NHH

INSTITUTT FOR SAMFUNNSØKONOMI

DEPARTMENT OF ECONOMICS

SAM 30 2015

ISSN: 0804-6824

December 2015

Discussion paper

Revisiting the Boston data set (Harrison and Rubinfeld, 1978): a case study in the challenges of system articulation

BY

Roger Bivand

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS

Revisiting the Boston data set (Harrison and Rubinfeld, 1978): a case study in the challenges of system articulation

Roger Bivand*

Abstract

In the extended topical sphere of Regional Science, more scholars are addressing empirical questions using spatial and spatio-temoral data. An emerging challenge is to alert "new arrivals" to existing bodies of knowledge that can inform the ways in which they structure their work. It is a particular matter of opportunity and concern that most of the data used is secondary. This contribution is a brief review of questions of system articulation and support, illuminated retrospectively by a deconstruction of the Harrison and Rubinfeld (1978) Boston data set and hedonic house value analysis used to elicit willingness to pay for clean air.

1 System articulation in Regional Science

In *Complex spatial systems*, Wilson (2000) distinguishes three dimensions which interact in urban and regional analysis: system articulation, theory, and method (see also Wilson, 2002, 2012). System articulation is in turn made up of three sub-dimensions, entitation, levels of resolution (sectoral, spatial, temporal), and spatial representation. He argues that all too little attention is paid in analysis to careful planning of the main dimensions, with system articulation typically treated in the least satisfactory way. His second and third chapters provide a succinct and enlightening review of why system articulation matters — pointing back to Paelinck and Nijkamp (1975). Perhaps a real challenge of as yet unknown size is the use of spatial data in regression discontinuity designs, natural experiments, and similar approaches. Hidano et al. (2015), Keele and Titiunik (2015) and Egger and Lassmann (2015) are among studies that probably constitute an important

*Department of Economics, Norwegian School of Economics, Helleveien 30, N-5045 Bergen, Norway; Institute of Socio-Economic Geography and Spatial Management, Adam Mickiewicz University, Dzięgielowa 27, 61-131 Poznań, Poland

innovative wave that is of significance for Regional Science, but which does not appear to be informed by legacy insights.

For example, Haggett et al. (1977) drew attention in *Locational Analysis in Human Geography* to the fact that in human geography — as in regional science — the vast majority of observations are taken from secondary, archival sources, very often of a non-areal nature, with consequences: "(i) locational analysis is using data which have been collected primarily for non-geographical purposes, and these data are usually oblique in varying degrees to the direct research needs ...; ... (iii) data are released in 'bundles' (i.e. for administrative areas) which are inconvenient and anachronistic, and pose extremely acute problems in mapping and interpretation."

Both the spatial level of resolution and the mode of spatial representation are involved in spatial scale (see also Dray et al., 2012). Scale is intimately connected to the pattern/process matching that is central to analysis, because certain causal effects may be present only at particular scales. If the spatial representation (driven by available data) misses this scale, real causal effects will be obscured. Usually, it is the micro-scale variation of a scale smaller than the observation units (or the distances between them) that is omitted — however, omitted large scale trends may be seen as autocorrelation rather than misspecification. Bivand (2008) provides an extended discussion of some of the consequences of passing perhaps too rapidly through system articulation for statistical inference from spatial data.

Support is the term used to describe the link between the observation and the spatial entity used for observation. Often the entities are not chosen to suit the data generation processes, but are those "to hand". Gotway and Young (2002) pointed to serious statistical questions involved in spatial data analysis. The underlying issue is the change of support problem, where the measurement may not capture the phenomenon under analysis well. This is endemic when integrating secondary data sources, as almost all "measurements" involve error processes, be they spatially structured or otherwise. For prediction (and particularly for putting confidence intervals on predictions), the uncertainty should be carried through.

Gotway and Young (2002, p. 634) start by defining *support* as: "the size or volume associated with each data value", but it "also includes the geometrical size, shape, and spatial orientation of regions associated with the measurements." They continue: "Changing the support of a variable ... creates a new variable (which) is related to the original, but has different statistical and spatial properties." These challenges are collectively known as the change of support problem.

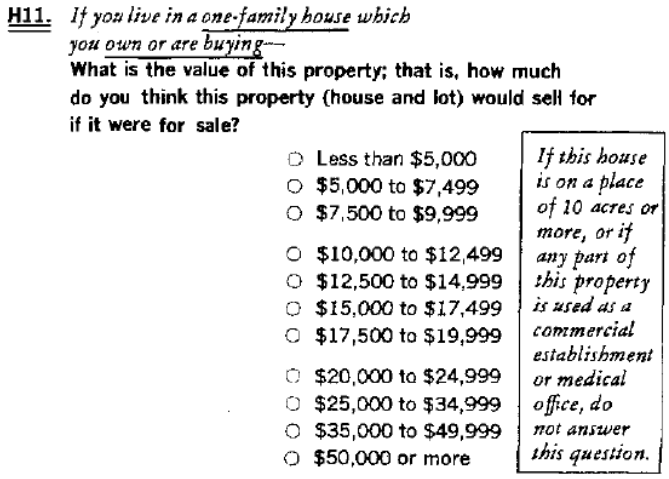
Gelfand (2010) shows where misaligned spatial data, the modifiable areal unit problem, and the change of support problem may take us (see also Haining, 2010). Wakefield and Lyons (2010) give a survey of the ecological fallacy in connection with spatial aggregation; the point of concern is the extension of aggregated inference to individuals within the aggregates. They motivate their survey by looking at county asthma disease counts and PM2.5

air pollution; of course, within county variability in the included variable is challenging, and inferring to the individual is hard. Haining (2010) also stresses that making statistical inferences about individuals based on aggregate data is flawed.

Scholars in "broad" Regional Science are increasingly using spatial and spatio-temporal data. Use in natural experiments, regression discontinuity designs and similar approaches is increasing rapidly, often using for example distance to borders or treatments to detect effects. These studies often appear to face the same challenges of relying on data collected for other purposes that have always troubled spatial analysts. We'll now turn to a specific example with well-known legacy data to explore some of the issues.

2 System articulation in Harrison and Rubinfeld (1978)

In order to approach willingness to pay for cleaner air, Harrison and Rubinfeld (1978) used a hedonic regression including air pollution levels with house values as the dependent variable. They use a data set for most of the Boston SMSA in 1970 at the census tract level of aggregation. The data were made available in text form by Belsley et al. (1980, pp. 229–261) in the form in which they appear to have been analysed. Pace and Gilley (1997) and Gilley and Pace (1996) found that there were errors in Belsley et al. (1980) and the statlib data file, and that the house value data were censored.


H11. *If you live in a one-family house which you own or are buying—*
What is the value of this property; that is, how much do you think this property (house and lot) would sell for if it were for sale?

- ○ Less than $5,000
- ○ $5,000 to $7,499
- ○ $7,500 to $9,999
- ○ $10,000 to $12,499
- ○ $12,500 to $14,999
- ○ $15,000 to $17,499
- ○ $17,500 to $19,999
- ○ $20,000 to $24,999
- ○ $25,000 to $34,999
- ○ $35,000 to $49,999
- ○ $50,000 or more

If this house is on a place of 10 acres or more, or if any part of this property is used as a commercial establishment or medical office, do not answer this question.

