# The Relationship Between Railroad Length and Cholera Caused Deaths in Prussia Between 1830-1875

Schwerpunktseminar

Winter Term 2022-2023

Epidemics in Economic History

Submitted by:
Ufuk Çavdar
M.Sc. Economics
12453471

# 1 Introduction

[1] Cholera was one of the deadliest diseases in the 19th century in Prussia, as it also was all around Europe. Before arriving in Europe, cholera first spread through Asia throughout the early 19th century. It then reached Moscow in 1830 and quickly spread to eastern Europe and Prussia (Ross III, 2015), partly due to the suitable environment for the disease to spread created by the unsanitary conditions of the big cities (Ross III, 2015). Starting from the early 1830s, several cholera outbreaks lead to the deaths of more than 350 000 people in Prussia by the year 1875 and it killed more than 100 000 people in its peak year 1866.

Early symptoms of the disease, caused by a bacterium named *vibrio cholerae*, contains dizziness and faintness, but it mostly proceeds without any major symptom of illness which causes the illness to go unnoticed before it is far advanced (Snow, 1856). Once it is advanced, cholera's most common symptoms are acute spasmodic vomiting and profuse diarrhea, which leads to extreme progressive dehydration also leads to death in the worst cases (Mari et al., 2012).

Besides the cholera outbreaks, Prussia also experienced a fast railroad expansion in the 19th century. With the establishment of Zollverein in 1834, trade between many provinces and counties became easier and cheaper, thus new and more efficient transportation systems between different regions of Prussia became more demanded. The first railroad in 19th century Prussia started operating in 1938 (Dunlavy, 1991). Even though Prussia started integrating railroad technology relatively late compared to the other European countries, by the year 1845, railroad length in Prussia surpassed France's railroad system (Hornung, 2015). Railroad expansion continued to increase exponentially in the following years. In the 25 years between 1850 and 1875, the total railroad length in Prussia increased fivefold.

Railroad expansion in Prussia is highly linked with the industrialization, urbanization, and economic growth of Germany. The *Stadtkreis* (urban) population in Prussia was approximately 770 000 according to the 1831 census and by the year 1875, it increased to more than 2 250 000. Between the same years, as opposed to an almost 200% increase in urban city population, the *Landkreis* (rural) population only increased by 54.7%. Several papers show the positive causal effects of railroad expansion on the urban city population. According to Hornung (2015), railroad-adapting cities had 1 to 2 percent higher

[1] 6762 words including table and excluding references

population growth compared to non-adapting cities. Braun and Franke (2022) show that adapting the railroad technology granted Kingdom of Württemberg 0.3 to 0.4 additional population growth.

In line with the railroad expansion, Prussia also experienced rapid industrialization during the 19th century. The demand that is created by railroad constructions, especially for coal and iron, was crucial for the increase in investment in mining and metallurgy in the 1850s (Tilly, 1966). Fremdling (1977) states that the industrialization of Germany can not be explained without railroad expansion between 1840 and 1880 and shows the positive backward linkage effect of railroad expansion on economic growth by looking German iron industry and comparing the period with the United States and Great Britain.

Other than the role of railroad expansion in economic growth, the role of railroads, and the role of transportation and human mobility in general, in the transition dynamics of epidemics is also an intriguing topic for many researchers. History of epidemics, and the recent COVID-19 pandemic, show us there is a clear relationship between human mobility and diffusion of the disease and the speed of spread. The literature is also in a consensus about confirming the effect of human mobility and public transportation on the spread of an epidemic.

In his paper, Ohadike (1981) highlighted that the 1918-1919 influenza epidemic in Nigeria reached the country after the war through British passengers on ocean liners. In addition to that, once the disease reached Nigeria, railroads, which were the dominant type of long-way transportation in Africa at the time, were one of the main elements of the transmission dynamics of the disease, as the first inland cities that were affected by the pandemic was the main ports of the railroad (Hogbin, 1985).

Another example of the effect of human mobility on the spread of diseases is the Covid-19 pandemic. Sirkeci and Yüceşahin (2020) show the causal relationship between human mobility and the severity of the Covid-19 pandemic by creating a model on migration corridor between China and travel intensity. The relationship between the number of trains in a day from Wuhan, the origin of Covid-19, and the number of Covid-19 cases in a city have been shown by Pang et al. (2023). According to their paper, the number of covid cases in a city increased by 1% with an additional train from Wuhan.

There is also a big literature on the spread dynamics of cholera and its relationship with human mobility, going back to Snow (1856). Snow discovered that cholera is transmitted through contaminated water by comparing the death rates in houses that supply water from different sources. In addition to short-range spread of cholera by contaminated water sources, some papers are focused on the long-range spread dynamics of the disease. Schweig (2022) argues that the Ottoman Anatolian Railroad, which started operating in 1892, played an important role in the spread of cholera in Anatolia in 1983, especially in Eskişehir, an important junction of the Ottoman Anatolian Railroad. In their paper, Mari et al. (2012) try to assess the effect of short-term human mobility in addition to the transmission through underground and surface water and find that human mobility is crucial to explain the inter-catchment and regional spread of cholera.

