Economics Bulletin

Volume 39, Issue 4

Estimating Treatment Effects With Artificial Neural Nets – A Comparison to Synthetic Control Method

Arne Steinkraus
Institute of Economics Braunschweig

Abstract

With the advent of big data in economics machine learning algorithms become more and more appealing to economists. Despite some attempts of establishing artificial neural networks in in the early 1990s, only little is known about their ability of estimating causal effects. We employ a simple forecasting neural network to analyze the effect of the construction of the Oresund bridge on the local economy. The outcome is compared to the causal effect estimated by the proven Synthetic Control Method. Our results suggest that neural nets may outperform traditional approaches.

Institute of Economics at TU Braunschweig, Spielmannstr. 9, 38106 Braunschweig, Germany
Citation: Arne Steinkraus, (2019) "Estimating Treatment Effects With Artificial Neural Nets – A Comparison to Synthetic Control Method", *Economics Bulletin*, Volume 39, Issue 4, pages 2778-2791
Contact: Arne Steinkraus - a.steinkraus@tu-braunschweig.de.
Submitted: July 30, 2019. **Published:** December 08, 2019.

1 Introduction

Machine learning methods such as artificial neural networks (ANN) were introduced in economics in the early 1990s, but did not become accepted by a wide range of economists (see e.g. Kaastra and Boyd 1996 and Lee et al. 1993). Their failure was not a cause of bad performance, instead classical methods of nonparametric estimation of treatment effects such as propensity-score-matching or kernel methods performed well due to the small number of available covariates. Moreover, these approaches also allow to estimate heterogeneous treatment effects, which had been hard to achieve with machine learning methods (see e.g. Crump et al. 2008). Today, more data becomes available and datasets are getting increasingly bigger. Especially the use of geocoded data, the need for image classification or the use of language data in economics cause classical approaches to fail due to e.g. the curse of dimensionality or the need for high non-linearity (see Jean et al. 2016, Wager and Athey 2017). Whereas there has been much effort to overcome this issue in computer science, machine learning techniques were largely ignored in economics. However, starting with Athey´s approaches to use machine learning techniques for estimating (heterogeneous) causal effects there is a gain of recognition (see e.g. Athey 2017 and Athey and Imbens 2016, 2017). Among others, Blazquez and Domenech (2018) as well as Obschonka and Audretsch (2019) provide excellent overviews of machine learning techniques, (un-)structured big data sources and their impact on economic analysis and decision making. Nevertheless, machine learning techniques remain fundamentally different from classical approaches of estimating causal effects such as Regression Discontinuity Design or Instrumental Variables because their goal is prediction – not causal inference (Athey and Imbens 2017).

We contribute to the causal inference and machine learning literature by assessing the practical relevance of ANN in ex-post treatment evaluation in comparison to the new quasi-gold-standard Synthetic Control Method (SCM) which was introduced by Abadie and Gardeazabal (2003). This comparison is of high relevance because according to Athey and Imbens (2017) SCM has been *"arguably the most important innovation in the policy evaluation literature in the last 15 years"*. However, SCM fails when applied to big datasets. We proceed in two steps. First, we describe the identification strategy of both methods: SCM and the nonlinear regression framework of simple forecasting ANNs. Afterwards, we evaluate the performance of ANN compared to the state-of-the-art SCM by estimating the causal effect of the construction of the Oresund-Bridge, that connects the metropolitan regions of Malmo (Sweden) and Copenhagen (Denmark). Our ANN finds treatment effects on local and regional level that are similar to those estimated by SCM. Therefore, it seems to be a suitable method in ex-post treatment evaluation.

2 Methods and Data

In our study, we aim to estimate treatment effects. Therefore, we need to have a closer look at the potential outcome framework (Rubin 1974). In the so called Rubin causal model, a treatment effect TR_i is defined as the difference between two potential outcomes Y_i^I and Y_i^N, where Y_i^I denotes the outcome of individual i in a state with treatment and Y_i^N is the potential outcome of individual i without treatment:

$$TR_i = Y_i^I - Y_i^N \qquad (1)$$

Unfortunately, we never directly observe TR_i because Y_i^N cannot be realized in a state where individual i is exposed to treatment (vice versa). This phenomenon is called "Hollands (1986) fundamental problem of causal inference". In order to calculate the treatment effect, we need to predict Y_i^N via ANN or SCM in a consistent manner.

2.1 Artificial Neural Network

Due to the fact that ANN belongs to the supervised learning regimes of machine learning labelled training data is mandatory. Therefore, we employ a forecasting version of an ANN that uses economic growth predictor variables of several regions at time t for training in order to forecast the regions outcome at time $t + 5$.[1] This approach also safeguards against possible simultaneity issues. Thus, our ANN learns how growth predictors forecast economic outcome. Afterwards, we employ the trained network on pre-treatment predictor data of Malmo and Södra Sverige. The forecasted outcomes serve as a proxy for Y_i^N.

This approach borrows the assets of three papers. Specifically, the identification strategy is inspired by Foster et al. (2010) – who aim to identify heterogeneous treatment effects along the covariate space X by estimating $\mathbb{E}(Y_i^I|X_i = x)$ and $\mathbb{E}(Y_i^N|X_i = x)$ separately via random forests – and Burlig et al. (2017) – who use machine learning technique and classical difference in differences estimation strategy to estimate the treatment effect of energy efficiency upgrades on electricity consumption. In addition, we build on Sokolov-Mladenović et al. (2016) who feed an ANN with trade data to predict economic growth.

