Eur. Phys. J. Spec. Top. (2022) 231:1735–1739
https://doi.org/10.1140/epjs/s11734-021-00378-5

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Spatial partial causality

Marcos Herrera-Gomez[1,a], Mariano Matilla-Garcia[2], and Manuel Ruiz-Marin[3]

[1] CONICET-IELDE, Facultad de Ciencias Económicas, Jurídicas y Sociales, National University of Salta, Av. Bolivia 5150, CP., 4400 Salta, Argentina
[2] Departamento de Economía Aplicada y Estadística, Facultad de Económicas, UNED, Paseo Senda del Rey, 11, CP., 28040 Madrid, Spain
[3] Department of Quantitative Methods, Law and Modern Languages, Technical University of Cartagena, 30201 Cartagena, Murcia, Spain

Received 9 October 2021 / Accepted 7 December 2021 / Published online 20 December 2021

Abstract This paper, on the one hand, proposes a statistical technique to detect potential causal relationships when the researcher has georeferenced data but not time dimension, and, on the other, applies this new methodology to the analysis of potential partial causal determinants of home prices. In particular, we find that the direction of causality for home prices in California goes from the income level to prices.

1 Introduction

The evolution of the social and economic dynamics related to the cities of the twenty-first century is closely linked to the evolution of the spatial configuration of cities. There are a number of determinants that help explain the constituency or district configuration. In this sense, home prices contain extraordinarily valuable information.

In this work, we are especially interested in the intrinsic and extrinsic determinants that can help shape the future of population settlements in the form of a city. In particular, we intend to discriminate by means of a semi-parametric approximation which are the elements that potentially cause the formation of prices associated with housing. There are many elements that undoubtedly affect the market price of a necessary good such as housing, discerning which of these elements have a causal link is a challenge on which we intend to shed some light. This is challenging to the extent that we will be using cross-sectional data that have no time component. Note that causal relationships are independent of the spatial or temporal configuration, although they have to develop (and therefore be detected) in time and space.

Central for this paper is that we consider space (location) as a critical element that is required to be taken into account to study any form of causality on this regard. The methodological approach to certain form of spatial causality is based on the Granger–Wiener concept of incremental information content. Causation means that the variable cause must provide additional information about the variable effect. Particularly, this information should be unique, meaning that, once we take into account the spatial structure inherent to the data, then variable x causes y when the information contained in x helps to reduce the uncertainty associated with y.

For another point of view, this approach might be partially understood as a consequence of the well-known Gibbons and Overman critique [1] to spatial econometrics. This critique advocate for an experimental methodological approach in spatial econometrics, as opposed to the dominant structural approach in which theory is the main source of identification in the model. Instead of using external variation to identification, we propose to use a semi-parametric approach. This approach is presented in Sects. 2 and 3 and it is illustrated in Sect. 4 where census information regarding houses in a given California district are studied.

2 Entropy measures and symbolic analysis

Given a random spatial process $\{X_s = (X_{1s}, X_{2s}, \ldots X_{ks})\}_{s\in S}$ (either univariate or multivariate), where S is a set of geographical coordinates that are given and fixed, one can measure the amount of uncertainty through its entropy $H(X)$ defined as

$$H(X) = -\sum_{(x_1,\ldots,x_k)\in\chi} P(X_1 = x_1, \ldots, X_k = x_k) \log\left(P(X_1 = x_1, \ldots, X_k = x_k)\right). \quad (1)$$

[a] e-mail: mherreragomez@conicet.gov.ar (corresponding author)

Based on this definition of entropy, given two spatial processes $\{X_s = (X_{1s}, X_{2s}, \dots X_{ks})\}_{s\in S}$ and $\{Z_s = (Z_{1s}, Z_{2s}, \dots Z_{ks})\}_{s\in S}$ we can define the conditional entropy as

$$H(X|Z) = H(X, Z) - H(Z), \tag{2}$$

and this conditional entropy is understood as the amount of uncertainty in $\{X_s\}_{s\in S}$ given knowledge about $\{Z_s\}_{s\in S}$.

Estimating the entropy value (uncertainty) of a spatial process, whose density function is unknown, is not an easy task. As an alternative to traditional plug-in density estimation, Sulewski [2], suggested to use equal-bin-width histograms when dealing with symmetric distributions, while equal-bin-count histograms should be preferred for asymmetric distributions. Nevertheless, in our analysis we follow the symbolic approach proposed by Herrera et al. [3] consisting in symbolizing the spatial process with a finite set of natural numbers (symbols), such that each observation X_s is associated with the number of neighbors of location s that coincides with X_s in being either above or below of the median of the spatial process $\{X_s\}_{s\in S}$. This symbolization procedure is trivially extended component-wise to a multivariate spatial process. Notice that the symbols gather a (rough) description of the spatial distribution of the process, and that the entropy associated with the discrete symbols' distribution measures its degree of disorder. This entropy is known as a form of symbolic entropy.

