Predicting the Geography of House Prices

Fingleton, Bernard
London School of Economics

February 2010

SERC DISCUSSION PAPER 45

Predicting the Geography of House Prices

Bernard Fingleton (SERC, SIRE and University of Strathclyde)

February 2010

This work was part of the research programme of the independent UK Spatial Economics Research Centre funded by the Economic and Social Research Council (ESRC), Department for Business, Enterprise and Regulatory Reform (BERR), the Department for Communities and Local Government (CLG), and the Welsh Assembly Government. The support of the funders is acknowledged. The views expressed are those of the authors and do not represent the views of the funders.

Predicting the Geography of House Prices

Bernard Fingleton*

February 2010

*SERC, SIRE and University of Strathclyde

Acknowledgements

Earlier versions of this paper were given at seminars at the LSE and at the Regional Science Association, British and Irish Annual Conference in Limerick. I am grateful to participants for their useful and constructive comments.

Abstract

Prediction is difficult. In this paper we use panel data methods to make reasonably accurate short-term ex-post predictions of house prices across 353 local authority areas in England. The issue of prediction over the longer term is also addressed, and a simple method that makes use of the dynamics embodied in New Economic geography theory is suggested as a possible way to approach the problem.

Keywords: new economic geography, real estate prices, spatial econometrics, panel data, prediction.
JEL Classifications: C21, C31, O18, R12, R31

Predicting the Geography of House Prices

Bernard Fingleton
University of Strathclyde

Introduction

Recently we have seen considerable interest in the causes of spatial variation in house prices, and the inclusion of real estate within contemporary spatial economics theory (Helpman, 1998, Hanson, 2001, 2005, Brakman et. al, 2004, Glaeser, 2008). As Behrens and Robert-Nicoud(2009) observe, writing in the context of the New Economic Geography (Fujita et. al., 1999), 'housing represents the single most important expenditure item and asset for households throughout the world'. This paper continues this theme by estimating a panel data model with network dependence (Anselin, 1988, Elhorst, 2003, Baltagi, 2005, Baltagi et. al. 2003, Anselin et. al. 2007, Kapoor et. al., 2007, Fingleton 2008b), checking the model using ex post prediction methodology (Baltagi and Li, 2006, Baltagi et. al, 2007, Fingleton 2009). The present paper maps out one simple way in which one might approach the problem of the long-run evolution of house prices that combines estimates from the panel data model with the dynamics embodied within New Economic Geography (NEG) theory.

Over the period from the mid-1980s to the start of the present millennium UK house prices[1] gradually increased relatively slowly and unevenly from an index value of 99.9 in 1983Q2 to 278.3 by 2000Q4. However from about 2000Q4 the index increased continuously and dramatically to reach a peak of 649.3 by 2007Q3 (see Figure 1). This changing pattern has a counterpart in an evolving spatial distribution, which is the main focus of the house price model elaborated below.

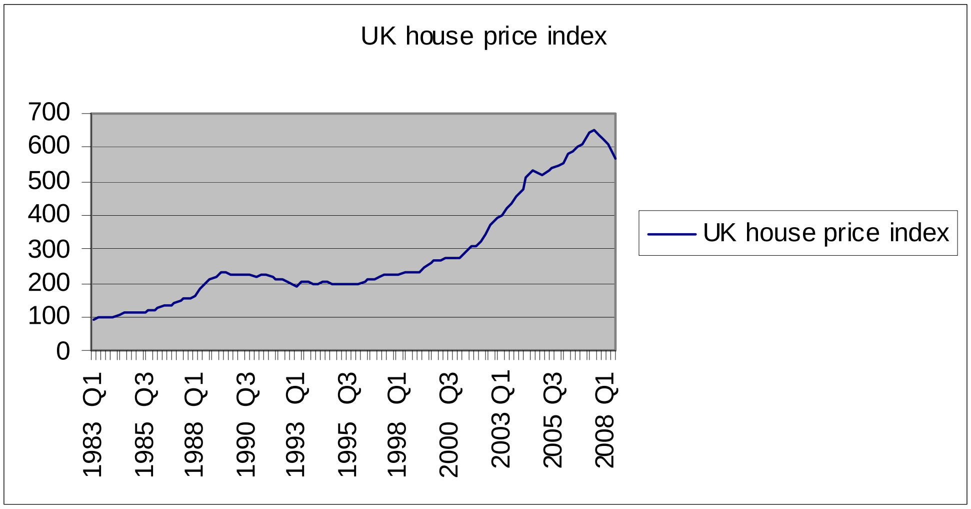

Figure 1 The evolution of UK house prices

The house price model

The estimated house price model is a simplified version Fingleton(2008a), predicting house prices p_{jt} for location $j = 1,...,R$ at time $t = 1...T$ on the basis of a reduced form derived from equilibrium housing supply and demand levels q_{jt}. On the demand side, assume that q_{jt} depends on income level Y^{c}_{jt};this is equal to the sum of income[2] within *j* plus income weighted by commuting distance (D_{jk}) between k (job) and j (home) summing across all R UALAD[3]'s, so that

[1] The Halifax house price index

[2] observed mean wage levels w^{o}_{jt} times number of workers λ_{jt}

[3] Unitary Authority and Local Authority Districts, which are small administrative areas of which there are 353 covering England.

$$Y_{jt}^{c} = \sum_{k=1}^{R} \exp(-\delta D_{jk}) w_{kt}^{o} \lambda_{kt}, D_{jk} = 0, j = k \tag{1}$$

It seems not unreasonable as a simplifying first approximation to assign a value $\delta = 0.05$, since this has the effect of giving approximately zero weight beyond 100 miles[4], and with $D_{jj} = 0$ within-area commuting is assumed to be costless.

