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

Nicolás Durán and J. Paul Elhorst

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

27 January 2017

Extended abstract

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, they may not be treated as being independent, which is the standard if a linear regression model is estimated by ordinary least squares (OLS). 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, 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 may be affected by a common factor, namely the business cycle.

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 et al., 2011), although according to Halleck Vega and Elhorst (2016) this terminology is misleading since it suggests that the latter is more important than the former, which is anything but the case.

Two statistics have been developed to test for cross-sectional dependence: the cross-sectional dependence (*CD*) test of Pesaran (2004, 2015a) and the exponent of cross-sectional dependence test of Bailey et al. (2015). Both tests are based on 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 defined as

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

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. This two-

sided test statistic has the limiting $N(0,1)$ distribution as N and T go to infinity, and thus -1.96 and 1.96 as critical values at the 5% significance level. One advantage of this test is that is not based on any (arbitrary) specification of a spatial weight matrix describing the spatial arrangement of the units in the sample. If the test statistic is significant, i.e. if it takes a value outside the interval (-1.96,+1.96), thereby corroborating the existence of cross-sectional dependence, the next question is whether it is weak, strong or both weak and strong. For this purpose, Bailey et al. (2015) developed the exponent α-test.

This test statistic takes a complicated mathematical form

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

where $\sigma_{\bar{x}}^2$ part of the second right-hand side term is defined as $\sigma_{\bar{x}}^2 = \frac{1}{T}\sum_{t=1}^{T}(\bar{x}_t - \bar{x})^2$ and $\bar{x} = \frac{1}{T}\sum_{t=1}^{T}\bar{x}_t$. These formulas state that firstly, the cross-sectional average ($\bar{x}_t$) needs to be determined in each time period, secondly, the time average $\bar{x}$ over these T cross-sectional averages and finally, the standard deviation $\sigma_{\bar{x}}^2$ of this time average.

Add description of the other variables in (2).

The exponent α-test of Bailey et al. (2015) can take values on the interval (0,1]; $\alpha \leq 1/2$ points to weak cross-sectional dependence only, while $\alpha = 1$ points to strong cross-sectional dependence. In between values indicate moderate to strong cross-sectional dependence and require additional research to discriminate between weak and strong cross-sectional dependence. Evidence in favor of weak cross-sectional dependence ($\alpha \leq 1/2$) 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 again. 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 (2015) paper. As part of this paper, these routines have been reprogrammed in Matlab. They will be made available in due time.

The disadvantage of the current versions of the CD test and the exponent α-test is that they have been developed for a balanced spatial panel only, i.e. for a cross-section of N units over T time periods. The application on housing prices presented in Bailey et al. (2016) as an empirical illustration of both tests also employs aggregated data only; 363 MSAs over the period 1975Q1-2010Q4 (T=144). However, most studies trying to explain housing prices are based on individual data. In this study, we have data on 163,323 housing transactions that

took place in the provinces of Groningen, Friesland and Drenthe located in the Netherlands over the period 2003-2014. The observed variables 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. These earthquakes are all relatively small in magnitude on the scale of Richter (smaller than 4), but when taken together they might have affected the transaction price. This is a topic for further research. First we need to find out whether global common factors and local spatial dependence are relevant extensions to a standard hedonic price model and need to be accounted for.

This data set is anything but a balanced panel. The majority of houses has only been sold once during the observation period. We can aggregate the data to a smaller sample of G geographical units over T=12 years based on the location of the houses, and then calculate the statistics based on these G times T observations, but this might lead to a considerable aggregation bias since the number of observations N_{gt} (g=1,..,G; t=1,…,T) in each unit and each year differs considerably. Pairs of units having more observations should get a higher weight in both statistics than pairs of units having less observations. The aim of this paper is modify the formulas of the test statistics such that they can also be calculated based on an unbalanced spatial panel, and if possible based on individual data.

The CD statistic

Pesaran (2015, Section 29.8) explains how to modify the CD test when having an unbalanced panel due to missing observations in the time domain

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

The following explanation is meant to provide a better understanding of the change made in equation (3) with respect to equation (1). The expression $\sqrt{2T/N(N-1)}$ is taken up before the 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; as a result 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 represent the impact of the lower triangular elements.