To be more general, a neural network is a nonlinear regression technique that is modelled by an unobserved set of so-called hidden nodes. Suppose X_i is a $(P + 1 \times 1)$ vector of P observed standardized predictor variables and one bias term that serves as the ANN equivalent of the intercept in a regression. This vector enters the $P + 1$ input nodes of our artificial brain. Weighted combinations of the input nodes are then transferred as inputs signals to the hidden nodes:

$$H_{i,input} = W_{input,hidden} \cdot X_i \tag{2}$$

where $W_{input,hidden}$ is a $(R \times P + 1)$ weight-matrix and $H_{i,input}$ denotes a $(R \times 1)$ vector that enters the R hidden nodes. At the hidden nodes input signals are transformed by a sigmoidal function:[2]

$$H_{i,output} = \frac{1}{1 + e^{-H_{i,input}}}. \tag{3}$$

The output of the hidden nodes $H_{i,output}$ is weighted by the $(1 \times R)$ vector $W_{hidden,output}$ and transformed by the sigmoidal function again. Thus, the standardized predicted outcome $\widehat{Y_i}$ is given by:

$$\widehat{Y_i} = \frac{1}{1 + e^{-(W_{hidden,output} \cdot H_{i,output})}} \tag{4}$$

Initially, the elements of the weight-matrices are randomly drawn from a normal distribution with mean zero and standard deviation $\sqrt{R}$ and 1 respectively. The subsequent training of the neural network employs the so-called back-propagation technique (BP) on the training sample that contains approx. 90 % of all available data (see Rumelhart et al. 1986). We chose BP because it is the standard approach in forecasting settings. In a first step, the squared prediction error E_{output} is calculated in a textbook like fashion as $\left(Y_i - \hat{Y}\right)^2$ and backward distributed across the hidden nodes in order to calculate the hidden error E_{hidden}:

$$E_{hidden} = W_{hidden,output}^{T} \cdot E_{output} \tag{5}$$

[1] We consider a five year forecast because they roughly span one business cycle.

[2] Other transfer-function such as arc-tan, step or linear exists but are seldom employed.

In a second step, we need to update the elements of both weight-matrices in order to minimize the error terms using gradient descent method. At this stage the advantage of the sigmoid function becomes obvious because of it is continuously differentiable and its deviation is easy to implement.[3] Thus the updated matrices are given by:

$$W_{hidden,output(\ m+1)} = W_{hidden,output(m)} + \varphi \cdot E_{ouptut} \cdot \hat{Y} \cdot \left(1 - \hat{Y}\right) \cdot H_{output}^{T} \tag{6a}$$

$$W_{input,hidden(m+1)} = W_{input,hidden(m)} + \ \varphi \cdot E_{hidden} \cdot H_{output} \cdot \left(1 - H_{output}\right) \cdot X^{T} \tag{6b}$$

Where the index m denotes the training stage and φ is the learning rate. Several approaches for choosing the optimal number of hidden nodes as well as the learning rate such as fixed, constructive and destructive exist. Most of them got in common that they aim to minimize the mean squared prediction error in the test sample. Therefore, we refuse from reporting these methods in detail and point to Kaastra and Boyd 1996 who provide a detailed overview.

The final matrices can be employed to forecast or predict the outcome variable of interest. In our case study we use economic growth predictor variables as input factors and Gross Domestic Product as outcome. Thus, the ANN learns how economic growth can be forecasted by finding the optimal non-linear combination of all predictor variables. Consequently, ANN provide just another but much more flexible way to solve the ordinary least squares (OLS) problem.

The superiority of ANN over classical econometric approaches results from the following fact. In usual regression analysis a so-called saturated model is needed to claim causality of treatment effects. However, when it comes to many variables the number of observations that is need to solve the closed-form saturated OLS regression increases rapidly. Therefore, researches often refuse from reporting saturated models and restrict their models to a small number of variables and interaction terms. Due to the fact that a closed-form solution is not necessary in case of ANN, any kind of non-linear variable interaction is incorporated in the model by construction. As a consequence, the resulting network can be interpreted as the machine learning equivalent to the econometric saturated OLS. Moreover, ANN can tolerate fat tails and noise, do not require strong assumptions regarding the error term, can adapt to new patterns and even incorporate observations with missing variables by including dummy nodes (Masters 1993, Kaastra and Boyd 1996, Tkáč and Verner 2016).

On the contrary, their black-box character remains immanent. Thus, it is hard to provide statistics such a standard errors or p-values.

2.2 Synthetic Control Method

The basic idea behind SCM is to build on classical difference in differences estimation but to select comparison units and assign weights based on a data driven approach. This feature is advantageous in comparison to ANN because results claim causality by construction. Therefore, in SCM, as introduced by Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015), it is assumed that $Y_{i,t}^{N}$ is given by the following factor model:

$$Y_{i,t}^{N} = \delta_t + \Theta_t Z_i + \lambda_t \mu_i + \varepsilon_{i,t}, \tag{7}$$

where δ_t denotes an unknown common factor, Θ_t is a $(1 \times r)$ vector of unknown parameters, Z_i is a $(r \times 1)$ vector of observed but unaffected predictors, λ_t denotes a time varying $(1 \times F)$ vector of unobserved common factors, μ_i is a $(F \times 1)$ vector of unknown factor loadings and $\varepsilon_{i,t}$ are zero mean unobserved shocks. Suppose that our sample consists of $i = 1, \dots, J$ units of

[3] The deviation is given by: $\frac{d\frac{1}{1+e^{-x}}}{dx} = \frac{1}{1+e^{-x}} \cdot \left(\mathbf{1} - \frac{1}{1+e^{-x}}\right)$

which only the first unit is exposed to the treatment and all other units serve as possible donors. All units are observed at dates $t = 1,..,T$. $[1, T_0]$ denotes the entire pre-treatment period so that $T_0 \in [1, T[$. Consider a $(J - 1 \times 1)$ vector of non-negative weights $W = (w_2, \dots, w_J)$, whose elements sum up to one. Each realisation of W represents a potential synthetic control unit. The resulting synthetic control units come with the following value of their outcome variable:

$$\sum_{j=2}^{J} w_j Y_{j,t} = \delta_t + \Theta_t \sum_{j=2}^{J} w_j Z_j + \lambda_t \sum_{J=2}^{J} \mu_j + \sum_{j=2}^{J} \varepsilon_{j,t} \tag{8}$$

Since SCM aims to assign weights to donors according to their similarity to the treated unit, we need to find the optimal set of weights W^* such that pre-intervention matching:

$$\sum_{j=2}^{J} w_j Y_{j,t} = Y_{1,t} \ \forall \, t = 1, \dots, T_0 \tag{9}$$

$$\sum_{j=2}^{J} w_j Z_j = Z_1 \tag{10}$$

is achieved at least approximately. Such a W^* does exist if $(Y_{1,1}, \dots, Y_{1,T_0}, Z_1')$ is not too far away from the convex hull of $\{(Y_{2,1}, \dots, Y_{2,T_0}, Z_2'), \dots, (Y_{J,1}, \dots, Y_{J,T_0}, Z_J')\}$. In this case, and under standard conditions $\sum_{j=2}^{J} w_j^* Y_{j,t}$ can be used as an estimator of $\boldsymbol{Y}_{\mathbf{1}}^{\boldsymbol{N}}$ for $T_0 < t \leq T$.