Figure 1: Scanned copy of 1970 Census question H11.

Harrison and Rubinfeld (1978) used median values in 1970 USD for 506 census tracts in the Boston SMSA for one-family houses. Here the values are not at the micro-level, but medians from census tracts from the 1970 US Census (for "owner-occupied one-family

housing"). The relevant question is H11,[1] reproduced in Figure 1. Consequently, the house value data have census tract support, and are median values calculated from group counts.

The published census tract tabulations shows the link between question H11 and the statlib-based data (after correction).[2] The tabulated median values can be reconstructed from the tallies shown in the Census results fairly accurately using the `weightedMedian` function in the **matrixStats** in R, using linearinterpolation, and midpoint values of USD 3,500 and USD 60,000 for the left- and right-censored intervals.[3]

Two tracts are entered as having a median house value below USD 5,000, and 15 have median values over USD 50,000, as was pointed out by Gilley and Pace (1996). One tract has a median of exactly USD 50,000, with 31 houses below the right-censored boundary, and 31 above. Having access to the Census value group counts by tract means that alternative aggregations of house value — the dependent variable in the analysis — may be constructed using the underlying data.

A further point, made by Harrison and Rubinfeld (1978), is that the number of housing units by tract varies greatly. They tried using weighted regression, using the logarithm of the counts of one family houses by tract, and noted some change in coefficient values. This step was taken to attempt to check the results for robustness to heteroskedasticity.

The data on 1970 air pollution concentrations were obtained from a meteorological model (TASSIM) of the Boston air shed. A mean air pollution concentration surface was generated by simulation of the Boston SMSA, then calibrated to values from monitoring stations. The calibrated model results were obtained for 122 zones, and assigned proportionally to the 506 census tracts. Many of the smaller tracts belong to the same TASSIM zones; this is a clear case of change of support, with very different spatial statistical properties under the two different entitation schemes. Harrison and Rubinfeld (1978, p. 86, footnote 14) do comment that "... the true correlation between NOX and PART is somewhat overstated because the TASSIM model generates data for 122 zones, not 506 census tracts. Translating zonal data into census tracts tends to overstate the correlation because relatively more census tracts are located in center city zones in which PART and NOX levels tend to be most highly correlated."

Since the data set from Belsley et al. (1980) does not include PART (particulate matter), nor the identifiers of the TASSIM zones underlying the assignation of copied values to census tracts, it is not possible from the data as they stand to retreive the zones with

[1] "If you live in a one family house which you own or are buying — What is the value of this property? That is, how much do you think this property (house and lot) would sell for if it were for sale?" `https://www.census.gov/history/pdf/1970_questionnaire.pdf`, p. 3.

[2] Census of Population and Housing-1970-Census Tracts: Part 3 Binghamton, N.Y.-PA.SMSA-Cedar Rapids, Iowa SMSA, `http://www2.census.gov/prod2/decennial/documents/39204513p3_TOC.pdf`, Sections 5 and 6, PHC(1)-29, table H1.

[3] `http://www2.census.gov/prod2/decennial/documents/00116813p1.zip`, Chapter 5, Census Users' Dictionary, p. 118.

full certainty. We can, however, aggregate contiguous census tracts with identical values of NOX, giving 96 approximated TASSIM zones, for which we can aggregate grouped house value counts, and calculate median values using the same procedure as that used at the census tract level of resolution.

Figure 2 uses colour fill to visualise the possible impact of using weighted regression, with the upper left panel showing — with the chosen class intervals for NOX — the actual relationship between house value and NOX, with the areas of the rectangles proportional to the counts of housing units. In the remaining panels, the areas are proportional to the counts of spatial entities with median house values falling into the input house value classes. NOX fill colours from salmon and darker represent higher levels of air pollution. Tracts and TASSIM zones with such higher levels have typically many fewer housing units.

Figure 3 shows clearly that the study of the relationship between NOX and house value will be impacted by "copying out" NOX values to census tracts, as noted by Harrison and Rubinfeld (1978) in the footnote mentioned above. Even if we were to use more class intervals in these choropleth maps, the visual impression would be the same, because the underlying data have the support approximated by the 96 TASSIM zones, not the 506 census tracts.

Figure 4 shows changes in the visual impression given by aggregating the H11 counts to approximate TASSIM zones before calculating interpolated weighted median values. It also shows the censored census tracts for which we have no reliable median values, as the values taken depend on the assumed under/over interval midpoints. Once we have aggregated to TASSIM zones, there are no longer any out-of-bounds median values.

Besides NOX, the other census covariates included in the hedonic regression to account for median house values are the average number of rooms per house (RM), the proportion of houses older than 1940 (AGE), the proportion low-status inhabitants in tract (LSTAT), and the Black proportion of population in tract (BB) — originally expressed as $(B-0.63)^2$, a broken-stick relationship, but here taken as a percentage. The crime rate is said to be taken from FBI data by town (CRIM), but on inspection of the data, it varies by tract. The distance from tract to employment centres (DIS) is derived from other sources, as is the dummy variable for tracts bordering Charles River (CHAS).

Other covariates are defined by town, with some also being fixed for all towns in Boston. The variables are the proportion of residential lots zoned over 25000 sq. ft (ZN), the proportion of nonretail business acres (INDUS), accessibility to radial highways (RAD), full-value property-tax rate per USD 10,000 (TAX), and pupil-teacher ratio by town school district (PTRATIO). These covariates are also "copied out" to tracts within towns, but do not coincide with the approximate TASSIM zones. Table 1 shows the descriptive statistics for the variables used in the 489 observation census tract data set omitting tracts with censored median house values.

These covariates were aggregated to approximate TASSIM zones using weighted averages, where the weights are the tract population counts. The Charles River dummy