This study tries to assess the relationship between the railroad expansion in Prussia between 1835-1875, urbanization and population growth, and cholera spread during the same period. There are two main analyses of the paper. The first analysis tries to estimate the relationship between population growth, specifically urban population growth, and railroad length. It starts with assessing the relationship between having access to different lengths of railroad in 1849 and population growth rates in different subperiods of 19th century. It shows the extent of urbanization during these subperiods and relationship between different sectors with annualized population growth. It then continues with a fixed effect regression for estimating the relationship between population growth, urbanization, and railroad length for the years where there is a census in Prussia.

The second analysis tries to estimate the relationship between the severity of cholera spread, railroad length, and cholera spread in neighbor counties with fixed effect regression. It contains years between 1831-1875 where there is at least 1 recorded cholera death in a total of all counties. There is a total of 22 years that fit this description and the years are shown in Figure 3. It then continues with two robustness checks that use the same empirical specifications, but different from the first analysis, use two additional sets of years. The first sample of years in the second analysis are the years where recorded cholera deaths exceed 1000 people and the second sample of years are the years where recorded cholera deaths exceed 5000 people throughout Prussia. There are only 16 years where there are more than 1000 cholera deaths and only 12 years where there are more than 5000 cholera deaths. The years that are contained in the first analysis but are not contained in the 1000 deaths sample are 1833, 1851, 1854, 1856, 1858, and 1868. In addition to these years, 1836, 1857, 1859, and 1874 are not included in the 5000 cholera deaths sample. The years that meet the descriptions can also be seen in Figure 3.

This paper is structured as follows. Section 2 describes the data that is used in the quantitative analysis in addition to describing the data imputation and variable creation processes. Section 3 conducts an explanatory analysis of the railroad expansion, the spread of the cholera epidemic, and possible factors that might have a relationship with the severity cholera epidemic in 19th-century Prussia. It also compares railroad length, railroad length density, and cholera death numbers in 1866 by using geographic information system (GIS). Section 4 shows the results of the empirical analyses. Finally, section 5 summarizes the results and discusses possible shortcomings and problems of the quantitative analyses.

# 2 Data

The original data contains the population only for the years where there is a census. There were 14 censuses in the period of interest in Prussia, and of these years only six of them correspond with the second analyses samples. I imputed population values for missing years by assuming all the years between two censuses has the same population growth rate. Growth rates between two census years are calculated in a way that the population increase rate equal to the calculated growth rate in each year between censuses will lead the population of the first census to the population of the next census in the next census year.

$$Pop_{cen_2} = Pop_{cen_1} \times (1 + g_{cen_2,cen_1})^{(cen_2 - cen_1)} \tag{1}$$

Equation 1 shows how the population imputations are conducted. $Pop_{cen2}$ is the population in a given census year. $cen_1$ is the year of the initial census and $cen_2$ is the year of the next census. $g_{cen_2,cen_1}$ is the growth rate in the years between $cen_1$ and $cen_2$. Besides population imputation, this growth rate is also used in the first analysis of the paper. $PopGrowth_{cen_2}$ is used as the dependent variable in both standard OLS regression for 1849 and fixed effect regression.

$$PopGrowth_{cen_2} = (g_{cen_1,cen_2})^{(cen_2 - cen_1)} \times 100 \tag{2}$$

To capture the effect of the severity of cholera spread in other counties on the cholera deaths in a given county, for every county I calculated the average of cholera deaths in other counties weighted by the distance between other counties and the county of interest.

$$WeightedAVG_{i,t} = \frac{1}{N-1} \sum_{j \neq i} \frac{CholeraDeaths_{j,t}}{Dist_{i,j}} \tag{3}$$

Equation 3 shows the formula for the calculation of the weighted average of cholera deaths in the rest of the counties. $WeightedAVG_{i,t}$ is the weighted average of neighbor cholera deaths in county $i$ in year $t$. $CholeraDeaths_{j,t}$ is the number of cholera-caused deaths in county $j$ in year $t$, and $Dist_{i,j}$ is the distance between county i and j in km. N is the number of counties in Prussia during the period, which is 343. The distances between counties are calculated by the QGIS software as the distance between centroids on the map. Table 1 shows the distribution of calculated weighted average variable thorough out the outbreak periods and in the year 1866.

