Research agenda, 2017–2021
Position paper

Do regional economists answer the right questions?

On the current discrepancy between the questions regional economists solve and the questions policy makers actually ask

Thomas de Graaff[1]*

Abstract
This position paper revolves around two main propositions: namely, (*i*) regional (or spatial) economists are very restrictive in the tool set they apply, and consequently (*ii*) their models do not always match with the type of questions policy makers are concerned about. To start with the latter, policy makers—whether national, regional or local—are oftentimes concerned about holistic approaches and future predictions. Exemplary questions are "Which policy instrument works best for my city", "What happens after the construction of this highway with housing prices and employment throughout the whole region" and "Given limited budget, which region should I first invest in". Regional economists—actually, most economists—usually isolate phenomena in order to, at best, explain the impact of a single determinant. Indeed, most regional economists feel very uncomfortable when asked to predict or give the best set of determinants for a certain phenomenon. This has its consequences for the tool set regional economists apply. Usually a parametric regression type of framework is applied isolating the determinant under consideration and controlling as much as possible for observables and unobservables, ideally in a pseudo-experimental framework. A direct consequence of this approach is that emphasis is very much on explaining the impact of an isolated determinant and not on predicting (non-marginal) changes in larger systems. For many applications that is definitely the right approach. However, as this paper ultimately argues, it is very much as well a selective approach that does not do well to deliver on some of the questions policy makers ask regional economists.

Keywords
Regional economics — predicting — causality — theory driven approach — data science

[1] *Department of Spatial Economics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*
***Corresponding author**: ✉ t.de.graaff@vu.n; thomasdegraaff.nl

Introduction: two different cultures

> The sexiest job in the next 10 years will be statisticians.
>
> Hal Varian, 2009

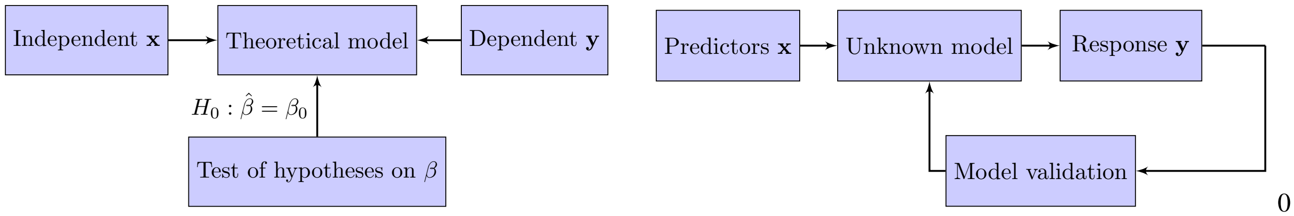
The quote above from Hal Varian is in one aspect wrong; nowadays, we do not call them statisticians but data scientists instead. Nevertheless, in the last two decades companies such as Google, Ebay, Whatsapp, Facebook, Booking.com and Airbnb, have not only witnessed enormous growth but to a considerable extent also changed the socio-economic landscape. Indeed, with the increasing abundance of (spatial) data and computer capacity, the ability to gather, process, and visualize data has become highly important and therefore highly in demand as well. And all the models and tools these data scientists within these companies use are very much *data driven* with often remarkable results.

In his controversial and path-breaking article, Breiman (2001) presented two different cultures in statistical science. One governed by a (probability) theory-driven modeling approach and one governed by a more (algorithmic) data-driven approach. These two cultures carry over to the econometric and ultimately the empirical regional economics domain[1] as well, where—commonly for all social sciences—the theory driven approach still very much dominates the landscape of the realm of contemporary regional economics.

Figure 1 is an adaptation from the one displayed in Breiman (2001) and describes the processes governing these two cultures. Figure (1a) is what I refer to as the modeling approach, where a statistical model is postulated and is central to this culture. This is the classical approach[2] where statistical probability theory meets the empiricism of Karl Popper. Usually the model assumed is stated as a linear model and in its most

[1] I use a wide definition for the regional economics domain, which consists of most aspects of regional science in general but for which the theoretical approach is always from an economic perspective. Topics such as, e.g, interregional migration, trade, transport flows and commuting on the one side and regional performance, regional clustering, population growth and specialisation on the other side fall all under this, admittedly, rather wide umbrella.