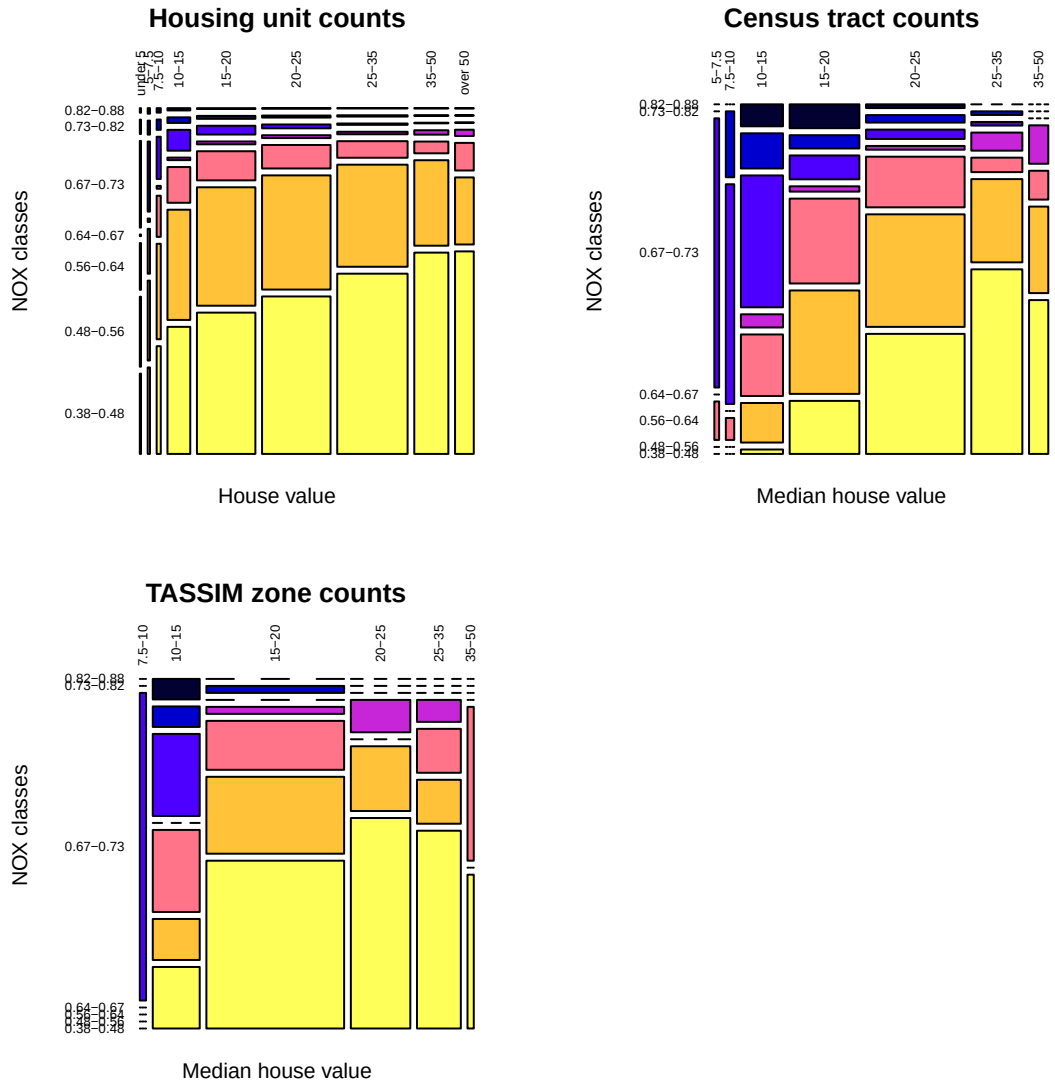


Figure 2: Mosaic plots by H11 classes (under 5, 5–7.5, 7.5–10, 10–15, 15–20, 20–25, 25–35, 35–50, over 50 thousand 1970 USD) and seven natural breaks NOX classes (also used for fill colours) — upper left panel: counts of housing units; upper right panel: counts of census tract median values; lower left panel: counts of TASSIM zone median values.

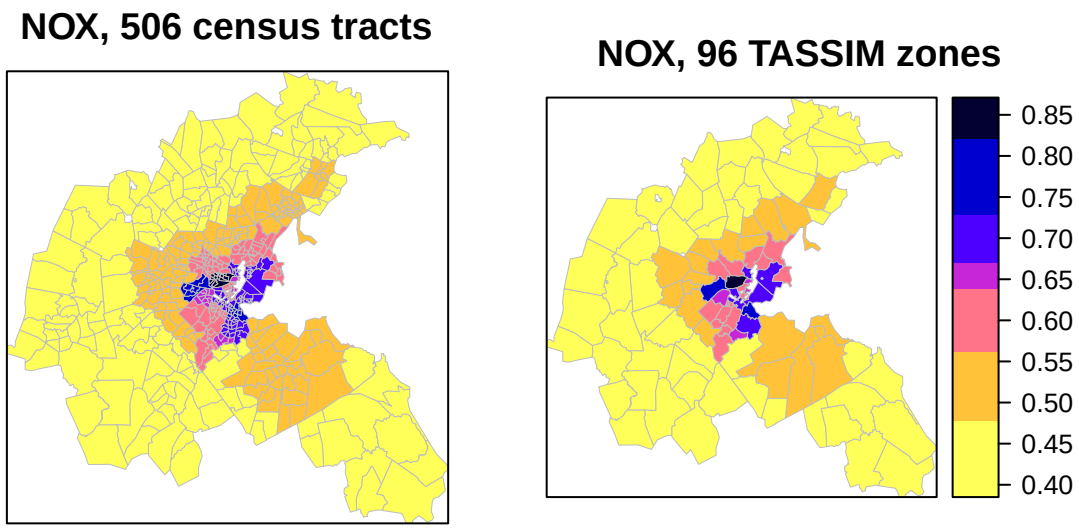


Figure 3: NOX values shown using the same class intervals and colour fill as Figure 2 for two entitations: 506 census tracts and 96 approximate TASSIM zones.

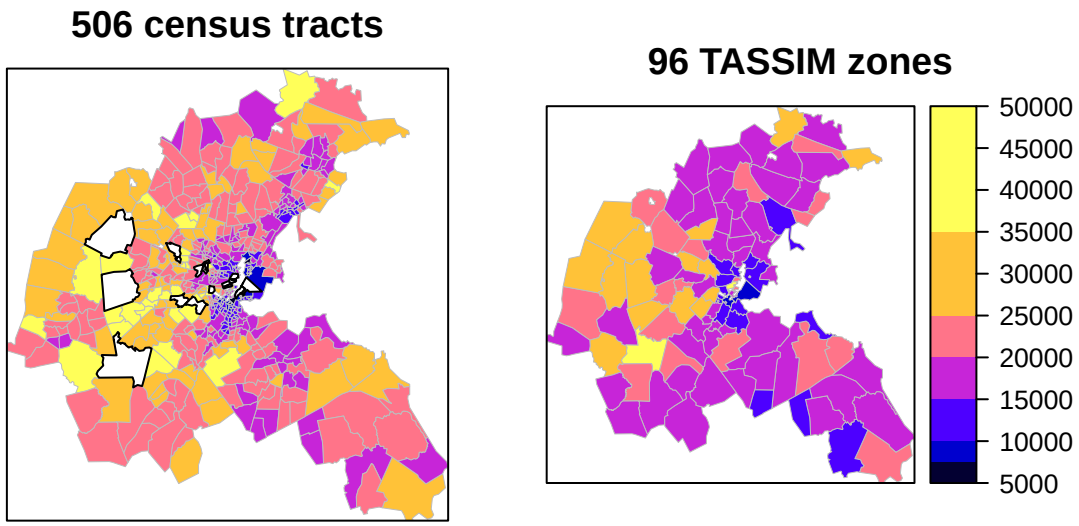


Figure 4: Median house values (USD 1970): 506 census tracts and 96 approximate TASSIM zones; in the left panel, the censored tracts are shown with black boundaries.

Table 1: Descriptives for variables included in the hedonic house value regression; 489 non-censored Boston census tracts

	Min.	Median	Mean	Max.
units	5.00	526.00	690.10	3031.00
log(median)	8.63	9.95	9.92	10.82
CRIM	0.01	0.25	3.45	88.98
ZN	0.00	0.00	11.13	100.00
INDUS	0.74	9.69	11.10	27.74
CHAS1	0.00	0.00	0.06	1.00
I((NOX * 10)^2)	14.82	28.94	32.04	75.86
I(RM^2)	12.68	38.35	39.46	77.09
AGE	2.90	76.70	68.21	100.00
log(DIS)	0.13	1.19	1.20	2.50
log(RAD)	0.00	1.61	1.86	3.18
TAX	187.00	330.00	407.50	711.00
PTRATIO	12.60	19.10	18.52	22.00
I(BB/100)	0.00	0.01	0.06	0.96
log(I(LSTAT/100))	-3.92	-2.15	-2.21	-0.97

was aggregated by taking the maximum value of any tract included in the approximate TASSIM zone. It would be possible to punch more census data for some of the covariates, but not all the variables used are present in the census tables available online. Table 2 shows the descriptive statistics for the variables used in the 96 observation approximate TASSIM zone data set.

