Residential Land Supply in 27 EU Countries: Pigovian Controls or Nimbyism?

Lasse de la Porte Simonsen*
and
Stephen Wright†

19th February 2016

Abstract

We use a new dataset of around a quarter of a million survey points that allows us both to derive estimates of residential land on a per capita basis for 27 EU countries, and to model its supply. There is a fairly strong negative correlation between residential land per capita and population density, despite the fact that residential shares are typically very low. In the national data there is also a striking *lack* of correlation between residential land and per capita consumption, but with no indication that this reflects any true economic scarcity value. We model the spatial distribution of residential land allowing both for spatial correlation and the impact on land supply of a consumption externality from nearby housing. We assume that planning policy restricts land supply to match its price to perceived marginal social cost. Our econometric results provide qualitative support for the model; but it is very hard to match our results to plausible structural parameters unless we assume a social planner who both gives a far greater weight to the impact of the externality than to the welfare gains from new housing, and perceives population density to be far larger then it actually is.

JEL Classification: *C21, D62, R31, R52*

Keywords: *Housing, Residential Land, Land Use Regulation, Spatial Economics*

*Department of Economics, Maths & Statistics, Birkbeck College, University of London, Malet Street, London W1E 7HX, UK. Email: `l.simonsen@mail.bbk.ac.uk`

†Department of Economics, Maths & Statistics, Birkbeck College, University of London, Malet Street, London W1E 7HX, UK. Email: `s.wright@bbk.ac.uk`

1 Introduction

Research into the housing market suffers from a paucity of data that allow direct inter-country comparisons of either quantity or absolute prices.[1] In this paper we focus on a key input to housing, residential land, that can be measured directly, thus allowing inter-country comparisons. We use a dataset taken from the European Land Use and Cover Area-Frame Statistical Survey (LUCAS) (Eurostat, 2012) that allow us both to derive national and regional estimates of residential land on a per capita basis, and model its spatial distribution and economic determinants, on a consistent basis in 27 EU countries.

LUCAS provides us with a dataset of around one quarter of a million points from a stratified grid covering most of the inhabited geographical area of the 27 EU countries in the survey (which was carried out in 2012). The motivation for LUCAS was primarily to survey land use in agriculture and forestry; but as a by-product it tells us whether any given point in the survey was used for residential purposes, as well as providing some classification of its physical properties. This dataset allows us to carry out two complementary exercises: in measurement and modelling.

1.1 Measuring residential land in 27 EU countries

In Section 2 we use the percentage of points classified as residential to derive area estimates of total residential land and some of its subcomponents at both national and regional (Nomenclature of Units for Territorial Statistics level 2 or NUTS2) levels. At a national level these estimates have relatively tight confidence intervals (even after we account for non-trivial degrees of spatial correlation). However precision falls at a regional level, and in smaller or more sparsely populated countries.

We focus on five key summary facts derived from the resulting estimates:

1. *Shares of residential land in total land are typically very low.* The median share at a national level is 2.4%, and at a (probably more representative) regional level it is 3.2%.[2] In only one country (Malta), and in fewer than 5% of the regions is the residential share above 20%. Shares of land that are actually built on are typically considerably smaller.

2. *Residential land on a per capita basis has a very wide range of cross-sectional variation*, both at the national and regional level. At a national level its cross sectional log standard deviation is similar to that of consumption per capita; while intra-country regional variations are typically even larger. Yet the *correlation of residential land*

[1] Datasets such as those of Federal Reserve Bank of Dallas (2015) *"International House Price Database"*, the Bank for International Settlements (2015) *"Residential property statistics"*, the International Monetary Fund (2015) *"Global House Price Index"* or the OECD (2015) *"Focus on house prices"* provide valuable information on price changes, but do not enable direct comparisons of quantities or absolute prices

[2] NUTS2 Regions: these are mostly more homogeneous in population terms than countries.

per capita with national consumption per capita is close to zero. Since land is a crucial input to housing consumption, and is fairly evidently an imperfect substitute for other inputs, this lack of correlation is quite striking.

3. Combining our land estimates with national accounts data, *estimates of housing expenditure per square metre of residential land display massive variation between EU countries.* For example, the average UK household has housing expenditure of around 22 Euros per square metre of residential land, roughly 10 times as much per square metre as an Hungarian or Polish households (for comparison, UK total consumption per capita is only around 3 times higher).[3]

4. We also combine LUCAS-based estimates of land used for non-residential purposes (predominantly agriculture and forestry) with estimates of value added for these sectors. This allows us to compare value added per square metre of land between housing and non-housing. While there is a modest positive cross-sectional correlation between the two, *the opportunity cost of land is extremely low in comparison to its value added in housing.* Thus it is very hard to explain the lack of correlation of residential land with consumption, noted in Fact 2, in terms of compeling economic demands.

5. *Regional variations in residential land per capita within most countries are very large*, and in some countries distinctly larger than variation between countries. This variation, is not only a results of the large discrepencies between highly urban regions (such as congested cities) and sparsely populated rural regions.

1.2 Modelling the spatial distribution of residential land: an implicit model of land supply

In the remainder of the paper we extend our analysis to the full dataset, by modelling the distribution of residential land at each of the roughly quarter million individual points in the survey. Our model incorporates both spatial correlation (captured by the share of neighbouring points that are classified as residential) and an implicit model of land supply, that is at least qualitatively consistent with a Pigovian equilibrium, that determines land supply (a framework similar in spirit to Cheshire and Sheppard (2002), but incorporating the impact of spatial correlation), which in turn determines the unconditional probability that any given point in a given country or region is residential.

In this framework, other people's consumption of the housing services provided by residential land imposes a consumption externality on those living nearby. This is accentuated by spatial correlation, which for a given share of residential land in total land, increases the conditional probability that housing will be consumed near other housing.

[3] We focus on comparisons where land estimates are reasonably precisely measured. The point estimate for Luxembourg is nearly 50 Euros per square metre, but the land estimates used in that calculation have a very wide margin of error.

We assume that planning policy deliberately restricts land supply to match its price to its perceived marginal social cost, in an attempt to mimic a Pigovian equilibrium.

In the absence of the externality, and with the opportunity cost of land in other uses assumed to be low and constant (consistent with Fact 4), the model would predict that consumption per capita and residential land per capita would rise at least proportionately. However, with Pigovian controls over land supply the impact of higher consumption will be much more limited. Residential land supply will be higher (and hence prices lower) where total land per capita is higher (since this lowers the unconditional probability of the consumption externality). Thus there will be a negative relationship between housing supply and population density (matching, at least qualitatively, the story told by Miles (2012)), but which would not arise in the absence of the externality. We estimate the model both in implicit form (as a determinant of the unconditional probability that a given point in any region will be residential, taking account of the impact of spatial correlation on the conditional probability) and directly using regional data. Results are very similar. Population density has a strong negative impact on land supply. We find that at a regional level consumption per capita does (as our theory would predict) also have a modest positive impact on supply of land, and hence on regional residential probabilities. However, this is largely obscured in data at a national level, due to large regional variations in population density.

While our econometric results provide qualitative support for our theoretical model, we show in Section 5 that it is very hard to match our reduced form results to plausible structural parameters unless we assume a social planner who gives a far greater weight to the impact of the externality than to the welfare gains from new housing.

1.3 Structure of the paper

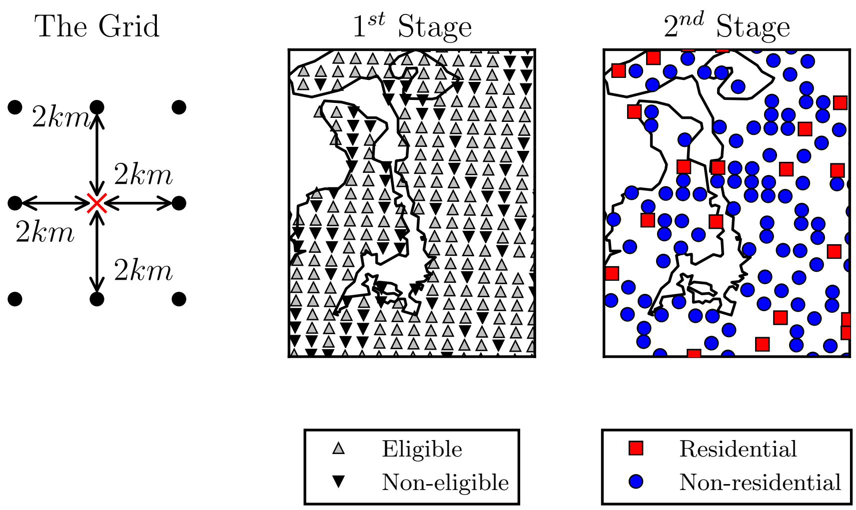
In Section 2 we describe our dataset and summarise its key features, which provides the basis for our Facts 1 to 5 as outlined above. Section 3 sets out our model of Pigovian land supply. Section 4 presents our estimation results. In section 5 we discuss the estimation of the reduced form parameters in relation to the structural parameters of the model. Finally Section 6 we draw conclusions from our analysis.

2 The dataset

2.1 The LUCAS Methodology

Eurostat's *"Land Use and Cover Area frame Statistical Survery"* (LUCAS) is a two phase sample survey. The first phase is an equally spaced systematic grid of 1,078,764 observations (in the 2012 sample) in 27 EU countries, separated by 2 km in the four cardinal directions. Each of the points in the first-stage sample are photo-interpreted and classified in terms of land cover,[4] as well as eligibility (based on accessibility) for the second stage of the survey.[5] Together these two classifications give the stratifications of the first stage sample. For the second a subset of 270,277 eligible points from the first-stage sample were visited in person by a surveyer. It is the dataset derived from this physical survey that we use in this paper.

Figure 1: The LUCAS dataset



For robustness we proceed with two different definitions of residential land, using land use and cover definitions from the LUCAS survey (Eurostat, 2013). The first, *"broad"* residential land, uses all survey points classified as residential by land use (LU *"U370"* in the dataset). The alternative "*narrow*" measure uses only the subset of residential points that are also classified by land cover as artificial structures (land cover A11 *"Buildings with one to three floors"* and A12 *"Buildings with more than three floors"*).

We augment the LUCAS dataset by using Eurostat data on population and gross value added at a regional level (from NACE) as well as Consumption, Actual Rent, Implied Rent and Maintenance at a national level from national accounts (NAMA).

2.2 Estimates of Residential Land and its Components

The key features of the estimates we derive from the LUCAS dataset are summarised in Table 1.

[4]Using the "CORINE" classification. 1: Arable, 2: Permanent Crop, 3: Grassland, 4: Woodland and shrubland, 5: Bareland, 6: Artificial, 7: Water and Wetland

[5]Only points that were both below 1,500m in altitude and accessible by road were included in the second stage.

