Mean Reversion of the German City System After WWII: To What Mean?[*]

Duc A. Nguyen
University of Groningen[†]

July 14, 2019

Abstract

This paper uses the synthetic control method to construct comparison units for West-German cities. We use these as counterfactuals to assess the long-run impact of the WWII bombings on the economic activity of 52 West-German cities, and the West-German city size distribution. We extend the literature in that we do not only look for whether cities following the WWII bombing shock are mean-reverting or not, but the use of counterfactuals allows us to distinguish whether individual cities experienced a positive or negative impact as well. We find a permanent impact on the majority of cities, where the ratio of positive to negative impact cities is around 5 to 6. Also higher populated metropolitan cities before WWII tend to either return or be positive impact cities, while negative impact cities mostly consist of lower populated places with the exception of Berlin. We also find a non-random decline of the majority of counterfactual cities around the 1960s. Overall, our findings support a hybrid theory, where the determinants of a city's size are underlying mutable locational fundamentals, and increasing returns.

Preliminary and Incomplete

1 Introduction

There is considerable interest in the study of the long-run impact of large and temporary shocks on a city's growth, whether they are caused by war or are a cause of nature. Researchers who try to explain economic activity across space exploit the findings of their long-run impact in order to gauge the relevance of some fundamental theories. For instance, the seminal paper by Davis & Weinstein (2002) looked at the WWII bombings on Japanese cities, while the focus of Brakman et al. (2004) and Bosker et al. (2008) were the WWII bombings on German cities. However, one issue researchers in the study of these shocks face is a potential change in the trend not caused by the shock, which may simultaneously occur within the aftermath of the shock. For instance, as Glaeser (2005) argues, there is no reason to remain in a city close to a productive natural resource once the demand for this particular natural resource declines. In a similar manner, the development away from a manufacturing to a service economy, the decline of the US rust belt or suburbanization of cities (Glaeser & Kahn, 2004) come to mind as underlying factors which could induce the change in trend. For Germany in particular, the starting decline of the coal industry in the German Ruhr area around the 1960s is one example of such a factor, which could have led to a change in the trend. If we do not account for this potential change, then we could otherwise wrongly attribute the decline of those cities to the WWII bombing shock; a decline which may have happened regardless of the shock through the decline of coal for example.

[*] The author thanks Steven Brakman, Harry Garretsen, Dimitrios Soudis, Tristan Kohl and Marcel Timmer for helpful comments and suggestions, and Maarten Bosker for the provision of the German city population data.

[†] Department of Global Economics and Management, University of Groningen, Nettelbosje 2, 9747 AE Groningen, The Netherlands. Email: D.A.Nguyen@rug.nl

The goal of this paper is to disentangle the effects of the WWII bombing shock on West-German cities from these other, underlying drivers of population growth which the literature so far did not explicitly account for. That is, in the evaluation of WWII bombings, we typically rely on methods which show whether city population is mean-reverting or not to the post-shock path as linearly extrapolated from the pre-shock path.

Hence, these methods rely on the 'as if WWII did not occur' counterfactual trend being linear throughout time. However, if the counterfactual is non-linear, then we run the risk of wrongly concluding no reversion to the mean which is driven by this non-linear development; yet we may attribute it to the shock itself even if the shock did not actually have any long-run impact. Vice versa, we may wrongly conclude mean reversion in the case of the shock actually having a long-run impact, which we do not detect as such because the non-linear counterfactual development counteracts this impact towards a linear trend. In other words, the question should not be whether cities are reverting to the mean, but whether they revert to a potential 'new mean'.

To achieve the disentanglement of the WWII bombing shock from other factors through the use of a 'new mean', we will employ the synthetic control method (SCM) by Abadie & Gardeazabal (2003) and Abadie et al. (2010, 2015) to construct the counterfactual/synthetic city for each sampled city, and use this synthetic city as the comparison unit for the bombed city. As compared to other methods which also use counterfactuals, the SCM is designed to ameliorate some issues of finding a suitable counterfactual: First, as the synthetic control method 'constructs' a comparison unit for the bombed city treatment unit by choosing a convex combination out of suitable non-treated comparison cities, we do in principle have infinitely many possible comparison cities.[1][2]This ameliorates the problem of not finding a suitable comparison unit, which is especially important when the treatment unit are cities at the upper tail of the city size distribution for which suitable single unit counterfactuals are rare. This is one reason why SCM is preferred over the typical difference-in-difference approach in our case since we can extend the number of potential counterfactual as we consider a convex combination instead of a single unit.

Second, we add objectivity to the counterfactual selection as this convex combination is selected from a data-driven procedure, where we minimize a loss function.

As we will exemplify later, the choice now is not about having enough comparison units to choose from, but the choice of matching covariates which will be minimized to obtain the most suitable counterfactual. Typical to the literature however is that the SCM is applied in a single treatment unit without much discussion on the choice of set of matching covariates. Yet the counterfactual and as such the results are sensitive to the inclusion or exclusion of certain covariates, with no apparent criteria as to which specification is the best. Since we apply the SCM with the very same set of covariates to many cities in the German city system, we implicitly add transparency to the whole covariate selection process. In this regard, the SCM method is especially suitable for our case.

Crucial for our purposes is that the synthetic control counterfactual allows for a non-linear city population path. That is because in the study of impacts stemming from large, temporary shocks on the location of economic activity, we need to consider a large time span due to the slow-moving nature of population; otherwise we cannot distinguish mean reversion from no mean reversion in the data.[3] However with a large time span considered, we increase the influence of underlying factors driving the city population path away from the supposed linear path. This implies that assuming a linear trend may not be appropriate for relatively long time spans and that for these purposes, an explicit estimation of the counterfactual path is more appropriate. The SCM provides one way of doing exactly that.

Suited in particular is the external validity inherent to matching methods such as the SCM, or rather, that we focus on the internal validity of each individual city. That is because the counterfactual for each treatment unit is estimated through its own individual SCM model. This stands in contrast to regression-based approaches for which we may, and for our case in particular, are only able to focus on the internal validity within a system of cities, where we seek an unbiased estimate based on many cities. It may in

[1] Of course, due to computation reasons, we only consider the finite case.

[2] That is, instead of considering a single city, or a single point in a convex hull as a combination chosen by the researcher, we now consider the entire convex hull as a possibility, where the first two convex combinations are simply a special case within the convex hull.

[3] See (Perron, 1991) for the importance of the time span against frequency in unit root tests.

principle be possible to construct a counterfactual through estimates based on regressions, but at most these estimates will suit the best for cities in general, but no city in particular.

Furthermore, we will replace the deterministic trend term in the unit root model with the synthetic control term, so that this simple correction is not very invasive in terms of changing fundamental testing principles. Thus, we retain some level of comparability with previous results in the literature (Davis & Weinstein, 2002; Brakman et al., 2004; Bosker et al., 2008; Miguel & Roland, 2011) as our results should not be driven solely by the different method itself.

We should also emphasize here that the explicit use of a counterfactual makes it possible to have a benchmark result to directly compare with, which was so far done implicitly in the literature in that the post-war benchmark would be extrapolated by the pre-war benchmark. This allows us to see, apart from finding mean reversion or not, whether the individual city population is above or below its synthetic German city counterpart after WWII. Furthermore, we can compare the actual city size distribution with an explicit benchmark, which we define here as the synthetic city size distribution. This is one advantage of replacing the typical deterministic trend with the term estimated by the SCM; and as it turns out, these adjustments matter:

Our results show that out of the cities for which we find a unit root, around 85% of those in terms of population end up above its corresponding synthetic city population after around 35 years after WWII. 38 out of 52 of our sample of German cities already in 1960 'returned' or were even above the synthetic population. Nevertheless, finding a positive impact at least for our sample cities, which mostly constitute the biggest German cities, is not what we would expect from a large negative shock such as the WWII bombings. This result stands in contrast to the more expected findings of Bosker et al. (2008), who find a structural break for 17 cities, of which 15 are negative.

This paper mainly contributes to the strand of empirical research on the lasting effect of a large, temporary shock on a system of cities (Davis & Weinstein, 2002; Brakman et al., 2004; Bosker et al., 2008; Miguel & Roland, 2011), with the aim of providing a stronger empirical method to distinguish the large, temporary shock from any other parallel development happening. The results of this particular strand also has implications on the relevance of some competing fundamental theories on the location of economic activity. (Davis & Weinstein, 2002) defines those theories as locational fundamentals theory (e.g. Rappaport & Sachs (2003), also called 'first-nature geography' in the literature), random growth theory (e.g. Gabaix (1999) and increasing returns theory (following Krugman (1991)).

This paper is also related to papers on multiple equilibria (Bosker et al. (2007), Davis & Weinstein (2008), Redding et al. (2011), Bleakley & Lin (2012)), although whether the German city system can be characterized by multiple equilibria or not is not the focus here.

To a lesser extent, this paper is related to the disaster literature (duPont IV & Noy, 2015; Siodla, 2015), with the difference that they look at natural disasters typically affecting a single city and not an entire city system.[4]

The empirical setup can be roughly categorized into two parts. First, we will construct the synthetic control for each German city in our sample, which we will describe in section 2. Second, we will replace the deterministic trend term with the synthetic control term in the typical augmented Dickey-Fuller test model in section 5.

In section 2, we will also postulate some principles that we will follow given this unique natural experiment. This very section will also serve as a guide to the data collection of section 3. In section 4, we will present an example city as well as establish the notion of an inverse U-shaped city population growth from the beginning of the German Reich of 1871 onwards, which serves as a justification for our empirical strategy. After we have constructed the synthetic control, sections 6 reports the impact of the WWII bombings on individual German city population, the city population share of the total German population respectively. Section 7 shows various robustness exercises. Section 8 is reserved for the robustness of the results. Section

[4] This distinction could be important in that shocks on the city system may induce relatively little migration given the magnitude of the shock if most cities experienced some form of destruction. This gives room for the conjecture that as long as the large, temporary shock on the city system is even, it does not induce migration whereas a shock on a single city in a city does lead to migration. This notion could explain why we overall find a permanent impact in the disaster literature and usually none where disasters affect the whole country.

8 will report some additional characteristics on a group of cities conditional on how the individial city was impacted, as well as some additional sensitivity analysis. Section 9 will bring the evidence established in the previous sections together, where we will discuss the aforementioned competing fundamental theories on the location of economic activity in the light of the new evidence. Section 10 concludes.

2 Synthetic City Construction

Before we turn to the data, it is important to justify the data. Hence, we turn to the construction of the synthetic control city first. We will follow closely the notation of Abadie et al. (2015) in the following. That is, the synthetic German city, indexed by i, is represented by a convex combination of donor pool US cities as indexed by j. Therefore, we have a sample of $I + J$ individual cities, with I number of 'treated unit' German cities, and J number of 'comparison units' US cities in the 'donor pool'. Hence, the donor pool is $j = I + 1, ..., I + J + 1$.

In the following, we will leave out the individual city superscript i for each variable for simplicity. The synthetic control for the individual city is a convex combination of cities in the donor pool i.e. it can be represented by the $(J \times 1)$ vector of weights $W = (w_{I+1}, ..., w_{I+J+1})$ for each city, where $0 \leq w_j \leq 1$ and $w_{I+1} + ... + w_{I+J+1} = 1$.

Let X_1 be a $(K \times 1)$ vector which contains the pre-treatment covariate of a treated city and let X_0 be a $(K \times J)$ matrix which contains the same pre-treatment covariate but from the donor pool.

We want to minimize the discrepancy between X_1 and X_0W i.e. minimize

$$\|X_1 - X_0W\| = \sqrt{(X_1 - X_0W)'V(X_1 - X_0W)} \tag{1}$$

which is done in a two-step procedure: Conditional on V, the synthetic control $W^*(V) = (w^*_{I+1}, ..., w^*_{I+J+1})$ is selected to minimize (1) subject to $w^*_{I+1} + ... + w^*_{I+J+1} = 1$ and $0 \leq w^*_j \leq 1 \; \forall j$. V is a $(k \times k)$ diagonal matrix with non-negative entries, where its values weights the relative importance of each pre-treatment covariate. The idea of this approach is that we want the covariates to have strong predictive power over the outcome of interest, so that relative strong predictors are given relatively larger weights. Equivalently, let X_{0jm} be the value for the m-th covariate for donor pool unit j and X_{1m} be the value for the treated unit covariate, where $m = 1, ..., K$, so that we choose W^* which minimizes

$$\sum_{m=1}^{K} v_m \left(X_{1m} - \sum_{j=I+1}^{I+J+1} w_j X_{0jm} \right)^2 \tag{2}$$

where v_m are the diagonal entries of V, i.e. $V = diag(v_1, ..., v_K)$.

One way to choose V, as done by Abadie & Gardeazabal (2003); Abadie et al. (2010, 2015) is based on minimizing the mean squared prediction error (MSPE) of the city population outcome variable S given the chosen W^* before:

$$\sum_{t=1}^{T_0} \left(S_{1t} - \sum_{j=I+1}^{I+J+1} w^*_j(V) S_{jt} \right)^2 \tag{3}$$

where T_0 is the treatment period. Summarizing, we first calculate $W^*(V)$ for any V in the first step, and then choose the V that gives us the $W^*(V)$ which minimizes the MSPE.

The literature follows the principle in the choice of the covariate in that it should be a good predictor of the variable of interest, so that often pre-treatment lagged variables of the outcome variable are included. However, what we are actually interested in are only good predictors of the outcome post-treatment. That is, we should not assume that predictors do not change in their relative importance over time. This needs to be especially stressed in our case, as we consider a long time span with various events in history which may have changed the determinants of individual city growth.

Hence, it is important to postulate some additional principles in the choice of covariates to be good predictors of post-WWII city population. First, we want the covariates to be as close as possible in time to the beginning of WWII as we think that the farther we go away from the treatment period, the worse the prediction power of the covariate becomes for the post-WWII period.

Hence, we should not be tempted to include all pre-treatment lags of 1870 to 1940 of the outcome as covariates, as we would otherwise give a relatively high weight towards US cities so that the synthetic control is a good counterfactual of city population of the German city of interest throughout 1870 to 1940, but a comparatively bad counterfactual for the post-WWII periods we are actually interested in. As mentioned before, this exclusion of earlier covariates can also ameliorate the issue of annexations of surrounding municipalities by the German city, where significant jumps in the population of some German cities occurred mostly during the 1920s-30s. In a sense, the likely increasing error going back in time from the extrapolation of the population data to account for the annexation will not be as problematic due to the exclusion of earlier covariates. Also, some geographical features may be the driver of pre-war growth, but also of post-war decline. Here, we can take coal as an example of such a feature. If we would not include it, it could lead to the SCM matching primarily high growth non-coal US cities with high growth German coal cities. If those high growth non-coal cities are also relatively high growth after the war, a decline of coal in the post-war years and a subsequent decline of coal cities would suggest that we would match, post-war, low-growth coal cities with high growth US cities. As such, through SCM we may wrongly and overly find a decline of those German coal cities. Again, the importance here is to find good post-treatment predictors, and not necessarily good predictors of the variable of interest in general.