[2] Sometimes as well referred to as the frequentists' approach. However, this typically concerns the debate between classical statistics and Bayesian statistics, where the two approaches I refer to are more concerned with wider frameworks, of which the Bayesian approach is just one of the elements.



(a) Theory or modeling approach

(b) Data driven approach

Figure 1. Two cultures of statistical/econometric modeling (inspired by Breiman, 2001)

simple form can be denoted as:

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon, \tag{1}$$

where in (regional) economics language, **x** is referred to as the independent variable, **y** as the dependent variable and ε as a residual term. In this setup, using the data at hand, one constructs a statistical test to which extent the estimated coefficient (denoted with $\hat{\beta}$) deviates from a hypothesized value of the coefficient (denoted with β_0)—typically the hypothesis $H_0 : \hat{\beta} = 0$ is used with as alternative hypothesis that $H_1 : \hat{\beta} \neq 0$. However, that is always within the context of the *postulated* model. So, when the null-hypothesis is rejected, it not necessarily means that the true β is unequal to zero, it might also be caused by errors in measuring **x** or even using the wrong *model*![3][4]

Figure (1b) yields a schematic overview of a more data driven approach. Here, we see an unknown model fed by predictors **x** that lead to one or multiple reponses **y**. The main objective here is not to test hypotheses, but to find the best model instead which able to explain the *in-sample* data and to predict the *out-of-sample* data. Usually, the models are evaluated by some kind of criterion (e.g., the mean squared error), which is not completely unlike the modeling approach. However, there are two main differences between the two approaches. First, the data driven approach considers several models in a structural approach. For instance, the question which variables to include is captured by an exhaustive sourse of all combinations in the modeling approach (e.g., with classification and regression trees or random forests), while in the theory driven approach, the choice of variables is based on the theory and a small number of variations in the specification. Second, measurements on model performance are done **out-of-sample** in the data driven approach and, typically, **in-sample** in the model approach. The latter is not that important for hypothesis testing, but for prediction this matters enormously, because adding parameters might increase the in-sample fit, but actually worsen the out-of-sample fit (a phenomenon called overfitting).

In economics in general, and in regional economics in specific, most of the tools employed are very much *theory or model driven* instead of data driven. My (conservative) estimate would be that at least 90% of all empirical work in regional economics revolves around postulating a (linear) model and testing whether (a) key determinant(s) is (are) significantly different from a hypothesized value—usually zero.[5] That is, *within* the context of the model assumed.

At best, this approach can be seen in a causal inference framework. If a determinant (such as a policy in the context of regional economics) x changes, does it cause then a change in the output y (most economists typically use some welfare measure).[6] This approach thus provides a rigid and useful approach to regional policy evaluation. If we implement policy x, does welfare measure y then improve? Note that this always considers a *marginal* change as x is usually isolated from other (confounding) factors.

However, policy makers oftentimes have different questions for which they need solutions. Usually, they revolve around questions starting with *"What determines performance measure A?"*, *"Which regions can we best invest in?"* or, more generally, *"What works for my region?"*. These types of questions require a different approach than the previous one. Namely, the former type requires an approach focused on **explaining** while the latter type requires an approach focused on **predicting**.

The remaining part of this position paper is structured as follows. Section 1 gives an overview of current modeling practices and describes the 'traditional' inference based approach as well as some data-driven approaches that have been used in the recent past (though by far not as often as the traditional

[3] One of the assumptions for regression techniques such as the one used here is actually no misspecification of the model, but—apart from some possible tests on the functional form *within* a specific regression form—usually little attention is give on the validity of the model used. More importantly, within this framework the model itself is usually not tested *a posteriori*.