Table 1: Distribution of Weighted Average of Neighbor Counties' Cholera Deaths

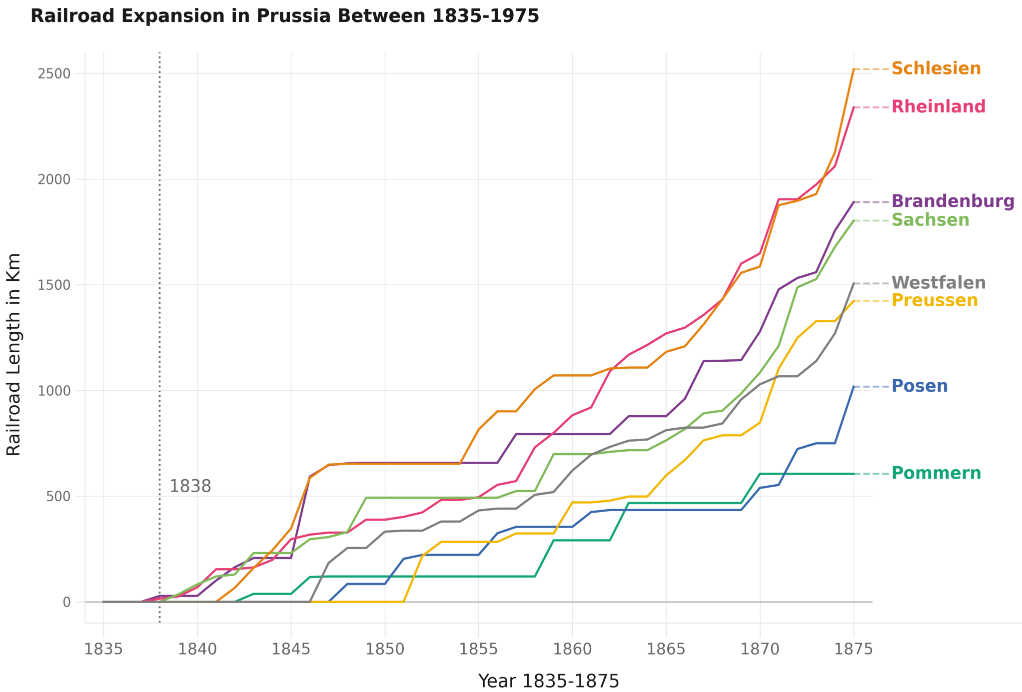
| **Weighted Mean of Neighbor Deaths** | **Min** | **Max** | **Mean** | **StD** | **Count** |
|---|---|---|---|---|---|
| For All Outbreak Years | 0.000 | 3.352 | 0.362 | 0.211 | 7546 |
| In 1866 | 0.653 | 3.352 | 1.480 | 0.512 | 343 |

Finally, the data contains railroad lengths for the years after 1837 for each county which allows us to capture the covariation between railroad lengths and population growth in the first analysis, and railroad length and cholera deaths in the second analysis. In both analyses, I used railroad measures in 100 kilometers.

# 3 Railroad Expansion, Population Growth and Cholera Outbreaks between 1830-1875

The first railroad of Prussia started operating between Berlin and Potsdam in 1838. Starting from the year 1838, Prussia experienced rapid growth in total railroad throughout the country. Figure 1 shows the magnitude of the railroad expansion throughout the century in the provinces of Prussia. The railroad length increased for all provinces, with increasing speed after 1850. The total railroad length in Prussia was approximately 2700 km and the railroad length per county was approximately 8 km in the year 1850. By the year 1875, the total railroad length increased to approximately 13 000 km, and per