Another potential issue is that German cities in our dataset are much older than US cities and we may give US cities a weight which did not even exist yet a few years earlier, so that they may not be as comparable to the historical German cities. One way to account for that is to only consider the historical US regions in the dataset, and exclude the regions which were not developed. We consider the Mississippi river as a good demarcation line on which to base this distinction; and as such we keep the principle that the bulk of covariates should be as close to the treatment period as possible. Therefore, we choose the bulk of covariates to start with the Weimar Republic in 1918 until the beginning of WWII in 1939. In section 7, we go further in excluding all US cities which did not exist in 1870. Having established the when and some of the which, the following data section will further elucidate which type of covariates we will consider.

3 Data

For the outcome of interest, city population, we extend the original German city dataset of Bosker et al. (2008) by including the years 1871 to 1925 if possible, and as such we exclude 10 cities due to insufficient data to a total of 52 German cities. Note that the excluded cities are relatively small, reflecting that bigger cities typically have complete population data throughout history. This data is obtained either from irregular censuses, or from updates of the population records which occurred annually. As the SCM constructs a convex combination out of donor pool cities, it is important to have each unit of observation to be of similar magnitude to the units in the donor pool. Therefore, we consider incorporated places as defined by the US census as the unit in the donor pool, as it is the most similar in magnitude to the German census cities, and covers the entire range in terms of population. This data was already partially obtained from the Spatial History Project of Stanford University for 1790-2010 (US Census Bureau and Steiner, E. (2018)), where we extend it by the decennial U.S. Census population data. For our purposes, we only consider incorporated places which had at least a population of 50,000 at some point in the census. There were throughout history many cases where one bigger incorporated place annexed a smaller one, which we need to account for. Fortunately, the annexed incorporated place will cease to be mentioned in the following census, so that we have simply added those up until the year 2010, so that we roughly have the same city boundary throughout the time analyzed. As for the German cities, we usually do not have the population data for the surrounding cities which were annexed by the German cities in our sample. To adjust for that, we adapt the series in the same manner as Bosker et al. (2008), where we assume the same growth rate for annexing and annexed places and extrapolate it on the additional population; this time with the city boundary fixed around WWII. However, this adaptation is likely to be less accurate the farther away we extrapolate from the time of annexation. We can account for that in the SCM by excluding the covariates which are far away from the WWII treatment period. Overall, we have decennial population data for the years 1870 until 2000. The other reason we choose US data is that the level of industrialization of East Coast US cities is very similar to that of Germany. As such, we are more confident that the, as if WWII did not happen, post-WWII population development are of similar nature between the German sample and the donor pool.

Since SCM applied to our case requires covariates which are good predictors of (post WWII) population, we will also consider a coal covariate. As discussed in the previous section, this inclusion is largely motivated by the decline of the coal industry in the Ruhr area and the coal industry in the US both happening around the 1960s. For that, we obtain the number of workers in the coal industry from Fischer (1989, 1995) on a Regierungsbezirk (administration area) level. Similarly, we have county level labor employed in the coal industry, taken from the historical U.S. Census (available from the US Census Bureau (1900-1930)). Here, we consider a city a coal city if it is within a significant coal county, producing at least 100,000 short tons or more in 1929, as it appears in the 1930 U.S. census. In order to have a similar level of magnitude between treatment and control covariate, we instead assign each of those coal cities the state level aggregate. Lastly, we include a river and sea-access dummy as simple first-nature determinants, where we consider a city a 'river' or 'sea' city if today's border is 1 kilometer away from either, as determined by the U.S. census bureau TIGER/Line and Census TIGER, and Google Maps. In data appendix A, the data is described in more detail.

4 Descriptive Statistics and Inverse U-shaped Growth

To justify our approach to alter the established methodology of the literature towards accomodating a possible non-linear counterfactual, we will show that the synthetic city population growth throughout time is found to be typically inverse U-shaped. An inverse U-shaped city population growth for German cities can be found by looking at the behavior of the synthetically estimated population growth of German cities over time. That is, we estimate the following simple quadratic model:

$$S_{i,t}^{synth} = \alpha_i + \beta_{1,i}T + \beta_{2,i}T^2 + \epsilon_{i,t} \tag{4}$$

with time trend T, and where $S_{i,t}^{synth}$ is the synthetically estimated city population for the corresponding German city i at time t. Based on the principles we have postulated in the second section, to construct the synthetic control for now, we consider the covariates to be 1900 and 1920 to 1940 lagged population, 1900 and 1920 number of coal workers, dummies for sea or river access, and lastly the simple 1940 urban potential measure, where we consider 80 West-German cities, and the 765 US places:

$$UP_i = \sum_j^n \frac{S_{i,1940}}{\tau_{i,j}^{\sigma-1}} \tag{5}$$

where we set $\sigma = 2$. We consider all US cities east of the Mississipi and Louisiana between 1870-2000 as comparison cities in the donor pool, as the regions west of the Mississipi were settled much later.[5] Hence we are left with 349 cities (out of 765 US cities with a population of more than 50,000 at some point in the census).

Table 1 presents the summary statistics of our 52 German sample cities and of our 349 US donor pool cities. Notice that the mean population is above the mean synthetic population after 1960. In section 6, we will see that this is not driven by a few single cities, but applies for the majority of cities.

Table 2 presents 3 arbitrarily chosen example cities, Aachen, Lübeck and Dortmund.[6] For Aachen, the discrepancy of the lagged actual and synthetic population is relatively low, while the discrepancy for all other

[5] We consider Louisiana simply because of New Orleans. Other states west of the Mississipi were not settled, thus are not considered to be similar to the old German cities.

[6] Respectively, the criterion of choice are: the city being alphabetically first, the city being the hometown of the author, and its football team being first in the Bundesliga as of January 2019.

covariates is relatively high. This does not necessarily mean that the relatively large discrepancy indicate a poor prediction power of those covariates, but it can also suggest that much of it is already taken away (or embedded) by the population lag covariates. In the case of coal workers, note that coal is a very localized resources and thus the number of cities which can be considered a coal city is sparse especially in the USA [7], which can explain the large discrepancy here, as between cities there is a large discrepancy by itself. That the coal covariate is above zero in this case still indicates that coal cities still match well with other coal cities, despite the sparseness of the resources or coal being embedded in the population lag covariates.

Lastly, given that historical events, such as the decline of coal or the beginning of suburbanization suggests an inverse U-shaped city population growth, which is supported by the results of the individual city regression of (4) being in favor of the inverse U-shaped city growth rate notion. For individual synthetic German cities from 1870 to 2000, we find $\beta_{1,i} > 0$ for all cities and $\beta_{2,i} < 0$ for 50 out of the 52 synthetic cities (96%), with the mean global maximum to be in 1948. If we restrict the time span from 1920 to 2000, we still find $\beta_{1,i} > 0$ for 48 out of 52, and $\beta_{2,i} < 0$ for 51 out of 52 synthetic German cities, with the mean global maximum around 1964 this time.[8]

These results indicate that there are underlying, systemic non-random factors which can influence the patterns of growth of a city system which are not initiated by a large, temporary shock on the city system itself. That is, the notion of first-nature geographies being constant throughout time is at odds with the notion established in this section.[9] Furthermore, it provides a justification for our approach to accommodate for a possible non-linear counterfactual.

Table 1: Descriptive Statistics

	mean (std. dev.)	min / max	mean (std. dev.)	min / max	mean (std. dev.)	min / max
Sample covariates	German cities		Synthetic cities		Donor pool	
population (1900)	206766 (394690)	11704 / 2712190	197717 (346306)	12569 / 2326866	51037 (224465)	0 / 3437202
population (1920)	296873 (563414)	54736 / 3879409	291657 (536690)	55035 /3653581	85288 (360833)	0 / 5620048
population (1930)	324798 (630359)	58300 / 4332834	325803 (632481)	58848/ 4351073	104314 (441805)	0 / 6930446
population (1940)	336567 (632019)	58713 / 4338756	340999 (663543)	58948 / 4593739	109192 (466640)	0 / 7454995
river access	0.788 (0.412)	0 / 1			0.791 (0.407)	0 / 1
sea access	0.019 (0.139)	0 / 1			0.166 (0.373)	0 / 1
coal worker (1900)	28920 (45031)	0 / 135717			2681 (13302)	0 / 92095
coal worker (1920)	56896 (80206)	0 / 219758			5382 (25137)	0 / 154992
urb. pot. (1940)	492988 (619403)	141020 / 4377748			242884 (485770)	34623 / 7627034
Non-covariates						
population (1950)	304948 (501294)	54100 / 3336026	366543 (706337)	59165 / 4877015	121509 (496310)	0 / 7891957
population (1960)	360505 (516693)	80200 / 3274016	354959 (691659)	55809 / 4788218	130726 (486614)	0 / 7781984
population (1970)	363550 (513521)	83300 / 3208719	343965 (691313)	49113 / 4815421	137225 (485596)	0 / 7894862
population (1980)	341610 (484494)	77096 / 3048759	302216 (611862)	44279 / 4283023	132048 (432183)	0 / 7071639
population (1990)	346544 (524886)	78633 / 3433695	295063 (616704)	39959 / 4350805	135037 436943	0 / 7322564
population (2000)	345103 (521048)	78565 / 3382169	295783 (653676)	40205 / 4646081	143370 (470672)	0 / 800827
Sample size	52		52		349	

[7] We consider 20 out of the 349 US cities here as coal city, as compared to the 42 out of 52 German coal cities.

[8] If we only consider cities for which both $\beta_{1,i} > 0$ and $\beta_{2,i} < 0$ are statistically significant at the 5% significance level, then for 1870-2000 we still find 47 cities fulfilling this requirement. Here, the mean global maximum is now around 1960. Similarly for 1920-2000, we now however only find 31 fulfilling this requirement, with the mean global maximum being almost unchanged around 1965 in this case.

[9] Although the notion of changing fundamentals has been explored empirically by Bleakley & Lin (2012) already, where they found that the relative decline of rivers due to the rise of train transportation did not affect portage sites being relevant even today, this development may not be considered a creative destruction process as initially thought by the authors. That is because of their economic importance accumulated over time, portage sites were also connected to railway lines, such that a creative destruction process may have taken place at the very same place first.

Table 2: Individual City Examples

covariates	Treatment Aachen	Synthetic Aachen	Treatment Lübeck	Synthetic Lübeck	Treatment Dortmund	Synthetic Dortmund
population (1900)	142427	133487	82098	82174	233566	249740
population (1920)	151731	154709	118709	118858	513419	481034
population (1930)	154682	158392	129842	130051	535087	544812
population (1940)	162164	158823	154811	155135	542352	544684
river access	0	1	1	1	1	1.001
sea access	0	0	1	0.939	0	0.018
coal worker (1900)	7989	276.285	0	0	135717	51389
coal worker (1920)	15426	464.976	0	0	219758	86486
urban potential (1940)	265614	278847	245207	245052	741768	650725
	Donor city	Weight	Donor city	Weight	Donor city	Weight
	Cincinnati, OH	14.6%	Charleston, SC	67.9%	Cleveland, OH	42.5%
	Fall River, MA	71.6%	Jacksonville, FL	12.7%	Johnstown, PA	55.8%
	Mckeesport, PA	0.3%	Jersey City, NJ	5.6%	New York City, NY	1.8%
	Troy, NY	13.5%	Miami Beach, FL	0.1%		
			Miami, FL	3.5%		
			New Orleans, LA	4.1%		
			Washington, DC	6.2%		

5 Empirical Strategy

To see whether the WWII bombings had any permanent impact or not on a city's population, we test for mean-reversion of the (relative) city population by testing for the absence of a unit root. Consider the typical augmented Dickey-Fuller estimable equation given below:

$$\Delta S_{i,t} = c_i + \delta_i t + \rho S_{i,t} + \sum_{k=1}^{p} \beta_{i,k} \Delta S_{i,t-k} + \epsilon_{i,t} \tag{6}$$

For the city size $S_{i,t}$ for city i at time t. Equation (6) had been estimated in Bosker et al. (2008), however by taking the natural log before. In the case of an inverse-U shaped counterfactual city development, taking the natural log is not necessary as it would not linearize the series. Also taking the log of the size and the log of the share of a city would make both equivalent with each other.[10] Next, we obtain the synthetically corrected (SC) form as the difference of the actual individual city population $S_{i,t}$ with its corresponding synthetic city population $S_{i,t}^{synth}$:

$$S_{i,t}^{SC} \equiv S_{i,t} - S_{i,t}^{synth} \tag{7}$$

Then, equivalent to 6, we write:

$$\Delta S_{i,t}^{SC} = \rho S_{i,t}^{SC} + \sum_{k=1}^{p} \beta_{i,k} \Delta S_{i,t-k}^{SC} + \epsilon_{i,t}^{SC} \tag{8}$$

We drop both the trend term δt and the constant c_i due to our correction in (7), as by construction they are assumed to be zero. Note here that since we do not detrend through OLS, the Dickey-Fuller critical values with trend should not apply as we do not use information after the war: We minimize the discrepancy of the actual and synthetic population in the years close to the war, but not after the war where we let actual and synthetic population diverge. Hence by construction, the values for $S_{i,t}^{SC}$ should be close to 0 before the war and are allowed to diverge from 0 after the war.

In other words, the correct critical values should be close to the Dickey-Fuller critical values with a trend if we were to only consider the pre-war period. However because $S_{i,t}^{SC}$ is allowed to diverge from 0 after the

[10] That is, when we subtract the population by the synthetic population for city i as estimated by SCM, $S_{i,t}^{synth}$.

war, the critical values should approach the significantly less negative critical values for the model without a constant or a trend term.

To account for that, we will also test for the presence of a unit root from 1940 instead of from 1920 onwards (cf. Bosker et al. (2008) test from 1925).[11]

To give us further justification on our approach, we can introduce the notion of the true counterfactual, where the error of our counterfactual estimation is simply the difference of the true counterfactual $S^{*}_{i,t}$ and the counterfactual estimate; in our case $S^{synth}_{i,t}$. For illustrative purposes, we can decompose this error into the sum of three differences; consisting of a common time component ζ which is the same across similar cities between similar countries, a country level component χ common across similar cities in the same country and an individual city component ψ common across similar cities. Note that whenever we say across countries here, it means between German treatment and the US donor pool in our approach specifically. As such, for the post-WWII years $t+K$, we can define:

$$S^{SC*}_{i,t+K} \equiv S^{*}_{i,t+K} - S^{synth}_{i,t+K} \equiv (\zeta^{*}_{i,t+K} - \zeta^{synth}_{i,t+K}) + (\psi^{*}_{i,t+K} - \psi^{synth}_{i,t+K}) + (\chi^{*}_{i,t+K} - \chi^{synth}_{i,t+K}) \tag{9}$$

Ideally, we want (9) to equal 0 for the post-war years $t+K$. [12]

The SCM however is designed mostly to reduce the individual city component error. That is, we may think of $\psi^{synth}_{i,t+K}$ as capturing some common drivers of different cities across countries. For instance, if coal is an important determinant of pre-war city growth, we match German coal cities with the US coal cities convex combination of a very similar size. That is, provided that the individual city characteristics across countries do not differ in impact post-WWII.