[4] There is another fallacy with this approach that is often overlooked and that is that the alternative hypothesis being true is a probability as well. Namely, most hypotheses researchers test are typically not very probable. Not taken this into account would actually lead to more null hypotheses to be rejected then should be (false positives).

[5] In a seminal contribution, Breiman, 2001 states that deep into the 90s 98% of the statisticians actually employed the theory driven paradigm and only 2% a data driven paradigm. With the advent of the availability of internet connectivity, large (online) data sources, and faster computers the statistical realm changed dramatically. However, this has not permeated yet in the social sciences (see as well Varian, 2014).

[6] Most of this research actually intends to mimic a *difference-in-difference* approach and gained enormous momentum with the textbook of Angrist and Pischke (2008).

methods). Section 2 sets out both a research and an education agenda as it addresses how to bridge the gap between the daily practices of regional economists and the demands of local policy makers. The final section shortly summarizes the main points raised in this position paper.

1. Regional economists turning the blind eye

Unmistakenly, in the recent decade the two major changes to economic empirical research in general are the advent of increasingly larger data sources and the large increase in computer power (Einav and Levin, 2014). The methods that mosts economists employ, however, have not changed. Linear regression or one of its close relatives (such as logistic, poisson or negative binomial regression), preferably in a causal framework, is still the most common tool. This also applies to regional economists, who—although coming from a tradition to use various methods from different disciplines—have increasingly used similar methods as in "mainstream" economics.

This focus on marginal effects and causality is certainly very worthwhile and brought us many important insights. However, it is also typically done within a very narrow framework and, below, I will lay out what we are missing both in research and in our educational curricula, when our *main* focus is on the framework above and as advocated so much as in Angrist and Pischke (2008).

1.1 The blind eye in research

The traditional model of a (regional) economist looks as follows:

$$y_i = \alpha + \beta x_i + \mathbf{z}_i \gamma + \varepsilon_i, \tag{2}$$

where y_i is referred to as the dependent variable, x_i is the main variable of interest, and $\mathbf{z}$ is a vector of other variables. α, β and γ are parameters, where we are especially interested in the value of β. Finally, ε_i is an identical and independent distributed error term.

Usually the main aim is to estimate β as unbiased as possible in a causal framework. So, ideally, we would like to control for unobserved heterogeneity bias, specification bias, measurement error, reverse causality, selection bias, and so forth. Econometric theory has produced some very powerful techniques to control for some of these biases, such as instrumental variables, diff-in-diff procedures and the use of fixed effects. However, these methods are not panacea for everything. First, they work wonders for only specific research questions that have to do with the preferably causal effects of marginal changes. Second, some of these techniques require very specific and strong assumptions which are possibly not always met, which leaves doubts upon the validity of the results.

Below, I will deal with instrumental variables, diff-in-diff and fixed effect techniques consecutively. I will specificially focus on some of the disadvantages. Some of the arguments are adaptions from Deaton (2010) and I refer to this reference for a more complete treatise on the disadvantages of using instrumental variables and diff-in-diff methods. For all the advantages not dealt with in this paper, read Angrist and Pischke (2008).

1.1.1 Exogeneity versus independence

Economists love instrumental variables, because a good instrumental variable can tackle reverse causality, measurement error and unobserved heterogeneity bias all at one. Originally, instrumental variables come from simultaneous economic models such as supply and demand models. A classical example in a regional context would be:

$$P_r = \alpha + \beta E_r + \mathbf{z}_r \gamma + \varepsilon_r, \tag{3}$$
$$E_r = \delta + \kappa P_r + \mathbf{w}_r \lambda + \nu_r, \tag{4}$$

where P denotes population, E employment and z and w are vector of other regional r characteristics. α, β, γ, δ, κ and λ are parameters to be estimated.

Obviously, one can not directly estimate (3) and (4) because of the intrinsic simultaneity. However, suppose one is interested in estimating the impact of employment on population growth, then one can use (4) and search for *exogeneous*[7] variation in employment to use it as an instrumental variable. A possible strategy could be to look into the population changes of surrounding regions (but within commuting distance), as they might not have an impact of the population change in the current region (see de Graaff et al., 2012a,b)

The main point[8], however, is that equations (3)–(4) constitute a full-blown economic *model* which has direct relations with underlying structural theoretical modeling frameworks (such as Roback, 1982). And the instrument then comes directly (is interal) from the model.