3 Spatial partial causality test

Under this setting, in a totally model-free framework, in [3] a causality in information test for spatial processes was proposed based on symbolic entropy. Concretely, given two real spatial processes $\{X_s\}_{s\in S}$ and $\{Y_s\}_{s\in S}$, and two association schemes W_x, W_y (spatial weighting matrix) for each one of them, the statistical test for the null hypothesis:

$$\begin{aligned} H_0 : X &\text{ does not cause } Y \\ &\text{under the spatial association schemes } W_x \\ &\text{and } W_y \end{aligned} \tag{3}$$

is given by

$$\delta_{X\to Y}(W) = h(Y|W_y Y) - h(Y|W_y Y, W_x X), \tag{4}$$

that is, if $W_x X$ does not add extra information about Y then $\delta_{X\to Y}(W) = 0$, otherwise the null hypothesis is rejected. The statistical significance is provided with a spatial block bootstrap procedure that breaks down the dependence structure between X and Y but preserves their own spatial structure. In [4], authors apply a similar approach for a spatial dependence tests. The statistical behavior of the causality test, empirical size and power, under different processes can be found in [3,5].

Now, we want to use the statistical test given in (4) to test for partial spatial causality, which consist in eliminating the effect of common inputs from latent variables when detecting the causal relationships among several process. To this end, we will make use of the Frisch–Waugh–Lovell (FWL) theorem, also known as the decomposition theorem [6,7]. Specifically, consider the following linear regression model:

$$Y = X\boldsymbol{\beta} + \mathrm{u}, \tag{5}$$

with an $N \times K$ matrix, X, of conditioning variables, including a possible causal variable X_1 that is our focus. Next, we decompose $X\boldsymbol{\beta}$ as

$$X\boldsymbol{\beta} = (\, X_1 \; X_2 \,) \begin{pmatrix} \beta_1 \\ \boldsymbol{\beta_2} \end{pmatrix} = X_1\beta_1 + X_2\boldsymbol{\beta_2}, \tag{6}$$

where $\boldsymbol{\beta_2}$ denotes the $(k-1)$-vector of all beta coefficients other than β_1. Using a direct consequence of FWL Theorem, if the orthogonal projection into the orthogonal complement, $X_2^{\perp}$, of X_2 is denoted by

$$M_2 = I_n - X_2 \left(X_2' X_2\right)^{-1} X_2', \tag{7}$$

so that by definition,

$$\begin{aligned} M_2' = M_2, \quad M_2 M_2 = M_2, \quad M_2 X_2^{\perp} = X_2^{\perp}, \\ \text{and} \quad M_2 X_2 = 0, \end{aligned} \tag{8}$$

then multiplying (5) and (6), it follows that

$$M_2 Y = M_2 X_1 \beta_1 + M_2 X_2 \boldsymbol{\beta_2} + M_2 u. \tag{9}$$

Therefore, by defining $\widetilde{Y} = M_2 Y$ and $\widetilde{X_i} = M_2 X_1$, we obtain the following expression:

$$\widetilde{Y} = \widetilde{X_1}\beta_1 + \widetilde{u}, \tag{10}$$

with $\widetilde{u} = M_2 u$. This equation can be understood as the reduced form of the relationship between Y and the potential causal variable, X_1.

Equation (10) can be estimated by means of ordinary least-squares (OLS): First, regress Y on X_2 and obtain residuals fitted values $\widetilde{u}_Y$. Second, regress X_1 on X_2 and obtain fitted values of the residuals $\widetilde{u}_{X_1}$. Finally, regress the residuals to obtain

$$\widetilde{u}_Y = \widetilde{u}_{X_1}\beta_1 + e. \tag{11}$$

The FWL Theorem states:

1. The OLS estimates of regressions (5) and (10) are numerically identical.
2. The residuals from regressions (5) and (10) are numerically identical.

An extensive revision with applications of this theorem is provided by Davidson and McKinnon [8]. Under a spatial setting, Smith and Lee [9] apply this theorem to discuss the relationship under two spatial variables.