Table 1: Summary table of the dataset

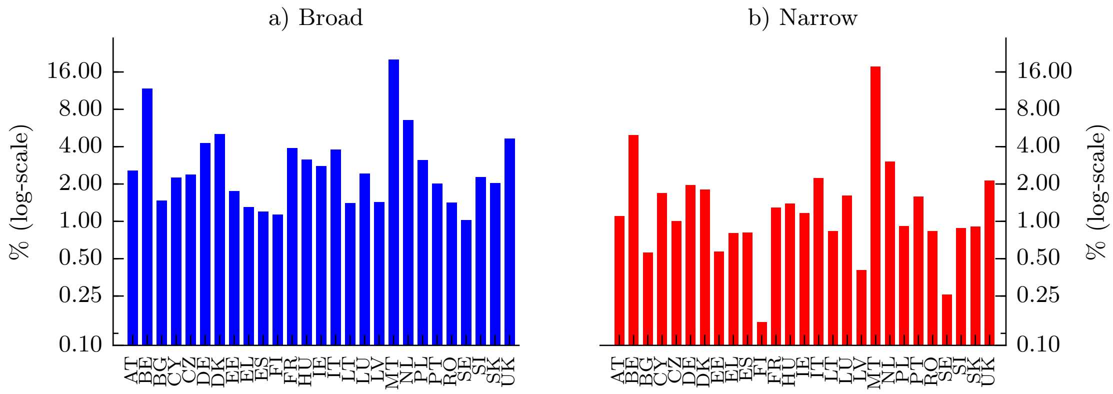
	Total Land (m^2 per capita)	Accessible Land (m^2 per capita)	Residential Broad Land (m^2 per capita)	Residential Narrow Land (m^2 per capita)	Housing share (%)	Value added ($\frac{EUR}{m^2}$)
AT	9,976	5,623	256	110	14.9	11.2
BE	2,752	2,295	323	137	18.1	9.2
BG	15,135	10,399	223	85	16.0	2.6
CY	10,732	6,691	241	181	15.0	9.7
CZ	7,507	6,152	178	75	18.2	7.6
DE	4,364	3,187	187	86	17.6	16.2
DK	7,687	5,569	390	138	20.6	10.7
EE	34,128	24,166	600	195	12.7	1.3
EL	11,863	7,959	155	96	18.8	14.7
ES	10,808	7,291	130	88	18.8	17.8
FI	62,658	48,492	713	97	22.5	5.7
FR	9,692	6,797	380	125	20.3	8.9
HU	9,366	6,519	295	131	12.9	2.3
IE	15,231	11,810	426	178	18.8	7.0
IT	5,073	3,807	192	114	17.4	14.5
LT	21,740	15,103	304	182	8.9	1.9
LU	4,927	2,959	119	80	18.5	47.3
LV	31,574	20,105	453	128	14.3	1.8
MT	757	757	153	134	8.1	5.8
NL	2,483	1,714	162	75	17.5	17.5
PL	8,113	6,628	252	74	10.9	2.5
PT	8,747	5,898	177	138	13.7	8.1
RO	11,863	8,067	168	99	16.8	4.0
SE	46,249	29,189	473	119	20.5	7.7
SI	9,863	8,068	225	87	12.0	5.4
SK	9,073	6,905	184	83	12.4	4.7
UK	3,914	3,002	182	83	21.1	20.4
Average	13,936	9,820	279	115	16.2	9.9
Median	9,692	6,691	225	110	17.4	7.7
CoV	1.02	1.04	0.53	0.31	0.23	0.94
Max	62,658	48,492	713	195	22.5	47.3
Min	757	757	119	74	8.1	1.3

CoV: Coefficient of Variation. Housing share is total housing expenditures divded by consumption. The value added is housing expenditures divided by the broad measure of residential land in m^2.

2.2.1 Estimated shares of residential land in total land

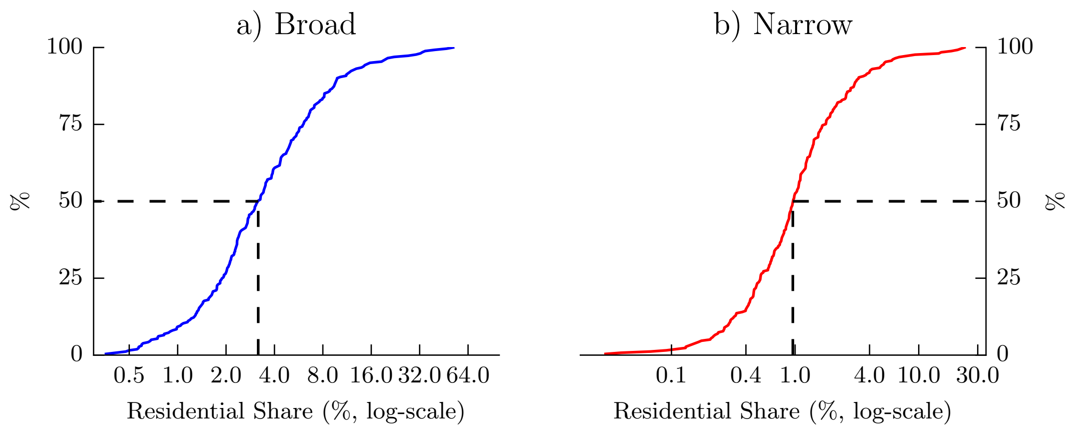
As documented in our Fact 1 and Table 1, Figure 2 show that for the great majority of EU countries residential land represents only a small proportion of the total land area. Only Malta and Belgium have residential shares in double figures.

Figure 2: Shares of residential land in total land in 27 EU countries



We can also do the same calculation for the 261 EU (NUTS2) regions covered by the survey. While individual regional point estimates are subject to nontrivial measurement error (discussed further in the next section) we can nonetheless derive some key features of the regional distribution. Figure 3 plots the empirical cumulative density function of residential shares at a regional level. This provides further substantiation of our Fact 1: in the great majority of regions of the EU residential shares are low, whether using our broad or narrow definition of residential land. The median regional residential shares are 3.2% on our broad definition, and less than 1% on the narrow definition. Only 10% of EU regions have broad residential shares in double figures; and only 4% of regions (all major urban areas) have shares above 20%.

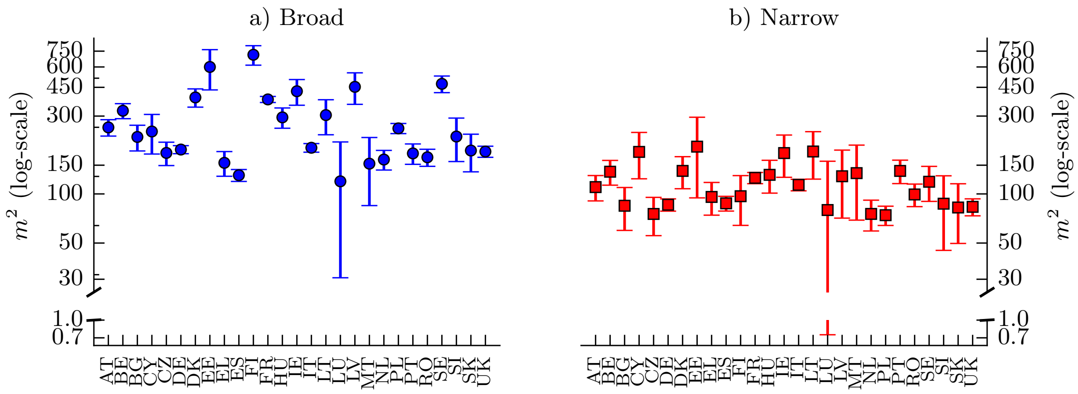
Figure 3: Residential Shares by EU Region: Empirical CDFs



2.2.2 Estimates of residential land per capita

To document our Fact 2 we can also express our land estimates in per capita terms, which again show very large differences both at a national and regional level. Figure 4 shows estimates of per capital residential land on both definitions at a national level, together with estimated 95% confidence intervals.[6] For most countries the range of sampling uncertainty in at least the broad measure of residential land is small in comparison with the large differences between countries. Table 1 shows that the coefficient of variation of broad residential land per capita across countries is around 50%: comparable to the degree of varation in consumption per capita within the 27 countries.

Figure 4: Estimates of residential land per capita in 27 EU countries, with 95% confidence intervals

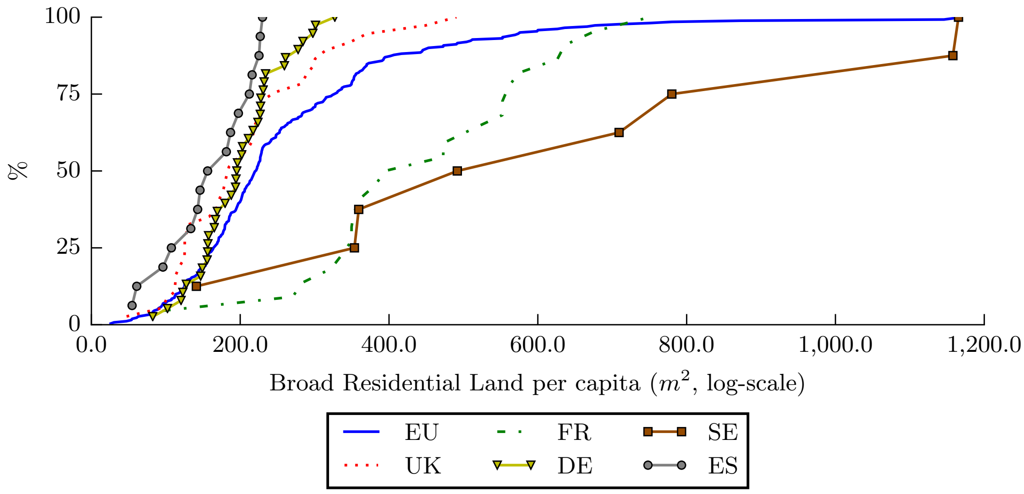


Since measurement error is decreasing in the number of points in any given area, which in turn is roughly proportional to its geographical size, precision of estimates of residential land falls off both for smaller countries and, a fortiori, for any given EU region. Nonetheless we can again derive some key features of the regional distribution, since these do not depend on the precision with which any given point is estimated. Figure 5 shows that the degree of variation in (broad) residential land per capita across EU regions, both within the EU as a whole, and within some individual countries, is very much greater than the variation between countries; but the chart also demonstrates the large differences between some countries even for the entire distribution across regions within that country.

Some features of these distributions are unsurprising. We would expect to see relatively low levels of residential land per capita in large cities, where many households live in apartment blocks that by their nature require little land per capita. And indeed the lowest levels of residential land per capita are typically found in cities in most EU countries. However, while this explanation is descriptively valid, in most cases it does not reflect any actual physical constraint, since, as shown in Figure 5, in only a handful of EU regions

[6]These are calculated using the binomial formula used in Gallego & Delince, 2010, equation 12.2. Given the evidence of spatial correlation that we present below, this typically understates measurement error. In a later draft we plan to provide confidence intervals that are robust to spatial correlation; however our preliminary investigations suggest that the degree of understatement of measurement error by the binomial formula is quite limited for most countries.