The common time component $\zeta^{synth}_{i,t+K}$ captures the drivers of city growth which are the same for each similar city across similar countries. We may think of this time component as technological change which determines city development among similar regions in terms of technology adoption at the same time; such as commuting cost reductions through cheaper transportation technology, or other general trends in city growth common across the world. As compared to the $\psi^{synth}_{i,t+K}$ component, the donor pool choice matters in this case and not the choice of covariates. Also, as previously discussed, we think that reducing the difference $(\zeta^{*}_{i,t+K} - \zeta^{synth}_{i,t+K})$ is best achieved by restrictung the US donor pool to states East of the Mississippi and Louisiana, as we think that they have been overall the most similar in terms of having an already developed city system comparable to Western Europe.

Lastly, we can think of the difference $(\chi^{*}_{i,t+K} - \chi^{synth}_{i,t+K})$ of the individual country level component as a difference in the overall development between countries which they do not have in common. Country specific policies which affect the growth of cities overall may lead to a divergence of this term, such as the separation of Germany into two parts, or a divergence due to overall German and US city development diverging post-WWII.

As such, the approach with SCM accounts for the first two components, so that we are more confident that the first two terms are closer to 0 post-WWII. The placebo exercise of figure 2 is consistent with this, as we use the SCM the control US cities as the treatment unit and as the donor pool, so that we have no cross-country differences here. However the SCM does not account for the last component explicitly. In other words, individual country events post-WWII, whether they are occuring in the US or in Germany, could drive (9) away from 0. [13]

Lastly, that (9) is equal to 0 is less likely the further we move away from the treatment period. That is, the SCM counterfactual is less likely to be close to the true counterfactual if we move away from the shock, either fowards or backwards in time, so that we consider a cutoff year i.e. the year were our analysis ends, at 1980, 1990 and 2000 in the following section.

[11] Or rather 1939, before the war since we set 1939 to 1940 for the German data to match it with the US data.

[12] Note that we do not necessarily want 9 to equal 0 for (all) the pre-war years as a result of the discussion we had about the principals of covariate selection and that covariates closer to the WWII bombing treatment are assumed to be better post-war counterfactual predictors

[13] Note that for the immediate pre-WWII case, if we assume the first two terms of (9) to be 0 then if there is a divergence of the last term, with a degree of freedom of 1 it starts from 0, since we can assume that due to the SCM, (9) overall is 0 for the immediate pre-WWII time case.

Nevertheless, the claim of our approach here is not whether the US city system is considered a good donor pool to estimate post-WWII German city counterfactuals. Taking the city system of other countries as a donor pool may be better or worse than previous methods in reducing $(\chi^{*}_{i,t+K} - \chi^{synth}_{i,t+K})$, which we do not know. The claim is that if we are willing to accept the decomposition above, especially in that SCM is able to reduce the difference of the first two terms appropriately, then we could be closer to the true counterfactual, as compared to the previous approach in the literature following Davis & Weinstein (2002); Brakman et al. (2004); Bosker et al. (2008), which does not control for any of the components.

Furthermore, consistent with the idea that there is a common time component $\zeta^{*}_{i,t+K}$, in the previous section, we have found an inverse U-shaped city growth over time for almost all synthetic, convex combinations of US cities. This supports our claim here that there are underlying factors of city growth which changed over time, consistent with the literature on transport costs and suburbanization for instance (Baum-Snow, 2007; Baum-Snow et al., 2017). Consequently, the same claim could be made for German cities over its entire counterfactual time span, at least for the upper distribution of cities, provided that the same development would have happened in Germany around the same time. In the robustness section 7, we will further test this claim by matching German cities with a 10 year lagged convex combination of US cities, as a case may be made that the US was an earlier adopter of transportation technology or was slightly more technologically advanced.

As we will see in the following, given that equation (8) does not depart much from Bosker et al. (2008) with the exception of the estimation of an explicit counterfactual in our case, our approach will give us significantly differerent results as compared to the previous literature. At the very least, it shows that the exercise above matters, regardless on whether we accept the approach of the previous literature, or this alternative approach.

6 Results

Next, we will report both the impact of WWII on the absolute and relative, to the total West German population, city population as both results were used to show the consistency of the theories in the face of this empirical evidence in Bosker et al. (2008). In other words, we do not change in which way the empirical approach evaluates the fundamental theories, apart from the SCM counterfactual. We will show in section 9 that with only a slightly altered framework, as a consequence, the interpretation of the fundamental theories of the results obtained here will be very similar as well.

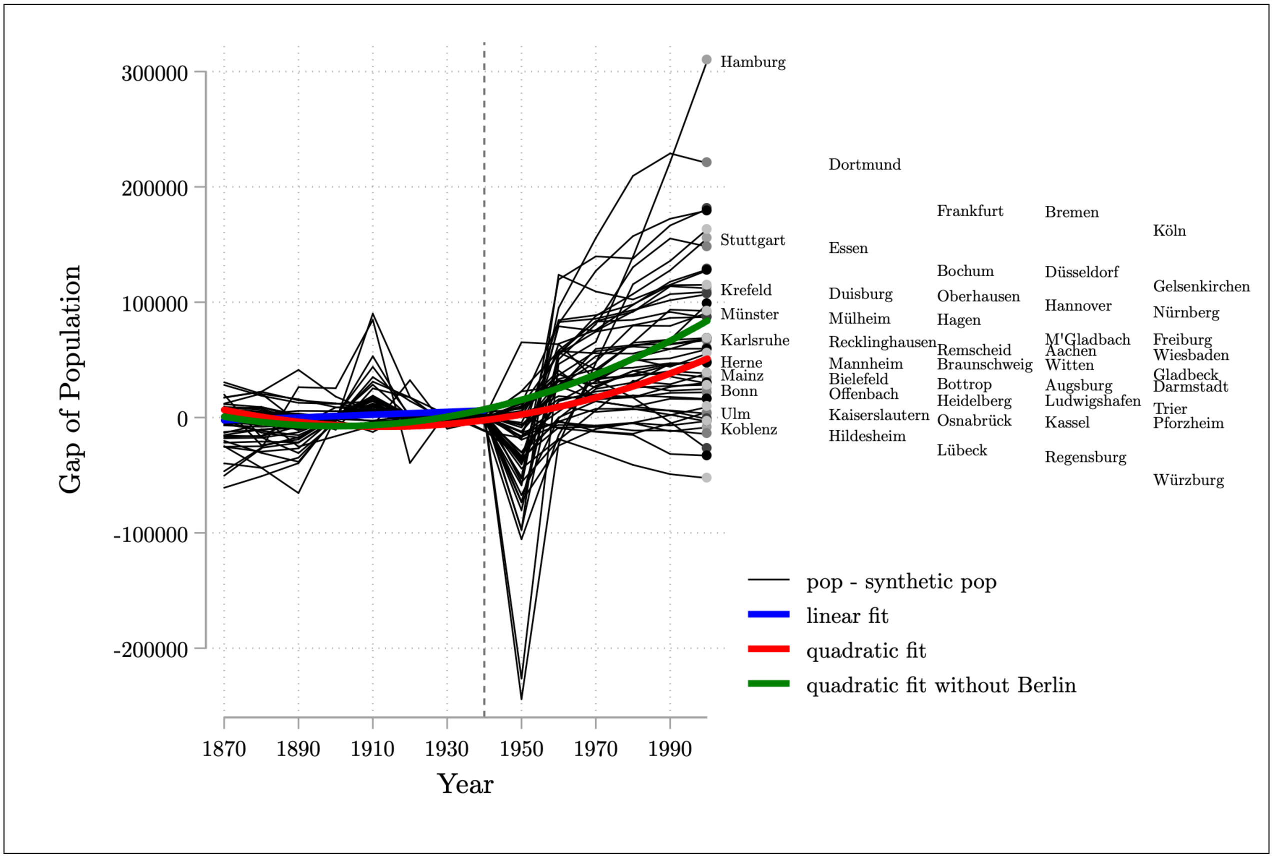
6.1 WWII Bombings Impact on City Size

We keep the same synthetic city as in section 4, where the covariates are 1900 and 1920 to 1940 lagged population, 1900 and 1920 number of coal workers, dummies for sea or river access, and the 1940 urban potential measure as above.

Figure 1 shows the gap of the population with the synthetic population for the individual city, where we have excluded Munich and Berlin as the gap exceeds an absolute value of 500,000 at some point in time. As expected, the gap is close to 0 for almost all cities between 1920-1940 by construction due to the choice of lagged covariates. Since we do not consider covariates far away in time from the treatment period, this manifests itself in figure 1 by the pre-1900 gap diverging from 0. If we consider the principles in section 2, this does not necessarily indicate that the control cities are not well constructed. Also notice that the linear fit line (red) until 1940 is close to 0, which reaffirms our exclusion of a trend term and a constant term of equation (8). Lastly, one thing to take away from the quadratic fits of figure 1 is that a significant number of cities ended up being better off in terms of population, as compared to its benchmark, synthetic population.

In figure 2 we show the gap of the 43 cities in the donor pool which have a weight of at least 0.05 as a comparison city for some German city. Here, we show the gap of these individual comparison cities with its synthetic control, where the donor pool is simply the same US donor pool as before, with the exception of the comparison city itself. That the quadratic fit is roughly a horizontal line around 0 is reassuring as this is what we should expect. That is, as the US did not experience a large, temporary shock on its city

Figure 1: Individual City Gap of Population and Synthetic Population



Note: We exclude cities with a gap larger than 500,000 at some point, which are Berlin and Munich. The linear (blue) and quadratic fit (red) do not exclude Berlin or Munich. The quadratic fit (green) excludes Berlin.

system and since the donor pool consists of the US cities itself, we should find on average no effect i.e. a horizontal line at 0. Although this does not necessarily mean that convex combinations of those comparison cities will behave similarly, it however weakly hints at the post-WWII upward trend in figure 1 to not be driven systemically by the post-WWII counterfactual going in the opposite direction.

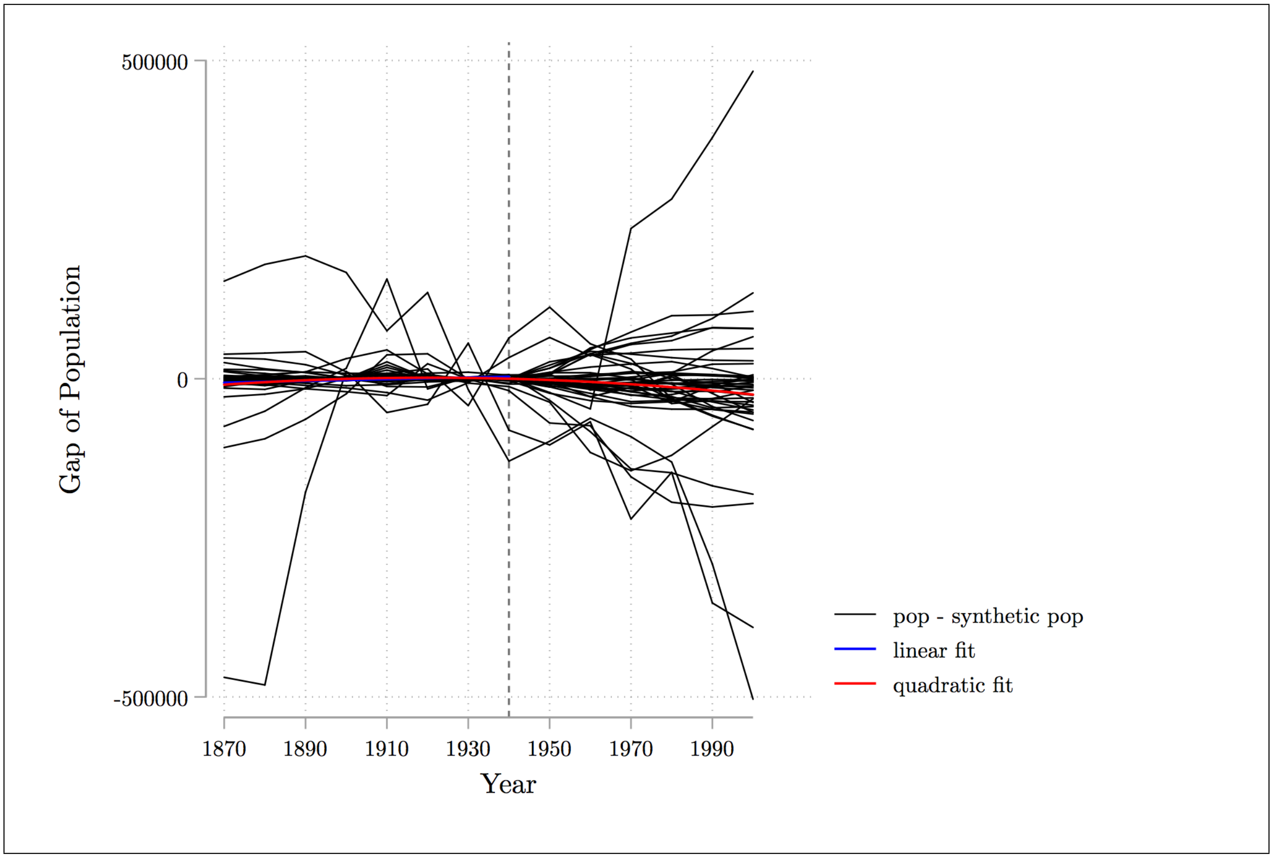
We present the results of the (panel) unit root tests of equation 8 on city size in table 3. The results of the individual city Dickey-Fuller test show that for about 8% to 12% of the cities in our sample with a cutoff year of 1980 to 2000, we find no unit root at the 5% significance level. This still leaves us with 88% to 92% of cities with a unit root. Also note that Bosker et al. (2008) fail to reject the null of a unit root for any significance level for all cities, so we repeat their exercise in b., as the results could be driven by us employing decennial data instead of annual data or the reduction of the number of cities from 62 to 52 among other reasons. This seems to be indeed not the case.[14]

Furthermore, notice that if we change the cutoff year to 1960, we find no unit root for a much larger proportion of cities, which is in line with figure 1 where we find that already in 1960 for 38 cities, the actual population is above the synthetic population. For one thing, this may help us explain the different conclusion of Brakman et al. (2004) of finding mean reversion, against Bosker et al. (2008), who find a unit root for most cities, as the cutoff year is set at 1963 and 1999 respectively.

In B., we additionally show the panel unit root tests. The one we prefer here is the Levin-Lin-Chu (Levin et al., 2002) test with a suppressed constant ($c_i = 0$). We have also included the Im-Pesaran-Shin test with

[14] The difference between taking the log and not is also marginal, as we find a 4%, 12% and 13% rejection of the null respectively at the 1%, 5% and 10% significance level for 1920-2000 if we take the log.

Figure 2: ≥ 0.05 Weight Comparison City Gap



Note: We exclude New York from this figure as there is no US city which can match New York in terms of population. The synthetic New York city is Chicago with a weight of 1, so that the gap is simply the population difference of New York and Chicago

.

a constant but allows for the test statistic to vary across groups. Yet, a suppressed constant is preferred if we consider the linear fit of figure until 1940 1. As for the former Levin-Lin-Chu, we reject the null that the panels contain a unit root against the alternative that the panels are stationary. As for the Im-Pesaran-Shin test (Im et al., 2003), we fail to reject the null of all panels containing a unit root against the alternative of some panels being stationary.

