this statistics based on the 511 zip codes for which no observations are missing at all (see Figure 2). It should be stressed that this causes only a small loss of observations relative to the whole sample: 7,922 on a total of 163,323 housing transactions, or 4.9%. The statistics reported in column (4) are based on imbalances in both the time and the cross-sectional domain and represent the method proposed in this paper based on the full data set. The CD statistic is calculated using (6), the correlation coefficients using the set of equations in (7), and the α-test using the set of equations in (8)-(10). Finally, column (5) of Table 1 duplicates column (4) when controlling for a common factor, to which we come back shortly.

<< Insert Table 1 and Figure 2 here >>

The results in columns (3) and (4) of Table 1 show that, even though the average pairwise correlation coefficients tend to be small ($\bar{\rho} < 0.10$), the CD statistic is highly significant, no matter which variable is being considered and no matter whether averaged data or the full data set are used. In spite of this, the CD statistic takes higher values when employing the full data set. There are two explanations for this. The first is that the average pairwise correlation coefficients tend to increase, from 0.048 to 0.053 for the transaction price per square meter, from 0.053 to 0.061 for the time on the market, and from 0.53 to 0.061 for the earthquake indicator. If the average correlation coefficient increases, so does the CD statistic since it determines a weighted average of all individual correlation coefficients. Only for the transaction price this correlation coefficient diminishes when switching to the full data set. The second explanation is that the number of observations available for each pairwise comparison resulting in positive correlations exceeds its counterpart resulting in negative correlations. To illustrate this we calculated the percentage of all pairwise correlations producing a positive and a negative outcome, i.e., $\bar{\rho}^{+}$ and $\bar{\rho}^{-}$, and reported the results in column (3). For the transaction price, the transaction price per square meter and the time on the market variable, these percentages range from 62.7 to 67.7% for $\bar{\rho}^{+}$ and from 32.3 to 37.3% for $\bar{\rho}^{-}$. Only in case of the earthquake indicator we obtain a different result; the

percentage of positive correlations amounts to 12.4%, while the percentage of negative correlations is 87.6%. By contrast, the average pairwise correlation coefficient in the first group of positive correlations is 0.406, the highest of all cased being considered, while its counterpart in the second group of negative correlations with -0.005 is close to zero. It indicates that the area in which houses are due to earthquakes is bounded and that the whole set of correlations ($\frac{1}{2}N(N-1)$=448,878) might be used to identify the boundaries of this area, for example, by means of cluster analysis. We consider this as an interesting topic for further research.

The results obtained for the α-test in columns (3) and (4) reconfirm the CD test results. When using averaged data, the degree of cross-sectional dependence ranges from 0.770 to 0.891. When repeating this analysis for a balanced panel of averaged data of 511 zip codes over 12 time periods having full data (see Figure 2), these numbers change to comparable values of 0.693 to 0.995. When using the full data set the degree of cross-sectional dependence take values ranging from 0.751 to 0.988; the outcomes are somewhat lower than those found in column (3) for the first two variables, the transaction price and the transaction per square meter, comparable for the earthquake indicator, and slightly higher and close to one, in line with the CD test statistic, for the time on the market variable.

An interesting outcome is that $\sigma_{\bar{z}}^2$ turns out to be smaller than $\sigma_{\bar{x}}^2$, i.e., the dominating term in the expression of the α-test statistic in equation (2). It says, not surprisingly, that $\bar{z}_t$ is a better estimate of the average housing price (or one of the other three characteristics) when taken over all housing transactions in a particular time period than $\bar{x}_t$, which in turn is taken over the average prices of all geographical units in a particular time period. By contrast, the estimate of the bias correction term u_v^2 turns out to have a greater upward effect on the estimate of α. This is because the number of units that is filtered out due an insignificant parameter γ_{i1} is much smaller when using the full data set instead of using average data. Apparently, just as $\bar{z}_t$ is a better estimate of the average housing price than $\bar{x}_t$, so is $\bar{z}_t^0$ compared to $\bar{x}_t^0$ since it is based on more observations.