Assume also that demand for housing in *j* depends negatively on *j*'s price level (p_t) and positively on prices 'nearby', since relatively high prices 'nearby' , at *k*, will suck demand into lower price *j*. Denote the weighted average of prices near to *j* as the relevant cell of the vector given by the matrix product $W_1 \ln p_t$, in which W_1 is a weighting matrix. This is defined by contiguity, so that $W_{1ij}^{*} = 1$ if i and j share a boundary, and $W_{1ij}^{*} = 0$ otherwise. The matrix W_1^{*} is subsequently normalised to give W_1 so that $\sum_j W_{1ij} = 1$.

Additionally, demand quantity depends on some covariates Δ_t and on some unmeasured variables represented by random disturbances φ_t. Hence, assuming that the relationship between prices and quantities is linear in natural logarithms[5], the demand function is

$$\ln q_t = a_0 + a_1 Y_t^{c} - a_2 \ln p_t + \nu W_1 \ln p_t + \Delta_t \beta + \varphi_t \tag{2}$$

The supply function

$$\ln q_t = b_0 + b_1 \ln p_t - \eta W_2 \ln p_t + \Lambda_t \gamma + \varsigma_t \tag{3}$$

assumes that the level of housing supply increases in p_t but is negatively related to 'nearby' prices $W_2 \ln p_t$ because relatively high *k* prices will pull supply out from *j* in to *k*,

[4] Fingleton(2008a) has a more elaborate procedure by which the decay function is specific to each of the 353 districts, as a result of calibration using commuting data. These data show few commuters travel in excess of 100 miles. For simplicity, in this paper we use an 'average' value for the distance decay parameter, which limits most commuting to within this distance.

[5] This produces a better fit to the data that a linear relationship, gives a constant elasticity and avoids negative prices.

with a weights matrix W_2 equivalent to W_1. Other covariates are represented by the n by k matrix Λ_t and unmodeled effects are captured by the disturbances ς_t.

Normalizing the supply function with respect to *p* thus

$$\ln p_t = \frac{1}{b_1}\ln q_t - \frac{b_0}{b_1} + \frac{\eta}{b_1}W_2 \ln p_t - \frac{\Lambda_t\gamma}{b_1} - \frac{\varsigma_t}{b_1} \qquad (4)$$

and substituting for $\ln q_t$ gives

$$\ln p_t = c_1[a_0 + a_1 Y_t^c - a_3 \ln p_t + \nu W_1 \ln p_t + \Delta_t\beta + \varphi_t] - c_0 + c_3 W_2 \ln p_t - \Lambda_t c_4 - \xi_t \qquad (5)$$

Simplifying by assuming that $W_1 = W_2 = W$ and writing in matrix terms gives

$$P_t = \rho W P_t + H_t d + e_t \qquad (6)$$

in which $P_t = \ln p_t$ and the endogenous spatial lag WP_t are *R* by 1 vectors, H_t is an *R* by *k* matrix with of covariates, *d* is a *k* by 1 vector of parameters, ρ is a scalar parameter and e_t represents unmodeled heterogeneity across *R* areas. Matrix *W* is a simple contiguity matrix standardised so that rows sum to 1.0.

Data

The house price data[6] give p_{rt}, which is the average selling prices (all property types) by UALAD (*r*) for each of the years (*t*) 2000 to 2007. Income by district for each year (Y_{rt}) is calculated by multiplying the mean wage rate (w_{rt}^o) by the employment level (λ_{rt}). The w_{rt}^o s were taken from the annual New Earnings Survey carried out by the UK's Office of National Statistics. These are workplace-based survey data of gross weekly pay for male and female full-time workers irrespective of occupation. These and employment levels are available on the Nomis [7] website.

[6] Provided by the UK's Land Registry.

[7] Nomis is a service provided by the UK's Office for National Statistics, ONS, to give free access to labour market statistics from official sources.

The covariates[8] Δ_r which are assumed to affect demand comprise indicators of amenity within each area, namely the geographical surface area (square km) per district $A_,$ the square of the distance of the area from London L, and the level of educational attainment S. Variable A is included to represent the effects of a spacious environment (rural amenity, lack of congestion) so we would expect a positive relationship between *A* and house prices. The variable *L* represents the effect of differential access to the amenities of London, represented as the square of km distance. We anticipate that the relationship between *L* and house prices will be negative. The variable S is a measure of the level of educational attainment (see Appendix for definition) with good schooling increasing housing demand in an area. We anticipate a positive relationship between *S* and house prices.

On the supply side, Λ_r simply comprises the variable *O* which is equal to the number of owner-occupier households in each district *r* as reported in the 1991 Census of Population[9]. We take the variable *O* as a measure of the stock of properties, a proportion of which are offered for sale. At any moment, one would expect the number of properties for sale to be larger in large cities than in small towns, thus increasing the supply and, ceteris paribus, reducing house prices.

Estimation

There are two issues of prime concern, whether we should used a fixed effects or random effects specification, and whether we should use 2SLS/GMM[10] or some other estimation method such as ML[11]. Regarding fixed versus random effects, we opt to control for individual (district) heterogeneity using (spatially correlated) random effects[12]. These

[8] These are constant over time.

[9] Local Base Statistics, Table L20 Tenure and amenities: Households with residents; residents in households. This is available in the Nomis database.

[10] Two Stage Least Squares combined with Generalised Method of Moments.

[11] Maximum Likelihood.