Furthermore, for the sake of comparison, we show the OLS detrended panel unit root test results of equation 6 in b. For one thing, the OLS detrended panel unit root test results change with regard to a earlier cutoff year of analysis. Even though the individual city OLS detrended unit root results do not change much with earlier cutoff year of 1990 and 1980, we fail to reject the null when the cutoff is 1980 for the Im-Pesaran-Shin test. This sensitivity of results is not found for the synthetically detrended panel unit root.

In any case, the results of the synthetic detrended unit root test stands in contrast to the OLS detrended results and Bosker et al. (2008), where a unit root is always found for all of their 62 sample cities.

We have seen already in figure 1 that for 43 out of the 52 cities in 2000, the actual population at some point after the war is above the synthetic population. In the following, we categorize cities as 'positive' ('negative') impact cities if we find unit root conditional on the actual population being above (below) the synthetic population in the given cutoff year. Table 4 presents the share of an individual city of all cities based on this categorization.

On a 5% significance level, around 70% to 80% of the cities in our sample show a permanent positive

Table 3: Results of unit root tests on city size

A. Individual City (augmented) Dickey-Fuller test

	a. Synthetic detrended ($c_i = 0$)				
Period	1920-2000	1920-1990	1920-1980	1920-1970	1920-1960
Significance level	% Unit root rej.	% Unit root rej.	% Unit root rej.	% Unit root rej.	% Unit root rej.
1%	2	2	2	4	42
5%	8	10	12	19	50
10%	15	17	21	31	60
	b. OLS detrended ($\delta_i t = 0$ or $\delta_i t \neq 0$)				
1%	0	0	0	2	12
5%	2	0	0	4	17
10%	6	8	8	15	19

B. Panel Unit root test

	a. Synthetic detrended				
	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)
Levin-Lin-Chu ($c_i = 0$)	-2.721 (0.003)	-3.274 (0.001)	-4.109 (0.000)	-5.492 (0.000)	-7.383 (0.000)
Im-Pesaran-Shin	1.280 (0.900)	1.898 (0.971)	1.646 (0.950)	1.602 (0.946)	-0.265 (0.396)
	b. OLS detrended (i.e. with trend term $\delta_i t$)				
Levin-Lin-Chu	-16.650 (0.000)	-16.964 (0.000)	-16.206 (0.000)	-22.483 (0.000)	-23.946 (0.000)
Im-Pesaran-Shin	-3.412 (0.000)	-2.180 (0.015)	-0.303 (0.381)	-2.151 (0.016)	-2.405 (0.008)

Note: The null hypothesis is a unit root in city size. The individual city unit root number of lags are chosen following Ng & Perron (1995), where the maximum lag length started at 2 for this procedure. Incidentally, the lag is chosen to be zero for all cities. $c_i = 0$ indicates a suppressed constant in the regression. $\delta_i t = 0$ or $\delta_i t \neq 0$ indicates an added trend in the regression. Given the individual city optimal number of lags, we set the lag to 0 for the panel unit root tests.

impact; a result which stays constant throughout the cutoff years. Furthermore, the share among cities with a unit root for which we find a positive rather than a negative impact in the given cutoff year is around 85%; and this share seems to be constant with changing significance levels or cutoff years.

One interpretation of the results of this section is that the bombing shock led to a creative destruction/migration process in cities for which we find a positive permanent impact. In section 7, we will look at the characteristics of those cities.

This also helps us explain the opposite findings between both the Levin-Lin-Chu and the Im-Pesaran-Shin test, as the latter allows for the heterogeneity of the ρ while the former does not. Judging from the mixed results on the individual city unit root test, a heterogenous ρ_i makes the most sense for us.

In order to determine the impact on the entire city system itself, Bosker et al. (2008) also looked at the evolution of the city size distribution. They found a more even city size distribution i.e. relatively more middle sized cities. However, they were only able to compare the post-war distribution to the pre-war distribution. We think that we are now in a better position for this endeavor, since we have an actual counterfactual in the synthetic control. That is, we can simply compare the actual city size distribution with the synthetic city size distribution over time. As such, we will perform the two-sample Kolmogorov-Smirnov test on the actual and the synthetic city size for each year. We report the results in table 5.

For each year, we fail to reject the null that both distributions are the same. Note that we more strongly reject the null between 1920 and 1940, which is by construction due to the synthetic control lag covariates we included in those years. Also note that the p-value declines after the war over time, which may indicate that the effect of the bombings on the city size distribution did not fully dissipate yet and that we may reject the null if we extend the cutoff year; and this would go against the very idea of having a cutoff year. Furthermore, given 38 cities out of 52 in our sample were above to the synthetic city population already in 1960, and heavily based on our estimate of this section, it would seem incredulous that it would take another 50 years for the effect of WWII to fully dissipate.

To conclude this section, despite the widespread bombings on German cities, the city size distribution did not diverge from its supposed counterfactual. Instead, based on the evidence in this section, it points towards

Table 4: Number of positive (negative) impact individual cities with a unit root

	Synthetic detrended ($c_i = 0$)		
Period	1920-2000	1920-1990	1920-1980
Significance level	% Unit root $\|S_{i,t} - S^{SC}_{i,t} > 0$ (% Unit root$\|S_{i,t} - S^{SC}_{i,t} < 0$)		
1%	81 (17)	83 (15)	83 (15)
5%	79 (13)	77 (13)	73 (15)
10%	75 (10)	71 (12)	69 (10)
	ratio of positive to positive and negative cities		
1%	0.82	0.84	0.84
5%	0.85	0.85	0.83
10%	0.89	0.86	0.88

Table 5: Actual and synthetic city equality of distributions

Kolmogorov-Smirnov			
Year	p-value	Year	p-value
1870	0.998	1940	1.000
1880	0.998	1950	0.734
1890	0.998	1960	0.570
1900	0.970	1970	0.291
1910	0.734	1980	0.195
1920	1.000	1990	0.195
1930	1.000	2000	0.195

Note: The null hypothesis is the equality of both distributions.

the distribution results of Bosker et al. (2008) to be driven by underlying factors which are not captured by their method. This however does not suggest that there is no permanent effect stemming from the WWII bombings, but that it did not induce a change in the distribution. Whether however the WWII bombings were indeed not 'enough' to induce a change in the distribution i.e. that this change in the distribution would even be possible in the first place, cannot be answered here.

6.2 Results City Share

Following Davis & Weinstein (2002), who looked at the impact of WWII bombings on the relative city size of Japanese cities, other papers (Brakman et al. (2004), Bosker et al. (2008)) followed suit in looking at the impact of WWII on the relative city size of German cities. That is with the total West German population[15] $s^{total}_{i,t}$, the relative city size is $s_{i,t}$

$$s_{i,t} = \frac{s_{i,t}}{s^{total}_{i,t}} \tag{10}$$

$$s^{synth}_{i,t} \equiv \frac{s^{synth}_{i,t}}{s^{total}_{i,t}} \tag{11}$$

We consider the synthetic relative city size to be the ratio of the same synthetic control as in the previous section, and the actual total West German population $s^{total}_{i,t}$. The synthetically detrended relative city size is

$$s^{SC}_{i,t} \equiv s_{i,t} - s^{synth}_{i,t} \tag{12}$$

[15] We consider the territory of 1957-90 and keep it constant for the total population between 1870-2000.

Then, equivalent to (1), we write:

$$\Delta s_{i,t}^{SC} = \rho s_{i,t}^{SC} + \sum_{k=1}^{p} \beta_{i,k} \Delta s_{i,t-k}^{SC} + \nu_{i,t}^{SC} \tag{13}$$

Table 6: Results of unit root tests on city share

A. Individual City (augmented) Dickey-Fuller test					
	a. Synthetic detrended ($c_i = 0$)				
Period	1920-2000	1920-1990	1920-1980	1920-1970	1920-1960
Significance level	% Unit root rej.	% Unit root rej.	% Unit root rej.	% Unit root rej.	% Unit root rej.
1%	4	4	4	8	40
5%	13	17	21	23	52
10%	23	29	29	38	63
	b. OLS detrended ($\delta_i t = 0$ or $\delta_i t \neq 0$)				
1%	13	12	6	4	2
5%	23	17	13	15	6
10%	33	27	17	21	17
B. Panel Unit root test					
	a. Synthetic detrended				
	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)
Levin-Lin-Chu ($c_i = 0$)	-4.758 (0.000)	-5.024 (0.000)	-5.677 (0.000)	-6.818 (0.000)	-8.071 (0.000)
Im-Pesaran-Shin	-0.166 (0.434)	0.583 (0.720)	0.571 (0.716)	0.610 (0.729)	-0.486 (0.313)
	b. OLS detrended ($\delta_i t = 0$ or $\delta_i t \neq 0$)				
Levin-Lin-Chu	-15.245 (0.000)	-17.923 (0.000)	-21.527 (0.000)	-23.274 (0.000)	-28.792 (0.000)
Im-Pesaran-Shin	-2.349 (0.009)	-2.920 (0.002)	-3.249 (0.001)	-3.393 (0.000)	-3.263 (0.001)

Note: The null hypothesis is a unit root in city size. The individual city unit root number of lags are chosen following Ng & Perron (1995), where the maximum lag length started at 2 for this procedure. Incidentally, the lag is chosen to be zero for all cities. $c_i = 0$ indicates a suppressed constant in the regression. $\delta_i t = 0$ or $\delta_i t \neq 0$ indicates an added trend in the regression. Given the individual city optimal number of lags, we set the lag to 0 for the panel unit root tests.

Table 6 reports the results of the relative city size. Overall, we now find a higher share of cities where we reject the null of a unit root as compared to the non-relative results of previous section. For the synthetically detrended unit root test on the relative city, we fail to reject the null of a unit root for 13% of cities at the 5% significance level if we set the cutoff year to 2000. This increases to 17% and 21% if we set the cutoff year to 1990 or 1980 respectively, and even to 52% if the cutoff year is 1960. This pattern stands in contrast with the OLS detrended series for which we more often find a unit root as we go to earlier cutoff years.

The most relevant Levin-Lin-Chu test with a suppressed constant ($c_i = 0$) also reports that we reject the null of the panels containing a unit root against the alternative of the panels being stationary for all cutoff years. As for the Im-Pesaran-Shin test on the synthetically detrended series, we reject the null of all panels containing a unit root against the alternative of some panels being stationary for all cutoff year.

Furthermore as compared to Bosker et al. (2008), we find a higher share of cities rejecting a unit root when we employ the OLS detrended series for the later cutoff years. This is likely not driven by the lower frequency as this would imply a lower power of the test, thus implying a higher incidence of rejecting the null hypothesis of a unit root. It could however be driven by not including lags due to the decenniality of the time series. Nevertheless, we find that looking at the relative city size increases the share of individual cities rejecting the null of a unit root, as compared to looking at the absolute city size.

Table 7 presents the share of individual cities with a unit root on the relative size, conditional on the actual population being above (below) the synthetic population in the given cutoff year. We again find that the share among cities with a unit root for which we find a positive rather than a negative impact in the given cutoff year is around 85%, which stays constant throughout significance levels or cutoff years.

Table 7: Positive (negative) relative impact individual city unit root

	Synthetic detrended ($c_i = 0$)		
Period	1920-2000	1920-1990	1920-1980
Significance level	% Unit root $\|s_{i,t} - s^{SC}_{i,t} > 0$ (% Unit root $\|s_{i,t} - s^{SC}_{i,t} < 0$)		
1%	81 (15)	81 (15)	81 (15)
5%	75 (12)	69 (13)	67 (12)
10%	67 (10)	62 (10)	62 (10)
	ratio of positive to positive + negative cities		
1%	0.84	0.84	0.84
5%	0.87	0.84	0.85
10%	0.88	0.86	0.86

Overall, the evidence based on the individual city and panel unit root test is mixed, but can be rationalized. First, notice the difference in inference between the Levin-Lin-Chu and the Im-Pesaran-Shin panel unit root tests. As in section 6, the mixed results of the individual city unit root test indicate that ρ should be heterogeneous i.e. the Im-Pesaran-Shin test being the more appropriate test in our case even if we do not suppress the constant. Nevertheless, the evidence from the individual city unit root test points towards a permanent effect especially after the 1970s.

7 Robustness

As mentioned, one caveat of the SCM as used in the literature is that the counterfactual, and as such the results itself, are sensitive to the choice of the covariates. Since we are applying the same set of covariates to all German cities, we are arguably more transparent as compared to if we were to use the SCM on only one treatment unit. Nevertheless, a systemic difference in results within the set of treatment units stemming from a different set of covariates may still occur. The purpose of this section is to find out whether there is such a systemic change in results.

7.1 Covariates Closer to Treatment Period

The approach here is to go towards better educated guesses in the choice of covariates and to restrict the number of comparison cities in the donor pool towards cities we think are even more likely to be similar to the sample German cities. As such, we remove all cities which had a population of 0 in 1870 since all German cities in the sample already existed so that we now have 220 cities in the donor pool. As a consequence, we exclude the population and the number of coal worker covariates of 1900 in order to shift even more towards covariates close to the beginning of WWII, while we keep the other covariates as before.

Figure 3 shows the gap of the population with the synthetic population given this alternative specification. As compared to figure 1, we see that a majority, 43 of the 52 cities in 2000 in terms of population are above its synthetic population. Figure 4 compares the first model with the second model. Most cities are located around the 45 degree line i.e. there is little systemic difference.[16] The initial look here suggests that the adjustments made here have little effect on the results.

Table 8, the individual city results of the second specification in I-III now show a higher rejection rate of the unit root null hypothesis as compared to the first specification. However, among the individual cities for which we find a unit root, we still find that around 85% of those are positive impact cities.

7.2 No Coal Covariates

We may 'wrongly' match US coal cities, which are located mostly around the less dense Appalachian basin, with the more dense Ruhr area cities. Hence, in the third specification, we further exclude the number of coal workers covariate. In figure 5, we again see that 43 of the 52 cities end up above the synthetic population in 2000. In figure 6, the mass is slightly left the 45 degree line towards the first specification, suggesting a less 'positive' overall effect but not very significant.

Table 8 IV-VI shows the individual city results of the no coal specification, and we slightly less often reject the unit root null hypothesis as compared to the previous second specification. In any case, the exclusion of coal does not change the results much.

7.3 Unit Root from 1940 onwards

By construction, the gap of the population is close to 0 in the 20 years before the war. As discussed before, this may result in critical values being too low, as we have not considered the artificial set trend before the war. Table 8, VII-IX reports the results of the unit root test on the specification of section 7.1, where we instead start in 1940 and thus sacrifice some statistical power. Overall, we reject the unit root null hypothesis for around 90% of the sample cities.

Interestingly, we now fail to reject the Im-Pesaran-Shin panel unit root null hypothesis for all cutoff years.

7.4 Forced Matches between Geographic Characteristics

In this exercise, we will force a match of river cities with other river cities, and cities with sea access with other sea access cities. That is, we split the donor pool up into 4 separate ones. In a sense, we will rely less on the lagged population as a predictor of post-war population and more on the geographic characteristics. For cities which have both river and sea access, this reduces the donor pool to 47 US places, and 1 German city, Lübeck. We did not count Hamburg or Bremen for instance as a sea-access city, as the part to the sea is technically still the Elbe/Weser river.[17]

As for cities with sea access, but no river access, there is only Flensburg in the larger sample which we did not consider in our sample. In any case, the donor pool would have consisted of 11 cities altogether i.e. sea access cities without river access are quite rare, which has probably something to do with fresh-water access.

