Synthetic Control for Time Series

Sebastián Martínez, University of Glasgow

July 8, 2020

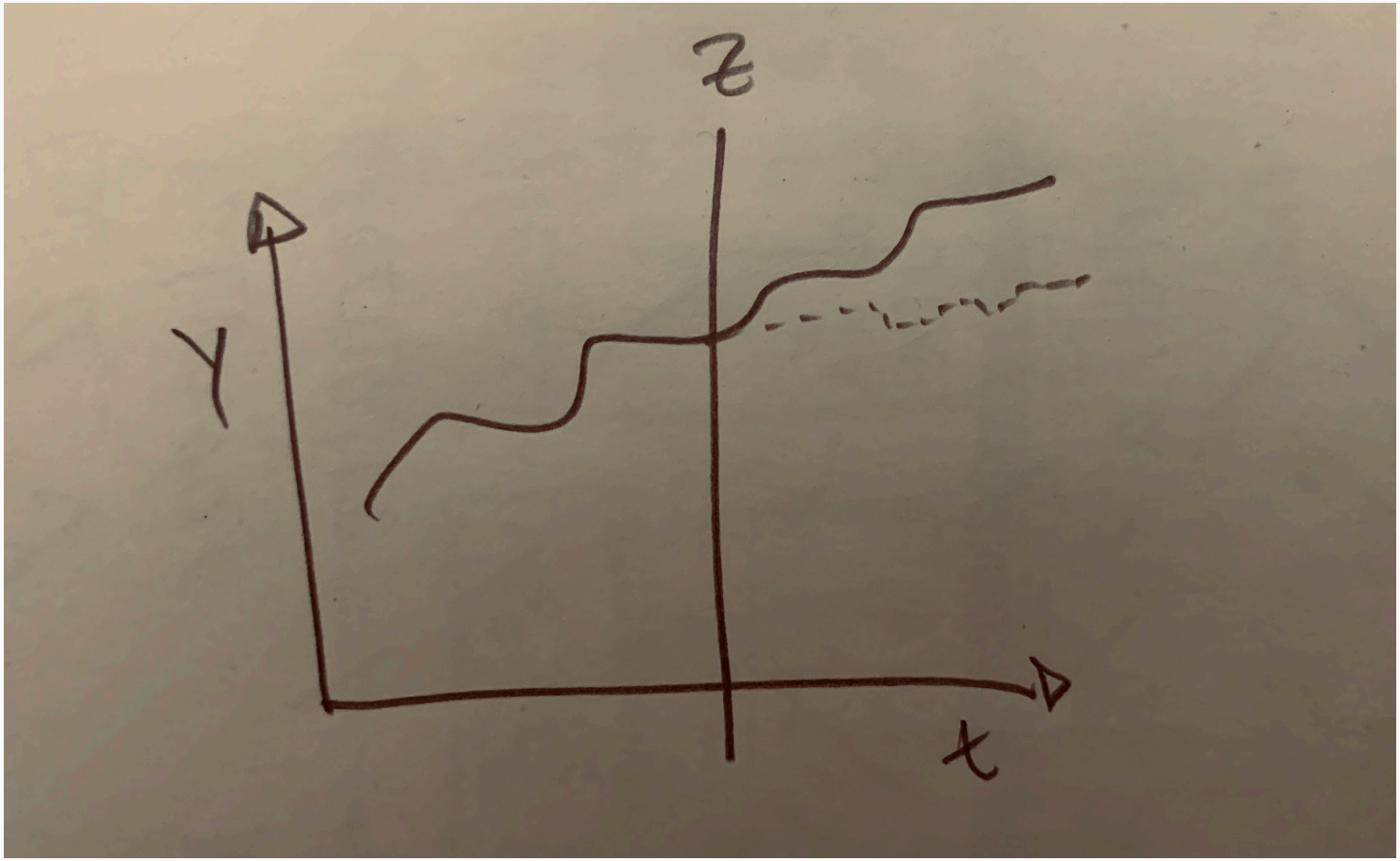
Berlin Time Series Analysis Meet-up

Table of contents

Causal Inference

Caveat

This could be part of a “causal inference with panel data” course, but I wanted to frame it in terms of time series because I think causal inference is important.



Causal Inference

The science of drawing *causal* connections between two events.

What would have happened if instead of doing A, we did B

Impossible!

We are looking for a valid comparison: a counterfactual.

How?

- Treatment (doing A or B) independent to outcome.
    - Randomised Control Trials (RCTs)
    - A/B Testing
    - Natural Experiments

Causal Inference

The science of drawing *causal* connections between two events.

What would have happened if instead of doing A, we did B

Impossible!

We are looking for a valid comparison: a counterfactual.

How?

- ~~Treatment (doing A or B) independent to outcome.~~
- Conditional independence = Good enough comparison
  - Propensity scores
  - Instrumental Variables
  - Inverse Probability Weighting

Causal Inference

The science of drawing *causal* connections between two events.

What would have happened if instead of doing A, we did B

Impossible!

We are looking for a valid comparison: a counterfactual.

How?

- ~~Treatment (doing A or B) independent to outcome.~~
- ~~Conditional independence = Good enough comparison~~
  - Regression Discontinuity
  - Synthetic Control

Causal Inference Resources

- Judea Pearl - Causality, Book of Why
- Donald Rubin - Potential Outcomes
- Miguel Hernán - Causal Inference in Epidemiology and Biostatistics
- Scott Cunningham - The Causal Inference Mixtape
- Richard McElreath - Bayesian data analysis and causal inference
- Susan Athey - Frontier in causal inference research

Motivating Question

How was West Germany's GDP affected by reunification?

What would have happened to West Germany's GDP if instead of reunificating with East Germany, it had not?

Classical question in the literature. Only here for pedagogical purposes. Don't shoot the messenger.