[12] Typically random effects are adopted when the data comprise a sample, but we can consider the data in this case to also be one of many realisations from a superpopulation since the spatial partitions giving the areal units are just one of a infinite number of possible sets that could have occurred.

are preferred to fixed effects[13] to allow the explicit presence of time constant covariates in the model, to allow more degrees of freedom, and to control for two distinct forms of spatial interactions involving the dependent variable and the disturbances. One, the endogenous spatial lag, is dictated by the theoretical specification of our house price model, but failing to additionally model a significant spatial error process can lead to biased inference because of bias in the estimated standard errors. One extra advantage of random effects estimation is that, because it takes account of permanent cross-sectional or between-variation, it picks up long-run effects, whereas within-unit fixed effect estimation focuses on short-run variation (Partridge, 2005, Baltagi, 2005, Elhorst, 2009).

With regard to the estimation method, there are some computational advantages in using 2SLS/GMM. Typically this approach is robust, for example according to Larch and Walde(2009) 'GMM estimators are preferable when the error distribution is not assured to be normal'. Although Kapoor et al(2007) only give formal large sample results, Larch and Walde(2008) demonstrate that the relative small sample performance of GMM is superior to ML when applied to skewed or fat-tailed distributions. Also ML can pose significant computational problems, for instance Kapoor et al (2007) note that 'even in its simplest form, ML estimation of Cliff-Ord type models entails substantial, and even forbidding, computational problems if the number of cross sectional units is large'. An ML estimator that takes account of both an autoregressive spatial lag and either an autoregressive or moving average error process is also not an achievable or practical option at this point in time. Moreover, For ML, one typically needs to compute the log determinant of the inverse covariance matrix, and in general this is time intensive and difficult for large N. However Le Sage and Pace (2009) do provide approximations, and large N is not such an issue in this particular study.

In our estimates in Table 1 and 2, we use the spatial lags of the exogenous variables as instruments for the endogenous spatial lag, as advocated by Kelejian and Prucha (1998). Variables *A*, *L* and *O* seem to be more clearly exogenous, but variable *S* arguably could be endogenous in that high house prices may cause good parents, teachers and schools to locate in an area. However we also treat *S* as exogenous partly because it relates to the start of the

[13] The assumption that the random effects estimates are not significantly different from comparable fixed effects estimates, which are assumed to be consistent, is supported by a Hausman test, as demonstrated subsequently.

period of analysis and can be shown[14] to be exogenous via Hausman tests of cross-sectional versions of this model estimated by 2SLS[15]. This may be because there are many causes of educational attainment variation, so that would weaken any direct feedback response due simply to house prices. One potential but unrealised disadvantage of 2SLS/GMM is that the parameter space for the spatial lag parameter ρ is unrestricted, leading to potential problems such as non-stationarity or spurious regression arising from estimates outside the known stable range given by the inverse of the maximum and minimum eigenvalues of the *W* matrix (Fingleton, 1999), whereas in the typical cross-sectional model estimated ρ is restricted to a stable continuous parameter space via the presence of a penalty term in the likelihood function.

The preferred model is therefore that of Kapoor et. al. (2007) extended to include both an endogenous spatial lag and a moving average error process (Fingleton, 2008b). Extending (6) to panel notation,

$$\begin{aligned} P &= \rho(I_T \otimes W)P + Hd + e = X\beta + e \\ X &= ((I_T \otimes W)P, H) \\ \beta' &= (\rho, d') \end{aligned} \tag{7}$$

P is a *TR* x 1 vector of observations obtained by stacking P_t for *t* = 1…*T*, *X* is a *TR* x (1 + *k*) matrix of regressors, comprising the *TR* x 1 vector ($I_T \otimes W$)*Y*, and *H* which is a *TR* x *k* matrix of regressors. Also β is the *k*+1 x 1 vector of parameters. In addition, given that I_T is a *T* x *T* diagonal matrix with 1s on the main diagonal and zeros elsewhere, and I_R is a similar *R* x R diagonal matrix, then $I_{TR} = I_T \otimes I_R$ is a *TR* x *TR* diagonal matrix with 1s on the main diagonal and zeros elsewhere.

While one form of spatial interaction is modelled by the endogenous spatial lag $\rho(I_T \otimes W)P$, the second source involves a spatial error process, either an autoregressive process or a moving average process. Equation (8a) is an autoregressive process, so that

$$e = (I_{TR} - \lambda I_T \otimes W_c)^{-1}\xi \tag{8a}$$

[14] Evidence supporting exogeneity is given in Fingleton(2008a).

[15] Two stage least squares.

in which λ is an unknown parameter, and ξ is an *RT* x 1 vector of innovations. The alternative moving average error process specification is

$$e = (I_{TR} - \gamma I_T \otimes W_c)\xi \tag{8b}$$

Both (8a) and (8b) entail time dependency, which is introduced into the innovations ξ via a permanent error component μ, thus

$$\begin{gathered} \mu \sim iid(0, \sigma_\mu^2) \\ \nu \sim iid(0, \sigma_\nu^2) \end{gathered} \tag{9}$$

$$\xi = (\iota_T \otimes I_R)\mu + \nu \tag{10}$$

in which μ is an *R* x 1 vector of area-specific errors. The component ν, the transient error component, comprises an *RT* x 1 vector of errors specific to each area and time. Also ι_T is a *T* x 1 matrix with 1s , and $\iota_T \otimes I_R$ is a *TR* x *R* matrix equal to *T* stacked I_N matrices. The result is that the *TR* x *TR* innovations variance-covariance matrix Ω_ξ is nonspherical. Also $\sigma_1^2 = \sigma_\nu^2 + T\sigma_\mu^2$.