In practice, however, researchers often take another approach. And that is to look for external instruments. Instruments that have no relation with a structural (simultaneity) model. And there are is (a large) potential pitfall when doing so and that is to end up with an instrumental variables that is not independent from the left-hand-side variable. As it seems, there is some confusion about terms as independence and exogeneity, so let's first clarify the exact assumptions a valid instrument should satisfy.

Suppose that somebody is interested in the impact of population on employment; so, one would like to identify κ in (4). To control for endogeneity researchers then search for an *exogenous* and *relevant* instrument, Z_r. The latter indicates that the instrument has an impact on the possible endogeneous variable (P_r) and the former indicates that the instrument does not affect the left-hand-side variable (E_r), only via P_r and other instruments. In formal notation: $E_r \perp Z_r | P_r, w_r$. Thus, exogeneity means that the instrument and the left-hand-side

[7] This is not really precise; I mean exogeneous to population variation. I will come back to the use of exogeneous later.

[8] Directly from Deaton, 2010.

variables are *independent* from each other conditional on the instruments.

Unfortunately, exogeneity is often used as an argument for variables that are external to the system, denote a sudden shock or for phenomena that are considered to be exogenous in other fields (such as in geography). And this usually leads to instruments that do not satisfy the independence assumption. I will give three examples below.

First, and very often use is the concept of deeplagging. So, in our case, we look for regional population say 100 years ago and use that as an instrument. It must be exogenous because we cannot change it, right? Well, it is definitely relevant, as regional population is remarkably resilient. Where people lived 100 years ago, they most likely live today. But, if we take model (3)-(4) seriously, then the population 100 years ago, must at least have affected employment 100 years ago, and if population is resilient then most likely employment as well (and we even do not consider yearly temporal dynamic between population and employment). So, in all likeliness, employment and population 100 years are not (conditionally) independent.

The second type of instruments people often use are regional characteristics (preferably deeplagged as well), and specifically accessibility measures as road, railroads and canals. For a large audience the following story typically seems very plausible at first sight. At the end of the 19th century the large scale introduction of the railways enabled households to live further from home and escape the heavilty polluted inner cities where the factories remained (making use of the same railroads intersecting in city centres). Railroads thus changed the location of population and not that of employment. While this story is entirely possible, what is often overlooked is the fact that factories and thus employment changed location as well, but only 20-30 years later, and typically along the same links as opened up by railway lines. So, the railway network 140 years ago and contemporary location of employment are not (conditionally) independent.

A last often used category of candidate instruments is geography-related variables. In our case that could be regions of municipalities. For instance, the Netherlands witnessed for a large period intensive population location policies. This entailed that the dutch government pointed out municipalities that were allowed to grow (in terms of housing policies). Using fixed effects of these specific municipalities then as instruments sound as a viable strategy. However, this requires strict assumptions. Namely, being a specific municipality will only have an effect on employment through being designated by the Dutch government; and by nothing else.[9]

Is this to say that instrumental variables is a bad technique? No, absolutely not. If the instrument is valid, this is one of the most powerful techniques in the econometric toolbox. The point made here is that good instruments are actually hard to find and that structural simultaneous models (typically, in the context of supply and demand) usually work better to find instruments than instruments that are completely external to your problem. And if you really need to use an external instrument, be very specific and open about the assumptions you need to make.

1.1.2 Local average treatment effects

1.1.3 Fixed effects and heterogeneity

An often used technique in applied econometrics is the use of fixed effects. They work brilliantly in removing unobserved heterogeneity but they come at a price which is typically overlooked. Namely, they remove valuable variation as well in both the dependent (predictor) x and the independent (reponse) variable y.