As we can see from Figures 3 and 4, some tracts in the study area considered by Harrison and Rubinfeld (1978) have either no one-family houses, or too few for tabulations to be published. It is unclear what should be done about these tracts, which were dropped from the analysis entirely, as they did not contain such housing (or results are suppressed), but were present downtown where air pollution was worst.

3 Consequences of changes in system articulation for inference

Pace and Gilley (1997), drawing on earlier work, felt that it should be worthwhile to check whether the original model was not spatially misspecified. They considered that the use of spatial aggregate units as observations might involve spillovers of some kind, chiefly in the housing values used — neighbouring census tracts may have similar values for a number of reasons. Had the included explanatory variables accounted for the similarities

Table 2: Descriptives for variables included in the hedonic house value regression; 96 approximate TASSIM zones

	Min.	Median	Mean	Max.
units	25.00	2926.00	3588.00	12410.00
log(median)	9.12	9.82	9.83	10.56
CRIM	0.01	0.08	1.96	18.13
ZN	0.00	0.00	25.89	100.00
INDUS	0.46	6.01	8.55	27.74
CHAS	0.00	0.00	0.15	1.00
I((NOX * 10)^2)	14.82	21.76	26.68	75.86
I(RM^2)	25.93	39.60	41.82	62.77
AGE	8.97	51.76	56.02	100.00
log(DIS)	0.14	1.54	1.42	2.50
log(RAD)	0.00	1.61	1.62	3.18
TAX	187.00	307.00	376.20	711.00
PTRATIO	12.60	18.25	17.93	22.00
I(BB/100)	0.00	0.01	0.04	0.78
log(I(LSTAT/100))	-3.52	-2.46	-2.42	-1.43

between neighbours, there might not have been any reason to go further, but the residuals turn out to be spatially highly patterned. So now we will turn to spatial econometrics methods to try to unravel the question of the "real" link between house values and NOX. We will be using row-standardised contiguity neighbours derived from the map of census tracts, omitting the censored tracts which leads to one tract having no neighbours, and from the map of merged census tracts constituting approximate TASSIM zones.

There are two reasons for choosing not to include the spatially lagged median house value dependent variable in the models considered. The first is based on LeSage (2014), and the probability that the aggregate nature of the dependent variable makes it seem more reasonable to consider local spillover specifications. The "copying out" of covariates across multiple tracts from different entitation schemes can arguably be seen as local rather than global spillovers.

The second reason is pragmatic, that weighted spatial regression code in the **spdep** package in R is so far only implemented for the spatial error (SEM) and by extension the spatial Durbin error model (SDEM). This code (`spautolm`) was originally written to replicate results in Waller and Gotway (2004, p. 378), but with unit weights gives the same results as the unweighted implementation (`errorsarlm`).

We will now present briefly the models used. Assuming that the variance of the disturbance term is constant, we start from the standard linear regression model:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim N(0, \Omega), \Omega = \sigma^2\mathbf{I}$$

where $\mathbf{y}$ is an $(N \times 1)$ vector of observations on a dependent variable taken at each of N locations, $\mathbf{X}$ is an $(N \times k)$ matrix of exogenous variables, β is an $(k \times 1)$ vector of parameters, and ε is an $(N \times 1)$ vector of disturbances.

The spatial error model (SEM) may be written as (Ord, 1975):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \qquad \mathbf{u} = \rho_{\text{Err}}\mathbf{W}\mathbf{u} + \varepsilon,$$

where $\mathbf{y}$ is an $(N \times 1)$ vector of observations on a dependent variable taken at each of N locations, $\mathbf{X}$ is an $(N \times k)$ matrix of exogenous variables, β is an $(k \times 1)$ vector of parameters, ε is an $(N \times 1)$ vector of disturbances and ρ_{Err} is a scalar spatial error parameter,

and $\mathbf{u}$ is a spatially autocorrelated disturbance vector with constant variance and covariance terms specified by a fixed $(N \times N)$ spatial weights matrix $\mathbf{W}$ and a single coefficient ρ_{Err}:

$$\mathbf{u} \sim N(0, (\mathbf{I} - \rho_{\text{Err}}\mathbf{W})^{-1}\Omega(\mathbf{I} - \rho_{\text{Err}}\mathbf{W}^{\top})^{-1}), \Omega = \sigma^2\mathbf{I}$$

In both cases, the Durbin forms (SLX, SDEM) are defined by augmenting the matrix of independent variables $\mathbf{X}$ with its spatial lag $\mathbf{WX}$, now using $[\mathbf{X}, \mathbf{WX}]$ instead of just $\mathbf{X}$. Also in both cases, the weighted versions are formed by altering $\Omega = \sigma^2\mathbf{I}$ by replacing the identity matrix by a diagonal matrix of the inverses of known case weights reflecting the relative "size" of the observations (Waller and Gotway, 2004). The variance term σ^2 is still estimated, but with the assumption of uniform variance replaced by variance proportional to the inverse of known case weights.

Figures 5 and 6 show the AIC values for fitted models for two sets of entities, with 489 census tracts and 96 approximate TASSIM zones, and approximately the same data. The best model fit is indicated by the lowest AIC value among comparable models. The fitted models for each of the data sets are either weighted using entity housing unit counts, or unweighted, and include lagged covariates (Durbin) or omit them. These models are then weighted or unweighted, Durbin or not, least squares or spatial error models. The Durbin versions of OLS and SEM will subsequently be termed SLX and SDEM (LeSage, 2014). For the census tract data set, the spatial error models out-perform the models without a spatially lagged error term, and the weighted models appear to outperform the unweighted models (although their comparability through including information in the weights is not taken into account in computing the AIC value). The weighted and unweighted OLS and SLX models were fitted by least squares, and the weighted and unweighted SEM and SDEM models were fitted by maximum likelihood.