Motivating Question

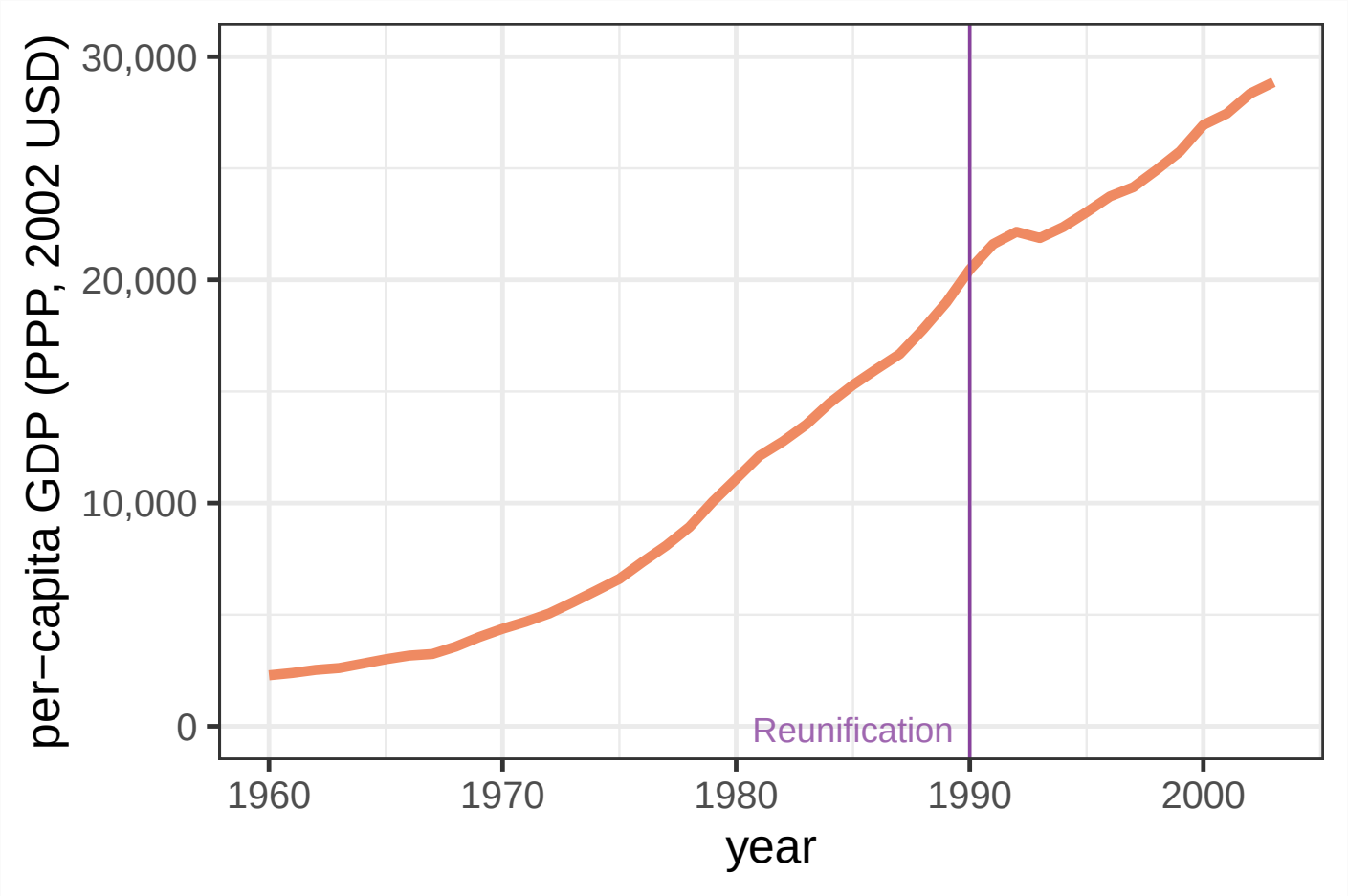


Figure 1: West Germany per-capita GDP

Potential Outcomes

Thought of by Splawa-Neyman et al. in the 1920s
formalised by Rubin in the 1970s

Consider treatment $Z = \{0, 1\}$.
Y_i^Z outcome of interest for unit i, under treatment Z.

Main potential outcomes assumption:

- Both Y_i^0 and Y_i^1 exist
- But we can only **observe** one

What do we do with the other one?

In our case: Build a *synthetic* version of it

Synthetic Control

Synthetic Control

- Comes from comparative case studies (subjective)
- Uses ideas from differences in differences

We don't deal with samples, but aggregates: countries, regions, institutions.

> *Synthetic control proposes a synthetic (fake) control unit as a weighted average of available control units that approximates the most relevant characteristics of the treated unit prior to the treatment.* Abadie and Gardeazabal (2003)

- A combination of similar units is better than a one-to-one comparison
- Optimally chooses set of weights which produce estimated counterfactual.
    - *What would've happened if...*
- Combination = weighted average

Synthetic Control - 4 advantages

1. Counterfactual is in convex hull of control units
    - No extrapolation but interpolation (King and Zeng, 2006)
2. No “peaking”. Researcher can build model without looking at outcome post treatment.
3. Explicit weights, which encourage conversation. Regression has *implicit* weights, i.e. not fun
4. Bridge between qualitative analysis and quantitative analysis

Synthetic Control - notation

Consider a set of units of observation indexed by $i = \{1, \cdots, N\}$.
$i = 1$ is our unit of interest, the one that was treated. $i \neq 1$ were not treated.
t_0 is the beginning of the observation period, $t_0 < T_Z < T$ is the moment the intervention took place.

- $Y_{i,t}$ is our outcome of interest
- $\mathbf{X}_{i,t} = \{X_{1,i,t}, \cdots, X_{m,i,t}\}$, m explanatory variables - unaffected by treatment (debatable assumption)
- $Z_i = \{0, 1\}$, the treatment
- $Z_1 = 1$ For all $t \geq T_Z$, 0 otherwise.

Synthetic Control - notation

Consider a set of units of observation indexed by $i = \{1, \cdots, N\}$.
$i = 1$ is our unit of interest, the one that was treated. $i \neq 1$ were not treated.
t_0 is the beginning of the observation period, $t_0 < T_Z < T$ is the moment the intervention took place.

- $Y_{i,t}$ is our outcome of interest
- $\mathbf{X}_{i,t} = \{X_{1,i,t}, \cdots, X_{m,i,t}\}$, m explanatory variables - unaffected by treatment (debatable assumption)
- $Z_i = \{0, 1\}$, the treatment
- $Z_1 = 1$ For all $t \geq T_Z$, 0 otherwise.

For $t \geq T_Z$, our unit of interest is:

- $Y_{1,t}^{Z=1} = Y_{1,t}$
- $Y_{1,t}^{Z=0}$ = ???

We are going to build a synthetic $Y_{1,t}^{Z=0}$ using all (or maybe some) other units $i = \{2, \cdots, N\}$

Synthetic Control Estimator

estimand

estimate

Ingredients

- 150g unsalted butter, plus extra for greasing
- 150g plain chocolate, broken into pieces
- 150g plain flour
- ½ tsp baking powder
- ½ tsp bicarbonate of soda
- 200g light muscovado sugar
- 2 large eggs

Method

1. Heat the oven to 160C/140C fan/gas 3. Grease and base line a 1 litre heatproof glass pudding basin and a 450g loaf tin with baking parchment.

