Journal of Real Estate Finance and Economics, 14: 3, 333–340 (1997)

Using the Spatial Configuration of the Data to Improve Estimation

R. KELLEY PACE
Department of Finance, E.J. Ourso College of Business Administration, Louisiana State University, Baton Rouge, LA 70803

OTIS W. GILLEY
Department of Economics and Finance, College of Administration and Business, Louisiana Tech University, Ruston, Louisiana 71272

Abstract

Using the well-known Harrison and Rubinfeld (1978) hedonic pricing data, this manuscript demonstrates the substantial benefits obtained by modeling the spatial dependence of the errors. Specifically, the estimated errors on the spatial autoregression fell by 44% relative to OLS. The spatial autoregression corrects predicted values by a nonparametric estimate of the error on nearby observations and thus mimics the behavior of appraisers. The spatial autoregression, by formally incorporating the areal configuration of the data to increase predictive accuracy and estimation efficiency, has great potential in real estate empirical work.

Key Words: spatial autocorrelation, SAR, hedonic pricing

In a well-known paper, Harrison and Rubinfeld (1978) investigated various methodological issues related to the use of housing data to estimate the demand for clean air. They illustrated their procedures using data from the Boston SMSA with 506 observations (one observation per census tract) on 14 nonconstant independent variables. These variables include proxies for pollution, crime, distance to various centers, geographical features, accessibility, housing size, age, race, status, tax burden, educational quality, zoning, and industrial externalities.[1]

Despite the inclusion of a wide variety of important economic variables, the Harrison and Rubinfeld model and data exhibit various problems common to many hedonic pricing or mass appraisal models.[2] For example, not all variables exhibit the proper sign. Specifically, the AGE variable is insignificant and positive. In addition, the residuals display a pattern across space, a result incompatible with the assumed independent and identically distributed (iid) error structure.

To resolve these empirical problems, this paper explicitly allows for the areal configuration of the observations through a spatial autoregression. By appropriate differencing of the observations, the spatial autoregression re-creates a more iid error structure, which greatly improves the results. Specifically, the estimated spatial autoregression yields a negative and significant coefficient for AGE while vastly improving the sample goodness of fit. The estimated sum-of-squares errors falls by 44% relative to the original ordinary least squares (OLS) results.

Section 1 discusses the spatial autoregressive estimator employed, section 2 estimates the resulting spatial autoregression, and section 3 concludes with the key results.

1. A Spatial Autoregressive Estimator

When errors exhibit spatial autocorrelation, a common estimator corrects the usual prediction of the dependent variable, $Y = X\beta + \varepsilon$, by a weighted average of the errors on nearby properties as in (1):

$$Y = X\beta + \alpha D(Y - X\beta) + \varepsilon \tag{1}$$

where D represents an n by n comparable weighting matrix with zeros on the diagonal (the observation cannot predict itself).[3] The rows of D sum to 1 as implied by (2). The nonzero entries on the ith row of D represent the observations whose errors interact with the error on the ith observation. We assume independent, 0 mean errors from a normal distribution. These assumptions appear in (2):

$$\begin{aligned}
&\text{(a)} \quad \underset{(n\,\text{by}\,n)}{D}\ \underset{(n\,\text{by}\,n)}{[1]} = \underset{(n\,\text{by}\,n)}{[1]} \\
&\text{(b)} \quad \text{diag}(D) = \underset{(n\,\text{by}\,n)}{[0]} \\
&\text{(c)} \quad 0 \le \alpha < 1 \\
&\text{(d)} \quad \varepsilon \sim \text{N}(0, \sigma^2 I)
\end{aligned} \tag{2}$$

In the spatial statistics literature, the model in (1) and (2) describes a simultaneous autoregression (SAR) with the log-likelihood function

$$L(\alpha, \beta, \sigma^2) = \tfrac{1}{2}\ \ln|B| - \tfrac{1}{2}[n\ \ln(2\pi\sigma^2) + \sigma^{-2}(Y - X\beta)'B(Y - X\beta)] \tag{3}$$

where B equals $(I - \alpha D)'(I - \alpha D)$.[4] The maximum likelihood (ML) method efficiently estimates the model asymptotically (given the assumptions hold).