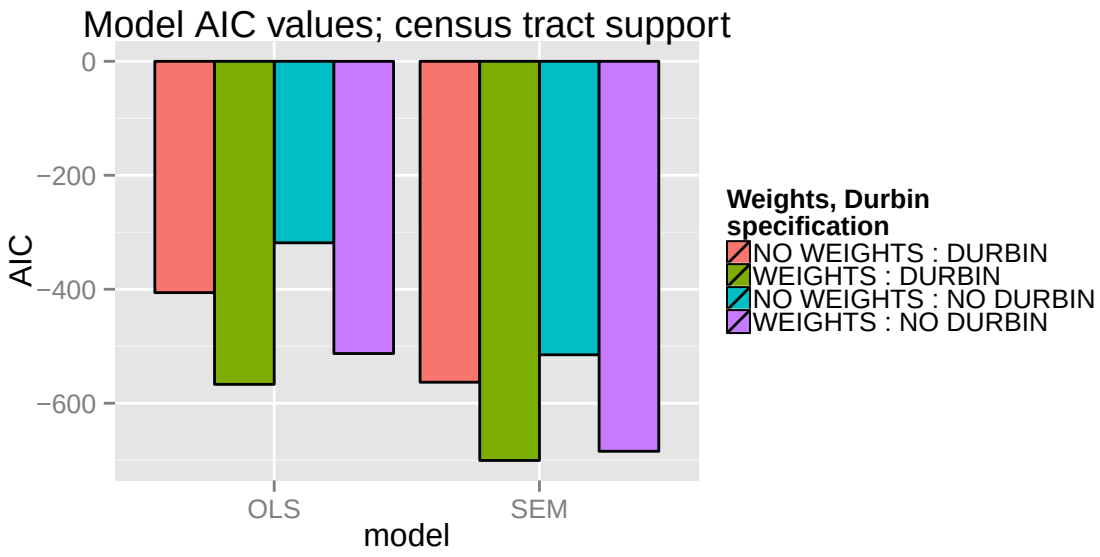


Figure 5: Model AIC values for observations on 489 non-censored census tracts

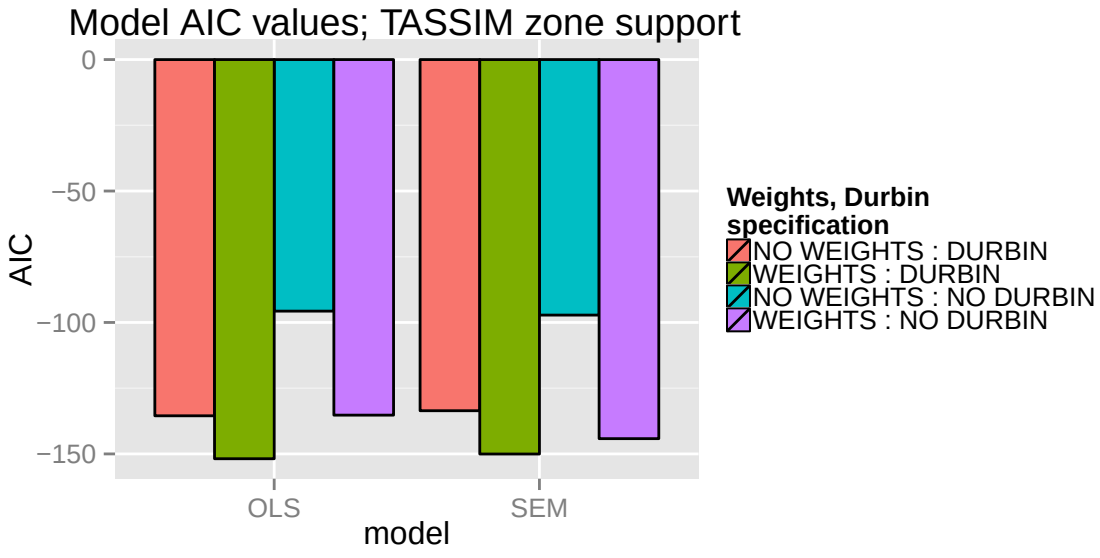


Figure 6: Model AIC values for observations on 96 approximate TASSIM zones

In Figure 6, the spatial error models no longer outperform the models without a spatially lagged error term, and the spatial coefficients of the spatial error models are not significant. The SLX models do outperform their non-weighted counterparts, subject to the remark above about the comparability of these models. If we prefer the census tract data set, we might conclude that the weighted SDEM model is to be preferred, but if we choose the approximate TASSIM zones, our choice would be the weighted SLX model, including the spatial processes in the Durbin term of spatially lagged covariates.

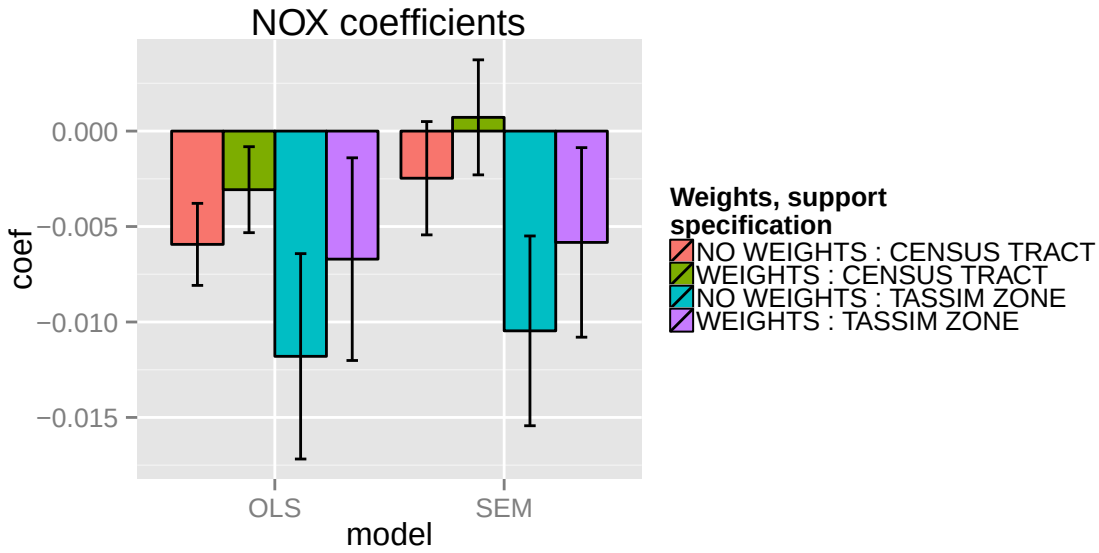


Figure 7: NOX coefficient values and ±2 standard error bars for OLS and SEM models, weighted and unweighted, for census tract and approximate TASSIM zone data sets.

While we cannot compare AIC values across the two sets of entities (census tracts and approximate TASSIM zones), we can compare coefficient values for the key variable of interest, air pollution, taken as the square of NOX in the original scaling. For brevity, and because our focus here is on the consequences of choices of system articulation for inference, other results are not given here, but may be obtained from the reproduction code. Figure 7 shows the coefficient values and ±2 standard error bars for eight models excluding spatially lagged covariates. The result for the unweighted SEM model for the census tract data set is not dissimilar from that given by Pace and Gilley (1997). The effect of residual spatial autocorrelation on the standard errors (and indeed on the coefficient values) is shown by comparing the OLS and SEM results for the census tract data set, regardless of whether weights are used.