The autoregressive specification for the disturbances *e* means that

$$e = (I_{TR} - \lambda I_T \otimes W_c)^{-1}\xi = (I_{TR} - \lambda I_T \otimes W_c)^{-1}((\iota_T \otimes I_R)\mu) + (I_{TR} - \lambda I_T \otimes W_c)^{-1}\nu \tag{11}$$

So that the permanent and transient error components have identical autoregressive processes (c.f. Baltagi, Eggar, Pfaffermayr, 2009, Baltagi and Li, 2006, Anselin, 1988). Both autoregressive and moving average errors lead to a simple forecasting equation (Baltagi, Bresson and Pirotte, 2007, Goldberger, 1962). Appendix C gives more detail of the estimation process.

Results

The autoregressive error process estimates for the 2000-2007 data given in Table 1 are appropriately signed, although one variable is insignificant using typical inferential rules for one-tailed tests. The variable *O* (housing stock) has a 0.05321 exceedence probability in the lower tail of the approximating N(0,1) distribution, which is greater than the usually acceptable Type I error rate of 0.05, however it is acceptable using a more liberal 0.10 rate for a one-tailed test. The endogenous lag ($W \ln p$), income within commuting distance(Y_j^C), schooling (*S*), rural amenity (*A*) and distance from London (*L*) are all significant using conventional inferential rules.

As stated above, a key assumption of random effects specifications is a lack of correlation between the unobserved effects and the observed variables. This assumption is required for consistency of random effects estimates. The standard Hausman test of consistency relies upon a comparison of random effects and fixed effects estimates. There is a very limited literature on the Hausman test applied to spatial random effects and consistent spatial fixed effects models, apart from Mutl and Pfaffermayr(2009), which is clearly at an experimental stage. The approach, as with Hausman per se, relies on using appropriate estimated covariance matrices from spatial models. In this paper we are interested in random effects with either an autoregressive or a moving average error process, both with an autoregressive spatial lag, $W \ln p$. We compare random effects estimates $(\hat{\beta}_r)$ given in Table 1 with presumably consistent maximum likelihood estimates $(\hat{\beta}_f)$ of a model with an autoregressive spatial lag (Elhorst, 2003), given in Table 3. The appropriate test statistic is

$$H = (\beta_r - \beta_f)'(\Sigma_r - \Sigma_f)^{-1}(\beta_r - \beta_f) \tag{12}$$

in which the Σs denotes the respective covariance matrices from our spatial models. In our case the $\hat{\beta}$ vectors comprise the estimates of parameters ρ and d_1 relating to $W \ln p$ and Y_j^C respectively, and the Σs are the relevant variances and covariance for these two parameters (other variables being constant over time and therefore not identified in the fixed effects panel). For the model with autoregressive errors, using the Table 1 and Table 3 estimates, the test statistic *H* is equal to 4.5533 which is insignificant when referred to the χ_2^2 distribution. We therefore do not reject the null of no difference between the parameters ρ and d_1 estimated via the fixed and random effects specifications. Under the moving averages error process, test statistic *H* is equal to 5.3966 which is quite close to the upper 5% point of χ_2^2 equal to 5.99. This is one reason to confine further analysis to the autoregressive errors specification.

Table 1 : parameter estimates AR error process

Dependent variable ln p		2000-2007		2000-2006	
regressor	β	estimate	t ratio	estimate	t ratio
constant	d_0	-0.0932611	-0.157237	-0.115493	-0.193142
W ln p	ρ	0.574458	5.53414	0.569511	5.41471
Y_j^C	d_1	0.212315	4.54955	0.212227	4.51994
O		-0.000887957	-1.6145	-0.000883734	-1.58684
S	d_2	0.563291	4.33377	0.571198	4.33047
A	d_3	0.105225	3.03076	0.105379	2.99923
L	d_4	-0.00112568	-2.62359	-0.0012043	-2.69181
	λ	-0.154777		-0.147882	
	σ_v^2	0.0271298		0.0252159	
	σ_1^2	0.412741		0.369946	
instruments		Y_j^C , *O, S ,A, L* WY_j^C *,WO, WS, WA, ,WL*			
RSS		187.921		164.842	
R-sq*		0.707993		0.699514	
Schwarz		4.97609		5.10499	
Akaike		4.90658		5.15908	

* squared correlation between fitted and actual
RSS = sum of squared residuals

Table 2 : parameter estimates MA error process

Dependent variable $\ln p$		2000-2007		2000-2006	
regressor	β	estimate	t ratio	estimate	t ratio
constant	d_0	0.0663453	0.0999311	0.00883076	0.01351
$W \ln p$	ρ	0.537082	4.81306	0.540883	4.85818
Y_j^C	d_1	0.230243	4.54798	0.225994	4.50978
O		-0.000799869	-1.40475	-0.000818268	-1.42591
S	d_2	0.567628	4.29107	0.573686	4.28234
A	d_3	0.113341	3.06913	0.112134	3.0309
L	d_4	-0.00118452	-2.45784	-0.0012561	-2.5695
	λ	-0.0413532		-0.00222417	
	σ_v^2	0.0211296		0.019604	
	σ_1^2	0.417313		0.375008	
instruments		Y_j^C , *O, S ,A, L* WY_j^C ,*WO,* *WS, WA, ,WL*			
RSS		195.325		169.372	
R-sq*		0.698913		0.692945	
Schwarz		5.01474		5.1321	
Akaike		4.94522		5.18618	

* squared correlation between fitted and actual
RSS = sum of squared residuals

Table 3 : Fixed effects estimates

Dependent variable ln p			
regressor	β	estimate	t ratio
$W \ln p$	ρ	0.756997	56.281634
Y_j^C	d_1	0.124556	2.154820
Time and individual dummies	$d_2 \ldots. d_{360}$	360 estimates	360 t ratios
	σ^2	0.0015	
diagnostics	RSS	4.30071	
	R-sq*	0.982524	

* squared correlation between fitted and actual
RSS = sum of squared residuals

The presence of a negative autoregressive error process $(\hat{\lambda} = -0.154777)$ is suggestive of an unmodeled variable(s), driven by 'alternating' urban and rural locations consistent with a central place model producing a 'checkerboard' pattern for the residuals.

