Annex 3: Counterfactual analysis

3.1 Methods' description and contextualisation of the case study

3.1.1 Method

The counterfactual methodology will help to quantify the effects of the high-speed rail (HSR) policy on the treated territories within the Milan-Bologna macro-regional corridor (i.e. Reggio Emilia AV Mediopadana) compared to other areas that did not receive directly the policy (i.e. Piacenza, Pavia). In particular, basing on a similar recent study that evaluated the transport investments in the Madrid-Barcelona high-speed rail corridor (Carbo *et al*., 2019), we are interested to evaluate the effects of the HSR program on various types of potential outcomes: economic growth (expressed in per capita gross domestic product); employment growth[1] (share of employed persons in the local firms) labour productivity (expressed in gross value added per employed person), patents intensity (number of patents weighted by the resident population), business demography (number of enterprises), transports intensity (difference in road freight transport between loading and unloading)[2].

Basing on a longitudinal data availability from national and European sources (ISTAT, EUROSTAT, OECD), data will be handled at NUTS-3 level, in order to capture the sprawl of the policy effects at a provincial scale, and two methods will be adopted:

i. the hybrid **Matching–Difference-in-Differences** (M–DID) estimator, which will be used to validate the research scenario and to produce a quantitative output (numerical values with the relative statistical significance) and it will use panel data

ii. the **Synthetic Control** (SC), which will produce a graphical representation of the effects of the policy

The DID estimator $\hat{\alpha}_{DID}$, in its general computation for panel data, is intended as follows:

$$\hat{\alpha}_{DID} = (\bar{y}_{i1}^{T} - \bar{y}_{i1}^{C}) - (\bar{y}_{i0}^{T} - \bar{y}_{i0}^{C}) \tag{1}$$

where $\bar{y}_{i1}^{T}$ is the *y* average on the treated at *t*=1; $\bar{y}_{i1}^{C}$ is the *y* average on the controls at *t*=1; $\bar{y}_{i0}^{T}$ is the *y* average on the treated at *t*=0; $\bar{y}_{i0}^{C}$ is the *y* average on the controls at *t*=0, which is equivalently the before/after estimator for the treated/control groups:

$$\hat{\alpha}_{DID} = (\hat{a}_{BA}^{T} - \hat{a}_{BA}^{C}) \tag{2}$$

thus, the potential outcome of the model in case of panel data will be:

---

[1] On the basis of the NACE classification of the economic activities in Italy (ATECO 2007) it will be possible to evaluate the change of employment for different sectors, i.e. knowledge intensive vs labour intensive sectors, tourism, manufacturing, creative industry, logistics and so forth.

[2] The list is not to be considered exhaustive, further desirable potential outcomes can be added basing upon stakeholders' requests.

$$Y_{it} = \gamma_i + \lambda_t + D_{it}\delta + \beta x_{it} + e_{it} \tag{3}$$

where Y_{it} is one of the possible aforementioned potential outcomes; γ_i is a specific geographical effect given by the panel structure of the data; λ_t is a time effect on the basis of the number of years considered into the analysis (pre- and post-program periods); $D_{it}\delta$ is the DID treatment estimator, βx_{it} is a set of additional covariates[3] that will be useful to build a Kernel propensity-score for taking into account the different extents of the territories included in the panel, and this will help to capture eventual spatial lags between the territories, controlling for possible exogenous effects, as well as this will allow to build a randomized quasi-experimental scenario with very low bias; e_{it} is an error term of the OLS. The final M–DID will be a non-parametric DID reweighted (Cerulli, 2014) on a Kernel function determined for the Matching procedure, where the final average treatment effect on the treated units (ATET) will be:

$$\widehat{ATET}_{M-DID} = \frac{1}{N_1} \sum_{i \in \{T\}} \left((Y_{i1}^T - Y_{i0}^T) - \sum_{j \in C(i)} h(i,j)(Y_{i1}^C - Y_{i1}^C) \right) \tag{4}$$

where *t*=0,1 is the time before/after the program; T/C is the treated/control set of units; $h(i,j)$ are the weights computed with the Kernel propensity-score; $C(i)$ is the *j*- non-treated neighbourhood of the *i*- treated unit on the basis of the selected observables (Cerulli, 2015).