Figure 1

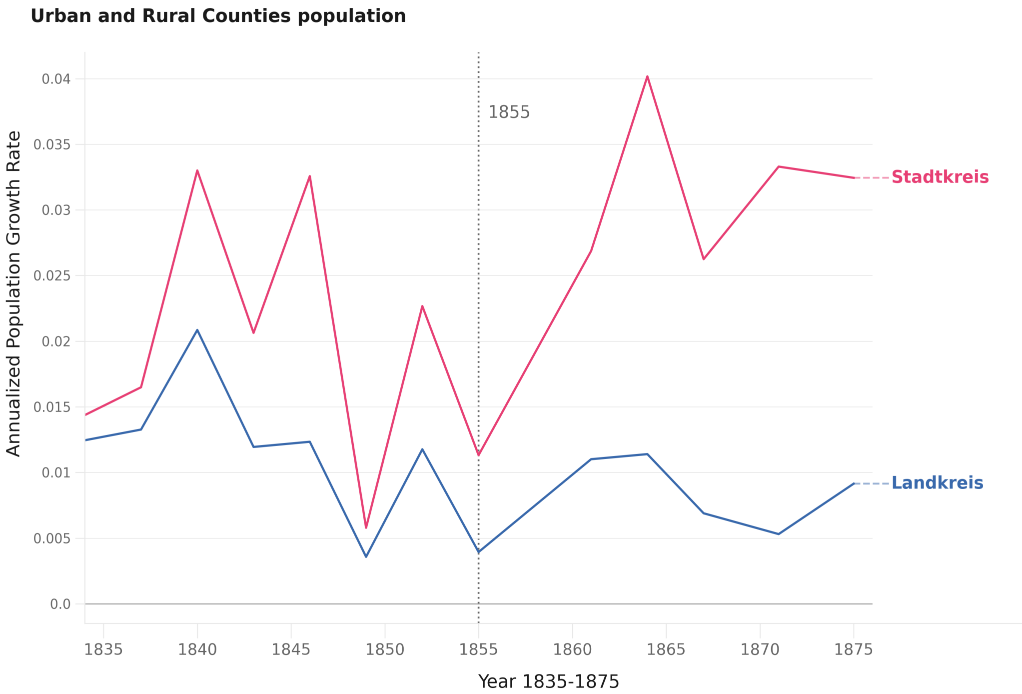



county, the railroad length increased to approximately 38 km.

Highly linked with railroad growth, Prussia also experienced rapid industrialization and urbanization during the 19th century. Figure 2 shows the annualized population growth between censuses as it is calculated according to Equation 2. For every census period, annualized population growth was higher than rural population growth. Especially after the year 1855, urbanization speed increased as annualized population growth rate increased as high as 4% between 1861-64. As opposed to increasing urban population growth over the years in the 19th century, rural population growth fluctuated around 1% growth level. As a result, the gap between urban population growth and rural population growth increased even further after 1855.

Several cholera outbreaks happened as well during the same period in Prussia. Figure 3 shows the number of deaths caused by cholera between 1830-1875. We can describe four different periods that Prussia experienced a cholera outbreak with varying duration and severity. The first period of cholera outbreaks started at the beginning of the 1830s and lasted until 1837, without any deaths in 1834 and 1835. The second period of cholera outbreaks started in the year 1848 and persisted until 1859. The second period of cholera outbreaks is the longest outbreak of our period of interest. The third period and the

Figure 2




deadliest one of all started in 1866 and vanished gradually until 1868. The year 1866 is the peak year of all four periods for cholera deaths in Prussia with almost 120 000 deaths. The final period of cholera outbreaks lasted only two years, 1873 and 1874.

1866 is a particularly interesting year for this study given the fact that it is both the peak year of cholera deaths and a year that Prussia has a significant railroad infrastructure present. Figure 4 shows the distribution of cholera-caused deaths and railroad length in counties in the year 1866. As it can be seen in Figure 4a, Berlin, Breslau, and Posen counties are isolated from the rest 340 counties with their high cholera death rates compared to their railroad length. The low railroad length in urban parts of Berlin and Breslau, two highly and densely populated cities, can be explained by their small areas, and high cholera death rates might be connected to their high population density instead of railroad length. The number of cholera-caused deaths in Posen, even though it is not as densely populated as Berlin and Breslau, might be connected to its high population.

Figure 4b shows the same distribution as Figure 4a, but without three outlier counties. Without the three, the slope of the linear regression line increased to 3.38 from 3.25, which might indicate a positive relationship between cholera-caused deaths and railroad length in 1866.