Our initial strategy is based on the framework proposed by Smith and Lee. That is, using the FWL Theorem, one can cancel out the effect of common inputs from confounding variables when detecting the causal relationships among spatial processes. Specifically, we test whether X_1 causes Y under spatial association schemes removing the effect of other $k-1$ variables, X_2, using the $\delta_{X\to Y}(W)$-test on the residuals $\widetilde{u}_{X_1}$ and $\widetilde{u}_Y$. The next section shows how to use the statistical procedure on a real data set.

4 Empirical application

This section analyses the relation between housing prices and income in 20,433 cross-sectional observations for the period 1990 from California census. The purpose is testing for causality between the two variables controlling for counfounders and, if so, detecting the direction of causation using the methodology introduced previously.

The data-set has been used in the second chapter of Aurélien Géron's book 'Hands-On Machine learning with Scikit-Learn and TensorFlow' [10]. The data pertain to the houses found in a given California district and some summary statistics about them based on the 1990 census data. The variables that we use are the follows:

- Ln(price): Logarithm of median house value.
- Income: Median income.
- Age: Housing median age.
- Rooms: Total room number.
- Bedrooms: Total bedrooms number.
- Population (within a block).
- Households (within a block).
- Geographical position (Longitude and Latitude).

Block groups are the smallest geographical unit for which the US Census Bureau publishes sample data (a block group typically has a population of 600–3000 people). This database has been used for prediction interest; however, according our knowledge, this is the first time used to detect causality. A summary of descriptive statistics is presented in Table 1.

Our interest is centered on the income and its spatial distribution as a determinant of housing price. To do this, we rely on a hedonic price model [11] which is a model that considers that prices are determined by internal factors (age, the number of rooms, baths, etc.) as well as by external factors (neighborhood and/or environmental factors). In general terms, hedonic price models assume that the price of a product reflects embodied characteristics valued by some implicit or shadow prices. Therefore, it is assumed that a house can be decomposed into characteristics such as number of bedrooms, size, distance to the city center. Particularly, the hedonic regression equation treats these attributes separately. These estimations estimate the extent to which each factor affects the market price of the property.

As external factor, the spatial relevance in hedonic house prices were first considered by Dubin [12,13] and Can [14,15]. If non-spatial factors are controlled, the remaining discrepancies in price will represent differences in the good's external surroundings. In this situation, the median of income could be a spatial conditioning or a surrounding factor. However, the hedonic literature considers the explanatory variables as conditioning, no causal variables. Then, we propose to advance in the identification of the income as a spatial causal variable.

Our methodology considers that the spatial support is relevant for each variable of interest, that is, distribution on space is not-randomly and give us information about the relationship. To show this relevance, the spatial distribution of both variables is presented in Fig. 1. We observe a coincidence of high values of prices and income, in special, near to the oceanic coast, with clustering in San Francisco, Los Angeles and surroundings.

Maps reveal the importance of geographical position between variables; however, this is only qualitative information. Then, additionally, we present in Table 2 the different tests that detects the spatial dependence for this variables. All spatial test requires to create a spatial weighting matrix that captures the neighborhood for each observation. In our case, the W was created using 14-nearest neighbors.[1]

The null hypothesis of the first tests (Moran's I tests) in Table 2 is that there is no spatial auto-correlation. This hypothesis is rejected for both original variables. The Bivariate Moran tests for the null hypothesis of no spatial correlation between Ln(price) and the spatial neighborhood of Income; and the hypothesis is also rejected. Tests ψ_1-test and ψ_2-test [4] are based on symbolic analysis and testing general form of spatial dependence, i.e., the tests are powerful against nonlinear spatial structures, with sharply contrasts with the Bivariate Moran test which is mainly focus on linear spatial structures. Both symbolic tests detect spatial dependence, of unknown form, into each variable (H_0 of ψ_1-test is rejected) and between variables (H_0 of ψ_2-test is rejected).

However, the relationship between Ln(price) and Income can be affected by other factors. Using the FWL Theorem, these omitted factors can be removed using linear models as:

[1] The k-nearest neighbors is a criterion that works in the following form: for each unit, the Euclidean distance from all the other units is calculated and sorted in an increasing order. The neighbors for each unit are then taken to be the nearest k of those units. In case two units are at the same euclidean distance, we take as neighbor the one with smaller angle in polar coordinates.