As robustness check we will run the same models but with ‘fake treatments’, by selecting as ‘treated’ units those provinces (Piacenza and Pavia) that are on the HSR but that did not benefit from the direct effects of the HSR stop. Briefly, to be satisfactory for our purposes, we expect that the impact of HSR program on Reggio Emilia will be positive and statistically significant, otherwise, we expect that the ‘fake treatment’ will not produce any statistical evidence on Piacenza and Pavia. In case of results that are different from these expected ones, it means that the HSR program was not an ‘activator’ of the growth processes, but that it would be considered (just) as an ‘accelerator’, meaning that other exogenous or endogenous processes led the development of the areas concerned.

The second method is the SC and it will be used to improve the results with graphical elaborations. It will be computed for the province of Reggio Emilia (where the Reggio Emilia AV Mediopadana stops) and for all the provinces will be considered as treated units (for instance, possible comparisons with other Italian corridors and with the related HSR stops).

The SC works when we are in presence of a single unit which have received the program (Reggio Emilia AV Mediopadana, for instance) and a series of control units that have not received the same policy; moreover, like the M–DID, the SC is effective when we have at our disposal a consistent number of year before and after the introduction of the program; other conditions needed to guarantee the validity of the method are that the program must not have

---

[3] Such covariates will cover several areas of human endeavour, such as social, economic, demographic, labour, health, institutional and environmental fabric of the geographical NUTS-3 provinces that will be part of the analysis.

been interrupted and that the control areas (NUTS-3 in this case) shall not have experienced similar programs.

The SC is based upon a system of weights $\omega = (\omega_1, \ldots \omega_j)'$ that are assigned to each of the *j*-control provinces in the following way:

$$\omega_j \geq 0 \; with \; \sum_{j=1}^{J} \omega_j = 1 \tag{5}$$

and the weights are selected in the way that the 'synthetic' outcome of the treated unit will look like as much as possible to the outcome before the HSR policy. The vector of weights is thus defined as follows:

$$D(\omega) = (x_1 - X_0\omega)'V(x_1 - X_0\omega) \tag{6}$$

where x_1 is the (K*1) vector of covariates previously mentioned predicting the growth in the different desirable potential outcomes, X_0 is a (K*J) matrix of values for the potential *j*- control provinces, *V* is a diagonal matrix with non-negative values which are able to reproduce the best synthetic on the basis of the predictors.

The weighted counterfactual of y (in absence of the HSR program) is:

$$y_1^* = y_0 \times \omega^* \tag{7}$$

where y_1^* is a (T*1) vector of real data about the potential outcome variables in 'treated' NUTS-3 province, y_0 is a (T*J) matrix of real data about the same outcome variables observed in the 'control' provinces.

3.1.2 Advantages and limitations

Among the main advantages of the counterfactual approach we find:

- Net evaluation of the program (HSR) avoiding biases deriving from possible exogenous or endogenous effects
- Reliable estimation concerning the economic/social/employment/innovative (and all of possible interest) impact of the policy
- High internal validity of the results due to the non-parametric approach

Among the possible disadvantages of these methods we note:

- Scarce external validity of the results, that would be very effective for the considered case studies but that would not be applicable to other contexts (even if very close in terms of features) in the role of policy recommendation
- Risk to be too generalist by using NUTS-3 level and not the municipality level

3.1.3 Expectations

Consistent with other very close case studies (for example, the evaluation of the HSR in the corridor Madrid-Barcelona, where Tarragona was directly involved into the policy as an in-between railroad stop), we

- suppose that the effects of HSR policy are higher for directly treated territories if compared to non-treated ones

- expect to find a significant positive effect of the HSR policy (in terms of economic/employment/innovative growth) on the NUTS-3 regions that directly received a gateway to the policy (for example, the railroad HSR stop, like in the case of Reggio Emilia AV Mediopadana, comparable to the Tarragona's case study)

3.1.4 Possible outputs

A possible example of the output of the counterfactual methodology is provided into the tables A1 and A2 (Di Matteo, 2020), where are reported, respectively, the number of observations computed in the M–DID, and the ATET of the SNAI program on a tourism-related potential outcome (expressed in nights spent), which is a tailored place-sensitive policy carried out in an Italian remote area which was directly treated by the program. The figure A1, instead, shows a possible example of the SC output, related to the same example.