Prediction methodology

Given the model $P_t = \rho W P_t + H_t d + e_t$, the variable $Y_{j,t}^C$, together with the other regressors $const., O_t, S_t, A_t, L_t$ are the columns of matrix H, and given $\hat{\rho}, \hat{d}, \hat{e}$ together with ϖ and $\hat{\Omega}$, we obtain the predicted log house price $\hat{P} = \ln \hat{p}$ via the prediction equation

$$\hat{P} = (I - \hat{\rho} W)^{-1}(H\hat{d} + \varpi' \hat{\Omega}^{-1}\hat{e}) \qquad (13)$$

In (13) $\varpi' \hat{\Omega}^{-1}\hat{e}$ is exactly the BLUP correction initially given by Goldberger(1962). In this the RT by RT error covariance matrix Ω is

$$\Omega = \Omega_{\xi}[(I_{TR} - \lambda I_T \otimes W_c)'(I_{TR} - \lambda I_T \otimes W_c)]^{-1} \qquad (14)$$

and ϖ_i is an RT by 1 vector of covariances of the prediction disturbance at location i with the estimate of the RT by 1 vector of residuals $\hat{e}$.

It turns out (Baltagi, Bresson and Pirotte, 2007) that

$$\begin{aligned} \varpi' \hat{\Omega}^{-1}\hat{e} &= \frac{\sigma_\mu^2}{\sigma_1^2}\left(\iota_T' \otimes l_i'\right)\hat{e} = \frac{T\sigma_\mu^2}{\sigma_1^2}\bar{\hat{e}} = \frac{T\sigma_\mu^2}{T\sigma_\mu^2 + \sigma_\upsilon^2}\bar{\hat{e}} \\ \bar{\hat{e}} &= \sum_{t=1}^{T}\hat{e}_t \Big/ T \end{aligned} \qquad (15)$$

So, the correction $\varpi' \hat{\Omega}^{-1}\hat{e}$ is simply equal to a proportion of the mean disturbance averaging over T periods.

Ex-post prediction

We use the parameter estimates for 2000-2006 data, also given in Table 1, to carry out ex-post predictions for out-of-sample house price levels for 2007. We carry out two predictions, one with the Goldberger correction (as in equation 13) and one without (using $\hat{P} = (I - \hat{\rho} W)^{-1} H\hat{d}$). The presence of the Goldberger correction is highly beneficial, as shown by Figure 2 and by the respective root mean square errors are 0.438914 (no correction) and 0.366963 (with correction).

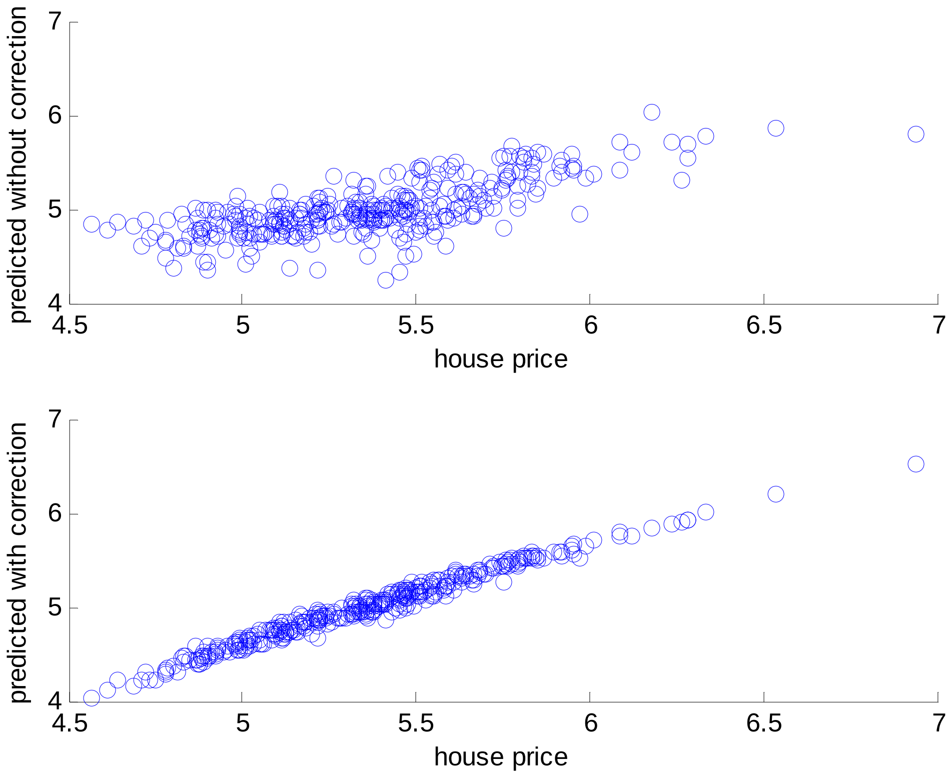

Figure 2 Ex post predictions of log house prices in 2007

Long-run prediction

While we have achieved reasonably accurate ex post predictions of house prices, the problem of ex ante long-run prediction is more difficult. We do not carry this out in practice in this paper, but simply suggest one simple way in which it might be accomplished. This relies on predicting the level of income within commuting distance, assuming everything else remains constant, basing our prediction on the dynamics embodied within the NEG model, as outlined below.