2. Put the butter and chocolate into a saucepan and melt over a low heat, stirring. When the chocolate has all melted remove from the heat.

estimator

[1] source: https://twitter.com/simongrund89/status/1085929122860359680?s=20

Synthetic Control Estimator

For $t > T_Z$, time of the intervention,
Estimand:

$$Y_{1,t}^{Z=1} - Y_{1,t}^{Z=0} = Y_{1,t} - Y_{1,t}^{Z=0}$$

Estimator:

$$Y_{1,t} - \sum_{j=2}^{N} w_j^* Y_{j,t}$$

where w_j^* are a set of optimal weights, such that $w_j^* \geq 0$ and $\sum w_j^* = 1$.

What do we mean by optimal? We need to use what we know about Y, i.e. $\mathbf{X}$

Calculating the weights

We want to minimise the discrepancy between

- $\mathbf{X}_{Z=1}$, the covariates of the treated units, and
- $\mathbf{X}_{Z=0}W$ the weighted covariates of the untreated units

$$\|\mathbf{X}_{Z=1} - \mathbf{X}_{Z=0}W\|_V = \|\mathbf{X}_1 - \mathbf{X}_0 W\|_V = \sqrt{(\mathbf{X}_1 - \mathbf{X}_0 W)' V (\mathbf{X}_1 - \mathbf{X}_0 W)}$$

V, $m \times m$ symmetric, positive semi-definite matrix that defines comparison metric.
Two optimisation problems:

- Choosing W: inner optimisation
  - Representing $\mathbf{X}_1$ with $\mathbf{X}_0$
- Choosing V: outer optimisation
  - Predictive value of each covariate

Calculating the weights

We want to minimise the discrepancy between

- $\mathbf{X}_{Z=1}$, the covariates of the treated units, and
- $\mathbf{X}_{Z=0}W$ the weighted covariates of the untreated units

$$\|\mathbf{X}_{Z=1} - \mathbf{X}_{Z=0}W\|_V = \|\mathbf{X}_1 - \mathbf{X}_0 W\|_V = \sqrt{(\mathbf{X}_1 - \mathbf{X}_0 W)' V (\mathbf{X}_1 - \mathbf{X}_0 W)}$$

V, $m \times m$ symmetric, positive semi-definite matrix that defines comparison metric.

- Straightforward V: diagonal $\{v_1, \cdots, v_m\}$
- This assumes no interdependence between variables
- More complicated versions exist, but not pretty
- v_k determines how important X_k is in the above discrepancy

Calculating the weights - continued

To get optimal W^*, find weights that minimise:

$$\sum_{k=1}^{m} v_k \left(X_{1,k} - \sum_{j=2}^{N} w_j X_{j,m} \right)^2$$

Calculating the weights - continued

To get optimal W^*, find weights that minimise:

$$\sum_{k=1}^{m} v_k \left(X_{1,k} - \sum_{j=2}^{N} w_j X_{j,m} \right)^2$$

- Importance of $\mathbf{X}_k$ in the difference
- Should reflect the predictive value of the covariates

Calculating the weights - continued

To get optimal W^*, find weights that minimise:

$$\sum_{k=1}^{m} v_k \left(X_{1,k} - \sum_{j=2}^{N} w_j X_{j,m} \right)^2$$

- Importance of unit j in reproducing the behaviour of outcome variable for the treated unit in the absence of the treatment

Calculating the weights - continued

To get optimal W^*, find weights that minimise:

$$\sum_{k=1}^{m} v_k \left(X_{1,k} - \sum_{j=2}^{N} w_j X_{j,m} \right)^2$$

BUT HOW?

- Subjective assessment of predictive power of explanatory variables
- regression
- minimise MSPE
- cross-validation
- Others

What works best? Kaul et al. (2018), Xu (2017)

Pragmatic solution

Look for vector $(w_2^*(V), \cdots, w_N^*(V))$ that minimise the Mean Square Prediction Error

$$\sum_{t=t_0}^{T_Z} \left(Y_{1,t} - \sum_{j=2}^{N} w_j^*(V) Y_{j,t} \right)^2$$

Before moving on - choices choices choices

Before moving on

- Make sure units $j > 1$ are comparable to $j = 1$
- Interpolation depends on linearity of covariates. Big assumption.
- Can adjust and penalise for nonlinearity (not our focus)
- Can use negative or > 1 weights: extrapolation
- Assumed that weighted observables also match unobservables

‘Synth’ package

- Created by the same people that came up with the methodology Abadie et al. (2010)
- *Predictor weight matrix V is chosen among all positive definite diagonal matrices such that MSPE is minimised for the pre-intervention period.* Abadie et al. (2011)
- Has all the tools you need to do this on your own