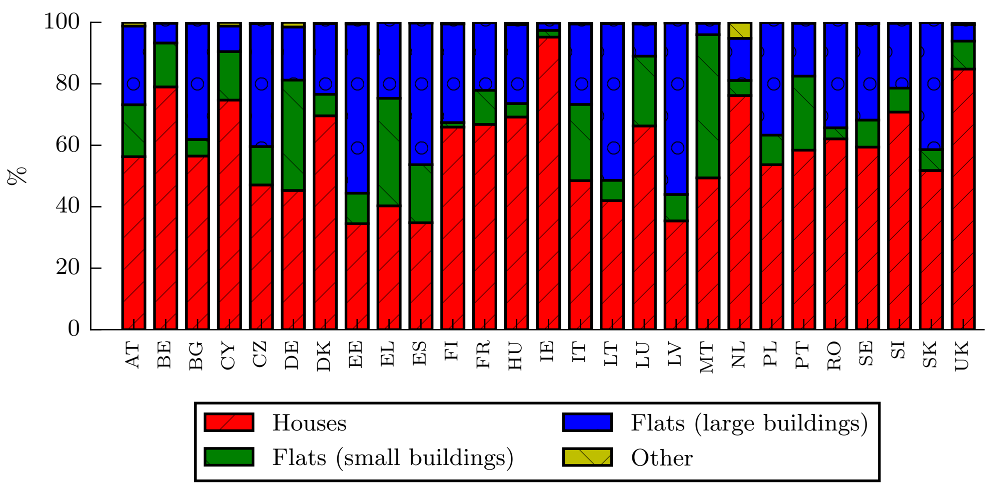
Figure 5: Residential Land Per Capita by Region: Empirical CDFs



are residential shares of total land sufficiently high that there is simply "no space" for more residential land. In the great majority of EU regions the amount of residential land per capita thus reflects a policy choice, not a physical constraint.

Figure 5 also shows some very striking differences between countries. Thus two relatively sparsely populated countries, France and Spain, which Table 1 shows have very similar amounts of total land per capita, (the reciprocal of population density) have extremely different regional distributions of land per capita. France has a regional distribution that almost spans the entire EU regional distribution; whereas the range of values across the regions of Spain is very much smaller, with clear-cut dominance by the French distribution. In contrast, the distributions for Germany and the UK (both with similar and distinctly higher rates of population density at a national level) cross, with a considerably larger range of variation in the UK, but around a very similar average value.

Figure 6: Composition of building stock

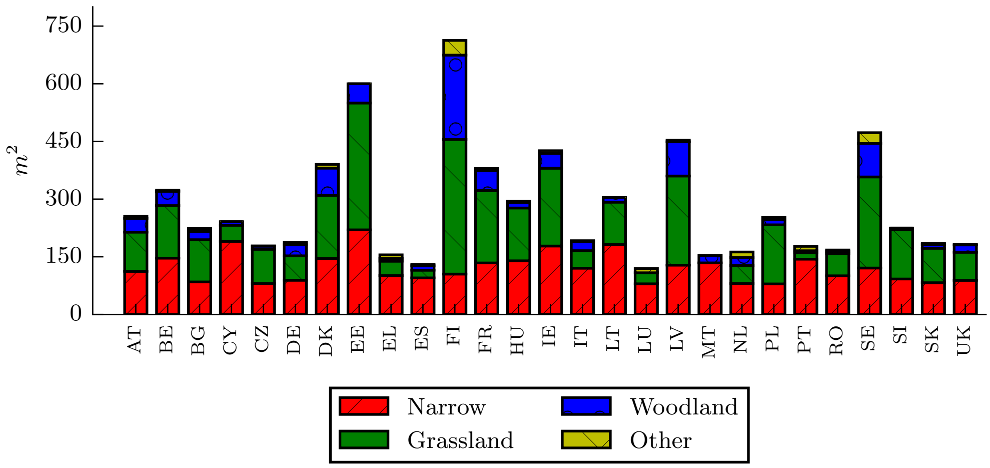


To investigate this difference further, figure 6 plots the composition of the residential building stocks across the European Union using a Eurostat dataset (see Appendix A. As can be seen there is a massive variation in the type of dwelling people live in, with, at one extreme, Ireland with 95.2% of the inhabitants living in single residence houses, whereas, for example, Spain has 65% of its residents living in apartment blocks. While the nature of residential buildings provides some insights into the differences in the distribution of residential land shown in Figure 6, it is, however, not of itself an explanation, since clearly the way in which residential land is utilised is also both endogenous to the price of land and to the restrictions on land supply and land use that planning policy imposes.

2.2.3 Composition of residential land

Figure 7 uses the land cover classification provided by LUCAS to show the breakdown of residential land into its main components.

Figure 7: Main components of residential land per capita in 27 EU countries



There are large differences in composition. At one extreme, in Malta amost all residential land consists of buildings; in contrast, in a number of Scandinavian and Baltic countries, a large proportion of residential land is made up of grass and woodland in gardens.

There is no clear-cut case for choosing between the broad measure of residential land, which includes green space, and the narrow definition, which only focuses on buildings. In practice both are clearly subject to regulation (most gardens would, potentially, have space for at least one, often two or more additional houses, but in most countries regulation would not actually permit this additional building). We thus proceed in parallel with both measures, noting, however, that the relatively low number of observations of the narrow measure means that, as shown in Figure 4, confidence intervals around any point estimates are distinctly wider, thus giving us more confidence in inferences that can be drawn from the broad measure.

2.2.4 Correlations

Figures 8 and 9 show the natureof the bivariate relationships between our two measures of residential land per capita and total land per capita (the reciprocal of population density).

Figure 8: Residential vs. Total land: National Data

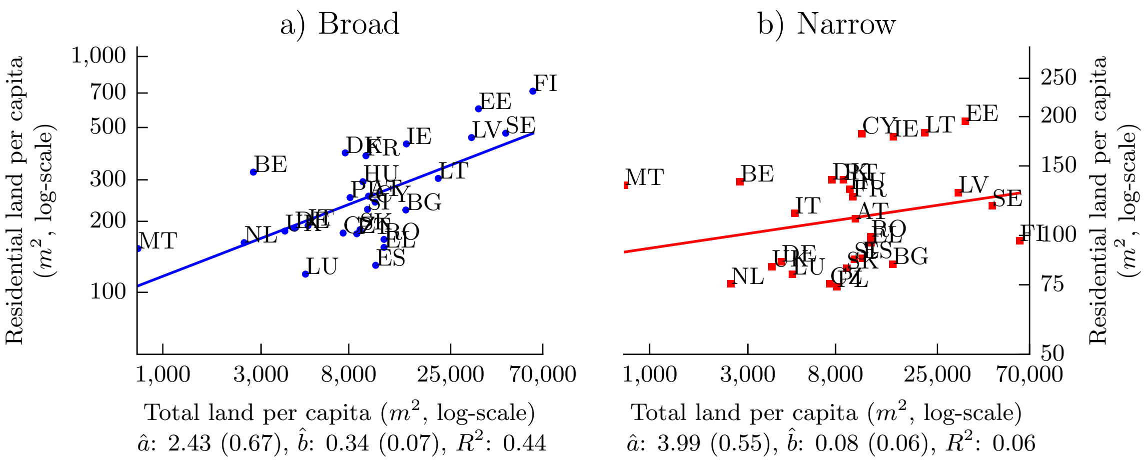
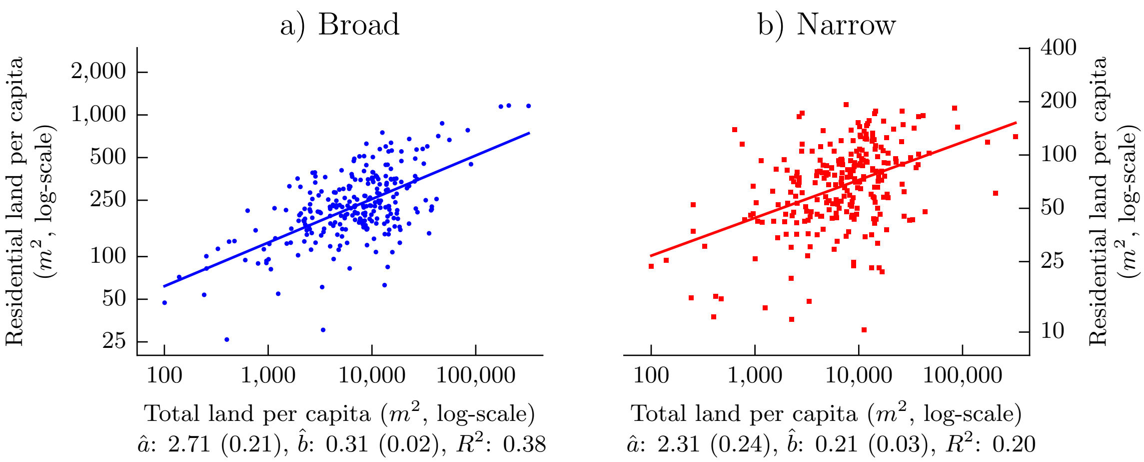


Figure 9: Residential vs. Total land: Regional Data

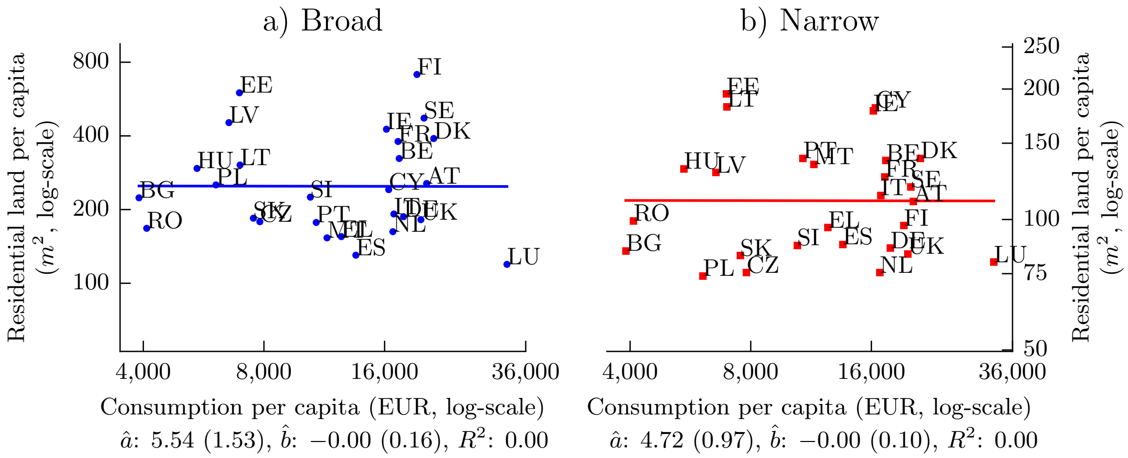


On the face of it, a positive association might seem unsurprising: more sparsely populated countries might be viewed as having *"more space"* for houses and gardens. However, a glance at Figure 2 should give pause for thought. For the overwhelming majority of countries residential shares are so small that this argument is distinctly less plausible. We therefore conclude that we need to look for other explanations in our modelling.

We now turn to the (lack of) relationship with consumption per capita, as in our Fact 2.

Figure 10 show a distinct lack of any apparent clear-cut bivariate relationship between residential land and consumption per capita across the cross section of 27 countries (our fact 2). This lack of relationship is itself very striking. As a key input to housing services, and indeed as a consumption good (*"space"*) in its own right, it appears unlikely on *a*

Figure 10: Residential land per capita vs. consumption per capita



priori grounds that consumption of the services of residential land is a borderline inferior good. Our model of Pigovian land supply set out below provides a rationale for such a weak correlation; our empirical results also show that the lack of a simple correlation across the cross-section of countries in our sample conceals some (albeit fairly weak) impact of national consumption at a regional level, once we factor in the impact of other determinants.