Figure 3

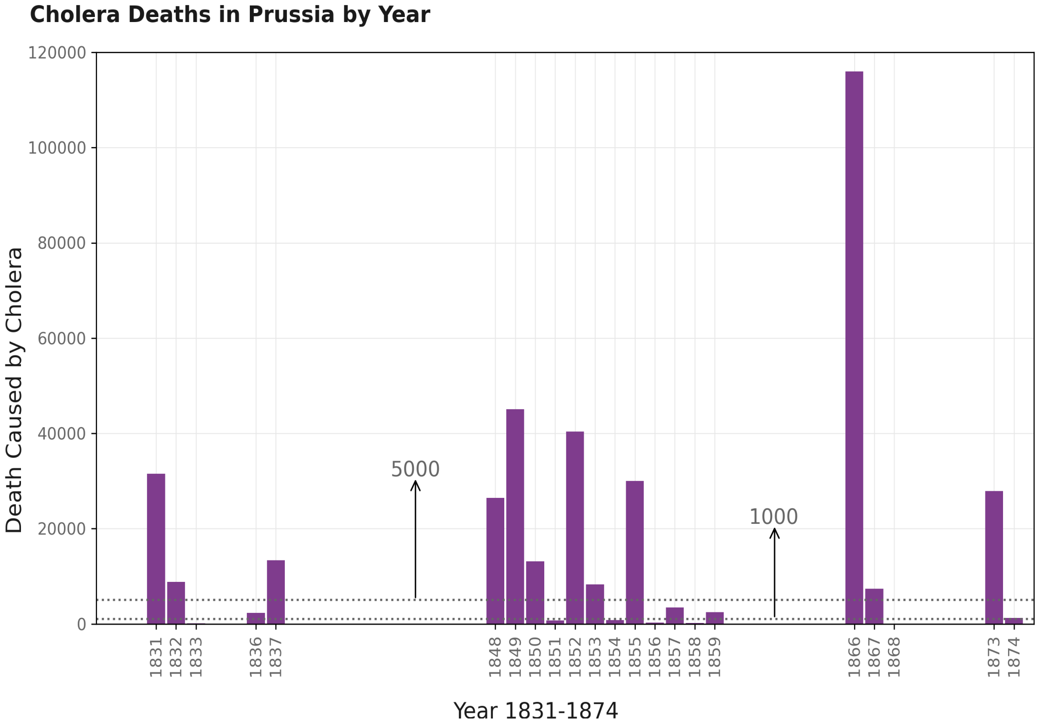
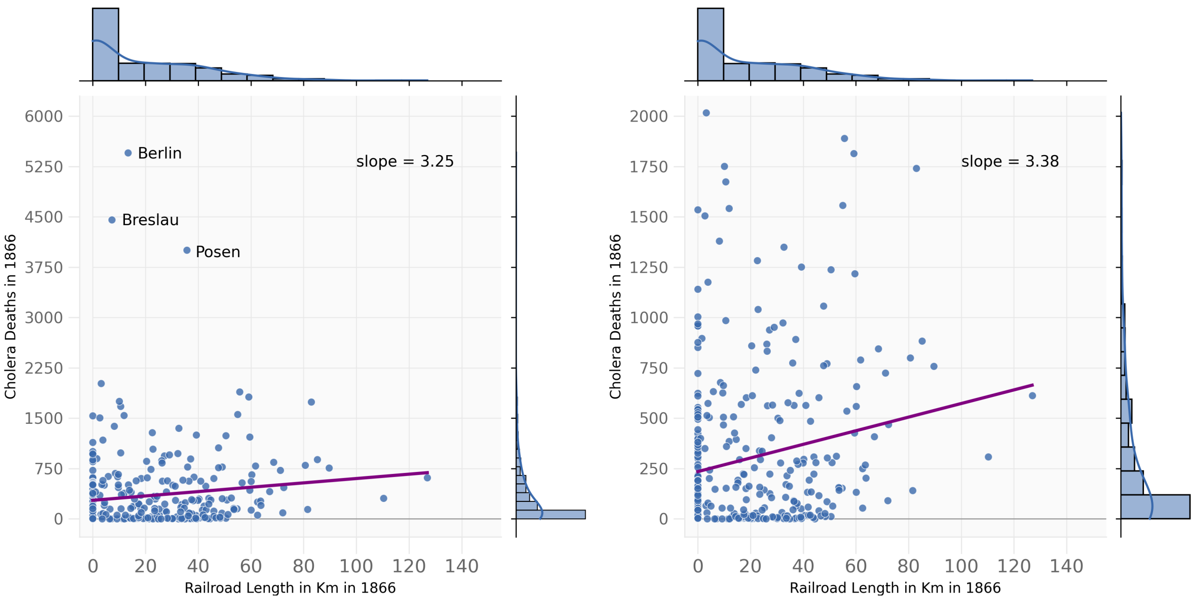



Figure 5a shows the comparison of cholera deaths in 5 counties with the highest weighted neighbor deaths average (Breslau, Niederbarnim, Obornik, Trebnitz, Neumarkt) and cholera deaths in 5 counties with the lowest weighted neighbor deaths average (Kreuznach, Bitburg, Saarburg, Trier, Saarlouis) in 1866. It can clearly be seen that, even by only looking at to top 5 counties, there is a variation of cholera deaths even within the high weighted average counties. On the other hand, the figure also indicates there is a positive relationship between the weighted neighbor deaths average and county cholera deaths, as the highest cholera deaths values are from the counties with high weighted average counties.

Similar to the previous figure, Figure 5b compares cholera deaths values in 5 counties with the highest population density (Königsberg, Barmen, Halle, Köln, Breslau), and cholera deaths in 5 counties with the lowest population density (Neidenburg, Deutsch Krone, Rummelsburg, Schlochau, Johannisburg) in 1866. The results are similar to Figure 5a as there is some variation in cholera death numbers in densely populated counties but the highest cholera death values are also from densely populated counties. Breslau is the only city that appears in both figures. Not surprisingly, Breslau has the second most cholera deaths in 1866, only second to Berlin with just almost 4500 deaths.