As for cities with river access, but no sea access, there are now 40 German cities which fits this criteria, and 229 US places in the donor pool. Note that with this restriction, we fit Berlin with Chicago with a weight of 1, as New York is thrown out of the sample here.

As for cities with neither river access or sea access, 11 German cities of this type are in the sample, with 62 US places in the donor pool.

Table 9, X-XII reports the results of this exercise. The conclusions drawn from previous specifications still hold here. As compared to the baseline model, we slightly more often reject the unit root null hypothesis, and find an overall higher rate of positive to negative impact cities. Also, given that we have split up the donor pool into 4 different once, which severely restricted the number of units in the donor pool for some cases, it did not change the results much. This is not too surprising given that the synthetic city for most cases are heavily weighted towards the given geographical characteristic in any case, which is simply even closer or as close as possible in this exercise. The corresponding figures are 7 and 8. Note here that we have included Bremen and Hamburg as if they were also sea-access cities with the grey squared symbol. Considering

[16] We have again excluded Berlin and Munich from this figure. The gap for Berlin for the first model is -1263912 and -1335310 for the second, so that the difference here is 71938. The gap for Munich is 525026.6 and 502199, so that the difference here is 22827.6.

[17] If they would be considered sea-access cities, then Bremen would now be considered a negative impact city, and not a positive impact city, except if the cutoff year is 1980 where we reject the null hypothesis of a unit root at the 5% significance level. As for Hamburg, we would now reject the null hypothesis of a unit root at the 5% significance level for the cutoff years 1980-2000, as compared to the X-XII case where we would reject it if the cutoff year is 1980, but not if the cutoff year is 1990-2000.

Bremen as a sea-access city in this case creates a significantly different synthetic Bremen post-WWII, which is not so much the case for Hamburg. Arguably, Hamburg is more so a sea-access place than Bremen, as Bremen has its designated Bremerhaven as its port while the Hamburg port is located in Hamburg. That is, considering Bremen as having de facto sea-access may have been wrong after all. Also, the one city which did not reduce the gap in figure 7 is Bochum, which is not surprising since Bochum is bigger than any of the 62 donor pool cities in the restricted pool of having neither sea or river-access. Fortunately, this seems to be the case for Bochum and Berlin alone, so that the results are not too distorted due to that.

Note also that a majority of cities are on the 45 degree line in figure 8, which simply reflects that in any case, the typical synthetic city in the baseline model was created out of US cities with the very same geographic characteristic, making the restriction not binding most of the time.

7.5 Matching German Cities with 10 Year Earlier US Cities

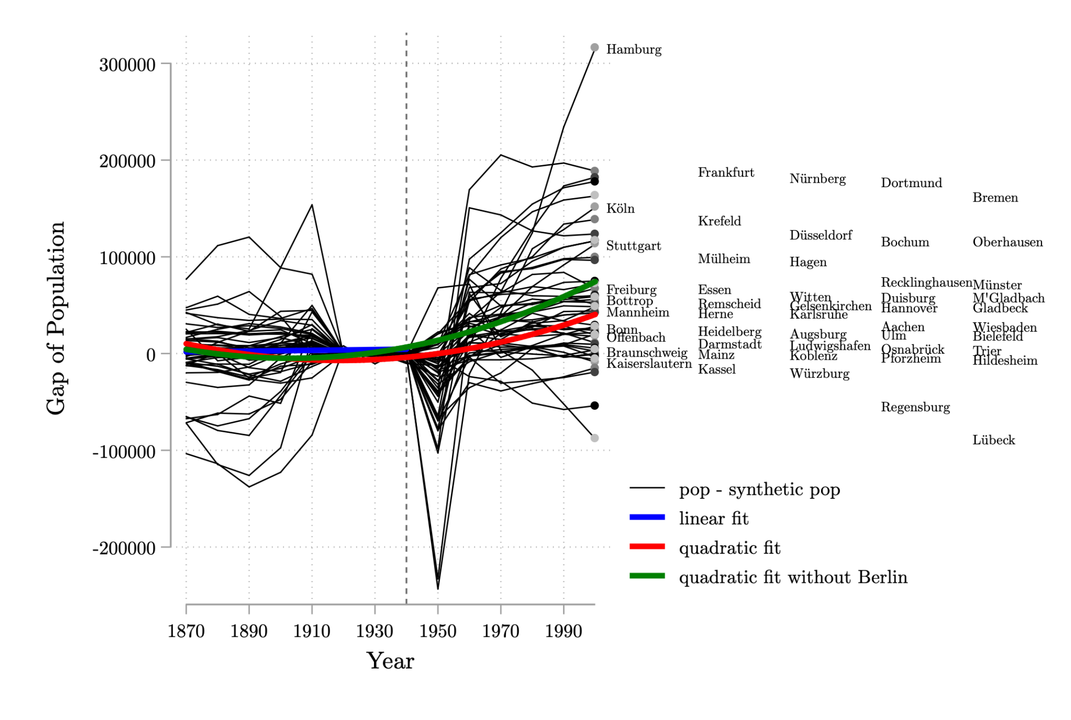
> ”Uncle August showed me around his self-directed factory, a small, simple iron foundry. They had around 15 to 20 workers and employees. But the fascinating thing was: There were as many cars in front as there were people employed, everyone owned a car! We did not dare to dream about such things in Germany; in fact the motorization of the ordinary Joe in our country was achieved on similar scale at first only in the 1970s.” - Helmut Schmidt (2011) about his first visit to the US in 1950 in Duluth, MN.

We may think that the US is more technologically advanced in certain aspects, leading for instance to an earlier urbanization and suburbanization due to earlier widescale adaptation of transportation technology as compared to the true counterfactual Germany. That is, there may had been a delay of technology adaptation which is, at least, not only caused by WWII itself. We will thus consider US incorporated places 10 years earlier as comparison units within the donor pool.[18] That is, we will match present German covariates with the US covariates 10 years earlier, including 10 year earlier coal covariates and urban potential measures. Table 9, XIII-XV reports the results, showing a lower rate of rejection of the unit root null hypothesis as compared to the baseline specification. The corresponding figures of this specification are 9 and 10.

Since we are also, more or less, building a synthetic control out of US cities a decade earlier, then it suggests that the inverse U-shaped development that we have described in section 4 should similarly occur earlier as well. That is, the decline around the 1960s of the synthetic city should now occur closer to the 1950s, which now coincides more with WWII. Thus, we should find that the gap of population to not be as negative in this specification as compared to the baseline specification right after WWII. The average gap of population in 1950 is -61595 in the baseline specification and -62534 in this specification. However, excluding Berlin here, we now find -32587 and -21121 respectively as conjectured. Interestingly in figure 10, the points are above the line indicating that for 1980, 41 out of 52 cities were above 45 degree line. To compare, this is the case for 33 out 52 German cities in figure 4, 35 out of 52 in figure 6 and 25 out of 51 in figure 8. Nonetheless, we would still draw a similar conclusion as previous specifications.

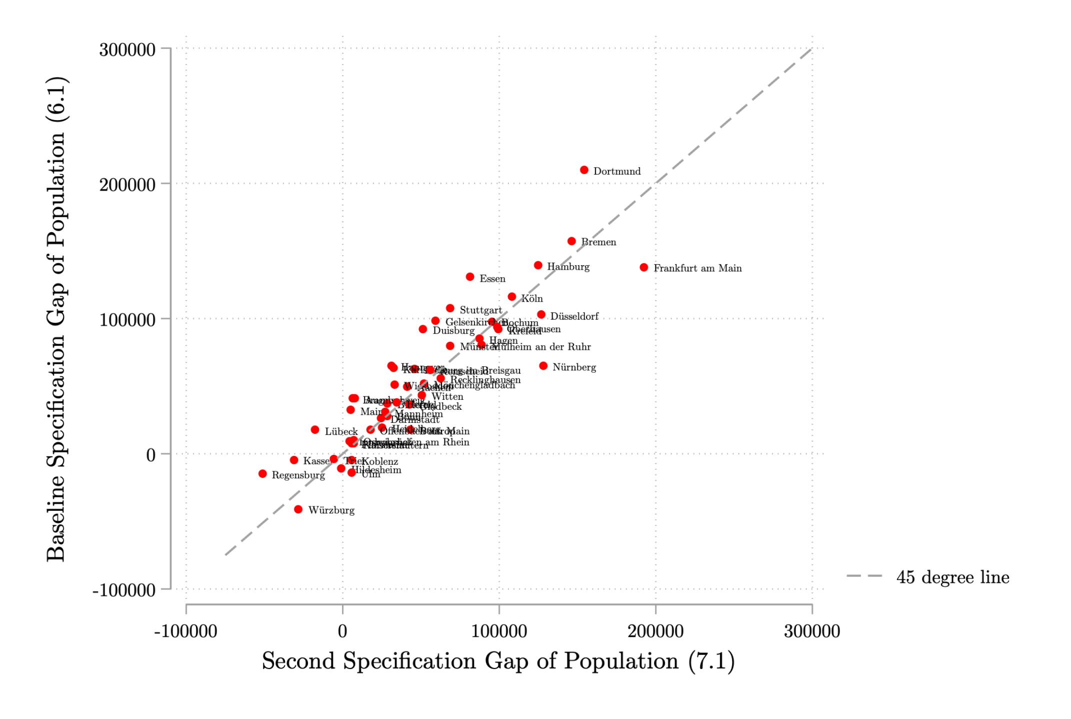
[18] We may interpret this approach as relaxing the first term in equation 9 and while trying to instead reduce the third term.

Figure 3: Individual City Gap of Population and Synthetic Population: Covariates Closer to Treatment Period



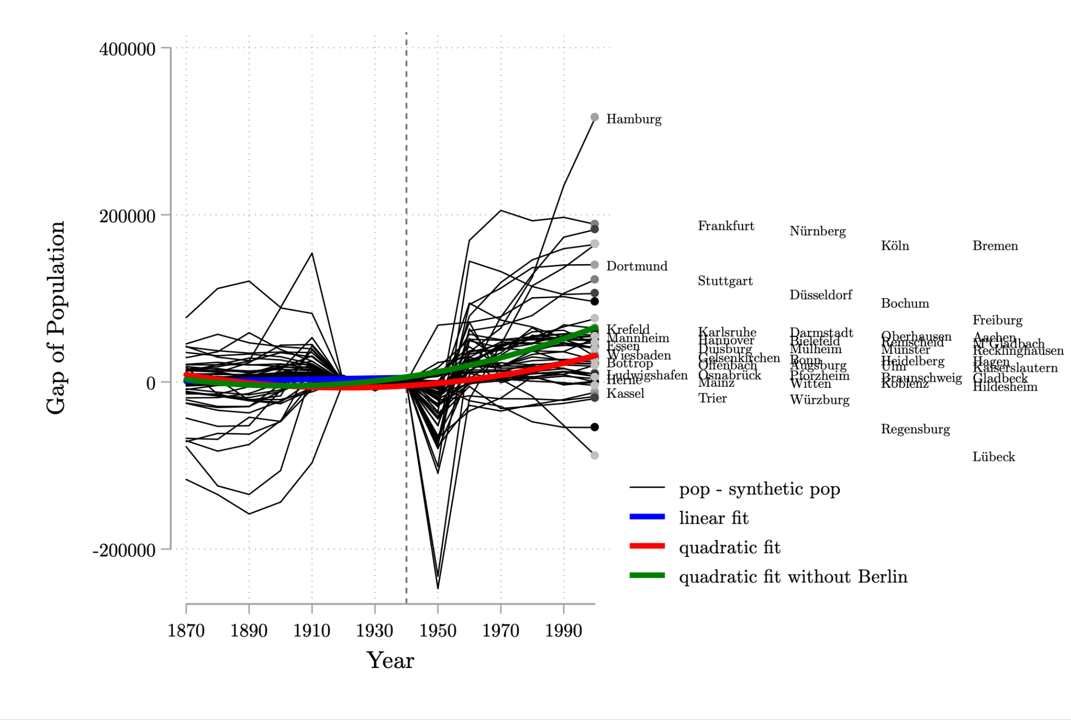
Note: We exclude cities with a gap larger than 500,000 at some point, which are Berlin and Munich. The linear (blue) and quadratic fit (red) do not exclude Berlin or Munich. The quadratic fit (green) excludes Berlin.

Figure 4: Difference in Gap of Figure 1 and Figure 3 in 1980



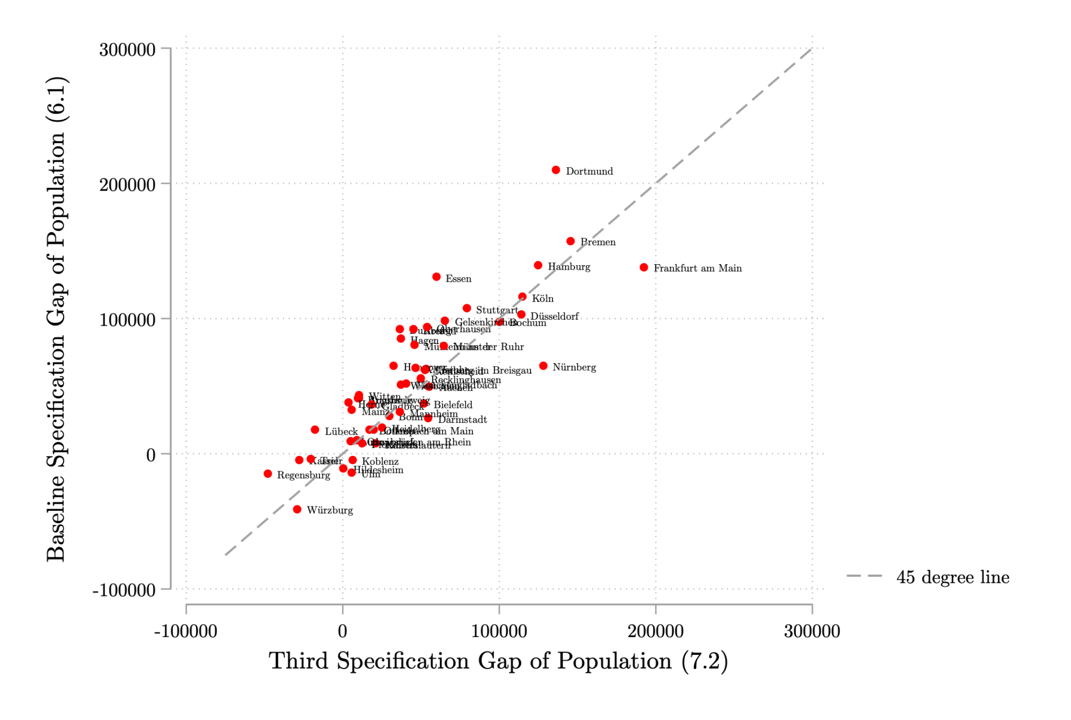
Note: We exclude cities with a gap larger than 500,000 at some point, which are Berlin and Munich.