2.3 Estimates of housing expenditure per square metre and the opportunity cost of housing

We can also combine our land estimates derived from LUCAS with data from the national accounts (see Appendix A) to derive estimates of housing expenditure, in Euros, per square metre of (broad) residential land, for each of the 27 countries in our sample. Table 1 shows the resulting figures, which, as noted in our Fact 3, show an extremely wide range of variation. As Figure 11 illustrates, in contrast to the lack of correlation of residential land with consumption, the single strongest explanatory factor for variation in housing expenditure is cross-sectional variation in total consumption per capita.

One possible supply-side-based explanation for the variation in housing expenditure (or, equivalently, value-added from the housing sector) per square metre of residential land might in principle be if there are also significant cross-sectional differences in the opportunity of land in other potential uses. Figure 12 shows that, in the majority of EU countries, agriculture (and to a lesser extent forestry) is the dominant alternative use. We therefore construct two measures of opportunity cost by dividing the valued added in forestry or agriculture (both measures from the national accounts, Appendix A), by the respective areas used for these purposes. In figure 13 we plot the scatter plots of these measures of the opportunity cost of land against our measure of valued adding from housing services, both measured on a comparable basis, per square metre of land.

Figure 13 shows that there is indeed a positive correlation between our measure of value

Figure 11: Housing Expenditure per m^2 vs Consumption per Capita in 27 EU Countries

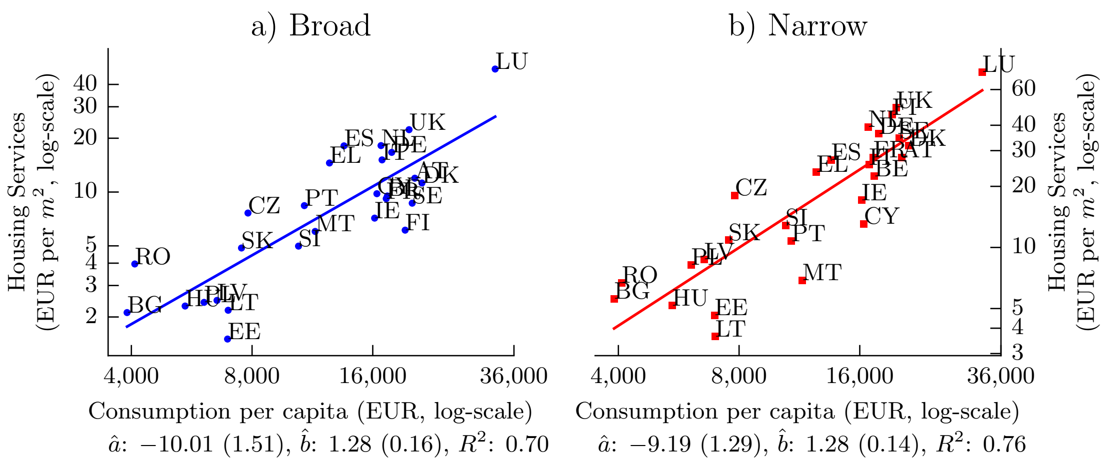
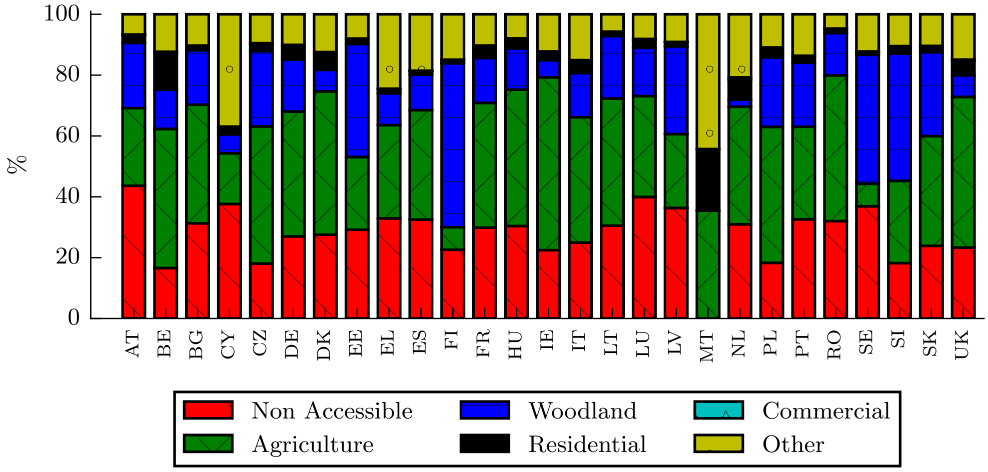


Figure 12: Composition of Total Land

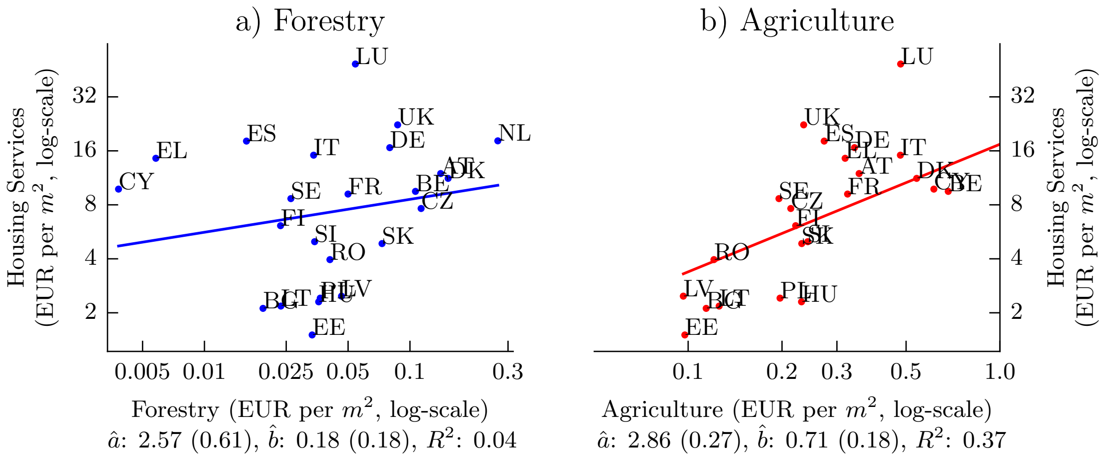


added from broad housing services and the opportunity cost of land used in agriculture (with a coefficient of 0.71 on the log values). However a closer look at the chart shows that this correlation cannot be viewed as representing any true economic relationship, since, as noted in Fact 4, the value added in the two sectors differ dramatically in magnitude. Where housing services has a value added per square metre of land between 1 and 50 euros, all our measures of value added from agricultural are below 1 euro. We speculate that the correlation might be driven by the historical importance of agricultural productivity of the land.

The correlation with the opportunity cost measure for forestry is not statistically significant and the differences in scales are even bigger.

Thus we conclude that there is no plausible explanation for the observed variation in housing expenditure per square metre in terms of the opportunity cost in competing uses of land.

Figure 13: Housing Expenditure per $m^2$and the Opportunity Cost of Land



3 A simple model of Pigovian land supply

3.1 Distribution of Households

A set $\mathcal{H}$ of households, $h = 1, \ldots, H$ live on a circle of points, $i = 1, \ldots, L \in \mathcal{L}$. Nature allocates households to an *"address"* (a point on the circle) via a random mapping from h to $i(h) \in \mathcal{I} \subset \mathcal{L}$.

We assume that the households are distributed clockwise around the circle by a Markov chain, with the transition matrix given by:

$$\mathbf{M} = \begin{bmatrix} \Phi & 1 - \Phi \\ 1 - \gamma & \gamma \end{bmatrix} \tag{3.1}$$

with the state vector

$$\mathbf{x}_i \equiv \begin{bmatrix} 1_{(i \in \mathcal{I})} \\ 1_{(i \notin \mathcal{I})} \end{bmatrix} \tag{3.2}$$

where the first element of x_i is equal to 1 if the point i is a household address and zero otherwise. This implies that the (notional) law of motion for the states around the circle (which in turn determines the distribution of addresses) is given by:

$$\mathbf{x}_{i+1} = \mathbf{M}\mathbf{x}_i + \mathbf{u}_{i+1}. \tag{3.3}$$

Let $l = L/H$, be per capita land, hence population density, $D = l^{-1}$. To ensure the correct steady state distribution we must have:

$$\Phi = D + (1 - D)\, s \tag{3.4}$$

$$\gamma = 1 - D\,(1 - s) \tag{3.5}$$

where $s \in (0,1)$ is a parameter determining spatial correlation, with $s = 0$ implying no spatial correlation. To simplify the algebra we assume that Nature repeats the allocation, until it reaches a finite sample with allocations that match the steady-state distribution.

3.2 Housing Technology

We assume a very simple housing technology that is intended to capture intensity of land use, which in turn determines the nature of the consumption externality. The housing technology at any address is given by the following recursion:

$$R_{i(h)+1} = \begin{cases} \frac{R_{h+1}}{i(h+1)-i(h)} & \forall i(h+1) - i(h) \leq R_{h+1} \\ 0 & \forall i(h+1) - i(h) > R_{h+1} \end{cases} \tag{3.6}$$

the technology imposes the restriction that households can only build houses to the left of their address, but also that the distance between two neighbouring addresses determines the nature of this housing. Thus, for example, if two addresses happen to be on adjacent points ($i\,(h+1) = i\,(h)+1$), then household $h+1$ is constrained to live in a house R points high by 1 point wide (a "high rise"). In contrast if the two households have addresses R or more points apart, household $h+1$ lives in a house 1 point high by R_h points wide (a "bungalow"). Intermediate values of the distance between the two addresses imply that household $h+1$ occupies a progressively taller building, and hence that there is progressively more intense use of housing in the neighbouring point. This in turn, as we show in the next section, determines the nature of the externality imposed on the neighbouring household.

3.3 Private Utility and Equilibrium

Each household is identical except in respect of the externality imposed by its neighbour. Household utility of household h (living at address $i(h)$) is given by

$$\max_{G_h, R_h} U_h = (1-\alpha)\ln(G_h - G^*) + \alpha \ln R_h + \ln(E - R_{i(h)+1}), \tag{3.7}$$

where G_h is nonresidential consumption, for which there may be a minimum subsistence level, $G^* > 0$. If so, R_h will be a superior good, with loglinearised income elasticity $\frac{C_h}{C_h - G^*} > 1$, where C_h is total consumption. We assume for simplicity that household h is indifferent between bungalows and high rises in terms of their own consumption: congestion of housing only matters to the extent that it may imply a greater intensity of housing in the adjacent point.[7]

The specification implies convex utility costs of housing in the neighbouring point,

[7] Allowing for this additional effect would complicate the algebra without changing the nature of the social planner's problem, as set out below, since it would simply accentuate the magnitude of the externality.

which can be rationalised in terms of claims on finite resources in the local *"environment"* (E). The stock of finite resources, E is treated as an endowment. We show below that the magnitude of E alone allows a sufficient parameterisation of the problem in terms of utility costs.