Figure 4: Distribution of Railroad Length and Cholera Deaths in 1866

(a) With Berlin, Breslau and Posen (b) Without Berlin, Breslau and Posen



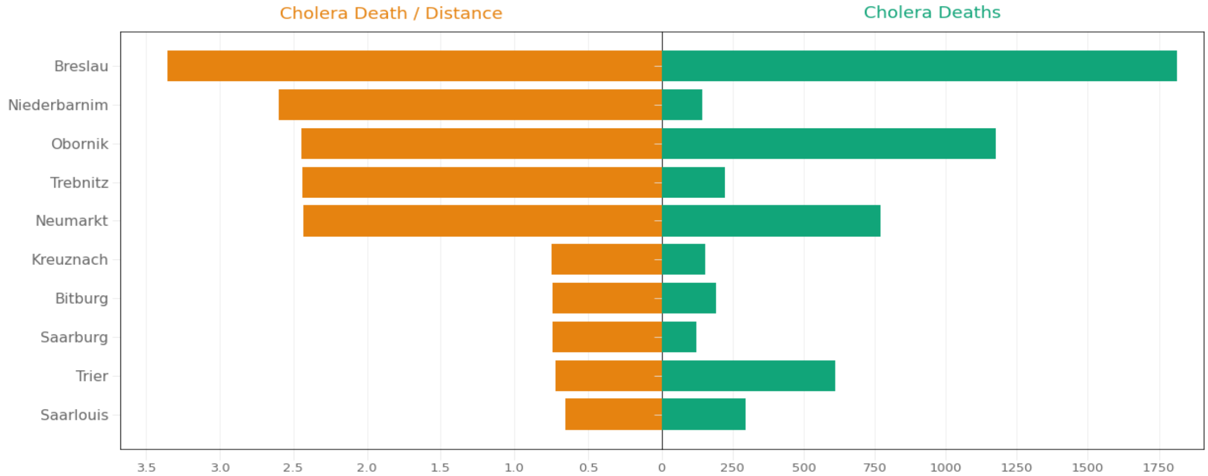

The three maps in Figure 6 show railroad length, the number of cholera-caused deaths, and railroad density for all counties. Color breaks are calculated by using Jenks natural breaks classification method at QGIS. Both Figure 6a and Figure 6c overlap with the cholera deaths map to a degree, which might be an indicator that there is a relationship between railroad length and cholera deaths in 1866. However, both railroad length maps do not overlap perfectly as both could not match the high death rates in Posen province.

Different provinces might have different attributes that have an effect on the transition dynamics of cholera outbreaks, which might lead to misinterpretation when comparing railroad length and cholera deaths at the country level. To overcome this issue, I created another set of maps. Maps in Figure 7 show the same variables as Figure 6, but instead of classifying counties according to Jenks breaks that are calculated at the country level, it classifies counties according to Jenks breaks that are calculated at the province level that the county is part of. This means different provinces have different Jenks breaks depending on the distribution of the variable inside the provinces. Provinces are separated by the colors of county borders. The color codes of the province are the same as in Figure 1.

Compared to the maps in Figure 6, Figure 7a and Figure 7c explains Figure 7b

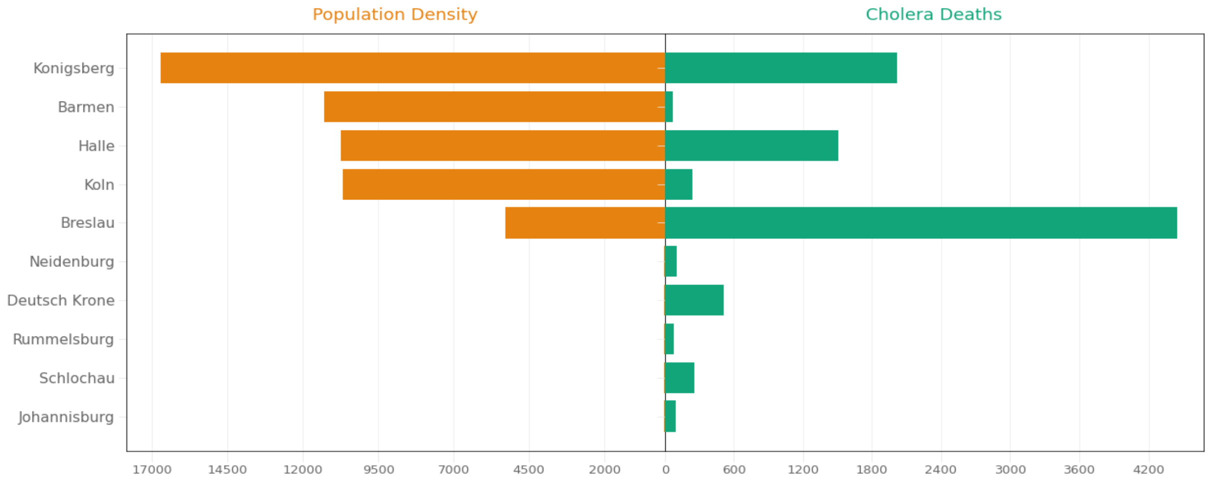
Figure 5

(a) Average Neighbor Deaths / Deaths caused by Cholera in 1866




The chart shows 5 counties with highest weighted average of neighbor deaths (Breslau, Niederbarnim, Obornik, Trebnitz, Neumarkt) and 5 counties with lowest weighted average of neighbor deaths (Kreuznach, Bitburg, Saarburg, Trier, Saarlouis).

