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

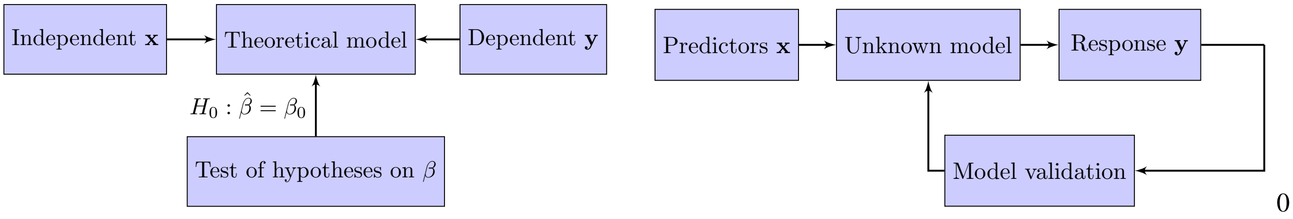
The quote above from Hal Varian is in one aspect wrong; nowadays, we do not call them statisticians but data scientists instead. Nevertheless, in the last two decades companies such as Google, Ebay, Whatsapp, Facebook, Booking.com and Airbnb, have not only witnessed enormous growth but to a considerable 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, $\mathbf{x}$ is referred to as the independent variable, $\mathbf{y}$ 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 $\mathbf{x}$ 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 $\mathbf{x}$ that lead to one or multiple reponses $\mathbf{y}$. 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.

The main aim is to estimate β as unbiased as possible. 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. 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 seach 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, 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).

1.1.2 Local average treatment effects

1.1.3 Fixed effects and heterogeneity

1.2 The blind eye in education

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).

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

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*.

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.