The overall significance of the CD and α-tests outcomes imply that the collected characteristics of housing transactions remain related to each other also when the distance separating two houses goes to infinity. Controlling for weak cross-sectional or local spatial dependence only, the standard approach in numerous empirical studies, will thus produce biased results. Controls for strong cross-sectional dependence are needed to begin with. The necessity to do so further increases when employing individual rather than aggregated data.

Column (5) of Table 1 reports the statistics when controlling for a common factor measured by $\bar{z}_t$. This calculation utilizes the full data set and accounts for imbalances in both the time and the cross-sectional domains. First, we run the regressions

$$z_{it} = \gamma_{0i} + \gamma_{1i}\bar{z}_t + e_{it} \qquad (11)$$

for every unit in the sample, where γ_{0i} and γ_{1i} are unit-specific parameters to be estimated, and e_{it} is a vector of the same length as z_{it} with independently and identically distributed error terms with zero mean and constant variance σ_e^2. Note that this procedure is exactly the same as the one that is used to determine u_v^2. Next, the raw data are de-factored by

$$\hat{e}_{it} = z_{it} - \hat{\gamma}_{0i} - \hat{\gamma}_{1i}\,\bar{z}_t \qquad (12)$$

In this way, aggregate fluctuations are extracted, such that the resulting de-factored variable. i.e., residuals, can be analyzed in a second stage for any remaining cross-sectional dependence.

The results show that the degree of cross-sectional dependence measured by the CD-test falls substantially when controlling for a common factor, i.e. the average price in the research area measured at time *t* (*t*=1,…,*T*) with unit-specific (heterogeneous) coefficients (γ). For three of the four variables (the time on the market variable is the exception), we still observe some degree of strong cross-sectional dependence, i.e., the CD test statistic still takes values outside the interval (-1.96,+1.96). This remaining cross-sectional dependence may be tackled by considering more than one single common factor. For example, one could split up the research area into different subareas and consider additional local common factors. For example, in addition to $\bar{z}_t$, Bailey et al. (2016a) considered local common factors $\bar{z}_{rt}$ for eight regions *r*. Another possibility is to make a distinction between urban and rural areas, which might be beneficial since the former tend to be characterized by increasing and the latter by decreasing population sizes during the sample period.

Generally, the α-test goes down and falls below 0.75 in column (5) of Table 1 when controlling for a common factor, although we need to be careful since it is only possible to identify and consistently estimate α for values of greater than or equal to 0.5 (Bailey et al., 2016b). Outcomes smaller than 0.75 after common factors have been controlled for indicate that weak cross-sectional dependence needs to be accounted for in addition to strong cross-sectional dependence, in this case when explaining housing prices. This finding does not come as a surprise since the application of the sales comparison approach is daily practice of NVM real estate agents in the Netherlands. A detailed description of this practice has been documented by Op't Veld et al. (2008). The idea to mix up both weak and strong cross-sectional dependence in one model has been picked up by Bailey et al. (2016a) using aggregated data on housing transactions, while Duran and Elhorst (2017) are working on a similar study using individual data.

6. Conclusion

We have modified and programmed the cross-sectional dependence (*CD*) test and the exponent of cross-sectional dependence test such that they can also be applied to unbalanced panel in both the time and cross-sectional domain. Due to its general form, the potential number of applications is unlimited. We have applied these tests on a microeconomic data set of individuals housing transactions over a twelve-year period and have found evidence in favor of both weak and strong cross-sectional dependence. Importantly, the real estate literature consists of numerous studies not accounting for any type of cross-sectional dependence, or only weak or only strong spatial dependence. We have decided not to mention any examples here to illustrate this, because it is difficult to be complete and any example would not be subjugate to others we could mention equally well. By contrast, studies

accounting for both are scarce, we mentioned two in the previous section, thereby, setting the research agenda for the next coming years.