Consider the following model in equation (5), which is at the moment a heavily researched issue in both regional and urban economics. The issue here is to what extent city density increases individual productivity.

$$\ln(w_{ic}) = \alpha + \beta \ln(d_{ic}) + \varepsilon_{ic}, \tag{5}$$

w_{ic} denotes here individual wages (as a proxy for productivity) and d_{ic} density of the city c individual i lives in. β is our parameter of interest and because of the log-log structure β denotes an elasticity. Obviously, direct estimation of model (5), would lead to a misleading parameter β if one is aiming to measure a causal effect.[10] Namely, β might be influenced by other (confounding) factors than only city density. One can think of factors such as skill level of the city population, accessibility of the city, sector structure of the city and city government. Moreover, a phenomenon called sorting might occur, where more ambitious, risk-seeking and high-skilled people migrate into larger and more dynamic cities.[11]

To answer the question to what extent density causes wages, researchers therefore resolved to using fixed effects. A baseline model can be seen in (6).

$$\ln(w_{ic}) = \nu_i + \xi_c + \beta \ln(d_{ic}) + \varepsilon_{ic}, \tag{6}$$

here, ν_i denotes individual i specific fixed effects and ξ_c city c specific fixed effects. So, everything that does not vary over time for individuals and cities is now controlled for. A more insightful way what exactly happens is to write model (6)

[9] A similar argument has been made by Deaton, 2010, who considers economic growth theory, where fixed effects of specific countries are used as instruments because they took part in specific agreements—i.e., the Camp David accord for Egypt. Then the heroic assumptions has to be made that being Egypt has no effect on growth, except for the Camp David accord.

[10] I specifically do not use the term *biased* here. Namely, model (5) is a perfectly fine model to measure the overall correlation (β) between city density and individual wages and mosts non-economists are perfectly fine with this model (see, e.g., Bettencourt and West, 2010). So, whether a parameter is biased depends ultimately upon the research question.

[11] Another issue we will not deal with here is reverse causality where it might be that higher wages lead to larger in-migration and thus larger density. This can, however, not be solved with fixed effects, but with tools as instrumental variables instead. We therefore leave it out of the discussion in this paragraph.

in changes, such as: $\Delta\ln(w_{ic}) = \beta\Delta\ln(d_{ic}) + \varepsilon_{ic}$.[12]. So our model (6) now identifies the *causal* effect by looking at the impact of a change in density on a change in wages *for the same individual within the same city*.

Multiple improvement have already been to this model including controlling for sector/task of work and migrating between cities. Including these fixed effects (and many more) has had a profound effect on the value of β. Directly estimating model (5) yields an elasticity of around 1.15, while estimating a model such as (6) including many fixed effects would yield an elasticity of around 1.02. So, there are economies of agglomeration, but they are not very large.

Is this now the end of the story? Alas, it is not. At least three remarks can be made that put the above into perspective.

First of all, note that we need changes over time—in our case in individual wages and city density. Now, if we take the extreme example of a subgroup of individuals who do not face wage changes and cities who remain relatively of equal size, than this subgroups will not be used for determination of β. Of course, not many observations will have these characteristics. Unfortunately, with more detailed data on sector structure and migration, we need individuals that move both residence and job for identification. All others are redundant. This increases the risk on what is called sample selection bias—identification is based on a specific subgroup with deviant characteristics. The point made here, is that with the use of many fixed effects, much is demanded from the data and one need always check whether the sample used for identification is not too restrictive.

Secondly, if there are unobserved factors that both relate to wages and density, then it is actually very likely that these unobserved factors are related to their *changes* as well. One particular example here is technological change, which might affect density (suburbs) and wages at the same time, and is definitely not time-invariant. If one thinks about it, most interesting socio-economic phenomena are not time-invariant, except perhaps longitude and latitude. For example, a specific argument to use fixed effects is to control for local attractivity. But what individuals and firms find attractive does change of time, especially within cities, but across cities as well. Before air-conditioning cities in Florida and Nevada were definitely not as popular as today. And malaria-rich areas such as wetlands and river banks were always avoided until recently.