Table 1 Descriptive statistics

Variables	Obs.	Mean	S.D.	Min.	Max.
Ln (price)	20,433	12.09	0.57	9.62	13.12
Income	20,433	3.87	1.90	0.50	15.00
Age	20,433	28.63	12.59	1.00	52.00
Rooms	20,433	2636.50	2185.27	2.00	39,320.00
Bedrooms	20,433	537.87	421.39	1.00	6445.00
Population	20,433	1424.95	1133.21	3.00	35,682.00
Households	20,433	499.43	382.30	1.00	6082.00

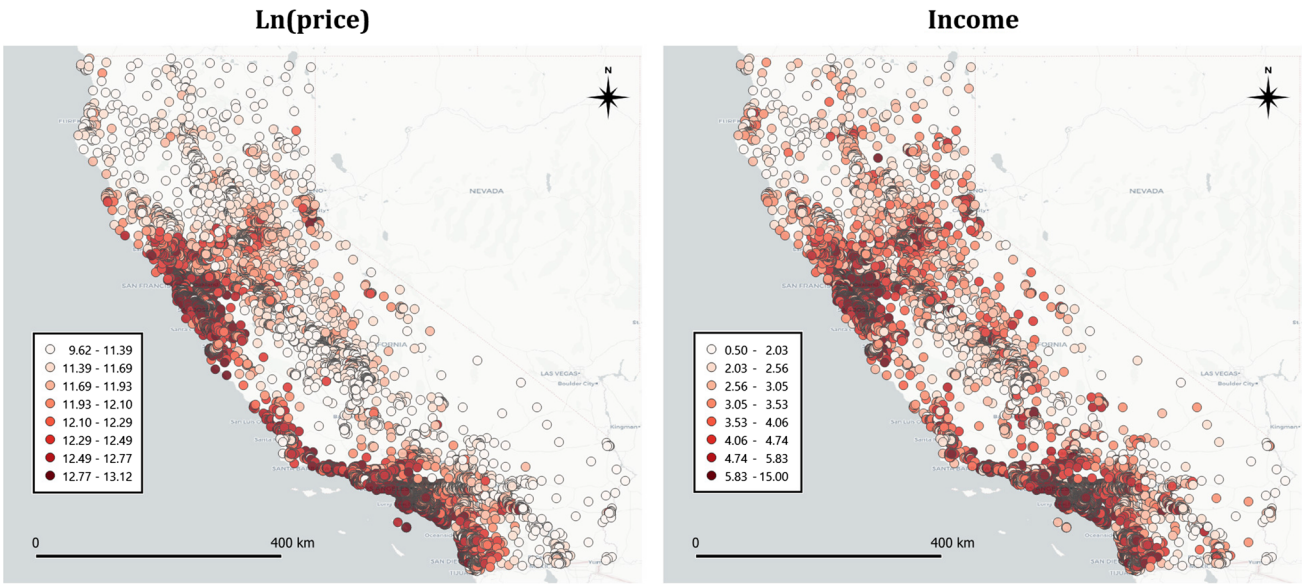

Fig. 1 Spatial distribution of Ln(price) and Income

Table 2 Spatial dependence tests

Test	Value	p-value	Conclusion
Original variables			
Moran-I Ln (price)	0.86	0.00	Spatial-autocorrelation
Moran-I Income	0.54	0.00	Spatial-autocorrelation
Bivariate Moran	0.51	0.00	Spatial-correlation
ψ_1-test	7.76	0.00	Spatial-dependence
ψ_2-test	0.07	0.01	Spatial-dependence
Residual variables			
Moran-I $\widetilde{u}_{\text{Ln (price)}}$	0.62	0.00	Spatial-autocorrelation
Moran-I $\widetilde{u}_{\text{Income}}$	0.19	0.00	Spatial-autocorrelation
Bivariate Moran	0.02	0.00	Spatial-correlation
ψ_1-test	7.48	0.00	Spatial-dependence
ψ_2-test	0.14	0.00	Spatial-dependence

In all cases, the W was generated under 14-nearest neighbors

$$\text{Ln(price)} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Rooms} + \beta_3 \text{Bedrooms} + \beta_4 \text{Population} + \beta_5 \text{Households} + u_{\text{Ln(price)}}, \tag{12}$$

$$\text{Income} = \gamma_0 + \gamma_1 \text{Age} + \gamma_2 \text{Rooms} + \gamma_3 \text{Bedrooms} + \gamma_4 \text{Population} + \gamma_5 \text{Households} + u_{\text{Income}}. \tag{13}$$

From Eqs. (12) and (13), we obtain the estimated residuals, $\widetilde{u}_{\text{Ln(price)}}$ and $\widetilde{u}_{\text{Income}}$, respectively. These residual variables have been used to test the presence of spatial correlation (Table 2, section: Residual variables). Similarly to the original variables, the tests detect the presence of spatial relationship between Ln(price) and Income.