Motivating example

OECD comparison

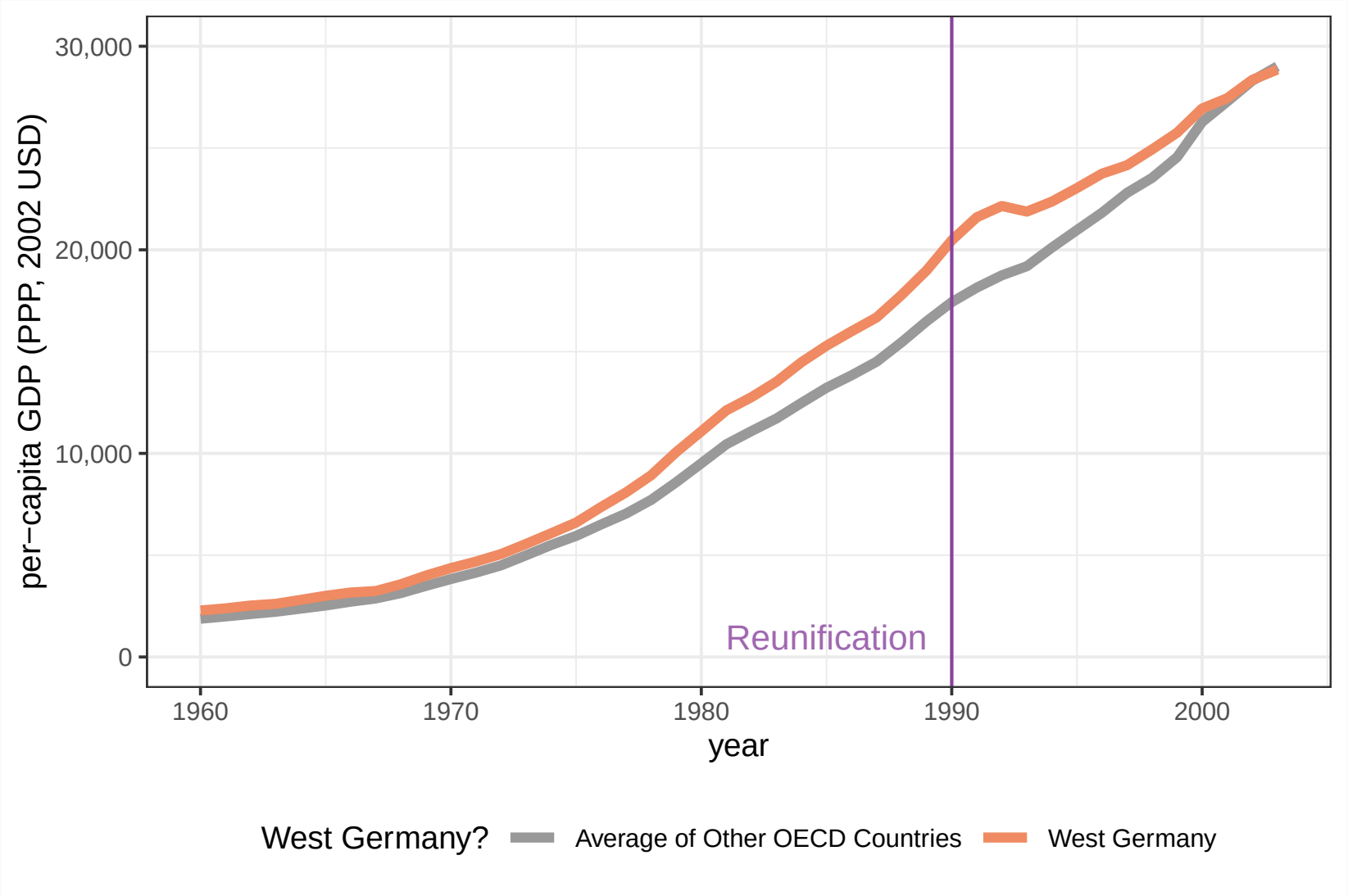


Figure 2: per-capita GDP, Germany and OECD average

Model selection

	Treated	Rest of OECD Sample
GDP per-capita	8169.8	8021.1
Trade openness	45.8	31.9
Inflation rate	3.4	7.4
Industry share	34.5	34.2
Schooling	55.5	44.1
Investment rate	27.0	25.9

Pool of controls

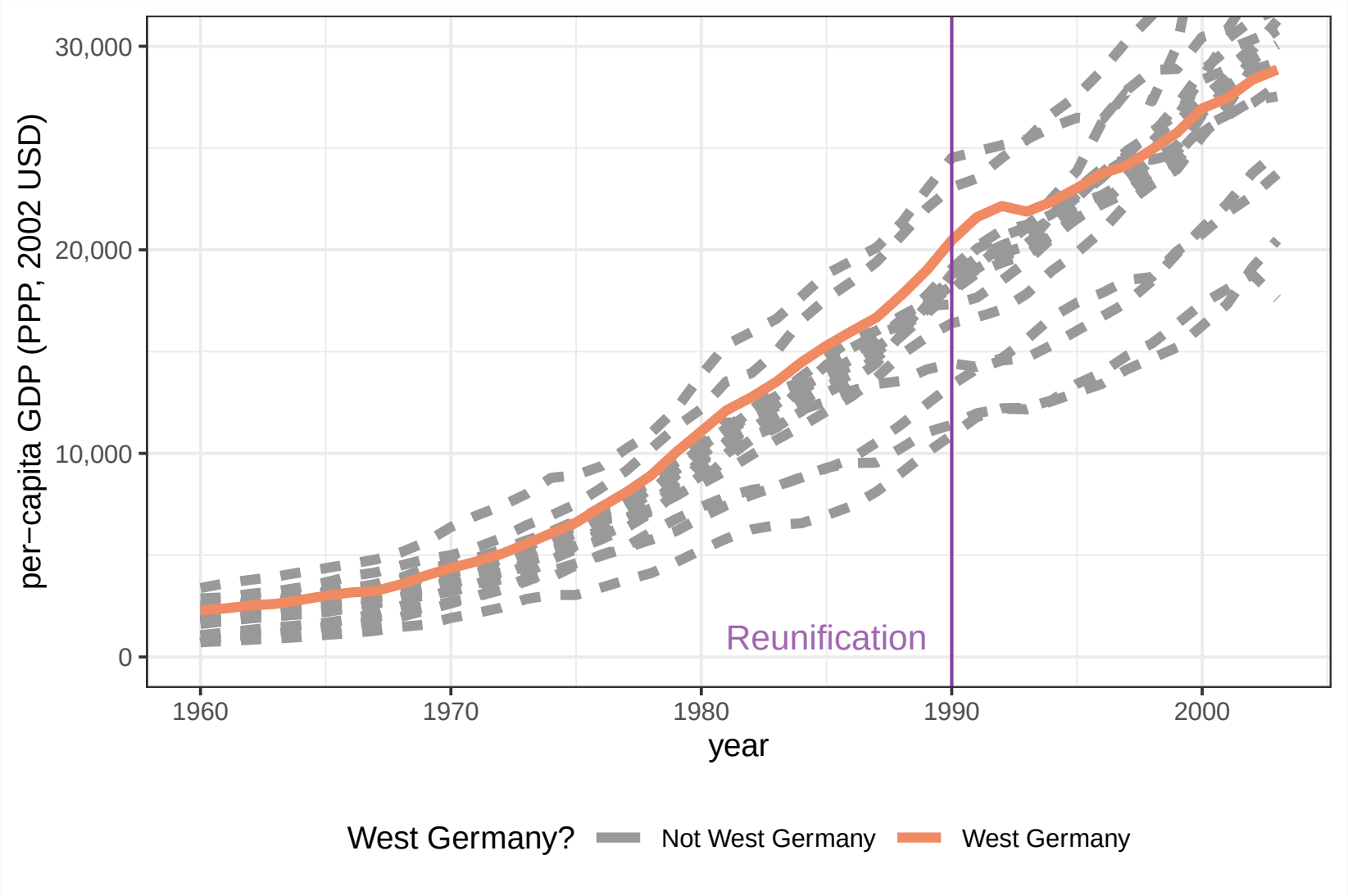


Figure 3: per-capita GDP, OECD countries

Synthetic Germany

	Treated	Synthetic	Rest of OECD Sample
GDP per-capita	8169.8	8169.8	8021.1
Trade openness	45.8	48.2	31.9
Inflation rate	3.4	5.3	7.4
Industry share	34.5	33.7	34.2
Schooling	55.5	53.9	44.1
Investment rate	27.0	26.3	25.9