Given the additive separable nature of the problem the externality has no direct impact on any household's choices (it will, as we shall show, have an indirect effect via the price of housing). The budget constraint faced by household h is given by:

$$G_h + QR_h = C_h, \tag{3.8}$$

where C_h is total consumption of other goods and housing services and Q is the price of housing services.

Optimising Equation 3.7 with respect to the constraint given in Equation 3.8 implies the following private demand function:

$$R_h = \alpha \frac{\widehat{C}_h}{Q}, \tag{3.9}$$

where $\widehat{C}_h = C_h - G^*$ is the surplus consumption.

Substituting back into the utility function and combine with the optimality conditions, gives the indirect utility for household h:

$$V_h = V\left(\widehat{C}_h, Q, R_{i(h)+1}\right) = \ln \widehat{C}_h - \alpha \ln Q + \ln\left(E - R_{i(h)+1}\right) + \mathcal{C} \tag{3.10}$$

where $\mathcal{C} = \ln\left((1-\alpha)^{1-\alpha}\alpha^{\alpha}\right)$ is the constant term. Hence we can straightforwardly calculate the consumption equivalent loss of utility (as a share of total consumption) for those with housing in neighbouring points using

$$V\left((1-\kappa)\,\widehat{C}_h, Q, 0\right) = V\left(\widehat{C}_h, Q, R_{i(h)+1}\right), \tag{3.11}$$

which if we combine the indirect utility expressions yields:

$$\kappa(R_{i(h)+1}) = 1 - \frac{E - R_{i(h)+1}}{E} \tag{3.12}$$

If we investigate the limits of this expression, we have:

$$\lim_{E \to R_{i(h)+1}} \kappa(R_{i(h)+1}) = 1 \quad \text{and} \quad \lim_{E \to \infty} \kappa(R_{i(h)+1}) = 0,$$

the consumption equivalent impact of the externality is monotonically increasing in $R_{i(h)+1}$ (the intensity of residential consumption at the next point) and decreasing in E, the scale of the exogenous "environment", which thus provides a sufficient parameterisation for the magnitude of the externality.

3.4 Social Welfare Maximisation

The social planner acts behind a veil of ignorance and chooses aggregate land supply ($R = \mathbb{E}_{\mathcal{H}} R_h$) to maximise a social welfare function given by the expected utility of a household chosen randomly from the set of $\mathcal{H}$. To simplify the analysis we asssume that all households are identical, except for the externality.[8] The expected utility of the randomly chosen household that the social planner maximises is then given by:

$$\max_{R,G} W = \mathbb{E}_{\mathcal{H}} U_h = (1-\alpha)\mathbb{E}_{\mathcal{H}} \ln (G_h - G^*) + \alpha \mathbb{E}_{\mathcal{H}} \ln R + \mathbb{E}_{\mathcal{H}} \ln \left(E - R_{i(h)+1}\right) \tag{3.13}$$

Substituting in the probability of the externality binding (and using the property of symmetric households), we get that:

$$\begin{aligned} W = &(1-\alpha)\ln (G - G^*) + \alpha \ln R \\ &+ \phi \ln (E - R) + (1-\Phi)(1-\gamma)\sum_{j=0}^{R-2} \gamma^j \ln \left(E - \frac{R}{j+2}\right) \\ &+ \left(1 - \Phi - (1-\Phi)(1-\gamma)\sum_{j=0}^{R-2}\gamma^j\right) \ln E, \end{aligned}$$

where, clearly, we must assume $E - R > 0$ to ensure a solution.

Defining a function for the expected marginal disutility of the externality as:

$$\mathcal{F}(R; \Phi, \gamma, E) = \Phi \left(\frac{1}{E - R}\right) + (1-\Phi)(1-\gamma)\sum_{j=0}^{R-2}\gamma^j \left(\frac{j+2}{E - \frac{R}{j+2}}\right). \tag{3.14}$$

Assuming a unit marginal rate of substitution between housing and non-housing (and non-rival environment, except via housing), we have (using Equation 3.8 and 3.9)

$$\frac{W_R'}{W_G'} = \frac{\alpha}{1-\alpha}\frac{\widehat{G}}{R} - \frac{\widehat{G}}{1-\alpha}\mathcal{F}(R; \Phi, \gamma, E) = 1 = \mathrm{MRT} \tag{3.15}$$

implying

$$\frac{R}{\widehat{C}} = \frac{\alpha}{1 + \mathcal{F}(R; \Phi, \gamma, E) \times \widehat{C}}. \tag{3.16}$$

Since we have from the private optimisation that $\frac{QR}{\widehat{C}} = \alpha$, this implies that the price is given by:

$$Q = 1 + \mathcal{F}(R; \Phi, \gamma, E) \times \widehat{C}, \tag{3.17}$$

[8] Although it is relatively easy to extend the model to have different incomes for the household, it will not add anything to the intuition of the model, as the social planner has no tools to deal with inequality. It will therefore only (unnecessary) complicate the mathematics.

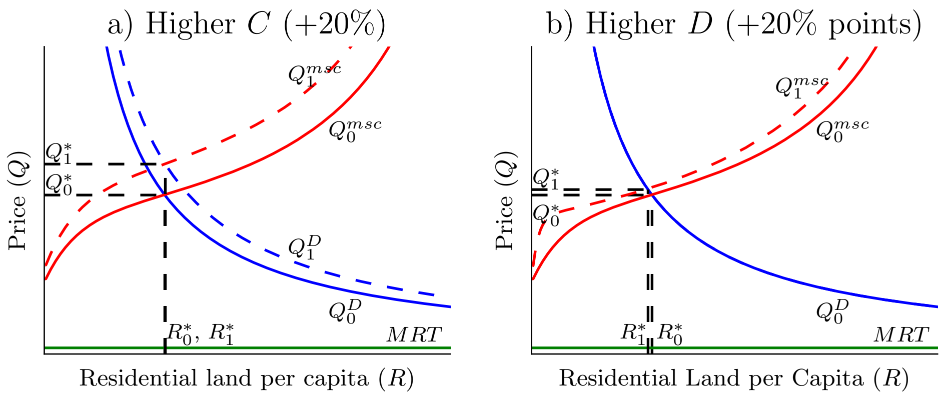
i.e., the expenditure share of housing in surplus consumption ($\widehat{C} = C - G^*$) is constant and equal to α (implying a rising share in total consumption) but Pigovian land supply implies that the share of real housing in surplus consumption (determined implicitly by (3.16) is decreasing in the externality term $\mathcal{F}(R; \Phi, \gamma, E)$, requiring an increase in the price of land, Q.

3.5 Geometry

The model is given by three simple elements above (1) the private demand function from Equation 3.9 (2) Marginal Rate of Transformation (set equal to one) and (3) The Pigovian marginal social cost, found from the planner problem in Equation 3.17.

Figure 14 plots comparative statics for higher consumption($\Delta C > 0$) and higher population density ($\Delta D > 0$):

Figure 14: Comparative Statics



In response to a shift in consumption both demand curve and marginal social cost functions shift, but the latter shifts by strictly less, so the new equilibrium implies increases in both price and quantity; higher population density shifts the MSC curve only, thus implies a higher land price but lower land supply per capita.

In contrast, in the absence of an externality the marginal social cost function would simply be an invariant horizontal line, so that a 1% shift in aggregate consumption would simply cause a $\left(\frac{C}{C-G^*}\right)$ % rise in residential land supply.

3.6 Log–Linearisation of the Model

In the Appendix we show that we can log-linearise the model as

$$\widetilde{r} = \lambda_c \widetilde{c} - \lambda_d \widetilde{d} \tag{3.18}$$

where $\widetilde{r}$, $\widetilde{c}$ and $\widetilde{d}$ are log deviations around an equilibrium where $\widehat{C} = \overline{\widehat{C}}$ and $D = \overline{D}$ are some mean values (eg across our cross-section) but land supply ignores the externality.

We then show that we can write

$$\lambda_c = \lambda_c\left(\overline{D}, s, \alpha, \kappa_\alpha\right) \quad \text{and} \quad \lambda_d = \lambda_d\left(\overline{D}, s, \alpha, \kappa_\alpha\right),$$

with λ_c,$\lambda_d \in [0,1]$, where κ_α is the consumption equivalent cost of the externality to household h if $R_{i(h)1} > 0$, in a a non-Pigovian equilibrium such that $Q = 1$ (which straightforwardly implies $\overline{R} = \alpha\widehat{\overline{C}}$). It is straightforward to show that we have

$$\lambda_c\left(\overline{D}, s, \alpha, 0\right) = \eta_c = \frac{\overline{C}}{\overline{C} - G^*} \quad \text{and} \quad \lambda_d\left(\overline{D}, s, \alpha, 0\right) = 0,$$

that is, in the absence of the externality population density would have no impact on land supply, and land supply would simply be determined by the income elasticity of residential land services.

In the next section we show that we can estimate the reduced form parameters λ_c and λ_d econometrically, as determinants of the probability that any given point in our dataset will be residential. In Section 5, we ask whether we can make sense of these reduced form estimates in terms of plausible structural parameters.

4 Estimation

We do not have directly measured data on residential land for our sample of 27 countries. Instead we have a sample of around a quarter of a million points from the LUCAS survey, which are classified both by land use and by land cover.

There are two alternative approaches to the use of the survey data. The first approach follows the methodology outlined in section 2, which translates shares of different point classications in the total number of points into area estimates. For relatively small shares (typically the case for residential land) these estimates are subject to error, but with a sampling error that is clearly reducing in the geographic size of the total area considered (since the number of points is at least approximately proportional to geographic area). For most, except a few small countries, Figure 4 showed that the resulting standard errors for residential land estimates are small, but non-negilible for some countries, especially after allowing for spatial correlation. More crucially, since geographical area is crucial to low standard errors, if we wish to make use of regional level data (for which we have at least some data), the resulting estimates of residential land shares at a regional level have to be viewed as having wide margins of error, and thus are unsuitable to direct econometric analysis.

The alternative approach, which we apply, is to use point level data. This allows us to use point–level regressors (in particular to capture spatial correlation) alongside regional and national regressors where available. It is straightforward to show that the loglinear approximation of Pigouvian supply can be manipulated to generate a probability that a given point is residential.

We can look at this in two stages. We first consider the impact of spatial correlation, which is generic to all our point–level estimation methods.

4.1 Capturing spatial correlation

We have a set of points, $\{S_{ijk}\},\ i = 1, \dots, 269, 328,$[9] which are zero (non-residential) or 1 (residential), in $j = 1, \dots, 261$ regions, and $k = 1, \dots, 27$ countries.

We can model the conditional probability that a given point is residential as

$$\mathbb{P}\left(S_{ijk} = 1 \mid \bar{r}_{ijk}\right) = \rho_{jk}\bar{r}_{ijk} + (1 - \rho_{jk})p_{jk}, \tag{4.1}$$

where $\bar{r}_{ijk}$ is the spatial autoregressive term, which can be interpreted straightforwardly as the local residential share in the neighbourhood of a given point. p_{jk} is the unconditional probability ($p_{jk} = \mathbb{P}(S_{ijk} = 1)$), which we will define more precisely later.