A suitable procedure to obtain W^* is described as follows. Define X_1 as a $(M \times 1)$ vector of pre-intervention values of predictor and outcome variables for the treated unit 1. Let X denote a $(M \times J - 1)$ matrix of the same variables for the $J - 1$ units from the donor pool. The optimal weights W^* are chosen to minimize the weighted distance between X_1 and X:

$$W^* = argmin_W \, (X_1 - XW)' V (X_1 - XW), \tag{11}$$

where V is a non-negative semidefinite $(M \times M)$ matrix whose diagonal elements reflect the importance of each considered predictor variable. At this step, the optimal $W^*(V)$ depends on the choice of the relative predictor importance. Among all possible matrices V, V^* is chosen to minimize the residual mean squared prediction error (RMSPE) of the outcome variable during the pre-intervention period:

$$V^* = argmin_V \, (Y_1 - YW^*(V))'(Y_1 - YW^*(V)), \tag{12}$$

where Y_1 is a $(T_0 \times 1)$ vector of pre-treatment outcomes for unit 1 and Y is a $(T_0 \times J)$ matrix that contains the pre-treatment outcomes for the donor units. Since there are infinitely many collinear solutions of V^*, the Euclidean norm of V^* is normalized to one. The optimal weights are given by $W^*(V^*)$. This synthetic control unit, which comes as a weighted average of units from the donor pool, is the best to reproduce unit 1´s trajectory in the absence of treatment. Thus, SCM provides a synthetic counterfactual as a convex combination of control regions.

Similar to ANN, the procedure of SCM allows us to perform so-called placebo studies to test the significance of the treatment effect (Abadie et al. 2015, and Abadie et al. 2010, Munasib and Rickman 2015).

However, when applied to real world data, SCM suffers from the following three shortcoming: First, Kaul et al. (2016) argue that pre-treatment matching is often achieved only if pre-treatment outcome variables are included as additional predictor variables. This procedure turns the other predictor variables irrelevant and reduces the credibility of the final results. Second, SCM suffers from severe reproducibility problems. Klößner et al. (2017) show that the choice of software package as well as the order of observations in the dataset are influential factors in determining the estimated treatment effect. Third, when it comes to big datasets or many control units, the execution of SCM is inefficient, causes standard computers to run into RAM limits and does not necessarily provide efficient estimates. By and large, these features diminish the practical relevance and plausibility of SCM and highlight the importance of adapting ANNs to the evaluation of treatment effects.

2.3 Data

To be consistent with the approach of Achten et al. (2018) we feed or ANN with data from the Cambridge Econometrics European Regional Database (ERD). Although it provides economic indicators over a period between 1980 and 2014, we limit the period of investigation to the year 2005 – five years after the opening of the bridge. We do so because our ANN gives us a five year forecast of Gross Domestic Product per Capita (GDPpC) and using post-treatment predictor variables would cause severe endogeneity issues. We employ investment share, sectoral shares of value added, population density and compensation of employees as fundamental predictors and the pre-treatment level of the GDPpC as technical predictor. These variables relate to Barro's (1991) economic growth predictors.

As in Achten et al. (2018), we also consider only those regions in our training and test sample for which data is available from 1980 onwards. If spillover effects are relevant in fact, the inclusion of Germany and Scandinavia would induce biased estimates. Therefore, and due to the fact that Jean et al. (2016), who focus on convolutional neural networks and satellite imagery, argue that the exclusion of the states of interest reduces predictive power only modestly, we exclude all German and Scandinavian regions. To make the outcome of SCM and ANN comparable and to guarantee causality, we also drop all observations that were exposed to large infrastructure investment during entire period. Our final sample consists of nearly 6.000 suitable observations. However, since we employ ten-fold cross-validation to avoid overfitting, our training- and test-samples contain 5.400 and 600 observations respectively

3 Results

Based on the pyramids rule (see Master 1993) and due to the fact that information is densified across the layers, we assume the optimal number of hidden nodes to be between 11 and 2. We employ a so called grid-search and to identify the optimal fixed learning rate and exact number of hidden nodes. Since this procedure examines (almost) all available combinations, we select the configuration that minimizes the mean of squared residuals in the test-samples using ten-fold cross-validation. Our final network contains 6 hidden nodes and is endowed with a learning rate of 0.5. At this stage, we need to mention that the BP algorithm – although it converges – does not necessarily find the global optimum solution and may stuck in local optima. However, we aim to overcome this problem, by using randomly chosen initial weights, repeating the algorithm several times and averaging the outcomes. Despite this shortcoming, we employed this algorithm because it is easy to implement, highly efficient and does not require strong assumptions regarding the error term. Similar to other machine learning techniques ANNs also tend to overfit. However, the number of weights in the network is small compared to the number of observations and we find good results in our cross-validation approach. Thus, refuse from pruning high weights and model complexity. In a robustness check, we also considered a deep neural network consisting of two hidden layers. We refuse from reporting the results, because it seemed to memorize instead of learning patterns so that it showed severe overfitting.