Synthetic Control

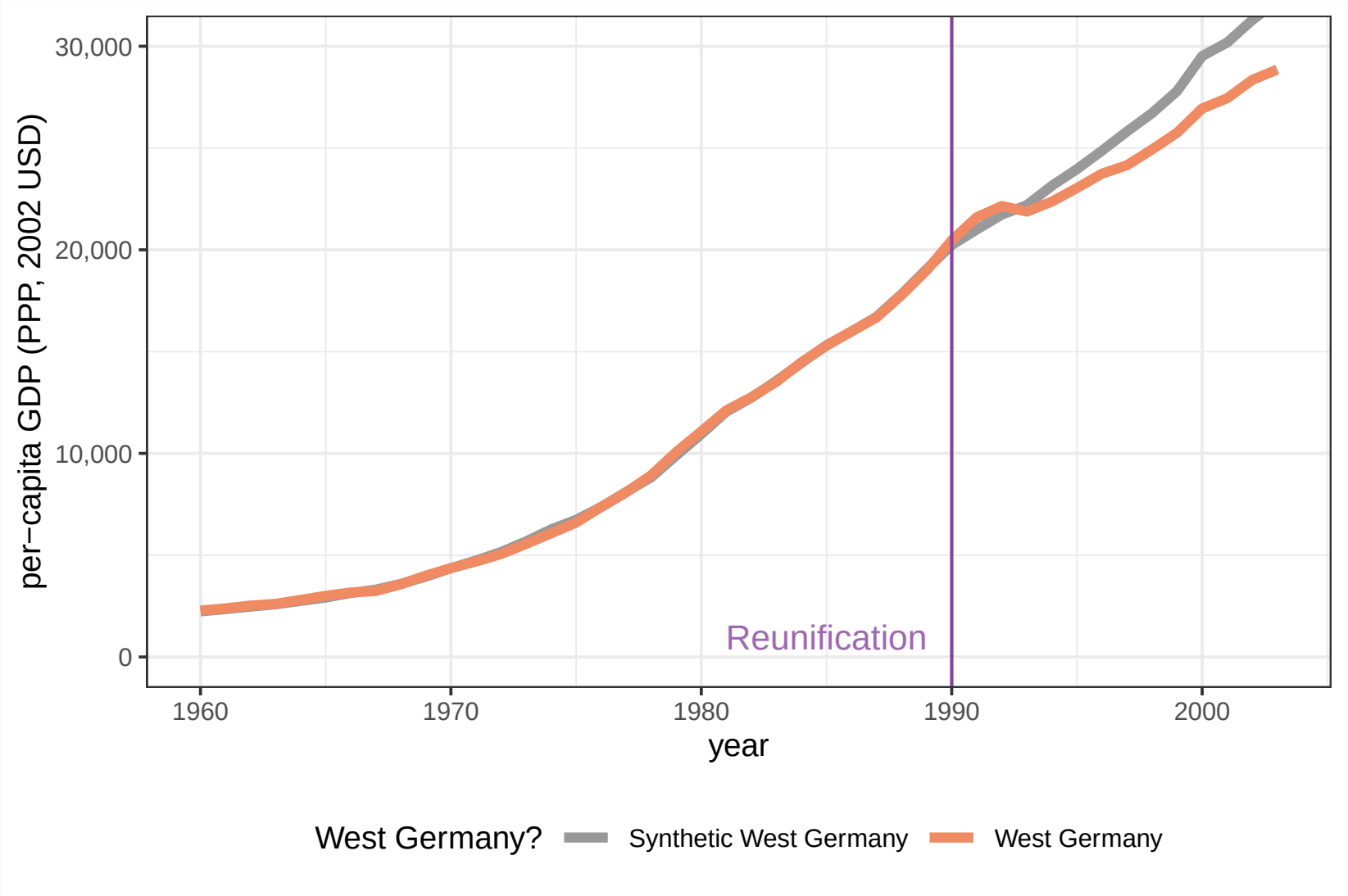


Figure 4: per-capita GDP, West Germany and Synthetic West Germany

Gap in GDP

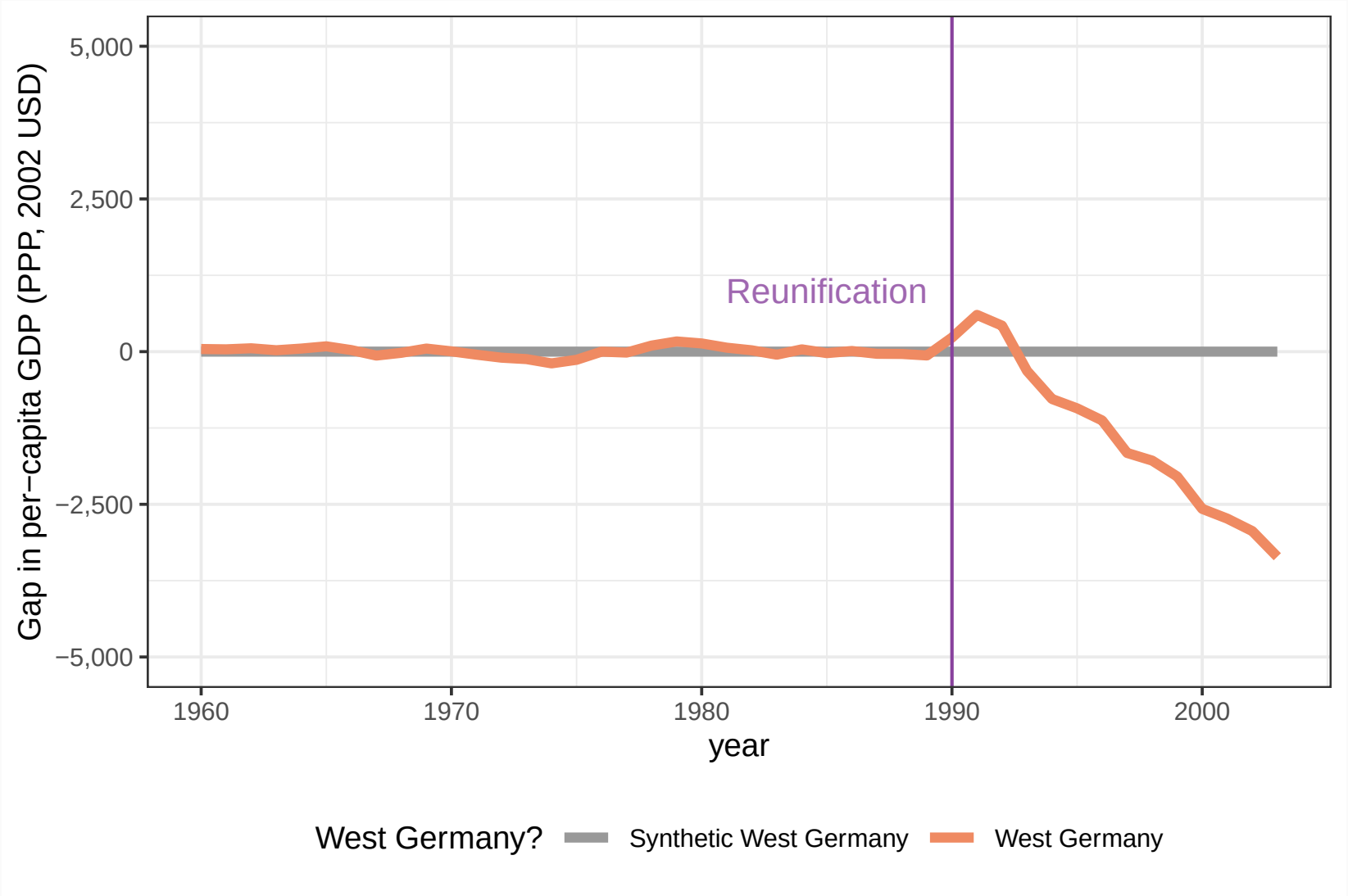


Figure 5: Gap in per-capita GDP, West Germany and Synthetic West Germany

Root Mean Square Prediction Error (RMSPE)

So far we have done the estimation part of the exercise
How do we know these are "statistically significant"?

Old fashioned exact p-values, i.e. Randomisation inference (Fisher, 1935)

Methodology:

1. Generate synthetic controls for all other units
2. Calculate RMSPE for pre- and post-treatment periods
3. Compute ration of post-to-pre-treatment RMSPE
4. Calculate unit's position in empirical distribution as