Assuming the existence of the ML estimate, one could predict Y via (4):

$$\breve{Y} = X\breve{\beta} + \breve{\alpha} D(Y - X\breve{\beta}) \tag{4}$$

Furthermore, (4) leads to the estimated errors in (5)

$$\breve{e} = Y - \breve{Y} = Y - X\breve{\beta} - \breve{\alpha} D(Y - X\breve{\beta}) = (I - \breve{\alpha} D)(Y - X\breve{\beta}) \tag{5}$$

Analogously, one could compute ex-sample errors by (6):

$$\breve{e}_{\text{ex}} = Y - \breve{Y}_{\text{ex}} = (I - \breve{\alpha} D_{\text{ex}})(Y_{\text{ex}} - X_{\text{ex}}\breve{\beta}) \tag{6}$$

2. Maximum Likelihood Sample Estimation of a Spatial Autoregression

This section illustrates the spatial autoregression estimator from section 1 using the augmented Harrison and Rubinfeld (1978) data. Section 2.1 discusses the data, section 2.2 presents the model, and section 2.3 presents the actual estimation results.

2.1. Data

In a well-known paper, Harrison and Rubinfeld investigated various methodological issues related to the use of housing data to estimate the demand for clean air. They illustrated their procedures using data from the Boston SMSA with 506 observations (one observation per census tract) on 14 nonconstant independent variables. These variables include levels of nitrogen oxides (NOX), particulate concentrations (PART), average number of rooms (RM), proportion of structures built before 1940 (AGE), black population proportion (B), lower status population proportion (LSTAT), crime rate (CRIM), proportion of area zoned with large lots (ZN), proportion of nonretail business areas (INDUS), property tax rate (TAX), pupil–teacher ratio (PTRATIO), location contiguous to the Charles River (CHAS), weighted distances to the employment centers (DIS), and an index of accessibility (RAD).[5] As mentioned previously, many authors have used the data to illustrate various points.

We manually collected the location of each tract in latitude (LAT) and longitude (LON) out of the 1970 census.[6] In the process of conducting this project, we rechecked the data against the original census data. We discovered eight miscoded dependent variable observations. We employ the corrected data in the estimation.[7]

2.2. Model

We fitted the following model from Belsley et al. (1980):

$$\begin{aligned}\ln(\text{Price}) = {} & \beta_1 + \beta_2\text{CRIM} + \beta_3\text{ZN} + \beta_4\text{INDUS} + \beta_5\text{CHAS} + \beta_6\text{NOX}^2 \\ & + \beta_7\text{RM}^2 + \beta_8\text{AGE} + \beta_9\text{DIS} + \beta_{10}\text{RAD} + \beta_{11}\text{TAX} \\ & + \beta_{12}\text{PTRATIO} + \beta_{13}\text{B} + \beta_{14}\text{LSAT} + \beta_{15}\text{LAT} + \beta_{16}\text{LON} \\ & + \beta_{17}\text{LAT*LON} + \beta_{18}\text{LAT}^2 + \beta_{19}\text{LON}^2 \end{aligned} \tag{7}$$

The quadratic expression involving latitude and longitude does not follow Belsley et al. However, the addition of these terms removes any ‘‘large-scale’’ locational factors from the conditional mean and follows a standard practice in the spatial statistics area. The addition of these variables raises the R^2 from 0.811 to 0.814, a very small amount.