Figure 5: Individual City Gap of Population and Synthetic Population: No Coal Covariates



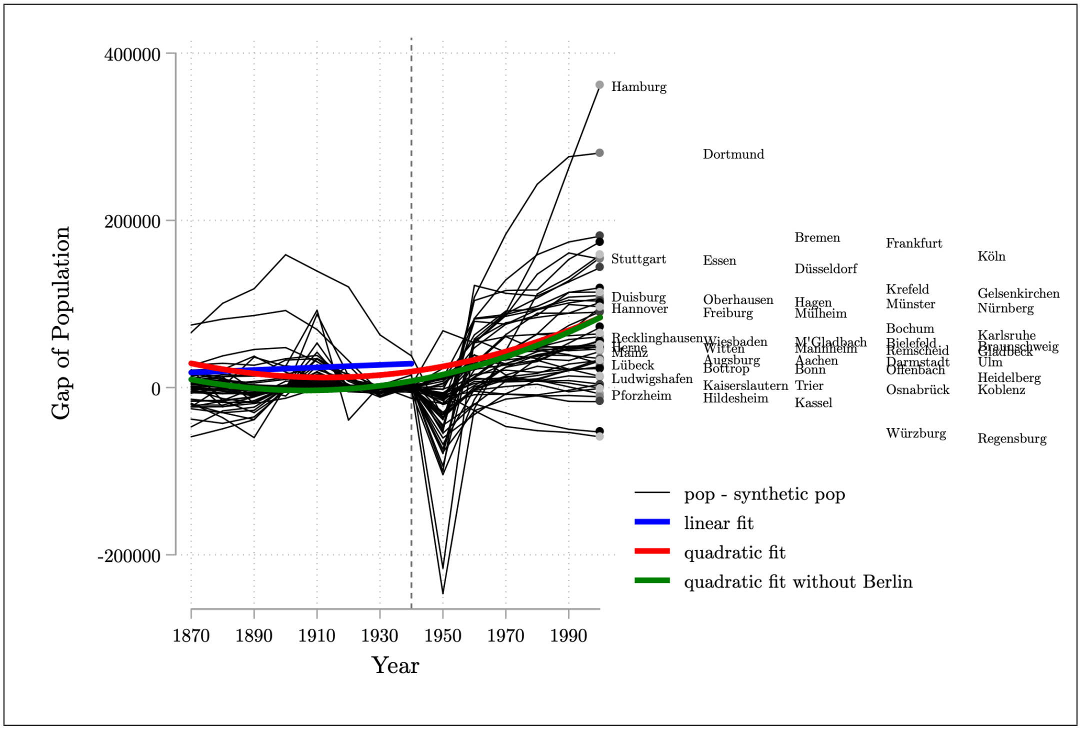
Note: We exclude cities with a gap larger than 500,000 at some point, which are Berlin and Munich. The linear (blue) and quadratic fit (red) do not exclude Berlin or Munich. The quadratic fit (green) excludes Berlin. As compared to figure 1, this figure represents the model which excludes any city which had a population of 0 at any time between 1870-2000, and excludes any covariate related to 1900 or to coal.

Figure 6: Difference in Gap of Figure 1 and Figure 5 in 1980



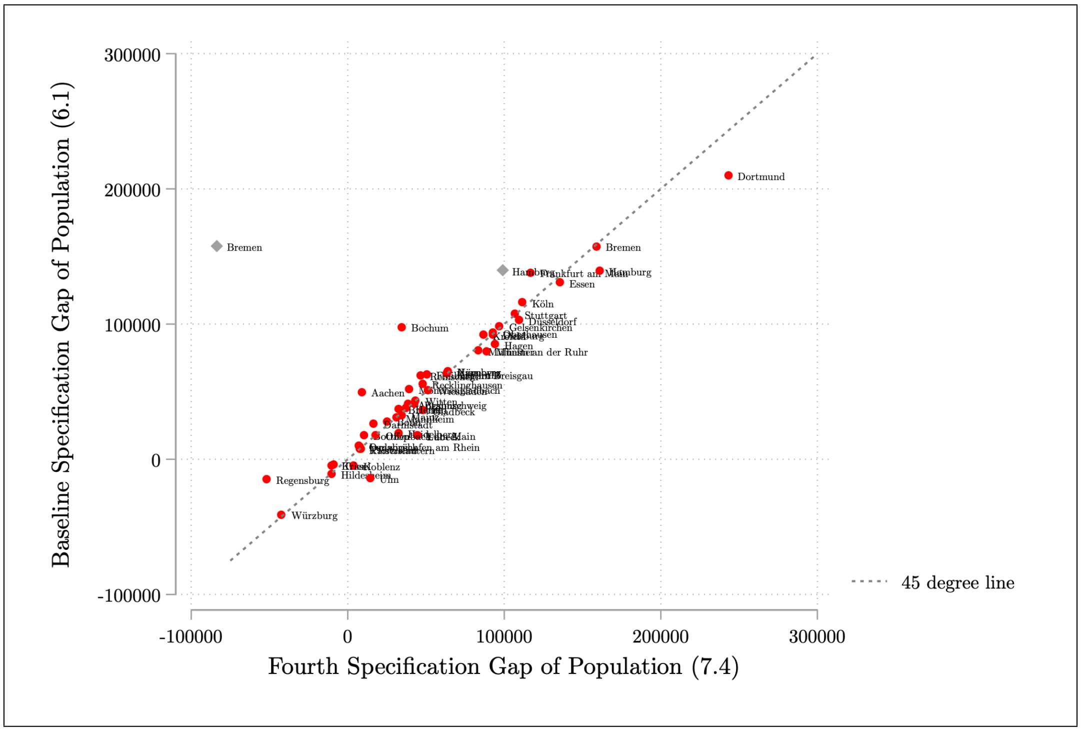
Note: We exclude cities with a gap larger than 500,000 at some point, which are Berlin and Munich.

Figure 7: Individual City Gap between Population and Synthetic Population: Baseline Model Forced Matches between Geographic Characteristics



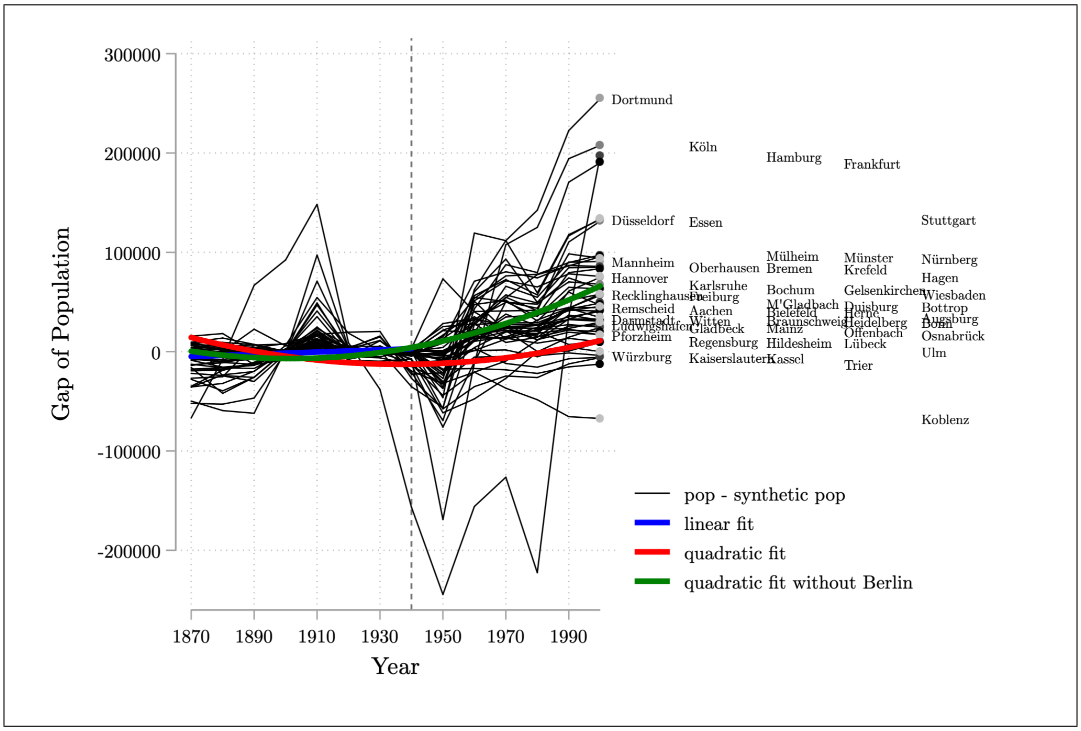
Note: We exclude Berlin and Munich. The linear (blue) and quadratic fit (red) estimation include Berlin or Munich. The quadratic fit (green) excludes Berlin. This specification considers the baseline specification of section 6, but where we match German cities towards 4 different donor pools whether it has sea and/or river access.

Figure 8: Difference in Gap of Figure 1 and Figure 7 in 1980



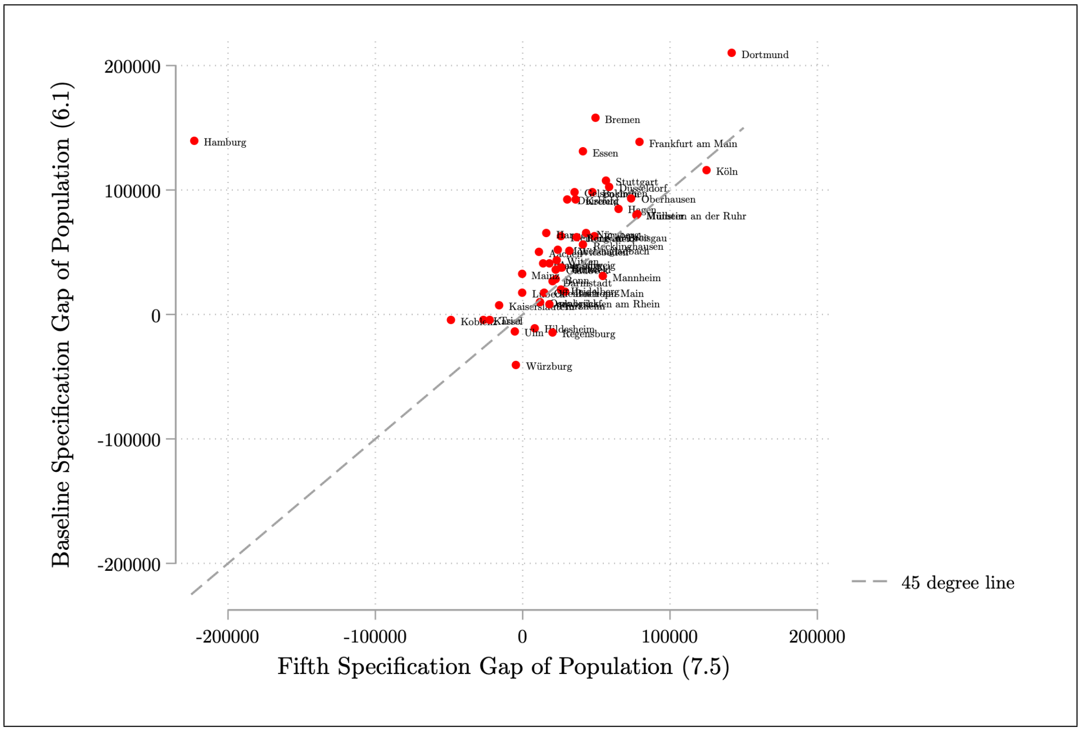
Note: We exclude cities with a gap larger than 500,000 at some point, which are Berlin and Munich. The grey diamond points are when we consider Hamburg and Bremen as also having sea-access, and not only river-access.

Figure 9: Individual City Gap between Population and Synthetic Population: Matching with 10 Year Earlier US Cities



Note: We exclude Berlin and Munich. The linear (blue) and quadratic fit (red) estimation include Berlin or Munich. The quadratic fit (green) excludes Berlin. We consider the baseline specification of section 6, but where we match German covariates with the US covariates 10 years earlier.

Figure 10: Difference in Gap of Figure 1 and Figure 9 in 1980



Note: We exclude cities with a gap larger than 500,000 at some point, which are Berlin and Munich.

Table 8: Robustness on city size I

	A. Individual City (augmented) Dickey-Fuller test								
	a. Synthetic detrended ($c_i = 0$)								
	Second Specification			Third Specification			Second Specification		
	I	II	III	IV	V	VI	VII	VIII	IX
Period	1920-2000	1920-1990	1920-1980	1920-2000	1920-1990	1920-1980	1940-2000	1940-1990	1940-1980
Significance level	% Unit root rejected								
1%	12	10	10	8	8	6	2	0	0
5%	17	21	25	15	17	23	12	10	10
10%	19	31	35	19	25	27	15	13	17
	B. Panel Unit root test								
	a. Synthetic detrended								
	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)
Levin-Lin-Chu ($c_i = 0$)	-3.055 (0.001)	-3.521 (0.000)	-4.149 (0.000)	-3.954 (0.000)	-4.232 (0.000)	-4.415 (0.000)	-2.611 (0.005)	-2.937 (0.002)	-3.266 (0.001)
Im-Pesaran-Shin	0.752 (0.774)	1.329 (0.908)	1.197 (0.884)	0.085 (0.534)	0.748 (0.773)	0.912 (0.819)	-5.222 (0.000)	-4.596 (0.000)	-5.942 (0.000)
	% Unit root $\|s_{i,t} - s_{i,t}^{SC} > 0$ (% Unit root $\|s_{i,t} - s_{i,t}^{SC} < 0$)								
1%	77 (12)	77 (13)	79 (12)	79 (13)	79 (13)	83 (12)	83 (15)	83 (17)	87 (13)
5%	73 (10)	69 (10)	65 (10)	71 (13)	71 (12)	65 (12)	77 (12)	77 (13)	79 (12)
10%	71 (10)	60 (10)	56 (10)	69 (12)	63 (12)	62 (12)	75 (10)	75 (12)	71 (12)
	ratio of positive to positive and negative cities								
1%	0.87	0.85	0.87	0.85	0.85	0.88	0.84	0.83	0.87
5%	0.88	0.88	0.87	0.84	0.86	0.85	0.87	0.85	0.87
10%	0.88	0.86	0.85	0.86	0.85	0.84	0.89	0.87	0.86
Donor pool N	220	220	220	220	220	220	349	349	349

Note: The null hypothesis is a unit root in city size. The lag is chosen to be zero for all cities. $c_i = 0$ indicates a suppressed constant in the regression. We set the lag to 0 for the panel unit root tests.