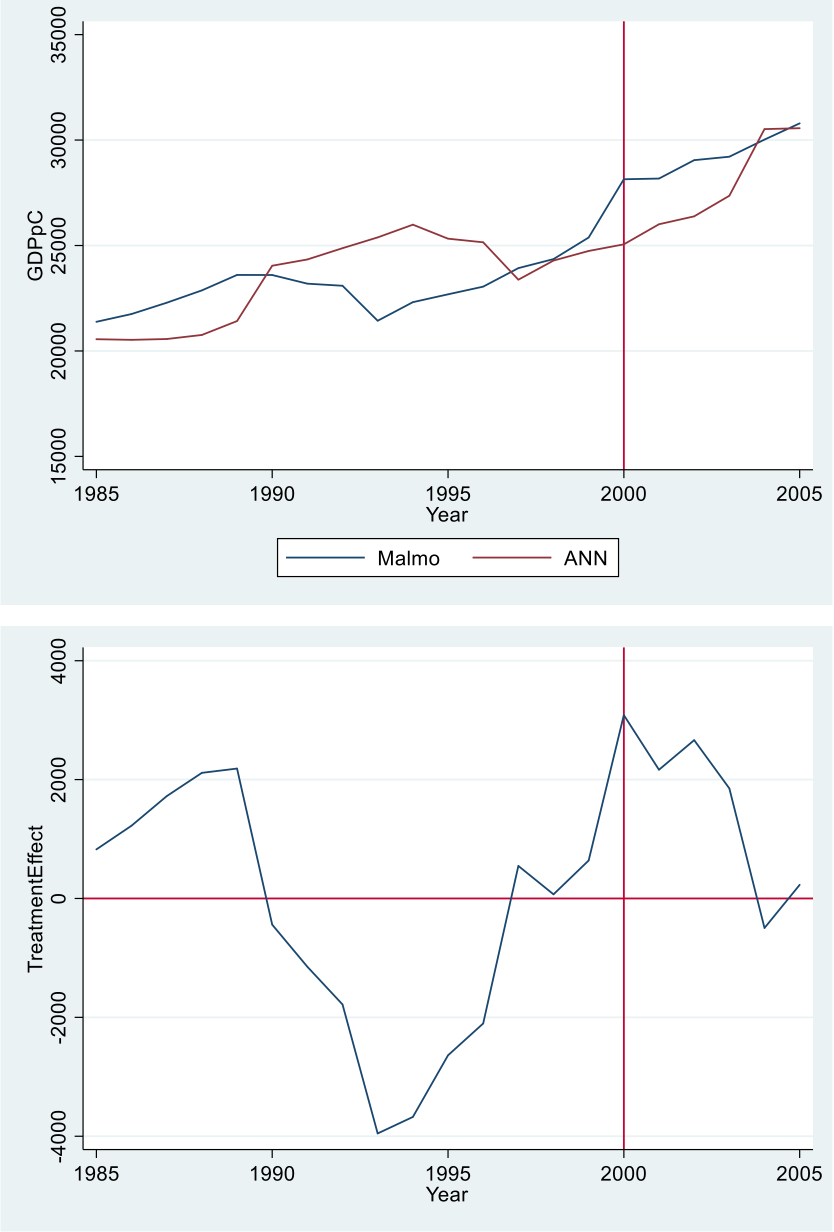

Figure 1: ANN forecast of Malmo and treatment effect

To avoid the above-mentioned local optima issue we repeat the forecasting of Malmo´s (local level NUTS[4] 2) and Södra Sverige´s (regional level NUTS 1) GDPpC trajectories one hundred times and calculate the averages. The results of our forecasting approach as well as the real trajectories are depicted in the upper plots of Figures 1 (Malmo) and 2 (Södra Sverige). To make a potential treatment effect become apparent, we also calculated the differences between the actual outcomes and the forecasted versions. The results are shown in the lower plots of Figures 1 and 2.

[4] Nomenclature of territorial units for statistics.

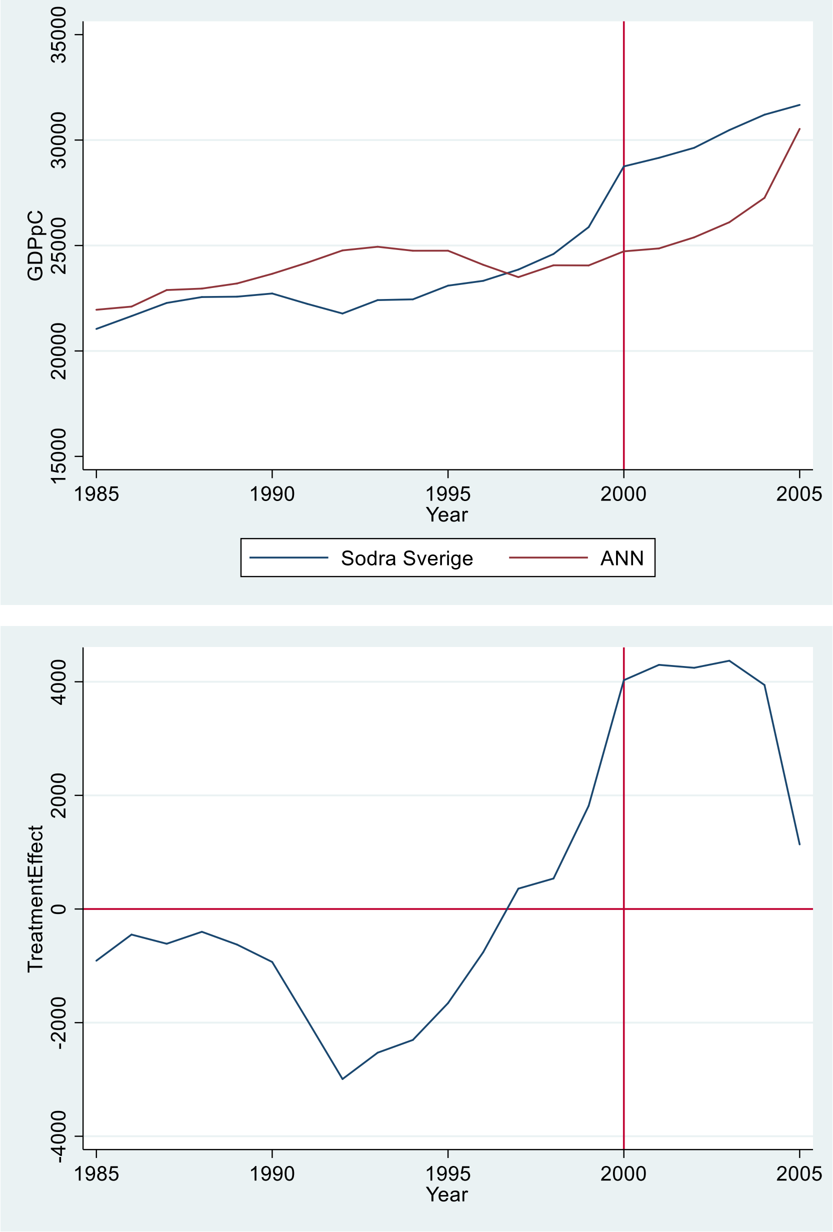

Figure 2: ANN forecast of Södra Sverige and treatment effect

For both regions, it turns out that the ANN estimation strategy reveals distinct positive treatment effects around the year 2000. To compare the results of our ANN with the state-of-the-art SCM approach, the outcomes of SCM are shown in Figures 3 (Malmo) and 4 (Södra Sverige).[5]

[5] Please note that the lower plots of Figures 3 and 4 also contain placebo treatment effects (grey).