The square root of T before the summation signs in (1) is moved after 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; if 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} ($T_{ij} = T_i \cap T_j$) denotes the number of observations each pair has in common. The calculation of the mutual correlation coefficients also needs to be changed, since they can only be determined over the number of observations each pair of units has in common. This gives

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

From this equation it can be seen that observations in unit *i* are still used to compute the average of x_i for that unit over time, even if it is not used to determine the correlation coefficient of unit *i* with another unit *j*. Furthermore, if a pair of units has more observations in common than another pair, the former pair also gets a higher weight in the determination of the CD statistic in (3). According to Pesaran (2015, p.793), this setup of the CD statistic "utilizes data in a more efficient way" (p.793).

To be able to weigh $\hat{\rho}_{ij}$ in the CD statistic when also having different numbers of observation in the cross-sectional domain, we first count the number of observations in different units in each year (N_{gt}). Zip-codes measured in four digits are used to determine the units. The three specified provinces in the Netherlands cover 948 of these zip-code areas. The CD statistics modified for the number of observations in each zip-code area and each year then takes the form

$$CD = \sum_{t=1}^{T}\sum_{i=1}^{G-1}\sum_{j=i+1}^{G}\sqrt{2N_{it}N_{jt}/(\sum_{k=1}^{G}\sum_{l=1}^{G}N_{kt}N_{lt} - \sum_{p=1}^{G}N_{pt}^2)}\,\hat{\rho}_{ij}\,. \quad (5)$$

In this modification not only the square root with respect to *T* but also with respect to *N* is moved after the summation signs. The mutual correlation coefficients can be calculated by (4), although x_{it} for every *i* and *t* needs to be replaced by $\bar{x}_{it}$, which denotes the average over all individual housing prices in a particular unit and a particular time period. This is because a correlation coefficient between two unequal series of individual housing observations does not exist or cannot be determined.[1] The principle to work with cross-sectional average within each unit in each time period is also used in Bailey et al. (2016).

The product term $N_{it}N_{jt}$ in (3) denotes how many observations in both unit *i* and unit *j* in year *t* underlie the correlation coefficient $\hat{\rho}_{ij}$. If both values of *N* are large (small), so will be this product term. If one *N* is large, but 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) rather than 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 missing observations in the time domain set out in equations (3) and (4).

The expression $\sum_{k=1}^{G}\sum_{l=1}^{G}N_{kt}N_{lt} - \sum_{p=1}^{G}N_{pt}^2$ denotes the total sum of these product terms, where the contribution of product terms with respect to the own unit is subtracted, since its correlation coefficient is also excluded from the summation. Since the number of observations is different in every year, we cannot multiply this expression by *T*, as in (1), but

[1] This question has been posed on internet several times, but an adequate answer has not been provided. Some experiments with alternative measures also exhibited unsatisfactory outcomes.

need to repeat this calculation for every year, which explains the addition of the summation sign with index *t*.

The results of this modified CD test for the four variables in our sample are reported in Table 1.

Table 1. Results of the modified CD-test

Variable	Modified test
Transaction price	620.2
Transaction price per square meter living space	2469.6
Time on market	1640.7
Earthquake indicator	2154.5

The exponent α-test statistic

When having an unbalanced panel both in the cross-sectional and the time domain, the term $\frac{1}{2}\frac{\ln \sigma_{\bar{x}}^2}{\ln(N)}$ in the right hand side of the α-test statistic can easily be modified.

K**ey words:** Housing prices, strong and weak cross-sectional dependence
JEL classification: C21, C23, R23

References

Bailey, N., G. Kapetanios, and M.H. Pesaran (2015) Exponent of cross-sectional dependence: estimation and inference. *Journal of Applied Econometrics*, doi: 10.1002/jae.2476.

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

Chudik, A., M.H. Pesaran, and E. Tosetti (2011) Weak and strong cross-section dependence and estimation of large panels. *The Econometrics Journal*, 14, C45-C90.

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 (2016) 85-95.

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(6-10), 1088-1116.

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