This implies that our model is a Spatial Autoregressive model (SAR) as in Vega and Elhorst (2013) and (Pesaran, 2015, pp.797–816) and the spatial autoregressive term can be writen as:

$$\bar{\mathbf{r}} = \mathbf{WS},$$

where $\mathbf{S}$ is the vector of individual points stacked (S_{ijk}) and $\mathbf{W}$ is the spatial lag matrix given by:

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & & \\ \vdots & & \ddots & \\ w_{n1} & & & w_{nn} \end{bmatrix}$$

with the restrictions that the individual elements in the matrix is given by:

$$w_{ij} = \begin{cases} 1/N & \forall i \neq j \text{ and j is part of nearest N points to i} \\ 0 & \text{Otherwise} \end{cases}$$

The unconditional probability, p_{jk}, in region j of country k are given by:

$$p_{jk} = \mathbb{E}\left(S_{ijk}\right) = \frac{R_{jk}}{L_{jk}} \tag{4.2}$$

where R_{jk} is residential land (not directly observable), and L_{jk} is total land in region j of country k (which we can take to be perfectly observable). Since R_{jk} is not directly observable, nor can p_{jk} be. We assume the expected value of the spatial correlation term, to be equal to the unconditional probabilities.[10]

[9] We lose a few numbers observations, from the points without any neighbouring points close it to

[10] This is only approximately true, as the adjacent points to other regions, will mbe slightly different, but only marginally change the results

To estimate Equation 4.1, we first model the unconditioanl probabilities as either constant or a national (regional) dummy, together with the effect of the spatial correlation and the country (and regional) heterogeniety:

$$S_{ijk} = \widehat{\rho}_{jk}\bar{r}_{ijk} + (1 - \widehat{\rho}_{jk})\widehat{p}_{jk} + \hat{u}_{ijk} \tag{4.3}$$

We estimate the spatial correlation term for both a homogenous spatial correlation term (at a national and regional level) as well as heterogenous and together with national (or regional dummies), we get the results reported in the table 1:

Table 2: The probability that any LUCAS survey point is classified as residential: Spatial Correlation and regional dummy variables

	Broad Residential						**Narrow Residential**					
	B1	B2	B3	B4	B5	B6	N1	N2	N3	N4	N5	N6
Conditional												
$\widehat{\rho}$	0.497	0.425		0.309			0.372	0.322		0.227		
	(0.01)	(0.007)		(0.007)			(0.01)	(0.007)		(0.008)		
$\bar{\widehat{\rho}}_k$			0.345		0.264				0.260		0.198	
			(0.179)		(0.173)				(0.207)		(0.204)	
$\bar{\widehat{\rho}}_{jk}$						0.155						0.061
						(0.377)						(0.419)
Unconditional												
$\widehat{p}$	0.035						0.015					
	(0.001)						(0.000)					
$\bar{\widehat{p}}_k$		0.043	0.043					0.022	0.022			
		(0.041)	(0.041)					(0.033)	(0.033)			
$\bar{\widehat{p}}_{jk}$				0.061	0.061	0.058				0.028	0.028	0.028
				(0.072)	(0.071)	(0.118)				(0.039)	(0.039)	(0.038)
R^2	0.022	0.025	0.026	0.031	0.032	0.035	0.010	0.012	0.014	0.017	0.018	0.023
SSR	8773	8744	8735	8691	8683	8655	3810	3801	3796	3782	3778	3762
AIC	−3.42	−3.43	−3.43	−3.43	−3.43	−3.43	−4.26	−4.26	−4.26	−4.26	−4.26	−4.27
BIC	−3.42	−3.42	−3.42	−3.41	−3.41	−3.39	−4.26	−4.26	−4.26	−4.24	−4.24	−4.22
Parameters	2	28	54	262	288	522	2	28	54	262	288	522

Observations: 269, 238. Standard errors in brackets. Estimation of Equation 4.3: $S_{ijk} = \widehat{\rho}_{jk}\bar{r}_{ijk} + (1-\widehat{\rho}_{jk})\widehat{p}_{jk} + \widehat{u}_{ijk}$, with the unconditional probabilities ($\widehat{p}_{jk}$) specified by either a constant term, national or regional dummy. S_{ijk} is equal to 1 if a given point is classified as residential (either broad or narrow), and 0 otherwise. Coefficients with a bar above are the mean group estimates (Pesaran et al., 1996) given by: $\bar{\widehat{\rho}}_k = \frac{1}{K}\sum_k \widehat{\rho}_k$ and $\bar{\widehat{\gamma}}_k = \frac{1}{K}\sum_k \widehat{\gamma}_k$.

As can be seen from the table, the best fit we can get of the data, requires quite a lot of parameters, to make the individual regions have both a heterogenous constant and a heterogenous interaction with the spatial lag term with 522 parameters to map the 261 regions. The Mean Group estimates of the dummy variables are calculated as outlined by (Pesaran et al., 1996, pp.155–156). As is to be expected, the more dummy variables in the model, and the greater the degree of heterogeneity allowed in the spatial correlation parameter, the lower the resulting mean estimate of the degree of spatial correlation. It seems likely however that such a highly parameterised model may understate the true degree of spatial correlation.[11]

4.2 Modeling the unconditional probability of residential land

In Section 4.1 we estimated regional and national residential probabilities directly, using dummy variables, to allow us to focus on the estimates of spatial correlation, but at the cost of a very large number of parameters.

We now attempt to estimate the determinants of these regional probabilities, using the reduced form of the model of Pigovian land supply set out in Section 3, which can be written as

$$\ln\left(\frac{R_{jk}}{H_{jk}}\right) = \beta + \lambda_c \ln\left(\frac{C_{jk}}{H_{jk}}\right) - \lambda_d \ln\left(\frac{H_{jk}}{L_{jk}}\right) \tag{4.4}$$

where H_{jk} is the population, C_{jk} is aggregate consumption, and R_{jk} is the unobservable total residential land (which the LUCAS dataset gives us an estimate of). This in turn can be tranformed straightforwardly to give the implied unconditional probability that any given point is residential:

$$\begin{aligned}\ln p_{jk} &= \ln\left(\frac{R_{jk}}{L_{jk}}\right)\\ &= \ln\left(\frac{R_{jk}}{H_{jk}}\right) + \ln\left(\frac{H_{jk}}{L_{jk}}\right)\\ &= \beta + \lambda_c \ln\left(\frac{C_{jk}}{H_{jk}}\right) + (1-\lambda_d)\ln\left(\frac{H_{jk}}{L_{jk}}\right)\end{aligned}$$

where now all terms on the right-hand-side apart from the error term are measurable, thus we have the implementable regression with regional/country regressors (a special case of our regression Equation 4.3):

$$p_{jk} = \exp\left(\beta + \lambda_c \ln\left(\frac{C_{jk}}{H_{jk}}\right) + (1-\lambda_d)\ln\left(\frac{H_{jk}}{L_{jk}}\right)\right) \tag{4.5}$$

Which we can estimate by nonlinear Least Squares. Note that this specification au-

[11] By analogy with temporal serial correlation, where a sufficiently large number of dummy variables will, in the limit, capture a large proportion of serial correlation, but typically only with the benefit of hindsight. In our case, while we treat regional dummies as exogenous, they are clearly not: the definition of regions postdates the emergence of population clusters.

tomatically imposes a lower bound of zero for the implied probability (since both terms are bounded below by zero).[12] There is no implied upper bound of unity, but in practice the implied estimate of the residential share is always so far below unity that this issue is immaterial. Thus while in principle we might need to follow Angrist and Pischke (2009) and estimate by restricted least squares, in practice there is no need to impose any restrictions in estimation. The specification in equation 4.5 does in principle allow for all regressors to be measured at a regional level. In practice at present we measure land, L_{jk} and population H_{jk} at a regional level and aggregate consumption $C_{jk} = C_k$ at a national level.

The error for the $\{ijk\}^{th}$ point is given by:

$$\varepsilon_{ijk} = S_{ijk} - \mathbb{P}(S_{ijk} = 1 \mid \bar{r}_{ijk})$$

We need to allow for heteroscedasticity in the errors, since the estimated conditional probability, and hence the variance of the point-wise errors, varies considerably across the sample.

Or if we write the model is matrix notation (and using the spatial lag matrix notation):

$$\mathbb{E}\left(\mathbf{S} \mid \mathbf{WS}\right) = \mathbf{WS}\rho + (1-\rho)\exp\left(\mathbf{X}\delta\right)$$

4.3 Estimation Results of Residential Points

We estimate the above model using non-linear least squares (Davidson and MacKinnon, 1993, Greene, 2012). To disentangle the effect of the model, we estimate Equation 4.5[13].

We estimate the models for first the national level, with constant spatial correlation, and then slowly add more heterogeneity in term of regional spatial correlation and as well as exclude various elements to see how much they contribute to (assuming there isn't a strong bivariate relationship with the explanatory variables). The results are summarised in table 3.

[12] Recall that our measure of spatial correlation is bounded between zero and one ($\bar{r}_{ijk} \in [0, 1]$)

[13] We have tested for an intercept term to the models, but all for all of the specifications, the intercept term was non-significant (in line with our theoretical model)

Table 3: The probability that any LUCAS survey point is classified as residential: Spatial Correlation and implicit model of regional land supply

	Broad Residential				**Narrow Residential**			
	B1	B2	B3	B4	N1	N2	N3	N4
Conditional								
$\widehat{\rho}$		0.380				0.271		
		(0.007)				(0.008)		
$\bar{\widehat{\rho}}_k$			0.331				0.253	
			(0.0)				(0.0)	
$\bar{\widehat{\rho}}_{jk}$				0.256				0.163
				(0.0)				(0.0)
Unconditional								
$\widehat{\beta}$	-9.33	-9.33	-9.33	-9.34	-10.15	-10.17	-10.19	-10.24
	(0.238)	(0.378)	(0.395)	(0.378)	(0.348)	(0.474)	(0.467)	(0.456)
$\widehat{\lambda}_c$	0.3600	0.3600	0.3599	0.3599	0.3580	0.3600	0.3598	0.3601
	(0.0255)	(0.0404)	(0.0423)	(0.0412)	(0.0372)	(0.0507)	(0.0499)	(0.0493)
$\widehat{\lambda}_d$	0.4404	0.4399	0.4398	0.4400	0.4326	0.4315	0.4296	0.4231
	(0.0073)	(0.0115)	(0.0116)	(0.0168)	(0.0105)	(0.0143)	(0.0144)	(0.0187)
R^2	0.017	0.028	0.029	0.032	0.010	0.015	0.016	0.020
SSR	8,817	8,723	8,713	8,685	3,810	3,792	3,787	3,773
AIC	-3.42	-3.43	-3.43	-3.43	-4.26	-4.26	-4.26	-4.27
BIC	-3.42	-3.43	-3.43	-3.41	-4.26	-4.26	-4.26	-4.24
Parameters	3	4	30	264	3	4	30	264

Observations: 269,238. Standard errors in brackets. Estimation of Equation 4.3: $S_{ijk} = \widehat{\rho}_{jk}\bar{r}_{ijk} + (1-\widehat{\rho}_{jk})\widehat{p}_{jk} + \widehat{u}_{ijk}$, with the uncondtional probabilities ($\widehat{p}_{jk}$) given by: $\exp\left(\widehat{\beta} + \widehat{\lambda}_c c_k + (1-\widehat{\lambda}_d)d_{jk}\right)$, as in Equation 4.5. S_{ijk} is equal to 1 if a given point is classified as residential, and 0 otherwise. $\bar{r}_{ijk}$ is the spatial correlation term, c_k is the log national consumption per capita, and d_{jk} is the log regional population density. Models B1 and N1 only provide estimates of the unconditional probabilities, ignoring spatial correlation. Models B2 and N2 impose homogenous spatial correlation terms, whereas models B3, N3, B4 and N4 allow spatial correlation terms to vary at a national and regional level. The table shows mean groups estimates (Pesaran et al., 1996) of heterogenous coefficient estimates.