The NEG model reduces to a small set of simultaneous equations. As in Fujita, Krugman and Venables (1999, Chapter 7), we assume one competitive sector (*C*) and one under monopolistic competition market structure (*M*), thus generalizing the basic specification to allow transport costs and wage and product differentiation for the *C* sector. There is rapid short-term adjustment to equilibrium by firms, but only in the long-run are *M* workers responsive to across-region variations in real *M* wages, and *C* workers are assumed to be immobile.

Preferences are the usual Cobb-Douglas form with a CES subutility function for *M* varieties. Hence $U = M^{\theta}C^{1-\theta}$, in which θ is equal to the proportion of expenditure of *M* goods out of total (*M* + *C*) expenditure. The quantity of the composite good *M* is a function of the *f* = 1…*x* varieties *m(f)*, where *x* is the number of varieties, so that

$$M = \left[\sum_{f=1}^{x} m(f)^{(\sigma-1)/\sigma} \right]^{\sigma/(\sigma-1)} = \left[\sum_{f=1}^{x} m(f)^{\frac{1}{\mu}} \right]^{\mu} = x^{\mu} m(f) \qquad (16)$$

This is because under monopolistic competition at equilibrium *m(f)* is a constant across all *f* varieties. Because of normalizations employed, θ is also equal to the equilibrium number of workers per firm and to the equilibrium output per firm.

We use the notation of Fingleton(2005), which extends Fujita, Krugman and Venables' (1999, Chapter 7) two-region representation to a multi-region specification, giving five simultaneous non-linear equations as a reduced form of the theoretical structural model. Equations (17) and (18) define *M* and *C* wages (w_i^M and w_i^C) for region i, equations (19) and (20) give *M* and *C* prices (G_i^M and G_i^C), and equation (21) gives income (Y_i). Nominal *M* wages and the *M* and *C* price indices determine real *M* wages (ω_i) as in (22). The elasticities of substitution are denoted by σ and η for *M* and *C* varieties respectively, and λ_r and ϕ_r which are the respective shares of the total supply of *M* and *C* workers for *r* = 1…*R*.

$$w_i^M = [\sum_r Y_r (G_r^M)^{\sigma-1} T_{Mir}{}^{1-\sigma}]^{\frac{1}{\sigma}} \qquad (17)$$

$$w_i^C = [\sum_r Y_r (G_r^C)^{\eta-1} T_{Cir}{}^{1-\eta}]^{\frac{1}{\eta}} \tag{18}$$

$$G_i^C = [\sum_r \phi_r (w_r^C T_{Cir})^{1-\eta}]^{\frac{1}{1-\eta}} \tag{19}$$

$$G_i^M = [\sum_r \lambda_r (w_r^M T_{Mir})^{1-\sigma}]^{\frac{1}{1-\sigma}} \tag{20}$$

$$Y_r = \theta\lambda_r w_r^M + (1-\theta)\phi_r w_r^C \tag{21}$$

$$\omega_i = w_i^M (G_i^M)^{-\theta} (G_i^C)^{\theta-1} \tag{22}$$

The solution to equations (17 to 22) gives the short run equilibrium, which is not sustainable in the long run because of regional differences in real wages, which ultimately cause labour migration and affect λ_r (C workers are immobile so that ϕ_r is constant). The dynamics in the original model (leading to one of several long-run equilibria) are a function of real wage differences, as follows

$$\begin{aligned} \varpi &= \sum_r \lambda_r \omega_r \\ \dot{\lambda}_r &= \kappa(\omega_r - \varpi)\lambda_r \end{aligned} \tag{23}$$

The rationale for (23) is described by Fujita, Krugman and Venables (1999, page 62). They simply assume that *M* workers move towards regions offering higher real wages and away from those offering below average real wages, so that the employment change $\dot{\lambda}_r$ is either positive or negative for region *r*, but $\Sigma\dot{\lambda}_r = 0$ so there are no new jobs generated or destroyed, but simply a reallocation. In the current house price context, the migration criterion is slightly more complex, being a composite variable taking account of both real wages and house prices.

House prices are treated separately from the prices of all other goods, which within the NEG model are given by (19,20), being simply a function of wages or prices at the source of production which are increased to allow for (iceberg) transport costs, summing across all sources of production and weighted by the emissivity of each source of goods (ϕ_r, λ_r). Clearly this price determination mechanism would be inappropriate in the case of house prices, since a house is not a good that is transported to market. Our model elaborated earlier suggests factors responsible for house price variations, and we retain this model specification below. Consider next an adaptation of the labour reallocation mechanism (23) in which

$$\bar{\omega} = \sum_r \omega_r^a p_r^{a-1} \lambda_r \qquad (24a)$$

$$\dot{\lambda}_r = \kappa(\omega_r^a p_r^{a-1} - \bar{\omega})\lambda_r \qquad (24b)$$