2.3. Specification of the Spatial Weight Matrix

The weight given to the census tracts for differencing depended on their proximity as measured by the latitude and longitude for each observation relative to all other tracts (using the Euclidean metric).[8] Initially, we weighted every observation j by its distance d_{ij} from the observation i as given by the function in (7):

$$w_{ij} = \max\left[1 - \frac{d_{ij}}{d_{\max}}, 0\right] \tag{7}$$

Naturally, this yields a weight of 1 for the tract itself ($d_{ij} = 0$) and 0 for each observation j more than $d_{\max}$ distance from observation i. Subsequently, in (9) we normalize the initial weights so that $\sum_{j=1/i\neq j}^{n} D_{ij} = 1$:

$$D_{ij} = \frac{w_{ij}}{\sum_{\substack{j=1 \\ i\neq j}}^{n} w_{ij}} \tag{9}$$

In addition, we set $D_{ii} = 0$, as assumed in (2), to prevent each observation from predicting itself. Depending on their areal configuration, some observations may remain undifferenced while others may become differenced with many nearby observations.

For example, suppose we have 506 observations. For the third observation, D might appear as

$$D_{3,1:506} = [0,\ 0.5,\ 0,\ 0,\ 0.3,\ 0,\ 0,\ 0.1,\ 0.05,\ 0.03,\ 0,\ 0.02,\ 0, \ldots, 0]$$

Note that the third entry of $D_{3,1:506}$ equals 0 while the row sums to 1.

2.4. Maximum Likelihood Sample Estimation

Table 1 contains the sample estimates from using OLS and the SAR maximum likelihood estimators. Based upon a two-dimensional grid search, the SAR maximum likelihood estimate of $\breve{\alpha}$ was 0.8 and $d_{\max}$ was 0.0099. For the SAR maximum likelihood estimate, the sample R^2 was 0.89571 while for OLS the corresponding R^2 was 0.81388, an increase in error of 79% over the corresponding SAR maximum likelihood estimated sum-of-squares errors. Note that the model contained numerous variables controlling for locational effects. It included a variable for distances to the various centers, a variable measuring accessibility to radial highways, a Charles River dummy, and a bivariate quadratic function of latitude and longitude. Despite a very reasonable effort to control for locational effects, the SAR maximum likelihood estimator greatly reduced overall errors.

Table 1. OLS and spatial autoregressive estimates.

	β_{OLS}	t_{OLS}	β_{SAR}	t_{SAR}
CRIM	−0.01186	−9.53	−0.00670	−6.83
ZN	−0.00021	−0.37	0.00091	1.81
INDUS	−0.00041	−0.17	−0.00101	−0.35
CHAS	0.08165	2.46	−0.01231	−0.45
NOXSQ	−0.59965	−5.07	−0.36895	−2.37
RM2	0.00593	4.50	0.00873	8.39
AGE	0.00009	0.17	−0.00162	−3.32
DIS	−0.21579	−4.40	−0.18685	−2.63
RAD	0.08882	4.53	0.07262	3.72
TAX	−0.00043	−3.50	−0.00041	−3.51
PTRATIO	−0.02709	−5.20	−0.01704	−3.09
B	0.00036	3.53	0.00067	5.99
LSTAT	−0.37763	−15.26	−0.24588	−11.35
LAT	−278.54000	−1.44	−262.61000	−1.38
LON	9.87540	0.03	555.95000	1.99
LAT*LON	−0.18337	−0.12	−1.60820	−1.19
LAT^2	2.01620	1.50	2.32520	1.71
LON^2	0.03862	0.01	−5.22980	−1.71
R^2	0.81388		0.89571	
δ			0.8000	
d_{max}			0.0099	

The explanation for this lies in the type of error. The spatial statistics literature draws a distinction between ''large-scale'' and ''small-scale'' variations.[9] All of the locational variables included in the Harrison and Rubinfeld data measure large-scale effects. However, as the activities of real estate appraisers attest, the small-scale neighborhood and very local influences may prove more important in the prediction of housing values. Differencing contiguous and other nearby tracts from each other cancels much of the error from unobservable local causes.[10] This lower error can increase the efficiency of parameter estimates, which in turn can aid accurate prediction.

Note the treatment by the two estimators of the AGE variable. OLS produces a positive but insignificant estimate of AGE while the maximum likelihood SAR produces a negative estimate with a *t*-statistic of −3.32. Furthermore, the zoning variable (ZN) under OLS has a negative but insignificant estimate. In contrast, the maximum likelihood SAR estimator yields a positive and significant estimate of the effects of zoning.