Table 9: Robustness on city size II

A. Individual City (augmented) Dickey-Fuller test						
	a. Synthetic detrended ($c_i = 0$)					
	Fourth Specification			Fifth Specification		
	X	XI	XII	XIII	XIV	XV
Period	1920-2000	1920-1990	1920-1980	1920-2000	1920-1990	1920-1980
Significance level	% Unit root rejected					
1%	0	0	2	2	2	4
5%	13	13	21	4	4	8
10%	21	25	27	6	10	15
B. Panel Unit root test						
	a. Synthetic detrended					
	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)	t-stat (p-val)
Levin-Lin-Chu ($c_i = 0$)	-2.443 (0.007)	-2.533 (0.006)	-5.313 (0.000)	-2.142 (0.016)	-2.507 (0.006)	0.461 (0.678)
Im-Pesaran-Shin	1.744 (0.960)	2.083 (0.981)	1.259 (0.896)	0.954 (0.830)	0.9135 (0.820)	2.1241 (0.983)
	% Unit root $\|s_{i,t} - s_{i,t}^{SC} > 0$ (% Unit root $\|s_{i,t} - s_{i,t}^{SC} < 0$)					
1%	87 (13)	88 (12)	88 (10)	87 (12)	87 (12)	81 (15)
5%	77 (10)	77 (10)	69 (10)	87 (10)	87 (10)	77 (15)
10%	73 (6)	65 (8)	56 (10)	87 (8)	81 (10)	71 (13)
	ratio of positive to positive and negative cities					
1%	0.87	0.88	0.90	0.88	0.88	0.84
5%	0.89	0.89	0.88	0.90	0.90	0.83
10%	0.93	0.92	0.89	0.92	0.89	0.84
Donor pool N	349	349	349	349	349	349

Note: The null hypothesis is a unit root in city size. The lag is chosen to be zero for all cities. $c_i = 0$ indicates a suppressed constant in the regression. We set the lag to 0 for the panel unit root tests.

8 City Characteristics Conditional on Impact

The purpose of this section is to show the 1940 characteristics of cities for which we find either a stationary series (below a 5% significance level), or otherwise a positive or a negative impact unit root of the relative city size with a cutoff year of 1980. For that, we use our baseline, first specification of section 6. In table 10, if we exclude Berlin, then the average size of cities which experienced a negative impact is around 62% smaller than the average positive impact city, and about 64% smaller than the positive impact cities excluding the Ruhr area cities. Remarkably, they are also about 79% smaller than those cities for which we find a stationary series. The same pattern holds for the urban potential measure in that stationary cities have a higher urban potential in 1940 as compared to positive impact cities, both with and without Ruhr area cities, and in particular a much higher measure than the negative impact cities barring Berlin. Similarly, only two cities within a metropolitan area are negative impact cities, Berlin and Hildesheim.[19] On the other hand, the 12 cities which are not considered within the EMR metropolitan area contribute 4 out of 6 the negative impact cities. Lastly, figure 11 shows the location of all sample cities within Germany, separated by impact. Note that the positive impact cities in figure 11a are highly localized in the Rhein-Ruhr area, and stationary cities in figure 11b both in the Rhein-Ruhr and the Rhein-Main/Rhein-Neckar area. Second, although there are only 6 negative impact cities in figure 11c, they are relatively spread out from each other. Given these patterns and that the Rhein-Ruhr area is the biggest and most dense German metropolitan region, it suggests that relatively bigger cities and metropolitan areas better deal with a shock on the city system. Furthermore, since stationary cities are on average much bigger, it points towards a natural limit to the size of German cities, and is consistent with models of sequential growth (Cuberes, 2011; Henderson & Venables, 2009).

Furthermore, the rate of internal refugees and displaced people (Vertriebene) due to WWII that are residing in this particular city in 1960 is about the same for negative and positive impact cities, but lower for stationary cities. As such, we think that on average, the discrepancy between positive and negative impact cities is not driven too much by a heterogeneous refugee choice of settlement. For instance, apart from Berlin for which we do not have data, the only cities which in 1960 had a rate of refugees relative to its total population below 10% were Trier and Ludwigshafen, at 6.9% and 9.6% respectively. Similarly, the only city above 25% is Lübeck, with 32.4%. If we were to adjust it 17.5±5%, we would add Aachen, Koblenz, Köln, Mainz and Mönchengladbach to the cities with a rate below 12.5%, and Bielefeld, Braunschweig, Hannover and Wiesbaden to the ones above 22.5%. Note that initially after WWII, the French occupied zone restricted the number of refugees. They only received 60,000 at the end of 1947, around 1% of the total population, as compared to the total 4.379 million in the Soviet (around 24.3% of the total at that time), 2.957 million in the US (around 17.7%) and 3.320 million in the British (around 14.5%) occupied zones (Volkmann, 1995), which more than compensated for the total German deaths during WWII. Yet, even though refugees were hindered initially to settle in the French occupied zone, we still find that the average rate is already 11% in 1960 for the 6 French occupied cities in our sample, Freiburg, Koblenz, Kaiserslautern, Ludwigshafen, Mainz and Trier.[20] Also, the mean rate of housing lost is lowest for positive impact cities at around 39%, but highest for stationary cities at 52%, with negative impact cities being in between. We would have expected that cities which lost relatively more housing would be negative impact cities, which would also suggest a more even city size distribution. Also note that the variation is significantly larger for housing lost rate as compared to refugee rate variation for each impact group. This indicates that the number of refugees are remarkably evenly distributed across Germany, given that the housing lost is not.

Nevertheless, it does not suggest that cities which were hit relatively harder are among the group of cities that did not recover, although this idea seem to at least still hold if we were to only compare positive and negative impact cities. If we were to go even further, given these patterns, we would tend to think of WWII more as an overall reset of the city system itself, and not as initial intuition might suggest, a readjustment of the city size system in favor of the city hit less severely by WWII. In that sense, one could think of the immediate aftermath of the war as a Stunde Null, an Hour Zero of the city system.

[19] The German metropolitan areas here are the Europäische Metropolregionen (EMR) as defined by the Ministerkonferenz für Raumordnung (MKRO).

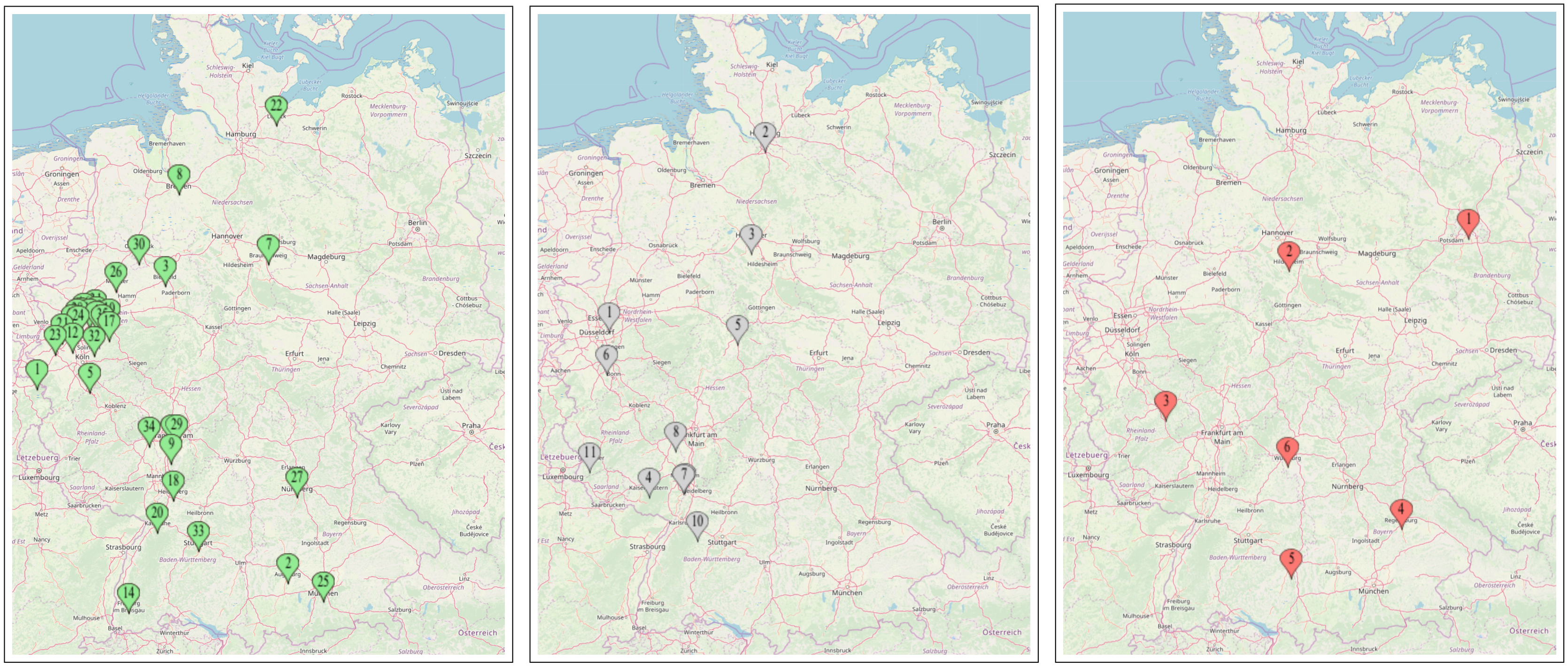
[20] In fact, the other city not making it below 12.5% are Freiburg and Kaiserslautern with 12.8% and 12.6%.

Table 10: 1939 city characteristics conditional on impact with cutoff year 1980

Impact	Positive	Positive no Ruhr	Stationary	Negative no Berlin	Negative
N	35	23	11	5	6
mean Pop	230233	242627	423962	88198	796624
Std. Err.	31169	41446	147991	6654	708447
mean Urb. Pot.	408402	360189	554845	172048	872998
Std. Err.	31831	39798	149540	10522	701003
mean Urb Pot. - Pop	178168	117563	130883	83850	76373
Std. Err.	16891	10996	21827	9775	10936
mean Coal Worker (1920)	78914	29352	17875	3	3
Std. Err.	14708	11525	12616	2	2
mean Vertriebenenrate (1960)	0.173	0.179	0.137	0.177	0.177
Std. Err.	0.007	0.010	0.013	0.016	0.016
mean housing lost	0.387	0.380	0.520	0.460	0.445
Std. Err.	0.026	0.031	0.038	0.113	0.093
Metropolitan regions					
Rhein-Ruhr (19)	17	5	2	0	0
Rhein-Main (5)	4	4	1	0	0
Rhein-Neckar (3)	1	1	2	0	0
Braunschweig (3)	1	1	1	1	1
München (2)	2	2	0	0	0
Hamburg (2)	1	1	1	0	0
Oberrhein (2)	2	2	0	0	0
Nürnberg (1)	1	1	0	0	0
Bremen (1)	1	1	0	0	0
Stuttgart (1)	1	1	0	0	0
Berlin (1)	0	0	0	0	1
Not Metropolitan	4	4	4	4	4
List of cities in the corresponding sample					
	Aachen	**Aachen**	Essen	Hildesheim	Berlin
	Augsburg	Augsburg	Hamburg	**Koblenz**	Hildesheim
	Bielefeld	**Bielefeld**	Hannover	**Regensburg**	**Koblenz**
	Bochum	Bonn	**Kaiserslautern**	**Ulm**	**Regensburg**
	Bonn	Braunschweig	**Kassel**	**Würzburg**	**Ulm**
	Bottrop	Bremen	Köln		**Würzburg**
	Braunschweig	Darmstadt	Ludwigshafen		
	Bremen	Düsseldorf	Mainz		
	Darmstadt	Frankfurt	Mannheim		
	Dortmund	Freiburg	**Pforzheim**		
	Duisburg	Heidelberg	**Trier**		
	Düsseldorf	Karlsruhe			
	Frankfurt	Krefeld			
	Freiburg	Lübeck			
	Gelsenkirchen	M'Gladbach			
	Gladbeck	München			
	Hagen	**Münster**			
	Heidelberg	Nürnberg			
	Herne	Offenbach			
	Karlsruhe	**Osnabrück**			
	Krefeld	Remscheid			
	Lübeck	Stuttgart			
	M'Gladbach	Wiesbaden			
	Mülheim				
	München				
	Münster				
	Nürnberg				
	Oberhausen				
	Offenbach				
	Osnabrück				
	Recklinghausen				
	Remscheid				
	Stuttgart				
	Wiesbaden				
	Witten				

Note: Metropolitan region Oberrhein is a trinational metropolitan region with Switzerland and France. The cities not in a metropolitan area are marked as **bold**. The full names of some cities are Freiburg am Breisgau, Frankfurt am Main, Ludwigshafen am Rhein, Mülheim an der Ruhr and Offenbach am Main. We consider a city stationarity in this table at a 5% significance level and a unit root if the significance level is above the 5% significance level.

Figure 11: City Location by Impact



(a) Positive Impact

(b) Stationary

(c) Negative Impact

9 Discussion

9.1 Link between the Evidence and Theory

Although this paper is meant to primarily be a methodological contribution, we can say some things about the links between the evidence and the fundamental theories. The original purpose of the seminal paper by Davis & Weinstein (2002) was to evaluate the three competing fundamental theories on the location of economic activity, namely locational fundamentals theory, increasing returns theory and random growth theory. One crucial element in that evaluation is whether cities are mean-reverting or not. In the following, we will have another look at the link between the theories and the empirical evidence. Then, based on the evidence we have found here, we will evaluate these different theories.

First, it is useful to present the set of stylized facts we have found so far: (1) In the absence of a large, temporary shock, city population between 1870-2000 behaves in an inverse U-shaped way. (2) We find a unit root for most cities after WWII. (3) The ratio of positive to negative impact unit root cities is around 5 to 6. (4) In 1940, cities for which we find reversion to the 'new' mean were on average the largest, while positive impact cities were comparatively smaller, and negative impact cities the smallest. (5) There is no statistically significant difference between the synthetic and the actual distribution, before or after WWII.

9.1.1 Locational Fundamentals

To start, we will first describe the typical characteristics and predictions of each mentioned theory, beginning with locational fundamentals theory (also called first-nature geography), which explain the location of economic activity through the spatial distribution of geographic features. This is probably most prominent if we look at the typical birthplace of ancient civilizations which are often found along rivers, such as the Indus Valley Civilization along the Indus river, ancient China along the Yellow river, Egypt along the Nile river, or the Mesopotamia between the Tigris and Euphrates rivers. Even today, these regions are heavily populated; and through that it is earning locational fundamentals its name.

Now, locational fundamentals theory would predict the WWII bombings to dissipate in effect in the long-run i.e. reversion to the mean. As our contribution here is that we consider a potential new mean defined by the synthetic control, being consistent with locational fundamentals theory in our case is equivalent to the the synthetic control detrended time series process being stationary even after a large, temporary shock.

In section 4, we have found evidence in favor of an inverse U-shaped city growth notion at least for the cities in our sample. Furthermore, the estimated turning point is on average around the year 1960. Together, it indicates that there are underlying non-random factors at play here, regardless of a large, temporary shock. These patterns can be explained by locational fundamentals being mutable throughout time, an idea which was already entertained by Michaels & Rauch (2017) and Bleakley & Lin (2012). That is, stylized fact (5) is consistent with locational fundamentals theory, but not (1), unless we consider locational fundamentals as mutable. Stylized fact (2), and by extension, (3) and (4), are however not consistent with it, at least not as a standalone theory.

9.1.2 Increasing Returns

Next, increasing returns theory (Krugman, 1991) explain agglomeration by the interaction and proximity of people. Theories in this manner leave the possibility for a permanent effect from a large, temporary shock, a shift in equilibria, and a change in the shape of the city size distribution open. This also means that if we find none of those possibilities, we are still consistent with increasing returns theory. In other words, it is not possible to falsify it, but only to show consistency with the empirical findings.