The next step is to determine the direction of spatial information or spatial causality. The results of this

Table 3 Results of partial causality tests

H_0 k-Nearest neighbors	$\widetilde{u}_{\text{Income}} \not\Rightarrow \widetilde{u}_{\text{Ln(price)}}$ p-value	$\widetilde{u}_{\text{Ln(price)}} \not\Rightarrow \widetilde{u}_{\text{Income}}$ p-value	Conclusion
13	0.00	0.16	Income ⇒ Ln(price)
14	0.02	0.42	Income ⇒ Ln(price)
15	0.01	0.37	Income ⇒ Ln(price)

"$\not\Rightarrow$" means 'does not cause', and "⇒" means 'causes'. Boots: 399, Blocks: 8

test are presented in Table 3. As sensibility analysis, we present the results for $k = 13, 14$, and 15 nearest neighbors. In all cases, we detect directionality of information from Income to Ln(price), after controlling by potential economic confounders.

5 Final comments

Home prices contain relevant amount of information. This information is the reflection of a series of geographic and economic determinants that help explain the configuration of cities. In this work, we have been especially interested in locating those determinants that can potentially have a causal relationship when explaining the behavior of prices in a given location. For this, we have developed an approach to partial spatial causality in terms of information.

A nonparametric statistical test has been developed that can be used in conjunction with the FWL theorem. We have illustrated the methodology by studying the price determinants of 20,433 California homes. The results suggest that there is a causal relationship (in terms of spatial information) from income to prices.

The strategy of causality proposed here is very useful for urban and regional studies where the spatial dimension is relevant and the information is non-experimental. Also, changing the symbolization procedure, an extension of the test could apply to spatiotemporal data.

Acknowledgements M. Ruiz Marín was supported by Ministerio de Ciencia, Innovacón y Universidades under grant number PID2019-107800GB-I00/AEI/10.13039/501100011033. This study was also part of the collaborative activities carried out under the programs of the region of Murcia (Spain): 'Groups of Excellence of the region of Murcia, the Fundación Séneca, Science and Technology Agency' project 19884/GERM/15. Mariano Matilla-García was funded by the Ministerio de Ciencia e Innovación under grant PID2019-107192GB-I00.

References

1. S. Gibbons, H. Overman, Mostly pointless spatial econometrics? J. Reg. Sci. **52**(2), 172–191 (2012)
2. P. Sulewski, Equal-bin-width histogram versus equal-bin-count histogram. J. Appl. Stat. **48**(12), 2092–2111 (2020)
3. M. Herrera-Gomez, J. Mur, M. Ruiz-Marin, Detecting causal relationships between spatial processes. Pap. Reg. Sci. **195**(3), 577–594 (2016)
4. M. Herrera-Gomez, M. Ruiz, J. Mur, Detecting dependence between spatial processes. Spat. Econ. Anal. **8**(4), 469–497 (2013)
5. M. Herrera-Gomez, Causality. Contributions to spatial econometrics, Ph.D. Thesis. University of Zaragoza, Spain (2011)
6. R. Frisch, F.V. Waugh, Partial time regressions as compared with individual trends. Econometrica **1**(4), 387–401 (1933)
7. M.C. Lovell, Seasonal adjustment of economic time series and multiple regression analysis. J. Am. Stat. Assoc. **58**, 993–1010 (1963)
8. R. Davidson, J. MacKinnon, *Econometric Theory and Methods* (Oxford University Press, New York, 2004), p. 768
9. T. Smith, K.L. Lee, The effects of spatial autoregressive dependencies on inference in ordinary least squares: a geometric approach. J. Geogr. Syst. **14**(1), 91–124 (2012)
10. A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media Inc, USA, 2019), p. 600
11. S.M. Rosen, Hedonic prices and implicit markets: product differentiation in pure competition. J. Political Econ. **82**(1), 34–55 (1974)
12. R.A. Dubin, Estimation of regression coefficients in the presence of spatially autocorrelated error terms. Rev. Econ. Stat. **70**, 466–474 (1988)
13. R.A. Dubin, Spatial autocorrelation and neighborhood quality. Reg. Sci. Urban Econ. **22**(3), 432–452 (1992)
14. A. Can, The measurement of neighborhood dynamics in urban house prices. Econ. Geogr. **66**(3), 254–272 (1990)
15. A. Can, Specification and estimation of hedonic housing price models. Reg. Sci. Urban Econ. **22**(3), 453–474 (1992)