The estimators differ in their estimates of the magnitude of other effects. For example, the SAR maximum likelihood estimate for the variable B, the effects of race, is 86% greater than the corresponding OLS estimate. However, the SAR maximum likelihood assigns other variables lower parameter estimates than OLS. Specifically, the coefficient on the pollution variable (NOXSQ) changes from −0.59965 under OLS to −0.36895 under the SAR maximum likelihood estimator. As the pollution variable was the main focus of the Harrison and Rubinfeld study, this highlights the importance of estimator choice.

3. Conclusion

One cannot judge estimators on the basis of a single sample. Nonetheless, the much higher degree of fit produced by the SAR maximum likelihood estimator relative to OLS ($SSE_{OLS}/SSE_{SAR} = 1.86$) should make it a candidate for real estate empirical work. In addition, the SAR maximum likelihood's negative and significant coefficient for AGE and positive and significant coefficient for zoning (ZN) coincides more closely with most individuals' priors than OLS, which produced insignificant parameter estimates with the opposite signs.[11]

The SAR maximum likelihood estimator can use the same variables as OLS to estimate a regression. However, the SAR maximum likelihood estimator, like an appraiser, uses the correlated errors on nearby properties to improve the overll prediction.

Ironically, the formal empirical tools currently employed in real estate make little use of the rich spatial information present in the data. Indeed, even the implementation of the SAR estimator here leaves substantial room for improvement. The present implementation assumes isotrophy (same variance-covariance structure over space) and does not take into account many of the factors appraisers might use. For example, we do not account for the road network or physical obstructions such as rivers.

The continual improvement of geographic information systems offers great potential for incorporating such types of spatial information in constructing the weight matrix, D (Clapp and Rodriguez, 1995). For example, using Census data, one could attempt the following refinement for transactions data, since the census attempts to group similar entities. Holding distance constant, one could give higher weights to transactions occurring in the same census block, slightly lesser weights to observations in the same block group, lower weights yet to those in the same tract, and the lowest weights to those in a different tract. As an additional example, one could program the geographical information system to change the weight given (holding distance constant) to an observation depending on traffic counts. The experience of appraisers over the years should lead to rich heuristics for specifying weights. The intersection of geographical information systems, appraiser heuristics, and spatial statistics has a great potential in sharpening the results from real estate data.

Acknowledgments

Both authors gratefully acknowledge the research support they have received from their respective institutions.

Notes

1. This data set has been analyzed extensively. For example, Belsley, Kuh, and Welsch (1980) used the data to examine the effects of robust estimation and reported their observations in an appendix. Krasker, Kuh, and Welsch (1983); Subramanian and Carson (1988); Brieman and Friedman (1985); Lange and Ryan (1989);

Breiman et al. (1993); and Pace (1993) have used the data to examine robust estimation, normality of residuals, nonparametric, and semiparametric estimation.

2. As Belsley et al. (1980, p. 239) noted in their analysis, ''Thus there appear to be potentially significant neighborhood effects on housing prices that have not been fully captured by this model.''
3. Essentially, the spatial autoregressive estimator adapts between using the usual parametric estimate and a nearest neighbor nonparametric estimate of local errors. See Pace (1996). As an alternative interpretation, the optimal combination of OLS and the grid estimators is isomorphic to a spatial autoregression. The spatial autoregression gains over OLS or the grid method by estimating the proper degree of differencing between the subject and comparable properties, which yields less-correlated errors and thus better estimates of β than OLS. See Pace and Gilley (1995).
4. For example, see Ripley (1981, pp. 88–97). See Can (1992) for an application. Note, B largely contains zeros. Hence, one can use sparse matrix techniques to greatly accelerate computations. See Pace (1995).
5. Belsley et al. (1980) reported the observations in an appendix. It also is one of the few moderate-sized hedonic data sets available on the Internet (via STATLIB).
6. The geographical information systems (GIS) literature uses the term *geocoding* for this activity, and the use of electronic means to do this would have greatly reduced the effort of collecting this data. However, the 1970 census antedates the availability of the relevant files. The widespread use of GIS makes it far easier to engage in spatial statistical analysis than in the past.
7. The goodness of fit as measured by R^2 rises from 0.806 to 0.811 when employing the corrected observations on the original Belsley, Kuh, and Welsch model without LAT and LON variables. Moreover, the magnitudes of the coefficients do not change much and the qualitative results from the original regression still hold.
8. We also tried the Manhattan metric, which costs less to compute but yielded a slightly poorer fit.
9. See Cressie (1993).
10. See Colwell, Cannaday, and Wu (1983) for a discussion of the adjustment grid estimator and unobservable errors. See Pace and Gilley (1995) for a discussion of the relation among OLS, the adjustment grid estimator, and the SAR estimator.
11. See Pace and Gilley (1993) and Gilley and Pace (1995) for a discussion of such priors.