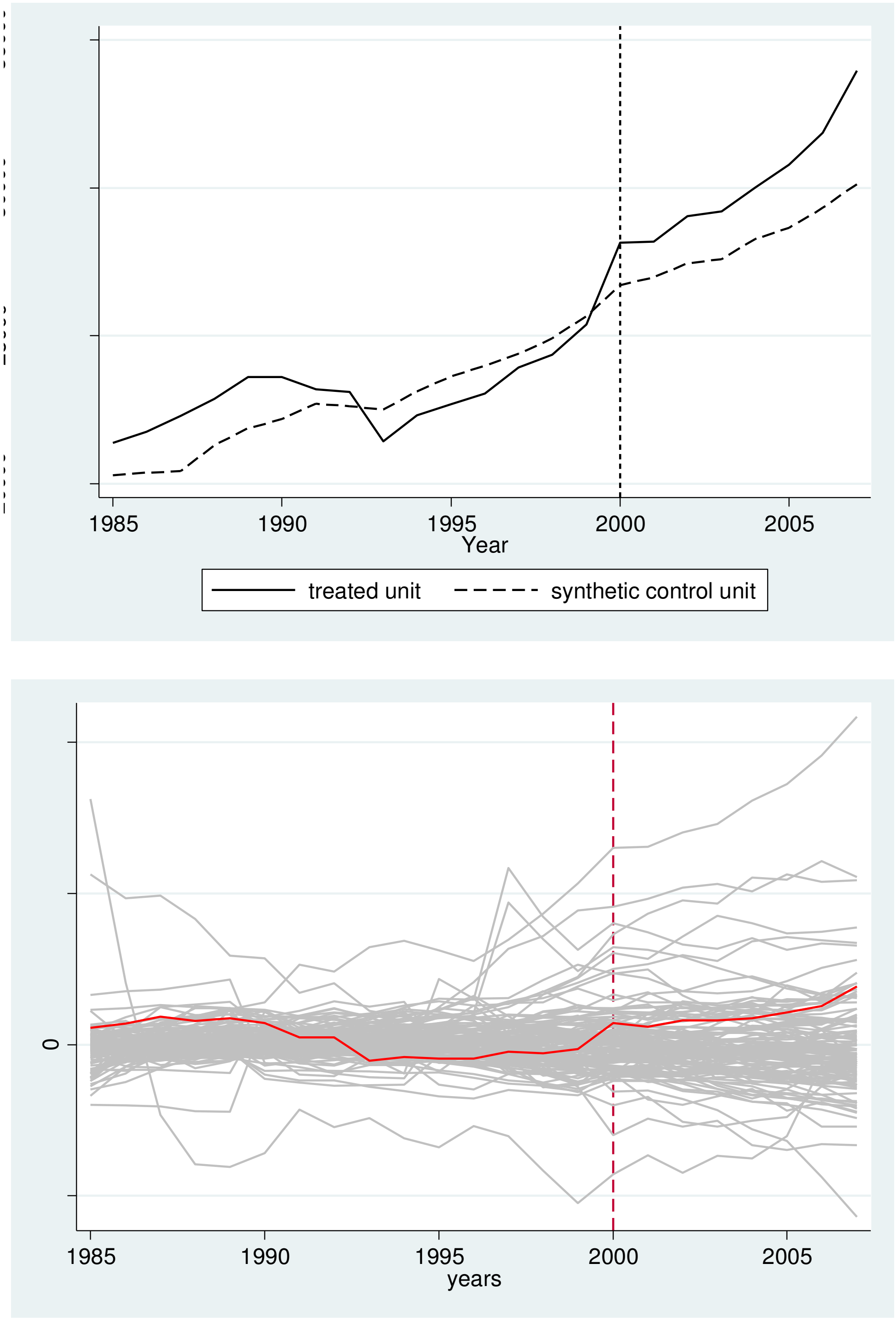

Figure 3: SCM results of Malmo - taken from Achten et al. 2018

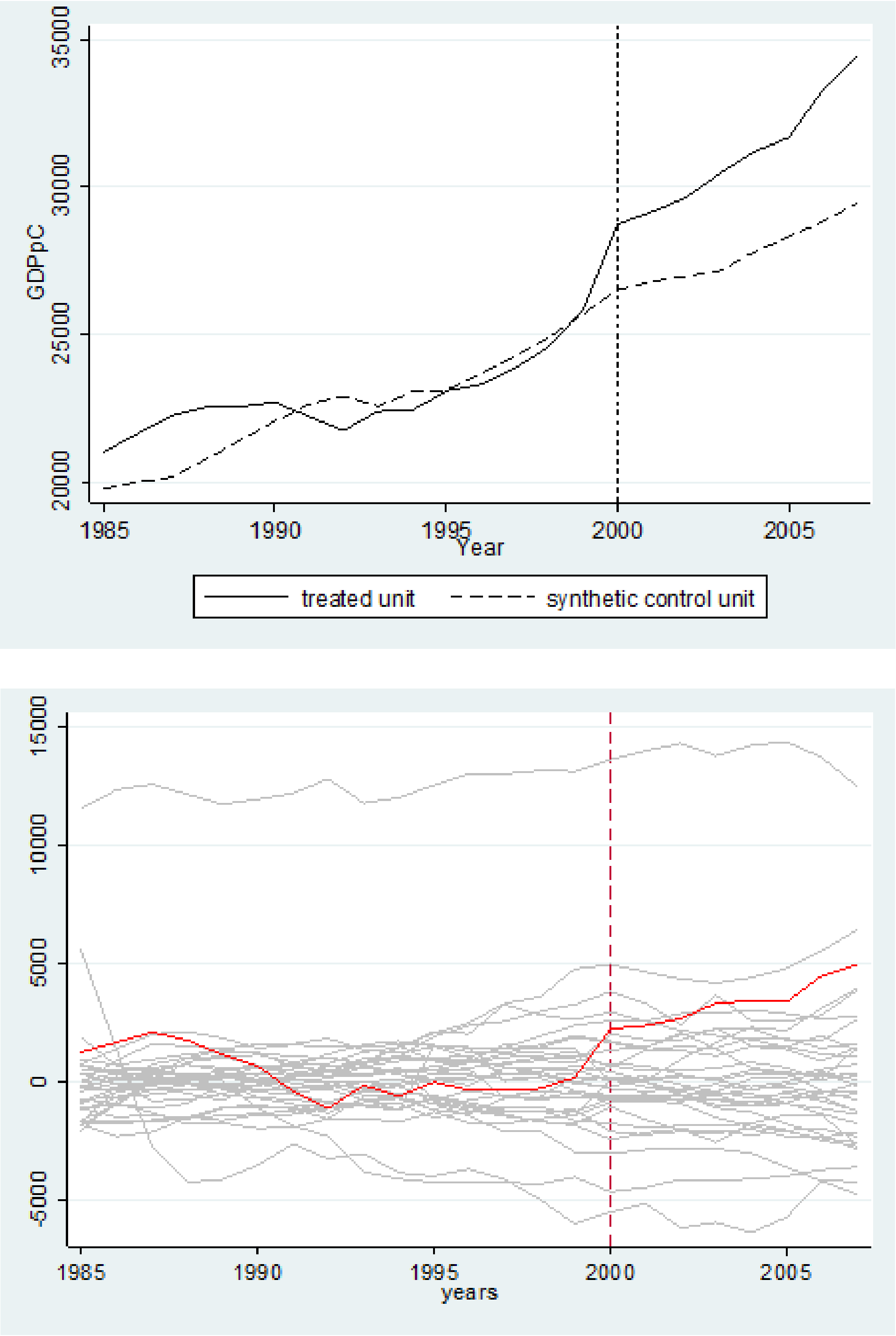

Figure 4: SCM results of Södra Sverige - taken from Achten et al. 2018

It becomes obvious that the SCM estimation strategies also reveals positive treatment effect in 2000. Additionally, the ANN confirms the previous SCM results of a more pronounced effect at the regional level (Södra Sverige) that arises from potential spill-over effects.

Since neither our ANN nor the SCM of Achten et al. (2018) do perfectly replicate the pre-treatment path of GDPpC, we apply the difference-in-differences treatment estimator TR as suggested by Bohn et al. 2014:

$$TR = \left(\overline{Y_{post}^{I} - Y_{post}^{N}}\right) - \left(\overline{Y_{pre}^{I} - Y_{pre}^{N}}\right) \qquad (13)$$

where Y_{post}^{I} is the actual post-intervention outcome , Y_{post}^{N} is the counterfactual post-treatment outcome, Y_{pre}^{I} denotes the actual and Y_{pre}^{N} the counterfactual pre-treatment outcome. Considering Malmo our ANN TR amounts to 1.489 € (in 2005 terms) and is very close to the SCM estimate (1.680 €). For Södra Sverige the ANN TR (4.133 €) is approximately 38 % larger than the SCM TR (2.988 €). However, the discrepancy may arise from the fact that the period of consideration in the SCM approach is 2 years longer and our ANN counterfactual indicates a decrease in treatment effect starting in 2005.

As a robustness checks, we exclude the year of prediction as input variable in our training procedure to prevent the network from memorizing Europe-wide exogeneous economic shocks. Instead, it now has to learn more general patterns from the economic predictor variables. The results are reported in the appendix and confirm our prior results.

Since both identification strategies reveal a distinct positive effect of the construction of Oresund, we are confident that this large-scale infrastructure investment did not only reduce transport cost but contributed substantially to market integration. The indication of spill-over effects also highlights the importance of the investment for entire Sweden (and for Denmark). Consequently, the estimated outcomes may serve as an argument for future projects such as the Trans-European Transport Network.

4 Conclusion

In this study we showed that ANNs are powerful tools for solving the least squares problem in a highly nonlinear fashion. Moreover, by applying a simple forecasting ANN to a case study, namely the construction of the Oresund bridge, we find that this single-layer feed forward neural network yields similar results as the much more advanced and appreciated SCM. Considering that ANNs are said to perform notably well in settings where there are many covariates relative to the number of observations, this result becomes even more astonishing. However, we need to mention two limitations of our approach: (1) the lack of hypotheses testing options. Since we cannot test whether our