When we move our attention to the Durbin models, including the spatially lagged covariates, we begin to be able to discern the consequences of the choice of entities for

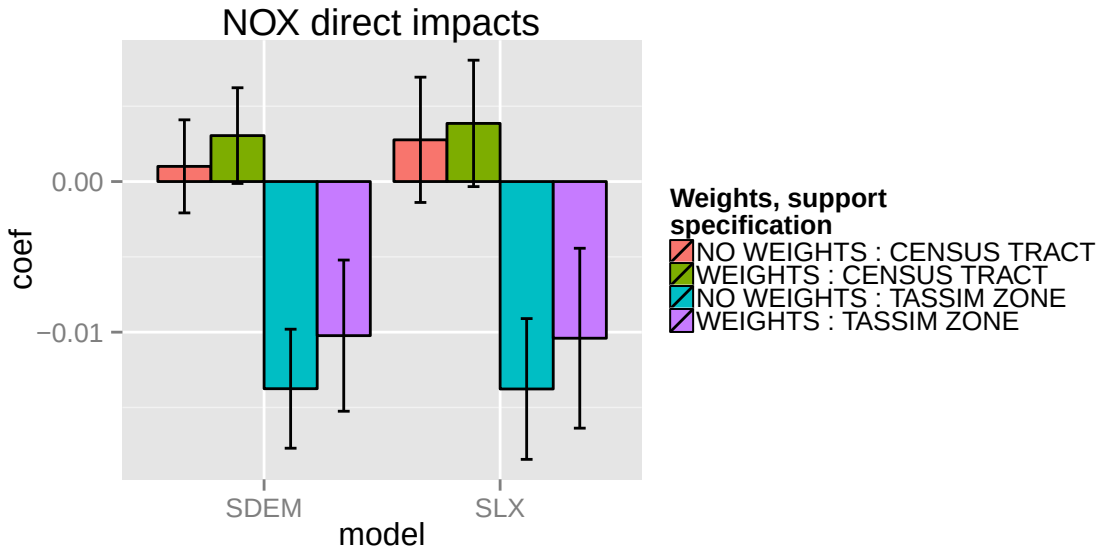


Figure 8: NOX direct impacts and ±2 standard error bars for models including spatially lagged covariates.

inference about the air pollution variable. Figure 8 shows the direct impacts, the NOX coefficient values from eight models. In the case of the 489 census tract data set, the coefficient values are positive and insignificant. For the 96 approximate TASSIM zones, the values are negative as expected and significant. It is entitation that makes the greater difference, larger than the inclusion or not of a spatial process in the disturbances, and larger than the inclusion or not of case weights to treat heteroskedasticity.

Figure 9 shows the indirect impacts, here the coefficients on the spatially lagged air pollution variable, and ±2 times their standard errors. All are negative, and here the 489 census tract data set models are all significant. The standard errors of the 489 census tract NOX coefficients are much smaller than those of the models fitted using the 96 approximate TASSIM zones data set. The NOX indirect impacts for the weighted models fitted using the 96 approximate TASSIM zones data set are at best marginally significant, so that with this data set and weighted regression, most of the "action" is in the direct impacts.

Finally, Figure 10 shows the total impacts for the eight models including the spatially lagged covariates, calculated using linear combination of the fitted model results for the NOX variable and its spatial lag. The total impact is simply the sum of the coefficient values, but the standard errors are calculated using the `estimable` function in the R**gmodels** package. The models fitted using the two entitations differ considerably, with strong residual spatial autocorrelation in both the weighted and unweighted SLX models. The

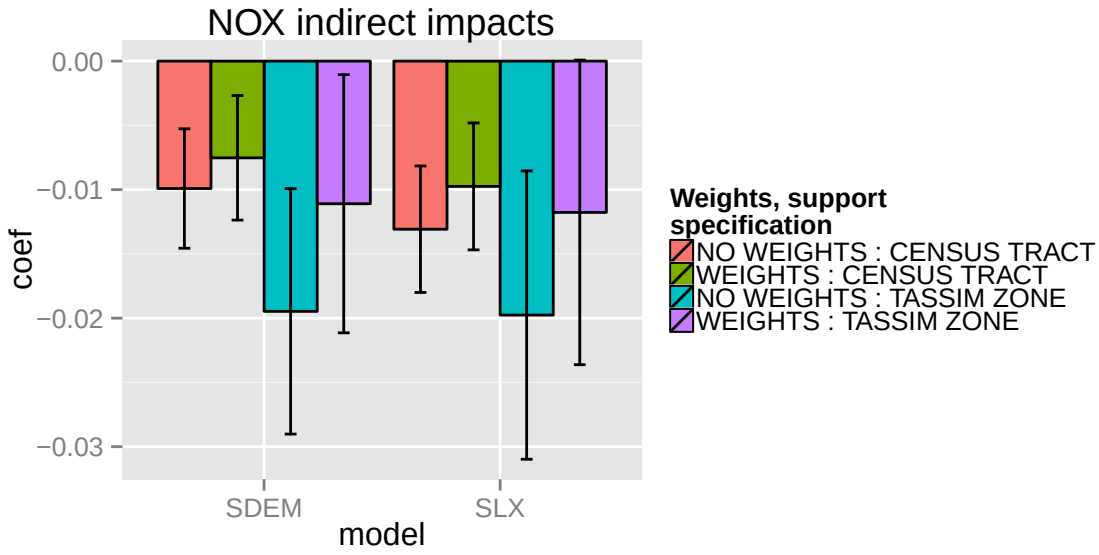


Figure 9: NOX indirect impacts and ±2 standard error bars for models including spatially lagged covariates.

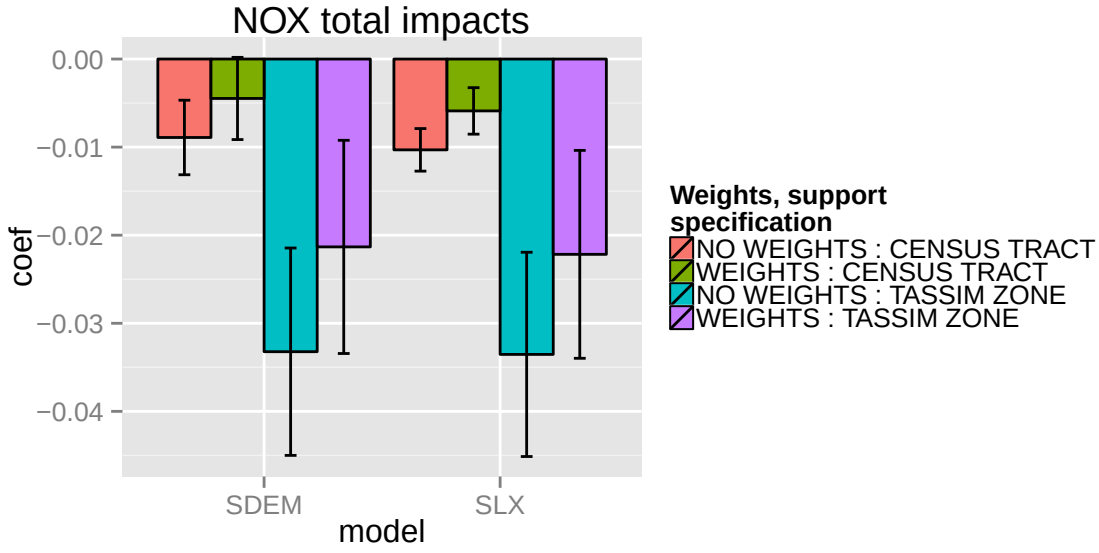


Figure 10: NOX total impacts and ±2 standard error bars for models including spatially lagged covariates.

SDEM spatial autoregressive coefficients ρ_{Err} in the SDEM models for the census tracts data set are 0.658 (standard error 0.0354) for the unweighted model, and 0.62 (standard error 0.0439) for the weighted model. The equivalent values for the SDEM models for the approximate TASSIM zones data set are 0.0562 (standard error 0.208) for the unweighted model, and 0.0964 (standard error 0.2) for the weighted model. The choice of entitation is driving the value of the spatial error coefficient, and inference on the appropriateness of its inclusion.