In (24) we have a composite variable $\omega_r^a p_r^{a-1}$ in which p_r is *r*'s house prices and $0 \le a \le 1$ determines the relative weight of real wages and house prices. We commence with an initial set of house prices p_r and real wages ω_r. The difference $\omega_r^a p_r^{a-1} - \bar{\omega}$ induces labour migration and hence via (24) a change $\dot{\lambda}_r$ in the distribution of *M* sector activity, again with $\Sigma\dot{\lambda}_r = 0$. If real wages outweigh house prices, then workers will be drawn to high wages regions. If house prices outweigh real wages, workers will move in the direction of low house prices. For some *a*, workers can be attracted simultaneously to regions either because they have low prices or because they have high wages. The parameter κ determines the overall magnitude of the $\dot{\lambda}_r$ s, which may be more or less sensitive to our real wage-house price variable relative to its mean. The solution to equations (17 to 22) for iteration *t* using resulting vector $\lambda_{r,t} = \lambda_{r,t-1} + \dot{\lambda}_{r,t-1}$ gives new sets of real wages $\omega_r = \omega_{r,t}$ and income $Y_r = Y_{r,t}$. In turn, this changes income within commuting distance $Y_j^C = Y_{j,t}^C$ and hence house prices (using equation (13)), leading to a new set of prices $p_r = p_{r,t}$. Applying the migration decision rule embodied in (24) using the iteration *t* values of ω_r, p_r and λ_r gives new $\dot{\lambda}_r$

and thus a new allocation of workers λ_r, so that we again need to iteratively solve (17 to 22) and consequently re-calculate prices using (13), and so on. Thus we can track the dynamics of real wages and house prices and the distribution of labour as they interact through time.

While this has not yet been explored empirically, it is possible that one or more long-run equilibria will emerge from this iterative process, and that such equilibria may differ from those based purely on real wage variations (or equivalently when *a* = 1). Defining equilibrium as $\omega_r^a p_r^{a-1} - \overline{\omega} = 0$, then $\dot{\lambda}_r$ also is equal to zero and the distribution of *M* activity λ_r is in steady state, but it is evidently possible that this equilibrium state may be achieved in the presence of <u>permanent</u> real wage and house price differences across regions. This is contrary to the long-run equilibria at which $\dot{\lambda}_r = \kappa(\omega_r - \overline{\omega})\lambda_r = 0$ implying equalized real wages across localities, given that that $\omega_r - \overline{\omega} = 0$ for all *r* under the dynamics suggested by Fujita, Krugman and Venables (1999). The existence of multiple equilibria under the set-up proposed in this paper is an additional possibility that warrants further investigation.

<u>Conclusion</u>

As Danish physicist Neils Bohr once said, ‘Prediction is very difficult, especially about the future’. The paper outlines one way in which long-run house price predictions may obtained based on the dynamics of an NEG model, and from this it is evident that any long-run equilibria need not necessarily imply undifferentiated real wage rates or house prices. Rather, steady state, if it occurs, is when there is equalization of a composite variable combining real wages and house prices. While we have mapped out a simple way in which we can integrate NEG dynamics and a static house price model, it is of course apparent that practical difficulties remain in turning these ideas into a realistic forecasting tool. At this juncture it would be more reasonable to treat model outcomes as simulations based on a set of assumptions, and to test the sensitivity of outcomes to different assumptions. Only with additional research will it be possible to establish whether we can make further progress with this line of analysis that satisfies our need for realism as well as for theoretical and econometric rigour.

References

Anselin L (1988) *Spatial Econometrics: Methods and Models* Dordrecht Kluwer.

Anselin L, Le Gallo J, and Jayet J (2007) Spatial Panel Econometrics, Chapter 19 in Matyas L. and Sevestre P. (Eds.), *The Econometrics of Panel Data*, *Fundamentals and Recent Developments in Theory and Practice* (3rd Edition). Dordrecht Kluwer.

Baltagi B H (2005) *Econometric Analysis of Panel Data 3rd Edition* Chichester Wiley.

Baltagi, B H, Song S H and Koh W (2003) Testing panel data regression models with spatial error correlation, *Journal of Econometrics*, 117 123-150

Baltagi B H and Li D (2006) Prediction in the Panel Data Model with Spatial Correlation: The Case of Liquor, *Spatial Economic Analysis* 1 175-185

Baltagi B H, Bresson G and Pirotte A (2007), Forecasting with Spatial Panel Data, *ERMES Working Paper* number 07-10, University of Paris II

Baltagi, B H, Egger P and Pfaffermayr M (2008) A Monte Carlo Study for pure and pretest estimators of a panel data model with spatially autocorrelated disturbances, *Annales d'Économie et de Statistique* 97-98 11-38

Behrens K, Robert-Nicoud F (2009) Krugman's *Papers in Regional Science* : the 100 dollar bill on the sidewalk is gone and the 2008 Nobel Prize well-deserved, *Papers in Regional Science* 88 467-489

Bowden R J and Turkington D A (1984) *Instrumental variables* Cambridge Cambridge University Press.

Brakman S, Garretsen H, and Schramm M (2004) The Spatial Distribution Of Wages: Estimating The Helpman-Hanson Model For Germany, *Journal Of Regional Science* 44 437–466

Elhorst J P (2003) Specification and Estimation of Spatial Panel Data Models *International Regional Science Review* 26 244 – 268

Elhorst J P(2010) Spatial Panel Data Models, Chapter C2 pp. 377-405 in Fischer M M and Getis A (eds.) *Handbook of Applied Spatial Analysis*, Berlin Springer-Verlag.

Fingleton B (1999) Spurious spatial regression: some Monte-Carlo results with a spatial unit root and spatial cointegration, *Journal of Regional Science*, 39, 1-19

Fingleton B (2005) Towards applied geographical economics: modelling relative wage rates, incomes and prices for the regions of Great Britain, *Applied Economics* 37 2417-2428

Fingleton B (2008a) Housing supply, housing demand, and affordability, *Urban Studies* 45 1545-1563

Fingleton B (2008b) A Generalized Method of Moments estimator for a spatial panel model with an endogenous spatial lag and spatial moving average errors *Spatial Economic Analysis* 3 27-44

Fingleton B (2009) Prediction using panel data regression with spatial random effects *International Regional Science Review* 32 173-194

Fujita M, Krugman P R and Venables A (1999) *The Spatial Economy : Cities, Regions, and International Trade* Cambridge Massachusetts MIT press

Glaeser E L (2008) Cities, *Agglomeration, and Spatial Equilibrium* Oxford Oxford University Press

Goldberger A S (1962) Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association* 57 369-375.