$$p = \frac{\text{percentile}}{\text{Number of units}}$$

Gap in GDP - All synthetic countries

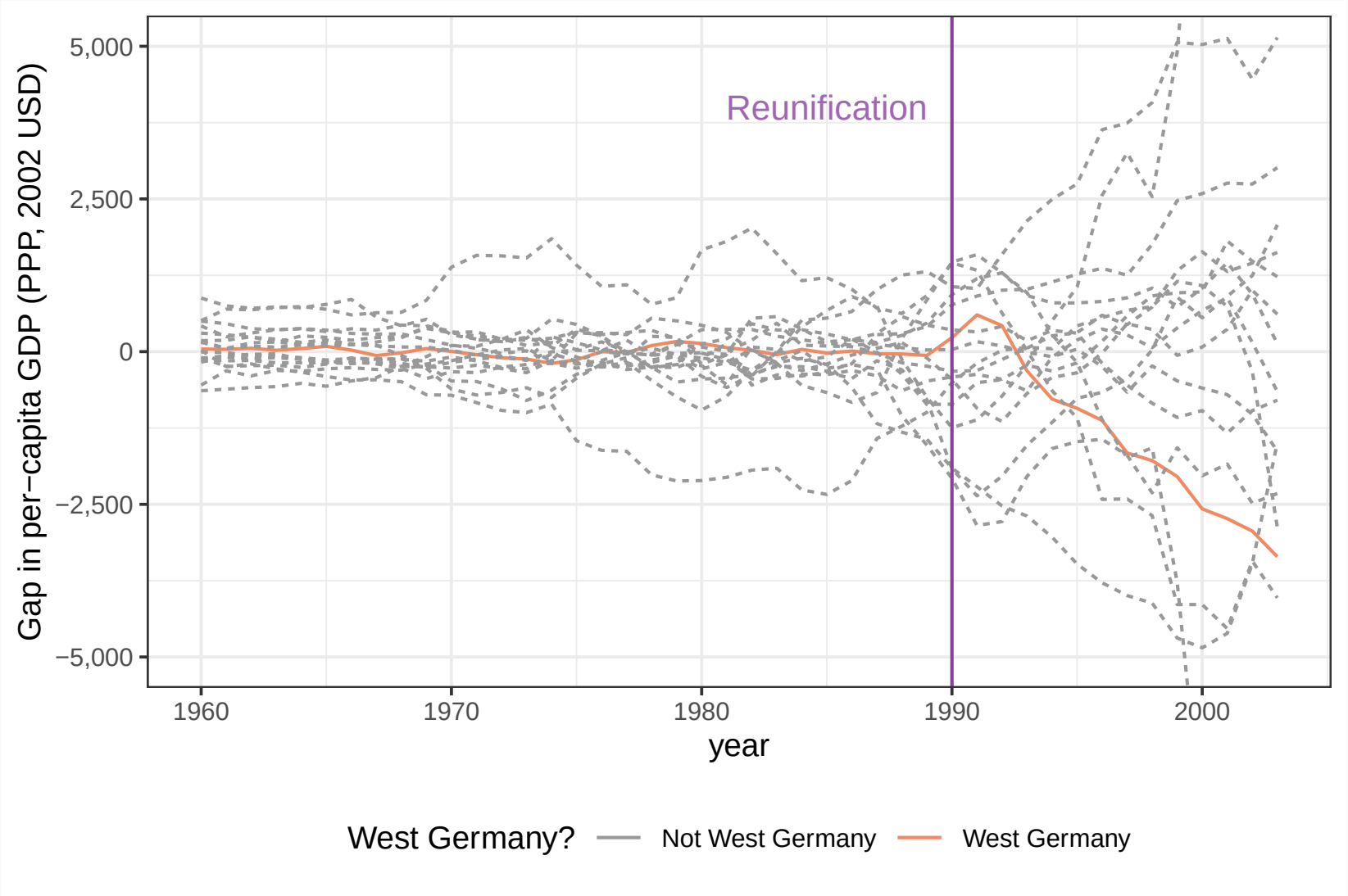


Figure 6: Gap in per-capita GDP, All synthetic nations

Gap in GDP - All synthetic countries

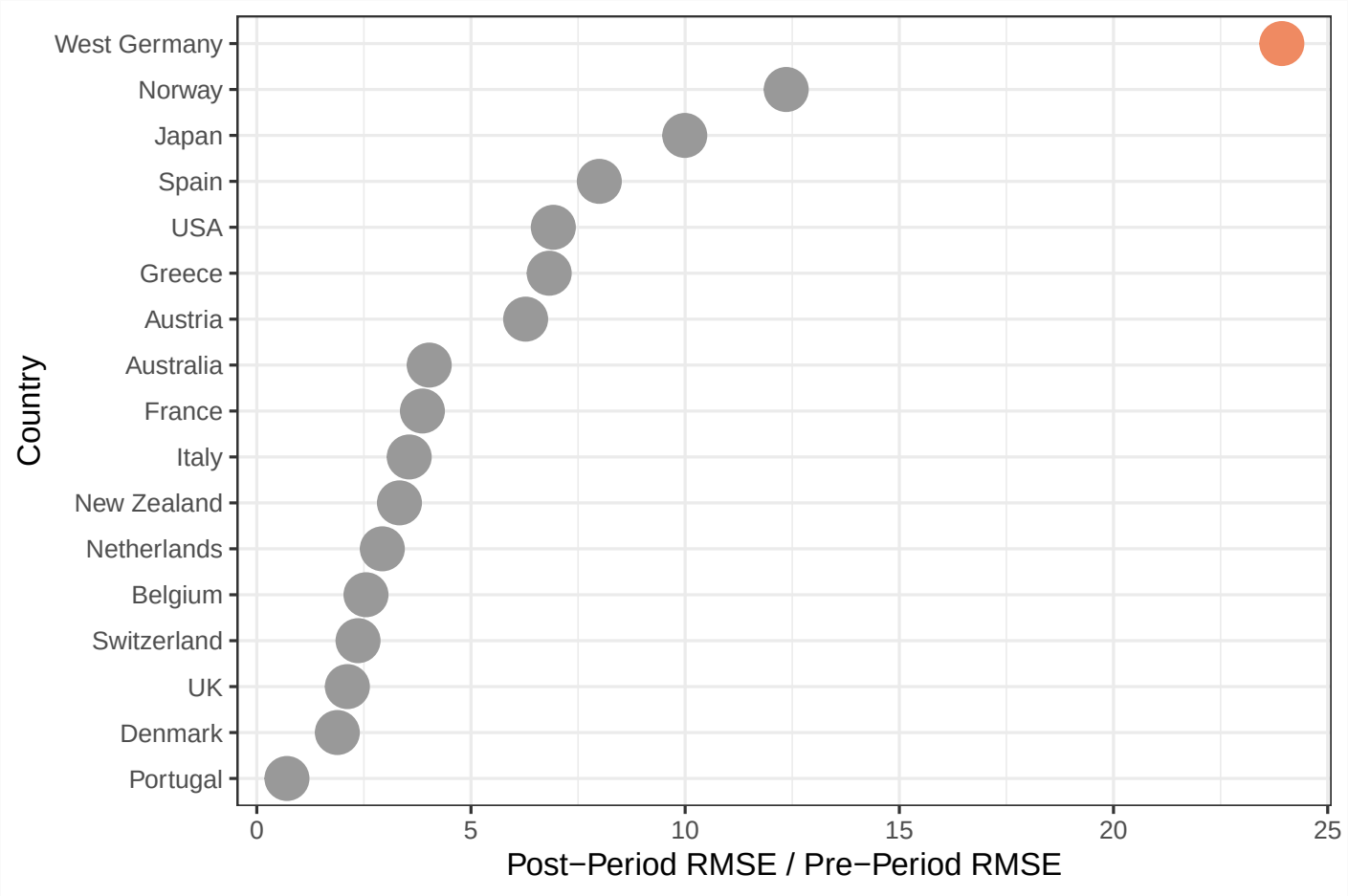


Figure 7: RMSE Ratio

Empirical *p*-value: 0.058

Weight comparison

Country	Synthetic Weights	Regression Weights
Australia	0.02	0.11
Austria	0.39	0.32
Belgium	0.02	0.12
Denmark	0.01	0.03
France	0.02	0.18
Greece	0.01	0.01
Italy	0.02	-0.15
Japan	0.04	0.33
Netherlands	0.02	0.18
New Zealand	0.01	-0.10
Norway	0.02	-0.06
Portugal	0.01	-0.16
Spain	0.01	0.02
Switzerland	0.14	-0.04
UK	0.02	-0.01
USA	0.23	0.22