treatment effects are statistically significant different from zero, our estimated outcomes might be stochastic noise purely. To overcome this issue at least partly, we relied on goodness of fit measures during our grid search for parameter tuning. (2) Our approach belongs to the co-called selection on observables identification strategies. Therefore, it is important to include all relevant variables in the predictor set. We are confident that our Barrow-style ANN suffices this request.

With the advent of big data in economics the importance of ANN will continue to rise. Specifically, since SCM and other popular approaches such as LASSO or elastic nets either become inefficient or hard to interpret as the number of covariates increases, there is a need in economics to borrow machine learning techniques from computer scientist in order to adapt them for policy evaluation or even prediction policy purposes. Specifically, since ANNs have strong prediction and forecasting abilities, they can be very useful in prediction policy problems. According to Kleinberg et al. (2015) prediction policy problems arise, if treatment effects depend e.g. on the expected future outcome of another variable which can be forecasted via ANNs. The link between the prediction policy problems and infrastructure investments is as follows. If we expect a future economic downturn, large infrastructure investments may be less fruitful because of lower spill-over and network effects. Thus, ANNs can be used to forecast future (economic) outcomes in a non-treatment scenario to ex-ante evaluate the efficiency of treatments. Consequently, future research may, among others, build on studies that identify heterogeneous treatment effects, predict future covariate outcomes and, thereupon, use both information to make recommendation regarding treatment allocation.

However, up to now, an ANN remains a black box without a well-understood sampling distribution what makes it hard to test hypotheses. In addition, other statistical properties such as unbiasedness and consistency of estimated effects are unknown and need to be analysed via Monte Carlo simulations in further studies. Although, there remains much work to establish causal ANN in economics we are confident that this paper excites economists and econometricians to explore the field of machine learning.

Literature

Abadie, A., Diamond, A., & Hainmueller, J. (2010) “Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program” Journal of the American Statistical Association, **105**, 493-505.

Abadie, A., Diamond, A., & Hainmueller, J. (2015) “Comparative politics and the synthetic control method” American Journal of Political Science, **59**(2), 495-510.

Abadie, A., & Gardeazabal, J. (2003) “The economic costs of conflict: A case study of the Basque Country” The American Economic Review, **93**(1), 113-132.

Aschauer, D. A. (1989a) “Does public capital crowd out private capital?” Journal of Monetary Economics, **23**(2), 177-200.