Greene W H (2003) *Econometric Analysis 5th Edition* New Jersey Prentice Hall

Hanson G H (2001) Scale Economies and the Geographic Concentration of Industry, *Journal of Economic Geography* 1 255–276

Hanson G H (2005) Market potential, increasing returns and geographic concentration, *Journal of International Economics* 67 1 –24

Helpman E (1998) The Size of Regions, pp. 33–54 in D. Pines, Sadka E. and Zilcha I. (eds), *Topics in Public Economics*. Cambridge Cambridge University Press

Larch M and Walde J (2009) Finite sample properties of alternative GMM estimators for random effects models with spatially correlated errors *Annals of Regional Science*, 43 473 – 490

Kapoor M, Kelejian H H and Prucha I (2007) Panel Data Models with Spatially Correlated Error Components *Journal of Econometrics*, 140 97-130

Kelejian H H and Prucha I (1998) A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances *Journal of Real Estate Finance and Economics*, 17 99-121

LeSage J and Pace K R (2009) *Introduction to Spatial Econometrics* New York CRC Press

Mutl J and Pfaffermayr M (2008) The Spatial Random Effects and the Spatial Fixed Effects Model: The Hausman Test in a Cliff and Ord Panel Model,   *Economics Series* 229 Institute for Advanced Studies, Vienna

Partridge M (2005) Does Income Distribution Affect U.S. State Economic Growth*? Journal of Regional Science* 45 363-394

Appendix

Estimating the house price model

The model is estimated by a combination of (robust) 2SLS and GMM, using an approach that is similar to that of Kapoor et al. (2007).

Simplifying to eliminate technical detail (see Fingleton, 2009), the first stage of estimation involves obtaining consistent estimates of β and hence the disturbances using appropriate instrumental variables. These comprise a linearly independent subset of exogenous variables to give the *TN* x $f \geq (k+1)$ matrix of instruments *Z*, and we assume matrices *X* and *Z* are full column rank with $f \geq (k+1)$. Following standard practice *Z* includes the spatial lags of the exogenous variables. Given consistent residual estimates we apply GMM, using nonlinear least squares, to estimate $\sigma_\mu^2, \sigma_\nu^2$ and λ, following the method of Kapoor et al. (2007), although for simplicity we do not use differential weighting[16].

Finally, we first use the Cochrane-Orcutt transformation in order to eliminate (autoregressive) spatial error dependence, given that an estimate of λ has been obtained, hence

$$\begin{aligned} P^* &= (I_T \otimes (I_N - \hat{\lambda} W_c))P \\ X^* &= (I_T \otimes (I_N - \hat{\lambda} W_c))X \\ \xi &= (I_T \otimes (I_N - \hat{\lambda} W_c))e \end{aligned}$$

We proceed using instrumental variables, but also take account of the non-sphericity of variance-covariance matrix Ω_ξ, which depends on σ_ν^2 and $\sigma_1^2 = \sigma_\nu^2 + T\sigma_\mu^2$, to obtain the final estimates of vector β Hence

$$\hat{\beta} = \left[(X^{*\prime} Z)(Z' \hat{\Omega}_\xi Z)^{-1} (Z' X^*) \right]^{-1} (X^{*\prime} Z)(Z' \hat{\Omega}_\xi Z)^{-1} (Z' P^*)$$

The estimated parameter variance-covariance matrix is given by

[16] Kapoor et al (2007) note that consistent estimates are obtained using equal weight to all six moments equations.

$$\hat{C} = \left[(X^{*\prime} Z)(Z'\hat{\Omega}_{\xi} Z)^{-1}(Z'X^{*}) \right]^{-1}$$

following Bowden and Turkington (1984) and Greene (2003). The quantities $\frac{\hat{\beta}_i}{\hat{C}_{ii}}$ are treated as ‘t-ratios’ for inferential purposes. For the moving average error process, the transformations are

$$P^{*} = (I_T \otimes (I_N - \hat{\lambda} W_c))^{-1} P$$
$$X^{*} = (I_T \otimes (I_N - \hat{\lambda} W_c))^{-1} X$$
$$\xi = (I_T \otimes (I_N - \hat{\lambda} W_c))^{-1} e$$

and the relevant moments equations are given by Fingleton(2008b).

D. Educational Attainment

This is based on the 1998 key stage 2 tests taken by 11-year-old pupils initially available for individual schools within smaller administrative areas nested within UALADs (these are known as wards, of which there are 8413 in England). The mean scores per Ward were then used to calculate mean scores for each of 353 English UALADs thus giving the regressor $S_.$

BERR
Department for Business
Enterprise & Regulatory Reform

Communities
and Local Government

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Llywodraeth Cynulliad Cymru
Welsh Assembly Government

Spatial Economics Research Centre (SERC)
London School of Economics
Houghton Street
London WC2A 2AE

Tel: 020 7852 3565
Fax: 020 7955 6848
Web: www.spatialeconomics.ac.uk

SERC is an independent research centre funded by the Economic and Social Research Council (ESRC), Department for Business, Enterprise and Regulatory Reform (BERR), the Department for Communities and Local Government (CLG) and the Welsh Assembly Government.