Thirdly, the use of fixed effects is based upon the assumption that all variation is based on variation of *levels*. That is, each fixed effects actually denotes a very specific constant (for each individual and city in our case). However, this really requires a very homogeneous sample except in levels. For illustration, assume that there are three individuals, where individual 3 has gigher wages than individual 1 and 2, because of, say, differences in skill levels (see as well Figure 2a). However, as Figure 2a clearly shows as well, apart from individual level variation, returns to density are similar for individuals 1, 2 and 3. So, each individual benefits equally from moving from a small village not a large metropolitan area. Now, assume that individuals are different with respect to the returns by living in large and denser cities. Then the impact β should also differ amongst individuals as is illustrated in Figure 2b. This is not an argument to say that using fixed effects is wrong. But if the sample might be heterogenous, i.e. that units respond differently to different predictors, then using fixed effect might not yield a complete pictures and in some specific cases even a distorted picture.

Fixed effect techniques is a must have for every empirical regional economists. However, the message I would like to convey here is that it does not remove time-invariant unobserved heterogeneity (of which there is more than most researchers realise), is not very suitable for tackling heterogeneity in your main effect and might lead in some cases to sample selection bias.

1.2 The blind eye in education

So, if the main instruments of regional economists are not always applicable and we miss tools in our toolbox to tackle, e.g., heterogeneity, prediction and non-marginal changes, how do we then fare in teaching? Are the students who now graduate equipped with the right toolbox that they use as well in their later careers? And do we have a consistent curriculum using similar or complementary tools running from the bachelor to the graduate studies? These types of questions are not frequently asked, and, if at all, not very well met. Mostly because of vested interests of departments and researchers.

In this subsection I will, however, try to answer partly some of these questions and identify what is missing in our curriculum. I will first look at the traditional applied econometrics approach and then to the (non-existence) of other courses, including the use of statistical software.

1.2.1 Applied Econometrics

1.2.2 Current day practice

2. Incorporating the data science culture

What do we need?

2.1 For research

2.1.1 Regional heterogeneity

(Thissen et al., 2016; de Graaff et al., 2012b,a)

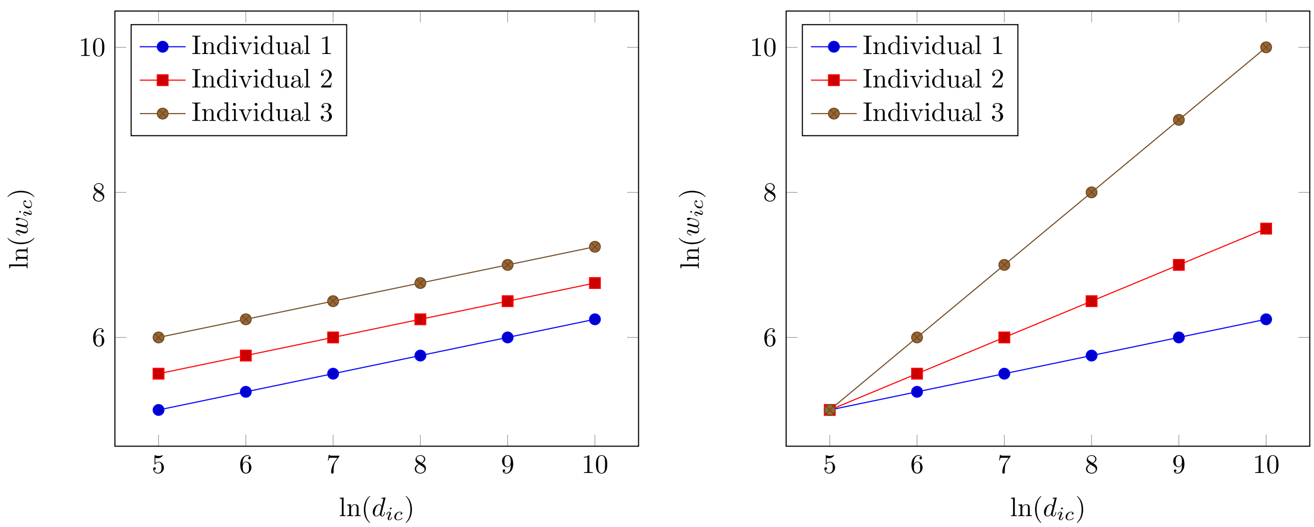
2.1.2 Conditional robustness

In regional science in general and in regional economics in specific, remarkably little attention has been given to reproducibility and robustness of results (with some exceptions as, amongst some others, by Rey, 2014; Arribas-Bel and de Graaff, 2015; Arribas-Bel et al., Forthcoming).