(b) Population Density / Deaths caused by Cholera Comparison in 1866

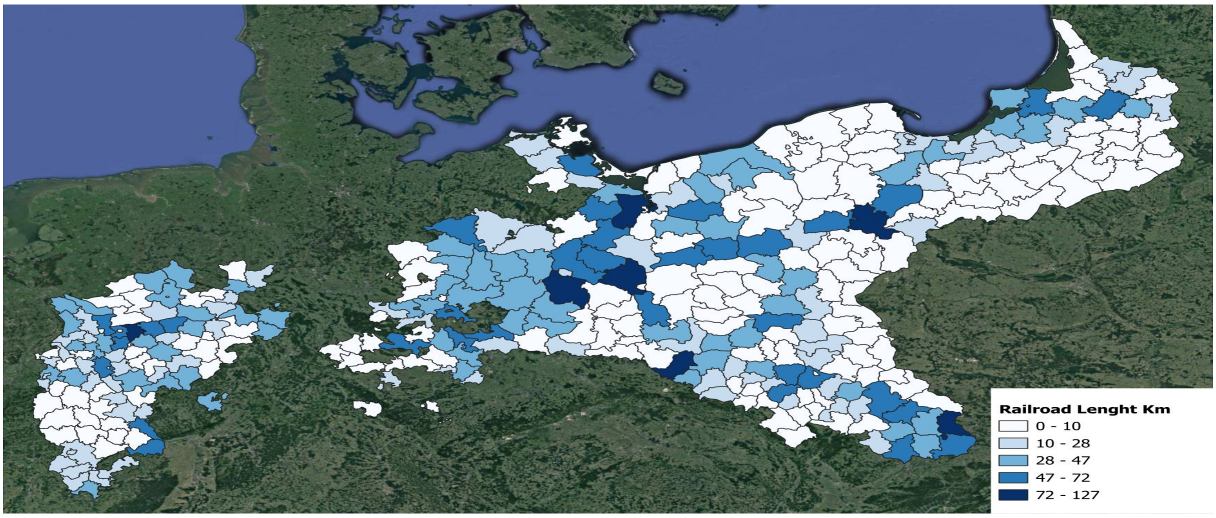



The chart shows 5 counties with highest population density (Königsberg, Barmen, Halle, Köln, Breslau) and 5 counties with lowest population density (Neidenburg, Deutsch Krone, Rummelsburg, Schlochau, Johannisburg).

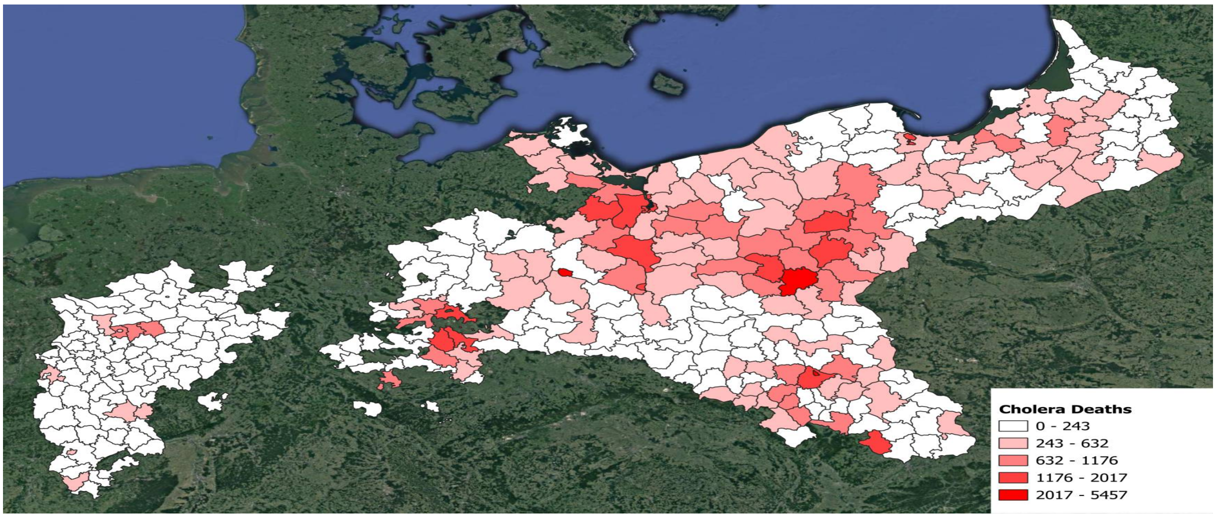
better. Especially area weighted railroad length map in Figure 7c matches the cholera deaths map in Figure 7b for every county significantly better than before. This implies that there might be some province-level effects that also define the relationship between railroad length and cholera spread.