Several studies control for cross-sectional (such as postcode or neighborhood dummies) and time-period fixed effects (Koster and Van Ommeren, 2015), which is one way to cover both types of cross-sectional dependence, but as Shi and Lee (2017) have recently pointed out, this is not more than a special case of a much wider class of models with spatial interactions and interactive fixed effects, which are alternative terms for weak and strong cross-sectional dependence.

References

Bailey, N., S. Holly, and M.H. Pesaran (2016a) A two-stage approach to spatio-temporal analysis with strong and weak cross-sectional dependence. *Journal of Applied Econometrics*, 31: 249-280.

Bailey, N., G. Kapetanios, and M.H. Pesaran (2016b) Exponent of cross-sectional dependence: estimation and inference. *Journal of Applied Econometrics*, 31: 929-960.

Bosker, M., H. Garretsen, G. Marlet, R. Ponds, J. Poort, R. van Dooren, and C. van Woerkens (2016) Met angst en beven: verklaringen voor dalende huizenprijzen in het Groningse aardbevingsgebied. Utrecht, Atlas voor Gemeenten.

CBS (2017) Woningmarktontwikkelingen rondom het Groningenveld. Den Haag, Centraal Bureau voor de Statistiek.

Chudik, A., and M.H. Pesaran (2015) Large panel data models with cross-sectional dependence. In: Baltagi B.H. (ed.) The Oxford Handbook of Panel Data, pp. 3-45. Oxford, Oxford University Press.

Duran, N., and J.P. Elhorst (2017) The effect of small earthquakes on housing prices in the north of the Netherlands: a spatio-temporal-similarity approach. Paper presented at XIth World Conference of the Spatial Econometrics Association Singapore.

Elhorst, J.P., M. Gross, and E. Tereanu (2017) Modeling cross-sectional dependence: where spatial econometrics and global VAR models meet. University of Groningen/European Central Bank. Paper presented at XIth World Conference of the Spatial Econometrics Association Singapore, 4th Conference of the International Association for Applied Econometrics Sapporo, and 23rd International Conference Computing in Economics and Finance, New York.

Halleck Vega, S., and J.P. Elhorst (2016) A regional unemployment model simultaneously accounting for serial dynamics, spatial dependence and common factors. *Regional Science and Urban Economics* 60: 85-95.

Holm, H. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6: 65–70.

Koster, H.R.A., and J. van Ommeren (2015) A shaky business: natural gas extraction, earthquakes and house prices. *European Economic Review* 80: 120-139.

Lee, L.-f. (2002) Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory* 18: 252-277.

Lee, L.-f. (2004) Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72: 1899-1925.

Op't Veld, D., E. Bijlsma, and P. van de Hoef (2008) Automated valuation in the Dutch housing market: the web-application 'MarktPositie' used by NVM-realtors. In: T. Kauko and M. d'Amato (eds.) Advances in Mass Appraisal Methods, pp. 70-90. Oxford, Blackwell Publishing, doi:10.1002/9781444301021.ch4.

Pesaran, M.H. (2004) General diagnostic tests for cross section dependence in panels. CESifo Working Paper Series No. 1229. Available at SSRN: http://ssrn.com/abstract=572504.

Pesaran, M.H. (2015a) Testing weak cross-sectional dependence in large panels. *Econometric Reviews*, 34: 1088-1116.

Pesaran, M.H. (2015b) *Time Series and Panel Data Econometrics*. Oxford, Oxford University Press.

Shi, W., and L.-f. Lee (2017) Spatial dynamic panel data model with interactive fixed effects. *Journal of Econometrics* 197: 323-347.