References

Belsley, D. A., E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Source of Collinearity.* New York: John Wiley. 1980.

Breiman, L., and J. Friedman. (1985). ''Estimating Optimal Transformations for Multiple Regression and Correlation,'' *Journal of the American Statistical Association* 80, 580–619.

Breiman, L., J. Friedman, R. Olsen, and C. J. Stone. *Classification and Regression Trees.* New York: Chapman and Hall. 1993.

Can, A. (1992). ''Specification and Estimation of Hedonic Housing Price Models,'' *Regional Science and Urban Economics* 22, 453–474.

Clapp, J., and M. Rodriguez. ''Using a GIS for Real Estate Market Analysis: The Problem of Spatially Aggregated Data,'' Working paper (1995), University of Connecticut.

Colwell, P. F., R. E. Cannaday, and C. Wu. (1983). ''The Analytical Foundations of Adjustment Grid Methods,'' *Journal of the American Real Estate and Urban Economics Association* 11, 11–29.

Cressie, N. A. C. *Statistics for Spatial Data*, rev. ed. New York: John Wiley. 1993.

Gilley, O. W., and R. K. Pace. (1995). ''Improving Hedonic Estimation with an Inequality Restricted Estimator,'' *Review of Economics and Statistics* 77, 609–621.

Harrison, D., and D. L. Rubinfeld. (1978). ''Hedonic Housing Prices and the Demand for Clean Air,'' *Journal of Environmental Economics and Management* 5, 81–102.

Krasker, W. S., E. Kuh, and R. E. Welsch. ''Estimation for Dirty Data and Flawed Models,'' In *Handbook of Econometrics*, Vol. 1. Amsterdam: North-Holland. 1983, pp. 651–698.

Lange, N., and L. Ryan. (1989). ''Assessing Normality in Random Effects Models,'' *Annals of Statistics* 17, 624–642.

Pace, R. Kelley. (1993). ''Nonparametric Methods with Applications to Hedonic Models,'' *Journal of Real Estate Finance and Economics* 7, 185–204.

Pace, R. Kelley. ''Performing Large-Scale Spatial Autoregressions,'' (forthcoming), Economics Letters.

Pace, R. Kelley. (1996). ''Relative Efficiencies of the Nearest Neighbor, Grid, and OLS Estimators,'' *Journal of Real Estate Finance and Economics* 13, 203–218.

Pace, R. Kelley, and O. W. Gilley. (1993). ''Improving Prediction and Assessing Specification Quality in Non-Linear Statistical Valuation Models,'' *Journal of Business and Economics Statistics* 11, 301–310.

Pace, R. Kelley, and O. W. Gilley. ''Optimally Combining OLS and the Grid Estimator.'' Manuscript 1995, University of Alaska.

Ripley, Brian D. *Spatial Statistics*. New York: John Wiley. 1981.

Subramanian, S., and R. T. Carson. (1988). ''Robust Regression in the Presence of Heteroskedasticity.'' In *Advances in Econometrics*, Vol. 7. JAI Press, 1988, pp. 85–138.