As can be seen from the above table, all the regressors are highly significant. As is to be expected from trying to predict a low probability event (bearing in mind the overall low share of residential points in the total sample), the R^2s are quite low, but with a simple economic model such as model B2 or N2 in Table 3 we are close to replicating the results given by the equivalent models in Table 2 for B2 and N2, which has a much higer degree of parameterisation.

4.4 Estimation of regional and national aggregates

A a cross-check we can estimate the following equation at a regionally aggregated level:

$$\ln\left(\frac{R_{jk}}{H_{jk}}\right) = \beta + \lambda_c \ln\left(\frac{Y_{jk}}{H_{jk}}\right) - (\lambda_d - 1)\ln\left(\frac{H_{jk}}{L_{jk}}\right) + u_{jk} \tag{4.6}$$

Where $R_{jk}\%$ are estimates of residential land constructed as outlined in Section 2.1. In the absence of spatial correlation (or if the aggregated spatial correlation term was orthogonal to regional regressors) this would be equivalent to a regional aggregation of the point-wise equation, thus reducing the dataset to 261 observations. We can also aggregate further to a national level, setting $R_{jk} = R_k$, which reduces to just 27 observations.

Table 4: Direct estimation of reduced form model of land supply on regional and national data

	Broad Residential					Narrow Residential				
	National data		Regional data			National data		Regional data		
	BN1	BN2	BR1	BR2	BR3	NN1	NN2	NR1	NR2	NR3
Intercept (β)	1.053	2.433	-0.043	0.411	2.708	3.486	3.798	0.755	1.647	2.311
	(1.464)	(0.672)	(0.627)	(0.594)	(0.214)	(1.448)	(0.652)	(0.710)	(0.680)	(0.238)
Consumption (λ_c)										
National	0.130		0.270			0.029		0.152		
	(0.123)		(0.058)			(0.122)		(0.066)		
Regional				0.199					0.057	
				(0.048)					(0.055)	
Pop. Density (λ_l)										
National	0.356	0.337				0.065	0.060			
	(0.074)	(0.073)				(0.073)	(0.071)			
Regional			0.329	0.343	0.308			0.226	0.224	0.214
			(0.024)	(0.025)	(0.024)			(0.027)	(0.029)	(0.027)
R^2	0.463	0.441	0.431	0.422	0.384	0.028	0.026	0.215	0.202	0.199
SSR	3.12	3.25	48.87	49.67	52.91	3.05	3.06	60.96	61.98	62.24
AIC	-2.16	-2.12	-1.68	-1.66	-1.60	-2.18	-2.18	-1.44	-1.42	-1.42
BIC	-1.43	-1.63	-1.55	-1.53	-1.51	-1.45	-1.69	-1.31	-1.29	-1.33
Parameters	3	2	3	3	2	3	2	3	3	2
Observations	27	27	261	261	261	27	27	257	257	257

Standard errors in brackets. Estimate of Equation 4.6: $r_{jk} = \widehat{\beta} + \widehat{\lambda}_c c_{jk} + (1 - \widehat{\lambda}_d) d_{jk} + \widehat{u}_{jk}$. Where r_{jk} is our LUCAS estimate of the log regional (or national) residential land per capita, c_{jk} is the log consumption per capita (regional or national), and d_{jk} is the log population density. BN1 and NN1 shows the estimation using national aggregates where BN2 and NN2 is at the national level but excludes consumption.

Table 4 shows that, using only national data, the lack of a cross-sectional bivariate correlation between consumption per capita and residential land per capita noted in Section 2.2.4 is also evident in a multivariate framework: the implied estimate of λ_c is positive but insignificant. However, on regional data, once we condition on land per capita (the reciprocal of population density) at a regional level, the estimate of λ_c is larger, and significantly different from zero, whether we use national data on consumption or a regional proxy, regional GDP per capita.

5 Pigovian Land Supply? Attempting to make sense of the econometric reduced form.

In setting out the our theoretical model in Section 3 we derived a log-linear reduced form from a model of Pigovian land supply, which we have shown is at least qualitatively consistent with our econometric estimates of reduced form coefficients. However, on closer inspection it proves distinctly harder to get even an approximate *quantitative* match that allows us to rationalise what we observe with a truly Pigovian equilibrium.

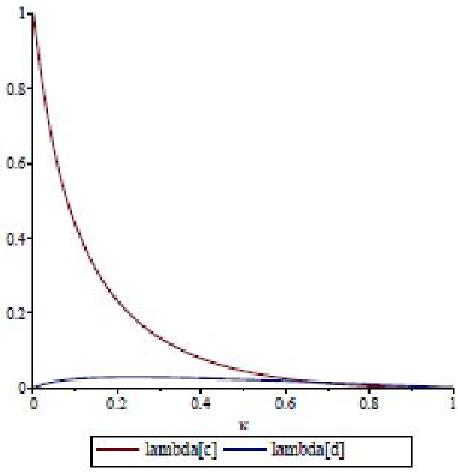
We have seen that the link between the theoretical model and the observable reduced form can be reduced to the impact of four key magnitudes: α, the weight of housing in total consumption; κ_α, the consumption equivalent value of the externality (evaluated in the absence of any attempt to mitigate it by Pigovian policies); D, population density; and s, a parameter determining spatial correlation.

Clearly a truly Pigovian policy would be required to trade off the cost of the externality, κ_α against the utility gain of higher housing, captured by α. We can get at least a ballpark value for α by looking at shares of housing expenditure in total consumption, as given in Table 1. These have a cross-sectional average around 16%. Under the maintained assumption of log utility (hence unit price elasticity) this magnitude will be invariant to the price of land; but must clearly be a significant over-estimate of α, since only a fraction of housing expenditure is on land *per se.* We start by setting $\alpha = 0.05$. Since κ_α is inherently un-knowable, we allow it to vary over its full range of $[0, 1]$.

Spatial correlation, for which we have shown there is nontrivial econometric evidence, matters in our model because, for any given value of population density, D, it increases Φ, the probability that the neighbouring point will be residential. In the absence of spatial correlation, this would reduce to the unconditional probability of a given point being residential: but we have seen that observed residential shares are so low that this would make it very unlikely that the externality will occur, thus reducing its impact on the social planner, who maximises the expected utility of a randomly chosen household. Thus higher spatial correlation accentuates the impact of the externality. Since, ceteris paribus, residential consumption would rise with aggregate consumption, this means that it will dampen the impact of higher consumption on land supply (ie, $\partial\lambda_c/\partial s < 0$).

However at the same time, higher spatial correlation will, ceteris paribus, make population density *less* important, simply by inspection of equation 3.2, which determines Φ, the Markov probability that point $i(h)+1$ will be residential, which can be re-written as $\Phi = s + D(1-s)$.

Figure 15: Calibrations: Realistic Calibration



Figures 15 and 16 illustrate the difficulties of reconciling the reduced form with plausible structural parameters.

In figure 15 we pick what appear, a priori, relatively plausible values of $\alpha = D = 0.05$, and $s = 0.5$, and then plot both reduced form parameters as a function of κ_{α}, the consumption equivalent value of the externality [14] We work in deliberately round values since our purpose is purely to illustrate the puzzle, rather than seek a precise match.

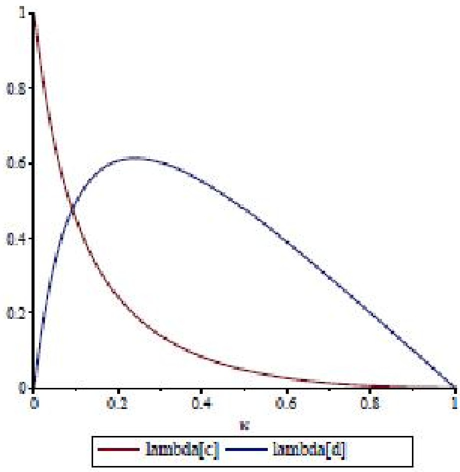
The two key features illustrated in this first calibration are, first, that the externality needs to be large (with an impact of the order of 10% to 20% of the consumption of affected households) to bring down the reduced form consumption elasticity to anything close to the observed value of around 0.35; but, second, more crucially, with this calibration, population density has only a very modest impact on land supply, whatever the consumption equivalent cost of the externality.

In Figure 16 we can get at least an approximate match for the two reduced form coefficients, at a relatively modest (but still high) value of κ_{α} but only by making two very significant changes.

The first is to pick an arbitrarily low value of α, which we set to one tenth of its value in Figure 15. We can crudely characterise this as a "Nimbyist" outcome, in which the social planner sets a very low weight on the utility gains from new housing. But this alone will not provide a match: we also need to assume (against the strong evidence in the data) that there is no spatial correlation (and thus set $s = 0$), which means that the conditional and unconditional probability of the externality are equalised. Without *both* of these features, we cannot even get close to matching both the reduced form coefficients.

[14] We assume that D is best captured by the residential share since this captures the probability of a given point having an address on it.

Figure 16: Calibrations: Unrealistic Calibration



We thus conclude that, at least on the basis of the simple model that we devise to analyse the spatial distribution of land, it is very hard to characterise land supply policies as truly Pigovian in nature. The very weak observed impact of higher consumption on land supply requires either that the externality be very costly, or that its costs are given excessive weight in the social planner's problem (i.e., Nimbyism). But the strength of the observed negative impact of population density is also a puzzle - despite the apparent intuition that less populous regions and countries will have "more space" for residential land. It is actually very difficult to rationalise the strength of this relationship, given that, as shown in Section 2, so little land is actually used for residential purposes.

6 Conclusions

In this paper we have analysed a new dataset of around $^1/_4$ million survey points, taken from the European Land Use and Cover Area-Frame Statistical Survey (LUCAS), covering 27 EU countries. This allows us both to derive national and regional estimates of residential land on a per capita basis, and model its spatial distribution and economic determinants, in light of a theoretical model in which restrictions on land supply attempt to mimic a Pigovian optimum.

Our econometric results show that supply of residential land per capita is affected rather weakly by higher consumption per capita, but somewhat more strongly (and negatively) by population density. While this is qualitatively in line with what would be predicted by a truly Pigovian land supply, we show that it is very hard to rationalise the magnitude of these effects with plausible structural parameters.