Table 1. Results of the modified CD-test and α-test

(1) Variable	(2) Statistic	(3) Only time unbalances based on averages $\bar{x}_{it}$ if available	(4) Time and cross-sectional imbalances based on full data set	(5) Time and cross-sectional imbalances applied to full set of residuals by controlling for a common factor
Transaction price	CD-test	59.0	118.4	3.3
	$\bar{\rho}$	0.032	0.026	0.000
	$\bar{\rho}^{+}$ (% > 0)		0.101 (62.7)	0.072 (45.4)
	$\bar{\rho}^{-}$ (% < 0)		-0.087 (37.3)	-0.055 (54.6)
	α-test	0.875 (0.920*)	0.790	0.629
Transaction price per square meter living space	CD-test	91.3	212.5	5.5
	$\bar{\rho}$	0.048	0.053	0.000
	$\bar{\rho}^{+}$ (% > 0)		0.112 (67.7)	0.074 (46.3)
	$\bar{\rho}^{-}$ (% < 0)		-0.098 (32.3)	-0.058 (53.7)
	α-test	0.890 (0.952*)	0.679	0.663
Time on market	CD-test	106.4	241.1	-0.04
	$\bar{\rho}$	0.053	0.061	0.000
	$\bar{\rho}^{+}$ (% > 0)		0.117 (67.0)	0.070 (43.9)
	$\bar{\rho}^{-}$ (% < 0)		-0.100 (33.0)	-0.054 (56.1)
	α-test	0.891 (0.955*)	0.988	0.459
Earthquake indicator	CD-test	111.4	294.6	14.8
	$\bar{\rho}$	0.056	0.094	0.005
	$\bar{\rho}^{+}$ (% > 0)		0.406 (12.4)	0.223 (7.4)
	$\bar{\rho}^{-}$ (% < 0)		-0.005 (87.6)	-0.016 (92.6)
	α-test	0.770 (0.693*)	0.751	0.506

Source: own calculations, * α-test when eliminating cells (zip code in a particular period) without any observations.

Figure 1. Average number of transactions per year in descending order for the zip codes in the sample

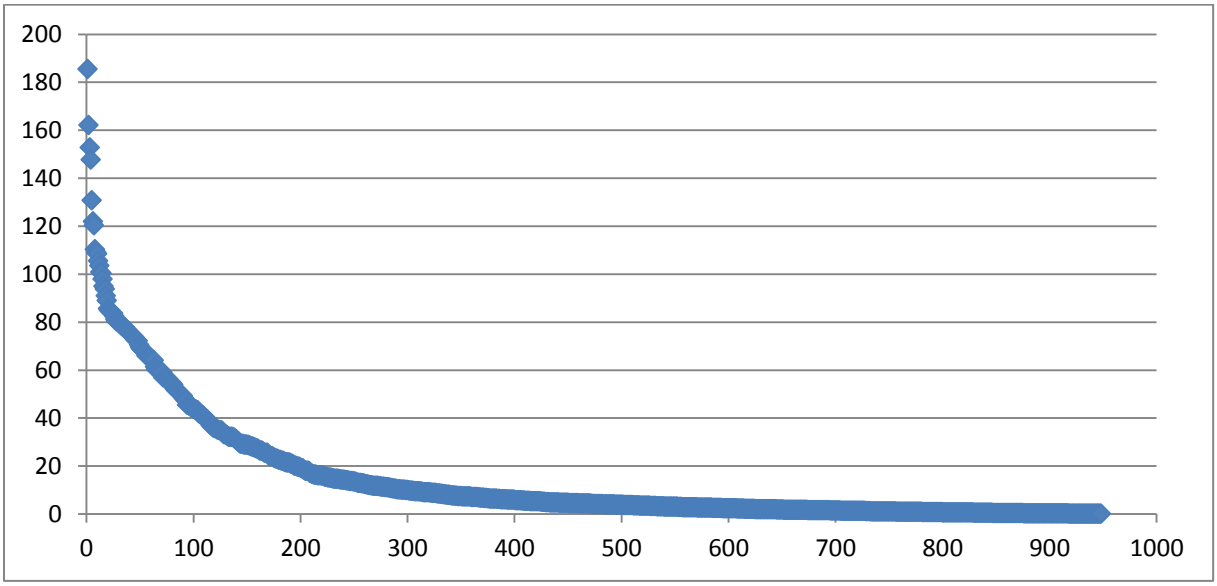

Figure 2. Histogram representing the number of zip codes (vertical axis) and number of time periods (horizontal axis) without any housing transactions