Aschauer, D. A. (1989b) ”Public Investments and productivity growth in the Group of Seven” Economic Perspectives, **13**(5), 17-25.

Athey, S., & Imbens, G. (2016) “Recursive partitioning for heterogeneous causal effects” Proceedings of the National Academy of Sciences, **113**(27), 7353-7360.

Athey, S. (2017) “Beyond prediction: Using big data for policy problems” Science, **355**(6324), 483-485.

Athey, S., & Imbens, G. W. (2017) “The state of applied econometrics: Causality and policy evaluation” Journal of Economic Perspectives, **31**(2), 3-32.

Barro, R. J. (1991) “Economic growth in a cross section of countries” The Quarterly Journal of Economics, **106**(2), 407-443.

Blazquez, D., & Domenech, J. (2018) “Big Data sources and methods for social and economic analyses” Technological Forecasting and Social Change, **130**, 99-113.

Bohn, S., Lofstrom, M., & Raphael, S. (2014) “Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population?” Review of Economics and Statistics, **96**(2), 258-269.

Buchmueller, T. C., DiNardo, J., & Valletta, R. G. (2011) “The effect of an employer health insurance mandate on health insurance coverage and the demand for labor: Evidence from hawaii” American Economic Journal: Economic Policy, **3**(4), 25-51.

Burlig, F., Knittel, C., Rapson, D., Reguant, M., & Wolfram, C. (2017) “Machine learning from schools about energy efficiency” National Bureau of Economic Research (No. w23908).

Cantos, P., Gumbau-Albert, M., & Maudos, J. (2005) “Transport infrastructures, spillover effects and regional growth: Evidence of the Spanish case” Transport Reviews, **25**(1), 25-50.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008) ”Nonparametric tests for treatment effect heterogeneity” The Review of Economics and Statistics, **90**(3), 389-405.

Foster, J. C., Taylor, J. M., & Ruberg, S. J. (2011) ”Subgroup identification from randomized clinical trial data” Statistics in medicine, **30**(24), 2867-2880.

Holland P.W. (1986) “Statistics and Causal Inference” Journal of the American Statistical Association, **81**(396), 945-960.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016) “Combining satellite imagery and machine learning to predict poverty” Science, **353**(6301), 790-794.

Kaastra, I., & Boyd, M. (1996) “Designing a neural network for forecasting financial and economic time series” Neurocomputing, **10**(3), 215-236.

Kaul, A., Klößner, S., Pfeifer, G., & Schieler, M. (2016) “Synthetic Control Methods: Never Use All Pre-Intervention Outcomes as Economic Predictors”.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015) „Prediction policy problems” American Economic Review, **105**(5), 491-95.

Klößner, S., Kaul, A., Pfeifer, G., & Schieler, M. (2017) „Comparative politics and the synthetic control method revisited: A note on Abadie et al.(2015)” Swiss Journal of Economics and Statistics.

Lee, T. H., White, H., & Granger, C. W. (1993) “Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests” Journal of Econometrics, **56**(3), 269-290.

Masters, T. (1993) “Practical neural network recipes in C++” Morgan Kaufmann.

Munasib, A., & Rickman, D. S. (2015) “Regional economic impacts of the shale gas and tight oil boom: A synthetic control analysis” Regional Science and Urban Economics, **50**, 1-17.

Obschonka, M., & Audretsch, D. B. (2019) “Artificial intelligence and big data in entrepreneurship: a new era has begun” Small Business Economics, 1-11.

Pereira, A. M., & Andraz, J. M. (2013) “On the economic effects of public infrastructure investment: A survey of the international evidence” Journal of Economic Development, **38**(4), 1-37.

Pereira, A. M., & Andraz, J. M. (2004) “Public highway spending and state spillovers in the USA” Applied Economics Letters, **11**(12), 785-788.

Pinotti, P. (2015) “The economic costs of organised crime: Evidence from Southern Italy” The Economic Journal, **125**(586), F203-F232.

Rubin, D. B. (1974) “Estimating causal effects of treatments in randomized and nonrandomized studies” Journal of educational Psychology, **66**(5), 688-701.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986) “Learning representations by back-propagating errors” Nature, **323**(6088), 533-536.

Sokolov-Mladenović, S., Milovančević, M., Mladenović, I., & Alizamir, M. (2016) “Economic growth forecasting by artificial neural network with extreme learning machine based on trade, import and export parameters” Computers in Human Behavior, **65**, 43-45.

Tkáč, M., & Verner, R. (2016) “Artificial neural networks in business: Two decades of research” Applied Soft Computing, **38**, 788-804.

Wager, S., & Athey, S. (2017) “Estimation and inference of heterogeneous treatment effects using random forests” Journal of the American Statistical Association (in press).