Synthetic Control - regression weights

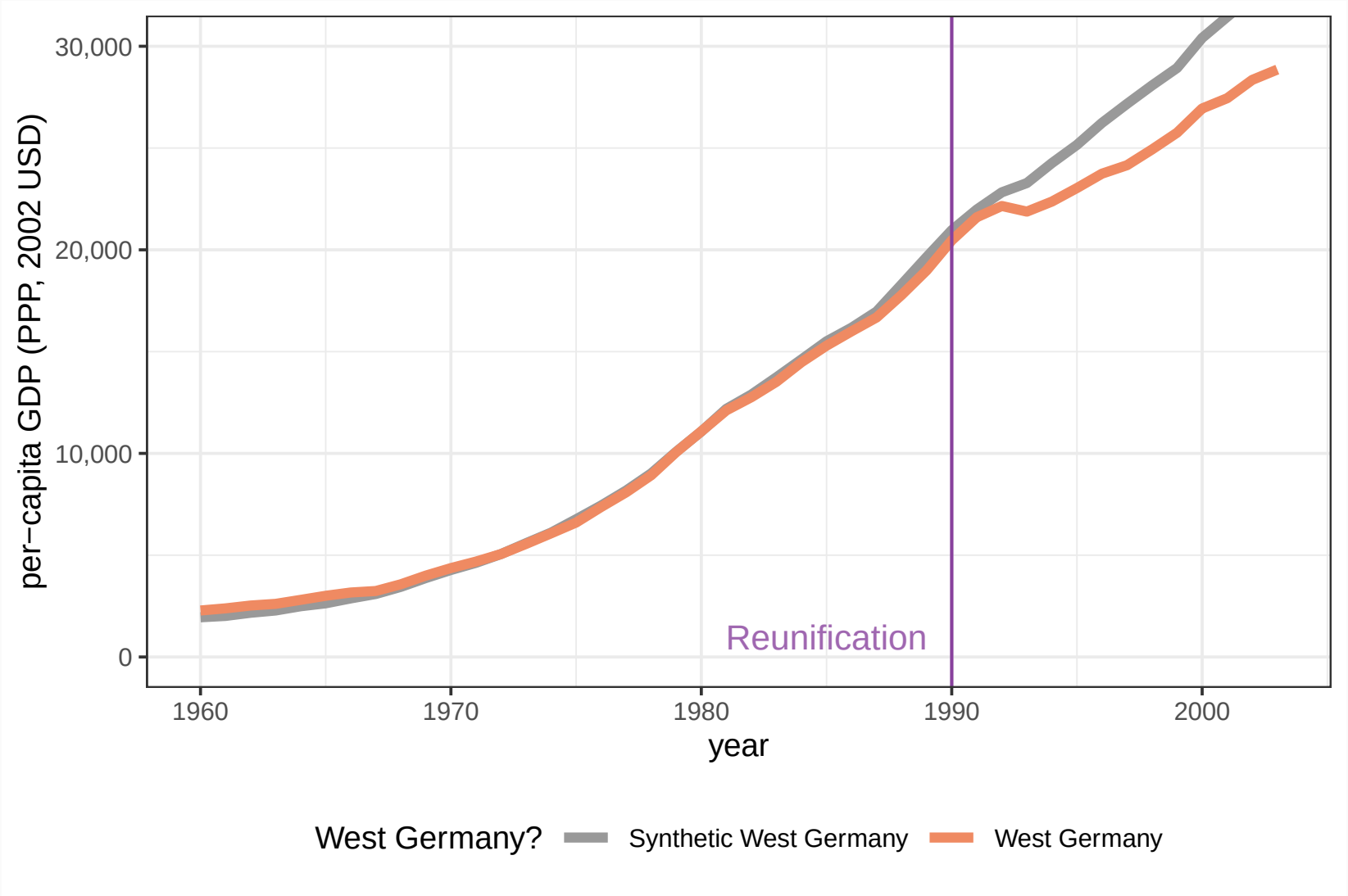


Figure 8: per-capita GDP, Germany and Synthetic West Germany, regression weights

Synthetic Control

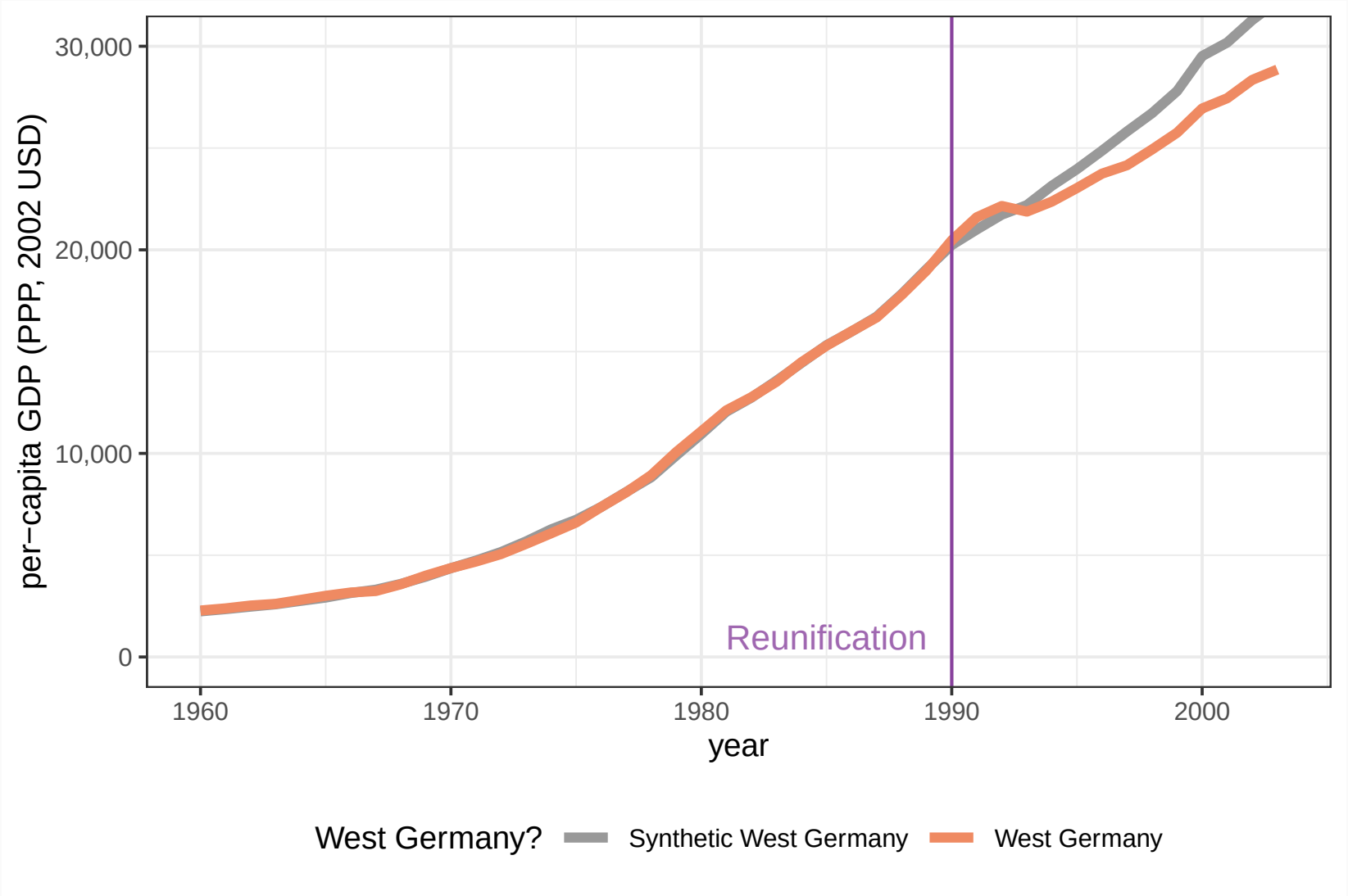


Figure 9: per-capita GDP, West Germany and Synthetic West Germany

Main Takeaways

Main Takeaways

- Synthetic counterfactual is only as good as your model (and even then it's an incomplete picture)
- Clear weights encourage conversation (but is this better?)
- Probably more efficient ways of getting weights (But unclear of what is more appropriate)
- Strong assumptions (Some might be hard to believe)

Extra Content

- Effectiveness of masks in against transmission of coronavirus
  https://www.iza.org/publications/dp/13319/face-masks-considerably-reduce-covid-19-cases-in-germany-a-synthetic-control-method-approach
- Measurement of marketing performance
  https://tech.wayfair.com/data-science/2017/08/using-geographic-splitting-optimization-techniques-to-measure-marketing-performance/
- Other R package that does synthetic control: 'MicroSynth':
  https://cran.r-project.org/web/packages/microsynth/vignettes/introduction.html

References

Abadie, A., Diamond, A., and 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(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42(13):1–17.

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *The American Economic Review*, 93(1):113–132.

Cunningham, S. (2020). *Causal Inference: The Mixtape*. Pre-print edition.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd, Edinburgh. OCLC: 2417943.

Kaul, A., Klossner, S., Pfeifer, G., and Schieler, M. (2018). Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together With Covariates. page 24.

King, G. and Zeng, L. (2006). The Dangers of Extreme Counterfactuals. *Political Analysis*, 14:131–159.

Rubin, D. B. (2005). Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472. Publisher: Institute of Mathematical Statistics.

Xu, Y. (2017). Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis*, 25(1):57–76.

THANK YOU

THANK YOU!

Source: Anthony Suau

mail@smartinez.co
@sbmrtnz