Given however that there was no large, temporary shock on the US city system at least in the 20th century, stylized fact (1) is not at all explainable by increasing returns theory. Rather, the opposite is the case in that one other prediction of increasing returns theory is the increasing concentration of economic activity over time or in history.[21] What that boils down to is that stylized facts (2) to (4) can be explained

[21] Of course, we may not find this prediction if we were to consider metropolitan areas definition instead of a city definition.

by increasing returns theory, but not (1) at least as a standalone theory. Stylized fact (5) is at least still consistent with increasing returns theory; although if we were to find the opposite case, it would be considered evidence in favor of increasing returns being at play here via the exclusion of the other two theories.

9.1.3 Random Growth

Lastly, random growth theory considers the urban growth process as a random process that will follow a random walk. Here, the prediction is that the shape of the city size distribution remains the same given any temporary shock, but that the rank of cities within the distribution is allowed to change. Unless we set the mean of this random process to follow a non-random inverse U-shaped development throughout the late 19th and across the entire 20th century, then we can also not explain this notion with random growth theory.

Now given the evidence on the impact of WWII itself, we find that the distribution does not change after WWII, which hinges on the Kolmogorov-Smirnov test. Furthermore, the unit root results on the city size suggests that individual city growth follows a random walk.

However, there are two caveats: First, we do not find a balanced number of cities with a positive and negative impact, but rather a ratio of about 5 to 6. Second, the higher populated places in metropolitan areas before WWII tend to be either not affected or positively affected, whereas lower populated places in non-metropolitan areas before WWII tend to be negatively affected by the WWII bombings. The case against random growth theory here is that randomness predicts the ratio to be close to 1, and the impact to be independent of the city size.

Hence, random growth theory is consistent with (2) through a random walk, but extensions (3) and (4) refute it. Furthermore it is consistent with (5), but cannot explain (1) at least not as a standalone theory.

9.1.4 Evaluation of the Theories Together

Serving as additional evidence and to be consistent with the previous literature, we have also considered the WWII bombings and the relative size of cities: For mutable locational fundamentals theory to be relevant alone, it would also require that the cities return to the relative size as predicted by SCM, apart from the distribution returning. Initially, most cities indeed seem to return if we consider the cutoff year of 1960, but we find more and more cities with a unit root if set a later cutoff year.

This brings us to the issue that the conclusion depends on the cutoff year, with the trade-off of a more likely correct counterfactual if the cutoff year is early, against a more likely WWII shock to dissipate fully if the cutoff year is late. A later cutoff year also introduces the risk of other shocks affecting the results, such as the division of Germany (Redding & Sturm, 2008) However, that we find more positive impact cities with a unit root suggests that choosing a later cutoff year matters; and this result seems to not be driven by the later cutoff year as otherwise, the ratio of positive to negative impact cities would then be around 1.

To summarize and to give an interpretional overview of the results in the previous section, we think that although the unit root results indicate a rejection of the basic locational fundamentals theory, the overall counterfactual, without any bombings involved, decline of cities around the 1960s indicate that locational fundamentals are mutable and relevant as an underlying factor, and can explain a great deal of the locational patterns of today. This conclusion is however by exclusion of all other alternative theories which are not capable of explaining this decline pattern without the need for a large, temporary shock. The previous literature on WWII bombings in a sense were unfortunate in that the decline happened in the immediate years after the war, so that two events coincided.

The evidence in favor of increasing returns theory against random growth theory is also not as straightforward to see, given that increasing returns theory cannot be falsified. For one thing, the figures on the gap of the population, figure 1, 3 and 5, would already indicate that the city size distribution changes towards even larger cities. The Kolmogorov-Smirnov results however indicate otherwise. Nevertheless, we come to a conclusion here through the non-random patterns which are not compatible with random growth.

In the end, our preferred theory is a hybrid theory similar to Davis & Weinstein (2002), with the difference that we consider locational fundamentals now as mutable, and that the source of evidence consistent with increasing returns theory is not alone stemming from the concentration of economic activity over time, but from the failure of mean-reversion and a non-random pattern of concentration after WWII.

10 Conclusion

We began by outlining the potential problem of and arising from separating the WWII bombing shock from other underlying factors, which can determine the city size. By creating a synthetic control for each of the 52 German cities in our sample, we obtain a benchmark to which we compare the corresponding city evolution after WWII, thus explicitly controlling for underlying factors. We find that the synthetic city population is typically characterized by an inverse U-shaped development i.e. there are underlying non-random factors at play which are not driven by a large, temporary, negative shock. We also find that German cities are permanently affected after WWII, of those the majority are actually better off in terms of population as compared to the benchmark. The immediate pre-war characteristics show that the most populated places recover fully, while the impact on slightly lesser populated, metropolitan places is mostly positive. The negative impact cities are usually the lesser populated, non-metropolitan places. We use this information to evaluate some fundamental theories on the location of economic activity. The non-random pattern of behavior strike out random growth theory as relevant. In the end, we support a hybrid theory as it is the most consistent with our findings: a mutable locational fundamentals theory, in which the importance of individual geographic characteristics can change, combined with increasing returns theory. However, this conclusion does not stem from a direct test on each individual theory or any combination of theories, but through the exclusion of given existing theories. This means that the introduction of other theories could simply nullify the conclusions we have made here, making the quest of constructing direct tests ever so important.

References

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*(490), 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. *American economic review*, *93*(1), 113–132.

Baum-Snow, N. (2007). Did highways cause suburbanization? *The Quarterly Journal of Economics*, *122*(2), 775–805.

Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A., & Zhang, Q. (2017). Roads, railroads, and decentralization of chinese cities. *Review of Economics and Statistics*, *99*(3), 435–448.

Bleakley, H., & Lin, J. (2012). Portage and path dependence. *The quarterly journal of economics*, *127*(2), 587–644.

Bosker, M., Brakman, S., Garretsen, H., & Schramm, M. (2007). Looking for multiple equilibria when geography matters: German city growth and the WWII shock. *Journal of Urban Economics*, *61*(1), 152–169.

Bosker, M., Brakman, S., Garretsen, H., & Schramm, M. (2008). A century of shocks: the evolution of the German city size distribution 1925–1999. *Regional Science and Urban Economics*, *38*(4), 330–347.

Brakman, S., Garretsen, H., & Schramm, M. (2004). The strategic bombing of German cities during World War II and its impact on city growth. *Journal of Economic Geography*, *4*(2), 201–218.

Cuberes, D. (2011). Sequential city growth: Empirical evidence. *Journal of Urban Economics*, *69*(2), 229–239.

Davis, D. R., & Weinstein, D. E. (2002). Bones, bombs, and break points: the geography of economic activity. *American Economic Review*, *92*(5), 1269–1289.

Davis, D. R., & Weinstein, D. E. (2008). A search for multiple equilibria in urban industrial structure. *Journal of Regional Science*, *48*(1), 29–65.

duPont IV, W., & Noy, I. (2015). What happened to Kobe? a reassessment of the impact of the 1995 earthquake in Japan. *Economic Development and Cultural Change*, *63*(4), 777–812.

Fischer, W. (1989). Statistik der Bergbauproduktion Deutschlands 1850 – 1914. *GESIS Datenarchiv, Köln. histat. Studiennummer 8448, Datenfile Version 1.0.0*.

Fischer, W. (1995). Statistik der Montanproduktion Deutschlands 1915 – 1985. *GESIS Datenarchiv, Köln. histat. Studiennummer 8400, Datenfile Version 1.0.0*.

Gabaix, X. (1999). Zipf's law for cities: an explanation. *The Quarterly journal of economics*, *114*(3), 739–767.

Glaeser, E. L. (2005). Reinventing Boston: 1630–2003. *Journal of Economic Geography*, *5*(2), 119–153.

Glaeser, E. L., & Kahn, M. E. (2004). Sprawl and urban growth. In *Handbook of regional and urban economics* (Vol. 4, pp. 2481–2527). Elsevier.

Henderson, J. V., & Venables, A. J. (2009). The dynamics of city formation. *Review of Economic Dynamics*, *12*(2), 233–254.

Im, K. S., Pesaran, M. H., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of econometrics*, *115*(1), 53–74.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of political economy*, *99*(3), 483–499.

Levin, A., Lin, C.-F., & Chu, C.-S. J. (2002). Unit root tests in panel data: asymptotic and finite-sample properties. *Journal of econometrics*, *108*(1), 1–24.

Michaels, G., & Rauch, F. (2017). Resetting the urban network: 117–2012. *The Economic Journal*, *128*(608), 378–412.

Miguel, E., & Roland, G. (2011). The long-run impact of bombing Vietnam. *Journal of development Economics*, *96*(1), 1–15.

Ng, S., & Perron, P. (1995). Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association*, *90*(429), 268–281.

Perron, P. (1991). Test consistency with varying sampling frequency. *Econometric Theory*, *7*(3), 341–368.

Rappaport, J., & Sachs, J. D. (2003). The United States as a coastal nation. *Journal of Economic growth*, *8*(1), 5–46.

Redding, S. J., & Sturm, D. M. (2008). The costs of remoteness: Evidence from German division and reunification. *American Economic Review*, *98*(5), 1766–97.

Redding, S. J., Sturm, D. M., & Wolf, N. (2011). History and industry location: evidence from German airports. *Review of Economics and Statistics*, *93*(3), 814–831.

Schmidt, H. (2011). *Menschen und Mächte*. Pantheon Verlag.

Siodla, J. (2015). Razing San Francisco: The 1906 disaster as a natural experiment in urban redevelopment. *Journal of Urban Economics*, *89*, 48–61.

US Census Bureau. (1900-1930). Mines and Quarries. *Bureau of the Census*.

US Census Bureau and Steiner, E. (2018). Spatial history project. *Center for Spatial and Textual Analysis, Stanford University*. Retrieved from `https://github.com/cestastanford/historical-us-city-populations`

Volkmann, H.-E. (1995). Ende des Dritten Reiches, Ende des Zweiten Weltkriegs: Eine Perspektivische Rückschau. *Piper*.

A Data appendix

Table 11: List of all sample cities

Aachen	Essen	Kassel	Oberhausen
Augsburg	Frankfurt am Main	Koblenz	Offenbach am Main
Berlin	Freiburg im Breisgau	Köln	Osnabrück
Bielefeld	Gelsenkirchen	Krefeld	Pforzheim
Bochum	Gladbeck	Ludwigshafen am Rhein	Recklinghausen
Bonn	Hagen	Lübeck	Regensburg
Bottrop	Hamburg	Mainz	Remscheid
Braunschweig	Hannover	Mannheim	Stuttgart
Bremen	Heidelberg	Mönchengladbach	Trier
Darmstadt	Herne	Mülheim an der Ruhr	Ulm
Dortmund	Hildesheim	München	Wiesbaden
Duisburg	Kaiserslautern	Münster	Witten
Düsseldorf	Karlsruhe	Nürnberg	Würzburg

Table 11 shows the West German cities in our sample. As compared to the 62 cities in the sample in Bosker et al. (2008), we have excluded 10 cities based on not having complete population data before 1920 for each 10 years until 1871. The excluded cites are: Bamberg, Flensburg, Fürth, Hamm, Kiel, Oldenburg, Solingen, Wanne-Eickel, Wattenscheid, Wuppertal.

Furthermore, we have excluded the following 18 cities, which we have also included for the urban potential measure as there is census data for 1939 for all the 80 cities: Bergisch Gladbach, Bremerhaven, Erlangen, Göttingen, Heilbronn, Ingolstadt, Leverkusen, Moers, Neumünster,Neuss, Paderborn, Reutlingen, Saarbrücken, Salzgitter, Siegen-Wittgenstein, Wilhelmshaven, Wolfsburg and Worms. As for the US incorporated places data, we have rechecked the dataset from the US Census Bureau and Steiner, E. (2018) with the original US census data from the US census bureau. This dataset had for every incorporated place the population if it was above 2500 inhabitants i.e. if the inhabitants were below 2500 in the particular year, this particular year had a missing value. We have filled those missing values for those years from the official census publications.

We have dealt with annexations of surrounding places by adding the incorporated places for which the population suddenly went to 0 for years before 2010 to the corresponding annexing incorporated place. That is, the disappearance of the incorporated place from the census indicates which place was annexed. Typically however, the annexations did not change population of US incorporated places not significantly.

This is different for the German cities in our sample. Typically, we find city fusions, such as Wuppertal from Barmen, Elberfeld, Ronsdorf, Cronenberg and Vohwinkel, mostly in the 1920s and to a lesser degree in the 1930s. The immediate pre-war fusions and annexations were also typically significant, and a significant distortion more frequent than the annexations and fusion of the Gemeindereform of the late 1960s and the 1970s. For instance in the post-war wave for the 62 cities dataset, we find that for 48 out 62 cities (or 44 out of our 52 sample cities), the added population is below 20%, and 10% for 32 out of 62 cities (or 28 out of our 52 sample cities).

In contrast, the pre-war annexations and fusions were typically more significant, where the added population is below 20% for 17 out of 62 (and 17 out of 52) cities, and below 10% for 9 out of 62 (and 9 out of 52) cities.

Also, we know very well how many people were added after an annexation or fusion, as we typically have records available. If this was not the case, we simply took the difference of the year before the annexation or fusion happened with the year after, which is likely a very good estimate if the population growth around the annexation or fusion year was not too far away from 0. Typically, the 1920s and 1930s, and the 1960s and 1970s have a low growth rate as compared to the pre-WWI periods, or the immediate post-WWII period, so that this simple estimation seems appropriate enough.

With this information, we can adjust for the annexation or fusion with the simple adjustment as done by Bosker et al. (2008). That is, we take the city boundaries during WWII as the reference point, in contrast to the city boundaries of 2010 we take for the US sample. That means that pre-WWII, we add to the given city the population of the added areas. On the other hand in the post-WWII period, we deduct the added population from the given city. Furthermore, we add the simplifying assumption that the population growth rate of the original city and the added areas are the same. Keeping the notation of Bosker et al. (2008), we can more formally write for the pre-WWII case:

$$\hat{S}_{iT} = S_{iT} \frac{S_{iT-k}}{S_{iT} - S_{inew}} \tag{14}$$

with S_{iT} being the population of city i with the added population at time T, S_{inew} is the (estimated) population of the new areas which we know, S_{iT-k} is the population in year $T-k$ before the change of the city's boundary. Finally $\hat{S}_{iT}$ is the adjusted population with the population of the newly added areas being extrapolated in the past.

As for the post-WWII case Gemeindereform happening at time T, we write:

$$\hat{S}_{iT} = S_{iT} - S_{inew} \tag{15}$$

so that we simply deduct the population from the new area post-WWII at time T. For each subsequent years, as we do not know the (estimated) population of the new areas, we will extrapolate according to:

$$\hat{S}_{iT+k} = \hat{S}_{iT} \frac{S_{iT+k}}{S_{iT}} \tag{16}$$

The assumption that the growth rate of the usual surrounding, newly added areas being the same as the core, original city is implemented in equations (13)-(15). Lastly, to compare the same years between the German and US sample, we set the German year of 1871 to 1870 and 1939 to 1940.