Appendix

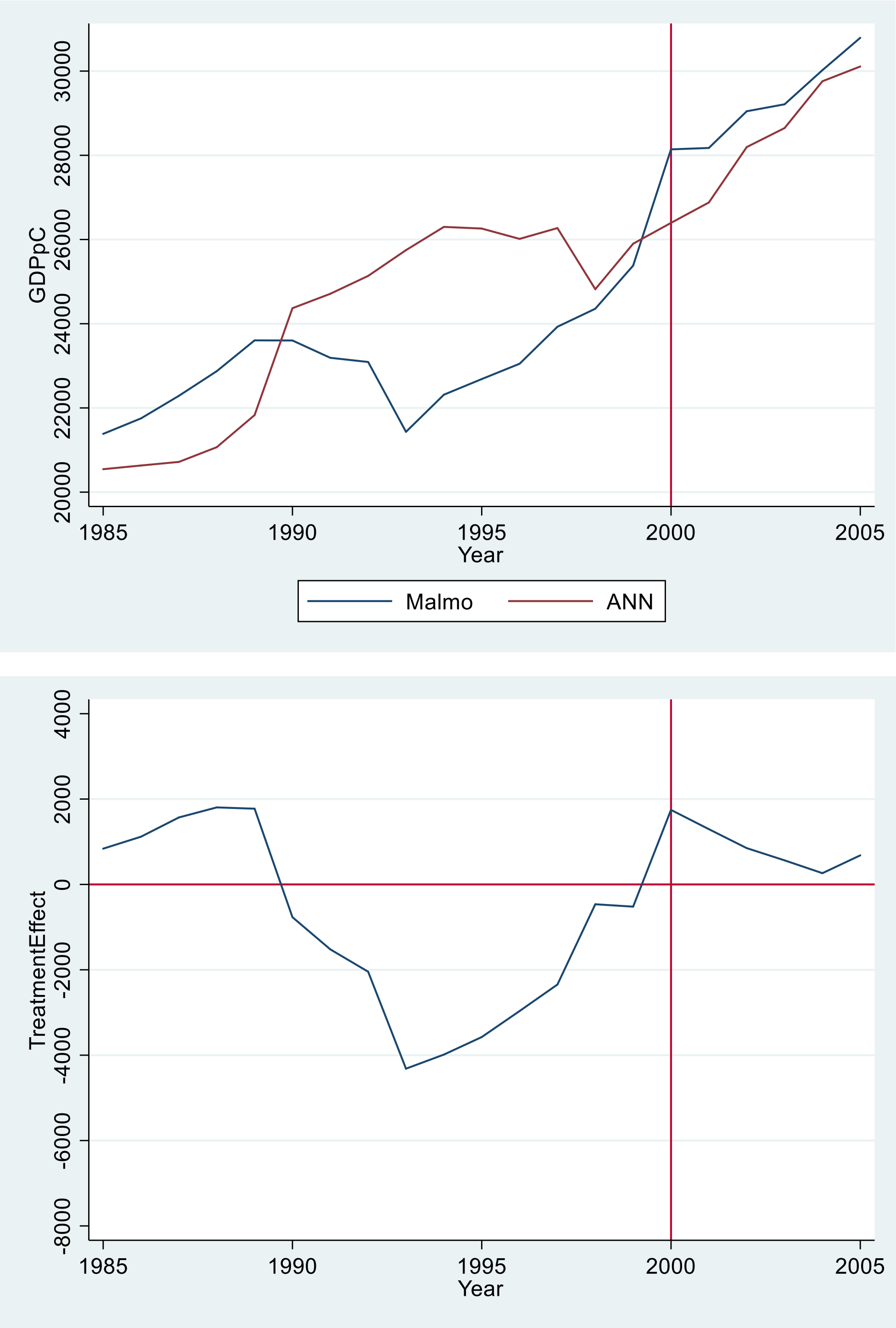

Figure A 1: ANN forecast of Malmo and treatment effect – Year excluded from set of predictors: TR: 1.583 €

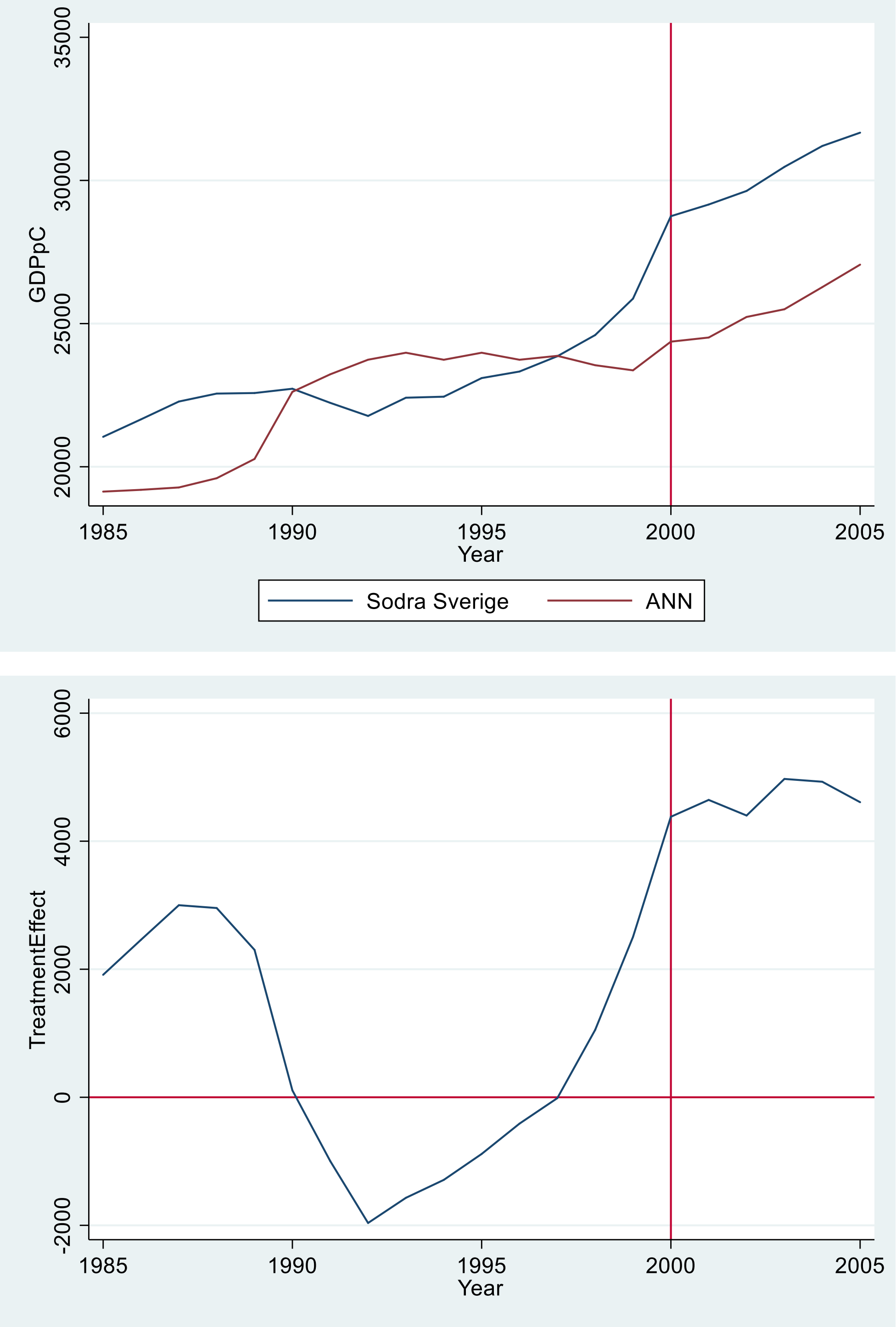

Figure A 2: ANN of Södra Sverige and treatment effect – Year excluded from set of predictors: TR: 3.863 €