If we choose the approximate TASSIM zones data set, and to drop the SDEM specification including ρ_{Err}, the spatial error coefficient, in favour of the SLX specification, we still need to choose whether to use the numbers of housing units as weights for the zones, or not to do so (upweighting zones with relatively fewer housing units, and downweighting those with many). Figure 6 shows that the AIC values differ, with that for the 96 observation weighted SLX specification being -152, and unweighted: -136. Although AIC values give some guidance, and are based on log likelihood values that take account of the given weights, the choice between the two models depends on the analyst's prior choice of weights. This suggests that Bayesian methods may well be relevant to permit better insight into this question. If we take the 96 observation weighted SLX specification, the total impact of NOX is -0.02217 (standard error 0.005901), with equivalent values for the unweighted case: -0.03353 (standard error 0.005798). These values are substantially larger in absolute terms when compared with those found in Harrison and Rubinfeld (1978), and could be interpreted as indicating a greater willingness to pay for clean air than in the original study.

4 Concluding remarks

System articulation is expressed through the choice of observational units (entitation), which may (or may not) manifest the spatial process and scale relevant for inference. Both theory and method impact choices related to the operationalisation of the response, which not infrequently is not directly observable. It is advantageous to avoid unforced resampling — involving change of support, and if it cannot be avoided, care should be taken to carry through uncertainty. Inference and the interpretation of results depend crucially on previous choices, suggesting that adequate hierarchical models may be required (possibly not just instrumenting).

The increasing availability of spatial data is occurring at the same time as increasing use of such data by "broad" Regional Science. We have opportunities to enhance communication of legacy Regional Science, Spatial Econometrics and Spatial Statistics insights into spatial processes to these new users of spatial data. For example, in spatial regression discontinuity designs, we have clear opportunities to propose adequate handling of spatial processes, but should engage beyond Regional Science "proper." Concerns about over-enthusiastic use of inappropriate secondary data appear justified, where the possibly

inappropriate nature of the use often relates to entitation and change of support.

References

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

Bivand, R. S. (2008). Implementing representations of space in economic geography. *Journal of Regional Science*, 48:1–27.

Dray, S., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P. R., Bellier, E., Bivand, R., Blanchet, F. G., de Cáceres, M., Dufour, A.-B., Heegaard, E., Jombart, T., Munoz, F., Oksanen, J., Pélissier, R., Thioulouse, J., and Wagner, H. H. (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs*, 82:257–275. pp. 19.

Egger, P. H. and Lassmann, A. (2015). The causal impact of common native language on international trade: Evidence from a spatial regression discontinuity design. *The Economic Journal*, 125(584):699–745.

Gelfand, A. E. (2010). Misaligned spatial data: The change of support problem. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, *Handbook of Spatial Statistics*, pages 517–539. Chapman & Hall/CRC, Boca Raton. pp. 23.

Gilley, O. W. and Pace, R. K. (1996). On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31(3):403–405.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97:632–648. pp. 17.

Haggett, P., Cliff, A. D., and Frey, A. (1977). *Locational Analysis in Human Geography, second edition*. Edward Arnold, London.

Haining, R. P. (2010). The nature of georeferenced data. In Fischer, M. and Getis, A., editors, *Handbook of Applied Spatial Analysis*, pages 197–217. Springer, Heidelberg. pp. 21.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

Hidano, N., Hoshino, T., and Sugiura, A. (2015). The effect of seismic hazard risk information on property prices: Evidence from a spatial regression discontinuity design. *Regional Science and Urban Economics*, 53:113 – 122.

Keele, L. J. and Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1):127–155.

LeSage, J. P. (2014). What regional scientists need to know about spatial econometrics. *Review of Regional Studies*, 44:13–32.

Ord, J. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126.

Pace, R. K. and Gilley, O. (1997). Using the spatial configuration of the data to improve estimation. *Journal of the Real Estate Finance and Economics*, 14:333–340.

Paelinck, J. H. P. and Nijkamp, P. (1975). *Operational theory and method in regional economics*. Saxon House, Farnborough.

Wakefield, J. C. and Lyons, H. (2010). Spatial aggregation and the ecological fallacy. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, *Handbook of Spatial Statistics*, pages 541–558. Chapman & Hall/CRC, Boca Raton. pp. 18.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Hoboken, NJ.

Wilson, A. G. (2000). *Complex spatial systems: The Modelling Foundations of Urban and Regional Analysis*. Prentice Hall, Harlow.

Wilson, A. G. (2002). Complex spatial systems: Challenges for modellers. *Mathematical and Computer Modelling*, 36:379–387.

Wilson, A. G. (2012). *The science of cities and regions: lectures on mathematical model design*. Springer, Dordrecht.

Issued in the series Discussion Papers 2014

2014

01/14 January**, Kurt R. Brekke**, Tor Helge Holmås, and Odd Rune Straume, "Price Regulation and Parallel Imports of Pharmaceuticals".

02/14 January, **Alexander W. Cappelen, Bjørn-Atle Reme, Erik Ø. Sørensen,** and **Bertil Tungodden, "**Leadership and incentives".

03/14 January, **Ingvild Almås, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø. Sørensen**, and **Bertil Tungodden**, "Willingness to Compete: Family Matters".

04/14 February, **Kurt R. Brekke**, Luigi Siciliani, and Odd Runde Straume, "Horizontal Mergers and Product Quality".

05/14 March, **Jan Tore Klovland**, "Challenges for the construction of historical price indices: The case of Norway, 1777-1920".

06/14 March, Johanna Möllerström, **Bjørn-Atle Reme**, and **Erik Ø. Sørensen**, "Luck, Choice and Responsibility".

07/14 March, Andreea Cosnita-Langlais and **Lars Sørgard**, "Enforcement vs Deterrence in Merger Control: Can Remedies Lead to Lower Welfare?"

08/14 March, **Alexander W. Cappelen, Shachar Kariv, Erik Ø. Sørensen**, and **Bertil Tungodden**, «Is There a Development Gap in Rationality?"

09/14 April, **Alexander W. Cappelen**, Ulrik H. Nielsen, **Bertil Tungodden**, Jean-Robert Tyran, and Erik Wengström, "Fairness is intuitive".

10/14 April, **Agnar Sandmo**, "The early history of environmental economics".

11/14 April, **Astrid Kunze**, "Are all of the good men fathers? The effect of having children on earnings".

12/14 April, **Agnar Sandmo**, "The Market in Economics: Behavioural Assumptions and Value Judgments".

13/14 April, **Agnar Sandmo**, "Adam Smith and modern economics".

14/14 April, Hilde Meersman, **Siri Pettersen Strandenes**, and Eddy Van de Voorde, "Port Pricing: Principles, Structure and Models".

15/14 May, **Ola Honningdal Grytten**, "Growth in public finances as tool for control: Norwegian development 1850-1950"

16/14 May, **Hans Jarle Kind**, Tore Nilssen, and **Lars Sørgard**, "Inter-Firm Price Coordination in a Two-Sided Market".

17/14 May, **Stig Tenold**, "Globalisation and maritime labour in Norway after World War II".

18/14 May, **Tunç Durmaz,** "Energy Storage and Renewable Energy"

19/14 May, **Elias Braunfels**, "How do Political and Economic Institutions Affect Each Other?"