Table A1. Number of observations in the M–DID

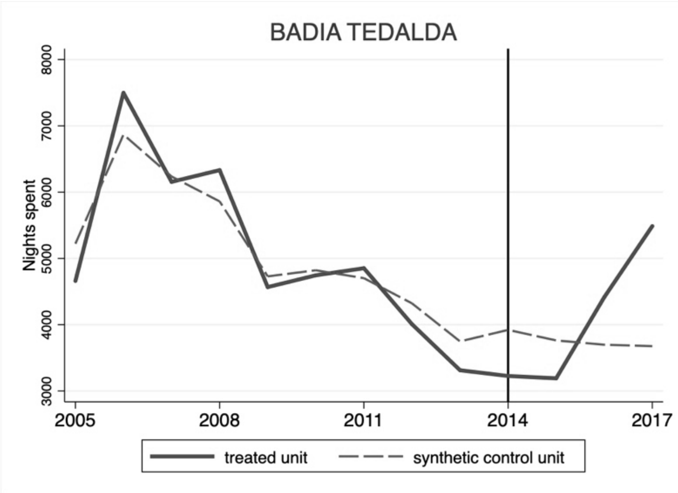
	Before (B)	After (A)	
Control (C)	1,019	545	*1,564*
Treated (T)	81	45	*126*
	1,100	*590*	1,690

Table A2. Results of the M–DID4

Outcome variable (nights spent)	Mean (y)
Before (B)	
C	*5,290.3*
T	*–3,838.8*
Diff (T–C)	–9,129.1*** [927.7]
After (A)	
C	*–3,100.0*
T	*–8,081.8*
Diff (T–C)	–4,981.8*** [1,230.1]
Diff-in-Diff (A–B)	4,147.3*** [1,391.7]
R^2	0.48

[4] Kernel propensity score matching Diff-in-Diff obtained through the set of Xi covariates; standard errors are in brackets; means and standard errors are estimated by linear regression; *** indicates inference at p<0.01.

Figure A1. Nights spent and synthetic nights spent in the municipality of Badia Tedalda (Tuscany)



For the case study taken as an example, the municipalities of the Casentino-Valtiberina Inner Area were included into the model as 'treated' units and all the other municipalities of Tuscany region (NUTS-2) were included as 'control' units. Besides the outcome variable (nights spent), several covariates and fixed effects controls were included in the model to build the optimal counterfactual (per capita GDP, resident population, accommodation capacity, degree of rurality/urbanisation, Walkscore®, TripAdvisor score). The M–DID was built upon a panel structure of data observed for the period 2005-2018 and the final number observations was 1690, among which 126 referred to the 'treated' units and 1564 were related to the 'control' units. General results tell us that the SNAI program had a positive impact on the local tourism performances, where the effect of the policy can be quantified in an approximate number of 4,000 nights spent more than it would have occurred if the policy were not implemented (which is around 5% more than the average of nights spent in the pre-treatment period in the considered pilot area).

As regards the figure A1 we take as reference the municipality of Badia Tedalda (1 out of the 9 treated municipalities), where we observe that the SNAI policy favoured a better performance in the outcome variable, which would have been lower in absence of the place-sensitive program.

Thus, getting back to the purposes of the Milan-Bologna urban corridor case study, the results of the counterfactual analysis will clarify the actual impact of the HSR policy in the directly treated areas and it will constitute an empirical proof to justify the need for an extension of the treated areas also to those which did not receive direct effects of the policy (Piacenza, Pavia).

The strengthen of the formerly non-treated areas in the urban macro-region will be a prerogative to foster a more organic development of the urban corridor Milan-Bologna, thus avoiding the risk of fuelling heterogeneity among in-between territories.

References:

Carbo JM, Graham DJ, Casas AD, Melo PC (2019) Evaluating the causal economic impacts of transport investments: evidence from the Madrid–Barcelona high speed rail corridor. *Journal of Applied Statistics*, 46:9, 1714-1723

Cerulli G (2014) treatrew: A user-written command for estimating average treatment effects by reweighting on the propensity score. *The Stata Journal*, 14:3, 541-561

Cerulli G (2015) *Econometric Evaluation of Socio-Economic Programs. Theory and Applications*. Springer-Verlag, Berlin Heidelberg

Di Matteo D (2020) Effectiveness of place-sensitive policies in tourism: an application to an Italian inner (pilot) area (*mimeo*)