7 Bibliography

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics – An Empiricist's Companion.* Princeton University Press, 41 William Street, Princeton, New Jersey 08540, US, first edition edition, 2009. 4.2

The Bank for International Settlements. Residential property statistics. `http://www.bis.org/statistics/pp.htm`, 2015. Accessed: 05-11-2015. 1

Paul Cheshire and Stephen Sheppard. The welfare economics of land use planning. *Journal of Urban Economics*, 52:242–269, 2002. 1.2

William G. Cochran. *Sampling Techniques.* Addison-Wesley, Reading, Massachusetts, third edition edition, 1977. B

Russell Davidson and James G. MacKinnon. *Estimation and Inference in Econometrics.* Oxford University Press, 198 Madison Avenue, New York 10016-4314, US, first edition edition, 1993. 4.3

Eurostat. Land cover/use statistics. `http://www.ec.europa.eu/eurostat/web/lucas/overview/`, 2012. Accessed: 05-11-2015. 1

Eurostat. *LUCAS 2012 (Land Use / Cover Area Frame Survey) Technical Reference Document: C-3 Land use and Land Cover Classification (revised).* European Commission. Eurostat. Directorate E: Sectoral and Regional statistics Unit E-4, January 2013. 2.1, B

The Federal Reserve Bank of Dallas. International house price database. `http://www.dallasfed.org/institute/houseprice/`, 2015. Accessed: 05-11-2015. 1

William H. Greene. *Econometric Analysis.* Pearson Education Limited, Edinburgh Gate, Harlow, Essex CM20 2JE, England, seventh edition edition, 2012. 4.3

The International Monetary Fund. Global house price index. `http://www.imf.org/external/research/housing/`, 2015. Accessed: 05-11-2015. 1

David Miles. Population density, house prices and mortgage design. *Scottish Journal of Political Economy*, 59(5):444–466, November 2012. 1.2

OECD. Focus on house prices. `http://www.oecd.org/eco/outlook/focusonhouseprices.htm`, 2015. Accessed: 01-12-2015. 1

Hashem H. Pesaran, Ron P. Smith, and Kyung S. Im. Dynamic linear models for heterogenous panels. In L. Mátyás and P. Sevestra, editors, *The Economics of Panel Data: A Handbook of Theory with Applications.* Dordrect: Kluwer, 1996. 2, 4.1, 3

M. Hashem Pesaran. *Time Series and Panel Data Econometrics*. Oxford University Press, Great Clarendon Street, Oxford, OX2 6DP, United Kingdom, first edition edition, 2015. 4.1

Solmaria Hallect Vega and J. Paul Elhorst. On spatial econometrics models, spillover effect and w. Paper presented at 53rd ERSA conference, Palermo, August 2013. 4.1

Appendix A Datasources

Table 5: Datasources

Data	Eurostat code
LUCAS	
Consumption	*nama_10_co3_p3_1_data*
Housing Expenditures	*nama_10_co3_p3_1_data*
Agricultural Output	*nama_10_a64*
Forestry Output	*nama_10_a64*
Regional Gross Value Added	
Population	*demo_r_d2jan*
Regional Area	*demo_r_d3area*
National Area	*demo_r_d3area*

Appendix B The LUCAS Methodology

Eurostat's *"Land Use and Cover Area frame Statistical Survery"* (LUCAS) is a two phase sample survey. The first phase is an equally spaced systematic grid of 1,078,764 observations (in the 2012 sample) in 27 EU countries, separated by 2 km in the four cardinal directions. Each of the points in the first-stage sample are photo-interpreted and classified in terms of land cover,[15] as well as eligibility (based on accessibility) for the second stage of the survey.[16] Together these two classifications give the stratifications of the first stage sample. For the second a subset of 270,277 eligible points from the first-stage sample were visited in person by a surveyer. It is the dataset derived from this physical survey that we use in this paper.

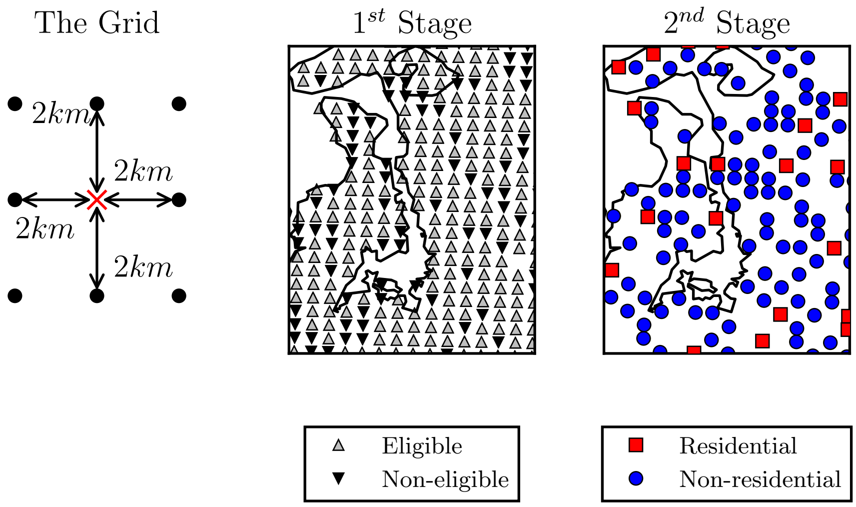
The individual points are then visited as in the figure below and interpretated by the photo taken:

The area estimates of land use and cover classification for the individual NUTS2 regions are then estimated following the methodology of Eurostat (following the methodology of Cochran (1977, chpt. 12)) as:

[15]Using the "CORINE" classification. 1: Arable, 2: Permanent Crop, 3: Grassland, 4: Woodland and shrubland, 5: Bareland, 6: Artificial, 7: Water and Wetland

[16]Only points that were both below 1,500m in altitude and accessible by road were included in the second stage.

Figure 17: The LUCAS dataset



$$L_{jk,c} = A_{jk} \sum_{h \in \psi} \frac{T_{jk,h}}{T_{jk}} \frac{t_{jk,hc}}{t_{jk}}, \tag{B.1}$$

where A_{jk} is the total area of region j in country k, $T_{ijk,h}$ is the number of points with stratum classification h and ψ is the set of stratum classifications (1 to 7) which is part of the second phase survey and T_{jk} is the total number of points in the first phase for region j and country k. t_{jk} is the total number of points for the region in the second phase survey and $t_{jk,hc}$ is the number of points with land use/cover classifcation c within stratum h.

For robustness we proceed with two different definitions of residential land, using land use and cover definitions from the LUCAS survey (Eurostat, 2013). The first, *"broad"* residential land, uses all survey points classified as residential by land use (LU *"U370"* in the dataset). The alternative "*narrow*" measure uses only the subset of residential points that are also classified by land cover as artificial structures (land cover A11 *"Buildings with one to three floors"* and A12 *"Buildings with more than three floors"*).

We augment the LUCAS dataset by using Eurostat data on population and gross value added at a regional level (from NACE) as well as Consumption, Actual Rent, Implied Rent and Maintenance at a national level from national accounts (NAMA).

Figure 18: Observation 44503638

$55°49'46.5''N$
$12°02'27.0''E$

Appendix C Log–Linearisation of the Model

In log terms, our equlibrium condition of the model in Equation 3.16 can be written as

$$\ln R - \ln \widehat{C} = \ln(\alpha) - \ln \underbrace{\left(1 + e^{\ln \mathcal{F}(R,D|s,E) + \ln \widehat{C}}\right)}_{\Gamma(R,D,\hat{C})}$$

Recall that the probabilities is a function of the population density, e.g.:

$$\Phi(D,s) = D + (1-D)s$$
$$\gamma(D,s) = 1 - D(1-s)$$

Which means that if we log-linearise the above equation, with respect to residential consumption (R), total consumption (C) and the population density (D) and defining $\tilde{r} = \frac{R-\bar{R}}{\bar{R}}$ for all the variables, we get the log-linearised model as:

$$\tilde{r} - \left(\frac{\bar{C}}{\bar{C} - G^*}\right)\tilde{c} = -\left(\frac{\Gamma_{\hat{C}} \times \bar{\hat{C}}}{\Gamma(R,D,\hat{C})}\frac{\bar{C}}{\bar{\hat{C}}}\right)\tilde{c} - \left(\frac{\Gamma_R \times \bar{R}}{\Gamma(R,D,\hat{C})}\right)\tilde{r} - \left(\frac{\Gamma_D \times \bar{D}}{\Gamma(R,D,\hat{C})}\right)\tilde{d}$$

Which we can write as:

$$\widetilde{r} = (1-\mu)\,\eta_c \widetilde{c} - \mu \eta_r \widetilde{r} - \mu \eta_d \widetilde{d}$$

where

$$\begin{aligned}
\mu &= \frac{\mathcal{F}(\cdot)\bar{\hat{C}}}{\Gamma(\cdot)} = \frac{\mathcal{F}(\cdot)\hat{C}}{1+\mathcal{F}(\cdot)\hat{C}} \\
\eta_c &= \frac{\bar{C}}{\bar{C}-G^*} \\
\eta_r &= \frac{\mathcal{F}_R(\cdot)\bar{R}}{\mathcal{F}(\cdot)} \\
\eta_d &= \frac{\mathcal{F}_D(\cdot)\bar{D}}{\mathcal{F}(\cdot)}
\end{aligned}$$

for some (reduced form equlibrium outcome)

$$\widetilde{r} = \lambda_c \widetilde{c} - \lambda_d \widetilde{d} \tag{C.1}$$

where

$$\lambda_c = \left(\frac{1-\mu}{1+\mu\eta_r}\right)\eta_c \quad \text{and} \quad \lambda_d = \left(\frac{\mu}{1+\mu\eta_r}\right)\eta_d \tag{C.2}$$

with $\lambda_c, \lambda_d \in [0,1]$. Thus far, in line with our empirics.

However, while we get a qualitative match, it is by no means so easy to get a quantitative match, particularly for the magnitude of the coefficient on population density.

To explore further, for simplicity linearise around an equilibrium where the key ratio $\overline{\frac{\widehat{C}}{E-R}}$ (determining μ, and hence the λ_i) is evaluated in an equilibrium where $\widehat{C} = \overline{\widehat{C}}$ is some mean value (eg across our cross-section) and land supply ignores the externality, such that $Q = 1$, hence $\overline{R} = \alpha\overline{C}$. Using (3.12) we have $\frac{R}{E-R} = \frac{\kappa}{1-\kappa}$, and hence

$$\mu = \frac{\Phi \frac{\overline{\widehat{C}}}{R}\overline{\frac{R}{E-R}}}{1+\frac{\overline{\widehat{C}}}{R}\overline{\frac{R}{E-R}}} = \frac{\frac{\Phi}{\alpha}\frac{\kappa_\alpha}{1-\kappa_\alpha}}{1+\frac{\Phi}{\alpha}\frac{\kappa_\alpha}{1-\kappa_\alpha}} = \mu\left(\Phi\left(D,s\right),\alpha,\kappa_\alpha\right)$$

where $\kappa_\alpha = \frac{E-\alpha\overline{\widehat{C}}}{E}$. Substituting into (C.2) we have, using $\eta_r = \frac{\kappa_\alpha}{1-\kappa_\alpha}$,

$$\begin{aligned}
\lambda_c\left(\Phi,\alpha,0\right) = &\quad \lambda_c\left(0,\alpha,\kappa_\alpha\right) = 1 \\
\lambda_d\left(\Phi,\alpha,0\right) = &\quad \lambda_d\left(0,\alpha,\kappa_\alpha\right) = 0
\end{aligned}$$

as expected. There is however a problem in finding parameter combinations that map to values similar to what we find in the data, for plausible values of κ_α. A high value of s, and hence Φ, lowers λ_c by enough to match our estimates but implies extremely low values for λ_d.