20/14 May, Arturo Ormeño and **Krisztina Molnár**, "Using Survey Data of Inflation Expectations in the Estimation of Learning and Rational Expectations Models"

21/14 May, **Kurt R. Brekke**, Luigi Siciliani, and Odd Rune Straume, "Hospital Mergers with Regulated Prices".

22/14 May, **Katrine Holm Reiso**, "The Effect of Welfare Reforms on Benefit Substitution".

23/14 June, Sandra E. Black, Paul J. Devereux, and **Kjell G. Salvanes**, "Does grief transfer across generations? In-utero deaths and child outcomes"

24/14 June, Manudeep Bhuller, Magne Mogstad, and **Kjell G. Salvanes**, «Life Cycle Earnings, Education Premiums and Internal Rates of Return".

25/14 June, **Ragnhild Balsvik**, **Sissel Jensen**, and **Kjell G. Salvanes**, "Made in China, sold in Norway: Local labor market effects of an import shock".

26/14 August, **Kristina Bott, Alexander W. Cappelen, Erik Ø. Sørensen**, and **Bertil Tungodden**, "You've got mail: a randomized field experiment on tax evasion"

27/14 August, **Alexander W. Cappelen, Sebastian Fest, Erik Ø. Sørensen**, and **Bertil Tungodden**, "The freedom to choose undermines the willingness to redistribute."

28/14 August, Marianne Bertrand, Sandra Black, **Sissel Jensen**, and Adriana Lleras-Muney, "Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labor Market Outcomes in Norway."

29/14 August, **Astrid Kunze**, "The family gap in career progression".

30/14 September, **Ragnhild Balsvik** and **Morten Sæthre**, "Rent Sharing with Footloose Production. Foreign Ownership and Wages Revisited".

31/14 October, **Nicola D. Coniglio** and Giovanni Pesce, "Climate Variability and International Migration: an empirical analysis"

32/14 November, **Kurt R. Brekke**, Armando J. Garcia Pires, Dirk Schindler, and Guttorm Schjelderup, "Capital Taxation and Imperfect Competition: ACE vs. CBIT"

33/14 November, **Jan I. Haaland** and Anthony J. Venables, "Optimal trade policy with monopolistic competition and heterogeneous firms".

34/14 December, Rolf Aaberge, **Kai Liu**, and Yu Zhu, "Political Uncertainty and Household Savings".

2015

01/15 January**,** Antonio Mele**, Krisztina Molnár,** and Sergio Santoro, "On the perils of stabilizing prices when agents are learning".

02/15 March, **Liam Brunt**, "Weather shocks and English wheat yields, 1690-1871".

03/15 March, **Kjetil Bjorvatn, Alexander W. Cappelen**, Linda Helgesson Sekei, **Erik Ø. Sørensen**, and **Bertil Tungodden**, "Teaching through television: Experimental evidence on entrepreneurship education in Tanzania".

04/15 March**, Kurt R. Brekke**, **Chiara Canta**, Odd Rune Straume, "Reference pricing with endogenous generic entry".

05/15 March, Richard Gilbert and **Eirik Gaard Kristiansen**, "Licensing and Innovation with Imperfect Contract Enforcement".

06/15 March, **Liam Brunt** and Edmund Cannon, "Variations in the price and quality of English grain, 1750-1914: quantitative evidence and empirical implications".

07/15 April, Jari Ojala and **Stig Tenold**, "Sharing Mare Nostrum: An analysis of Mediterranean maritime history articles in English-language journals".

08/15 April, **Bjørn L. Basberg**, "Keynes, Trouton and the Hector Whaling Company. A personal and professional relationship".

09/15 April, Nils G. May and **Øivind A. Nilsen**, "The Local Economic Impact of Wind Power Deployment".

10/15 May, **Ragnhild Balsvik** and Stefanie Haller, "Ownership change and its implications for the match between the plant and its workers".

11/15 June, **Kurt R. Brekke**, **Chiara Canta**, Odd Rune Straume, "Does Reference Pricing Drive Out Generic Competition in Pharmaceutical Markets? Evidence from a Policy Reform".

12/15 June, **Kurt R. Brekke**, Tor Helge Holmås, Karin Monstad, and Odd Rune Straume, "Socioeconomic Status and Physicians'Treatment Decisions".

13/15 June, **Bjørn L. Basberg**, "Commercial and Economic Aspects of Antarctic Exploration - From the Earliest Discoveries into the 19th Century".

14/15 June, **Astrid Kunze** and Amalia R. Miller, "Women Helping Women? Evidence from Private Sector Data on Workplace Hierarchies"

15/15 July, **Kurt R. Brekke**, Tor Helge Holmås, Karin Monstad, Odd Rune Straume, «Do Treatment Decisions Depend on Physicians Financial Incentives?"

16/15 July, **Ola Honningdal Grytten**, "Norwegian GDP by industry 1830-1930".

17/15 August, **Alexander W. Cappelen**, Roland I. Luttens, **Erik Ø. Sørensen**, and **Bertil Tungodden**, «Fairness in bankruptcy situations: an experimental study».

18/15 August, **Ingvild Almås, Alexander W. Cappelen, Erik Ø. Sørensen**, and **Bertil Tungodden**, "Fairness and the Development of Inequality Acceptance".

19/15 August, **Alexander W. Cappelen**, Tom Eichele,Kenneth Hugdah, Karsten Specht, **Erik Ø. Sørensen**, and **Bertil Tungodden**, "Equity theory and fair inequality: a neuroconomic study".

20/15 August, Frank Jensen and **Linda Nøstbakken**, «A Corporate-Crime Perspective on Fisheries: Liability Rules and Non-Compliance".

21/15 August, Itziar Lazkano and **Linda Nøstbakken**, "Quota Enforcement and Capital Investment in Natural Resource Industries".

22/15 October, **Ole-Petter Moe Hansen** and Stefan Legge, "Trading off Welfare and Immigration in Europe".

23/15 October, Pedro Carneiro, Italo Lopez Garcia, **Kjell G. Salvanes**, and Emma Tominey, "Intergenerational Mobility and the Timing of Parental Income".

24/15 October, David Figlio, Krzysztof Karbownik, and **Kjell G. Salvanes**, "Education Research and Administrative Data".

25/15 October, **Ingvild Almås, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø. Sørensen,** and **Bertil Tungodden**: «Fairness and family background».

26/15 November, Lars Ivar Oppedal Berge, **Kjetil Bjorvatn**, Simon Galle, Edward Miguel, Daniel Posner, **Bertil Tungodden**, and Kelly Zhang "How Strong are Ethnic Preferences?".

27/15 November, **Agnar Sandmo**, "The Public Economics of Climate Change".

28/15 November, **Aline Bütikofer** and **Kjell G. Salvanes**, "Disease Control and Inequality Reduction: Evidence from a Tuberculosis Testing and Vaccination Campaign".

29/15 December, **Aline Bütikofer**, Katrine V. Løken and **Kjell G. Salvanes**, "Long-Term Consequences of Access to Well-child Visits"

30/15 December, **Roger Bivand**, "Revisiting the Boston data set (Harrison and Rubinfeld, 1978): a case study in the challenges of system articulation".

Norges Handelshøyskole

Norwegian School of Economics

NHH
Helleveien 30
NO-5045 Bergen
Norway

Tlf/Tel: +47 55 95 90 00
Faks/Fax: +47 55 95 91 00
nhh.postmottak@nhh.no
www.nhh.no