2.1.3 Regional sorting models

As in Bayer et al. (2004) and Bayer and Timmins (2007) and recently by Zhiling et al. (2016) and Bernasco et al. (Forthcoming).

[12]Using changes (first differences) to remove fixed effects is a viable but often overlooked technique for dealing with fixed effects.



(a) Fixed effects **(b)** Heterogeneity in β

Figure 2. Heterogeneity in levels versus slopes.

2.2 For education

Schwabish (2014)

3. Into the abyss

References

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arribas-Bel, D. and T. de Graaff (2015). "WooW-II: Workshop on open workflows". In: *REGION* 2.2, pp. 1–2.

Arribas-Bel, D., T. de Graaff, and S. Rey (Forthcoming). "Looking at John Snow's cholera map from the XXIst Century: a practical primer on reproducibility and Open Science". In: *Regional Research Frontiers: The Next 50 Years*. Ed. by R. Jackson and P. Schaeffer. Berlin: Springer.

Bayer, P. and C. Timmins (2007). "Estimating Equilibrium Models Of Sorting Across Locations". In: *The Economic Journal* 117.518, pp. 353–374.

Bayer, P., R. McMillan, and K. Rueben (2004). "An Equilibrium Model of Sorting in an Urban Housing Market". NBER working paper: No. w10865.

Bernasco, W., T. de Graaff, J. Rouwendal, and W. Steenbeek (Forthcoming). "Social Interactions and Crime Revisited: An Investigation Using Individual Offender Rates". In: *Review of Economics and Statistics*.

Bettencourt, L. and G. West (2010). "A unified theory of urban living". In: *Nature* 467.7318, pp. 912–913.

Breiman, L. (2001). "Statistical Modeling: The Two Cultures". In: *Statistical Science* 16.3, pp. 199–231.

De Graaff, T., F. G. van Oort, and R. J. Florax (2012a). "Regional Population-Employment Dynamics Across Different Sectors of the Economy". In: *Journal of Regional Science* 52.1, pp. 60–84.

— (2012b). "Sectoral heterogeneity, accessibility and population-employment dynamics in Dutch cities". In: *Journal of Transport Geography* 25, pp. 115–127.

Deaton, A. (2010). "Instruments, Randomization, and Learning about Development". In: *Journal of Economic Literature* 48, pp. 424–455.

Einav, L. and J. Levin (2014). "Economics in the age of big data". In: *Science* 346.6210, p. 1243089.

Rey, S. J. (2014). "Open regional science". In: *Annals of Regional Science* 52.3, pp. 825–837.

Roback, J. (1982). "Wages, rents, and the quality of life". In: *Journal of political Economy* 90.6, pp. 1257–1278.

Schwabish, J. A. (2014). "An economist's guide to visualizing data". In: *The Journal of Economic Perspectives* 28.1, pp. 209–233.

Thissen, M., T. de Graaff, and F. G. van Oort (2016). "Competitive network positions in trade and structural economic growth: A geographically weighted regression analysis for European regions". In: *Papers in Regional Science* 95.1, pp. 159–180.

Varian, H. R. (2014). "Big data: New tricks for econometrics". In: *The Journal of Economic Perspectives* 28.2, pp. 3–27.

Zhiling, W., T. de Graaff, and P. Nijkamp (2016). "Cultural Diversity and Cultural Distance as Choice Determinants of Migration Destination". In: *Spatial Economic Analysis* 11.2, pp. 176–200.