Figure 6: Prussia in 1866

(a) Railroad Length in Prussia in 1866




(b) Cholera Deaths in Prussia in 1866




(c) Railroad Length weighted by Area in Prussia in 1866

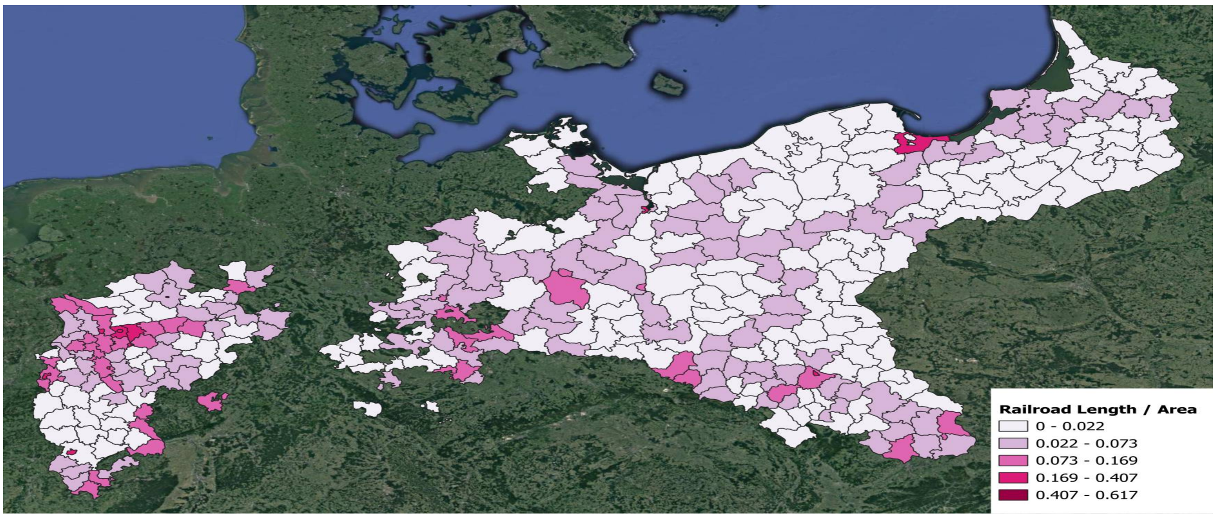
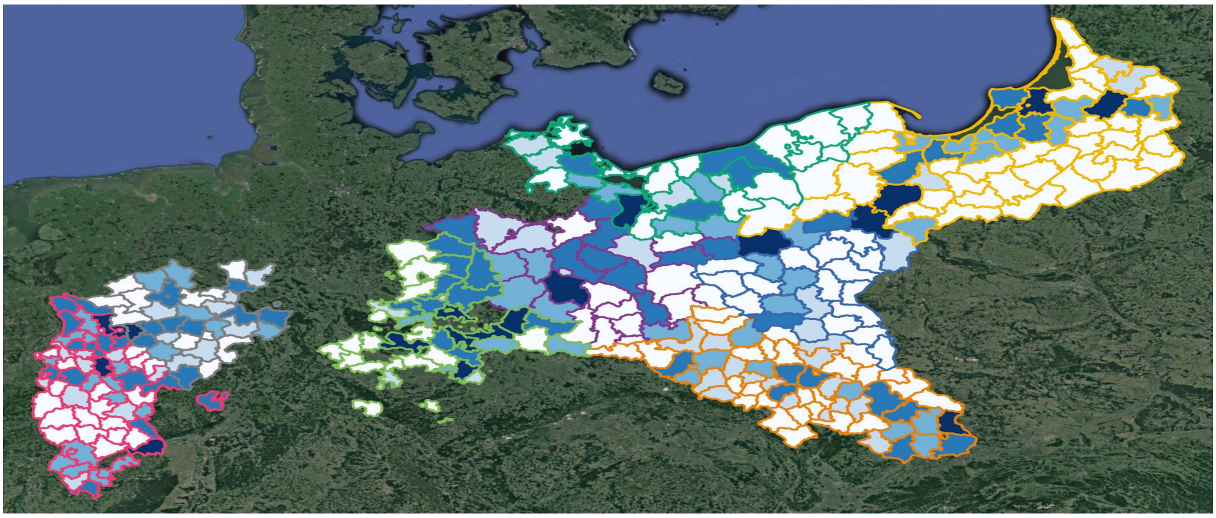
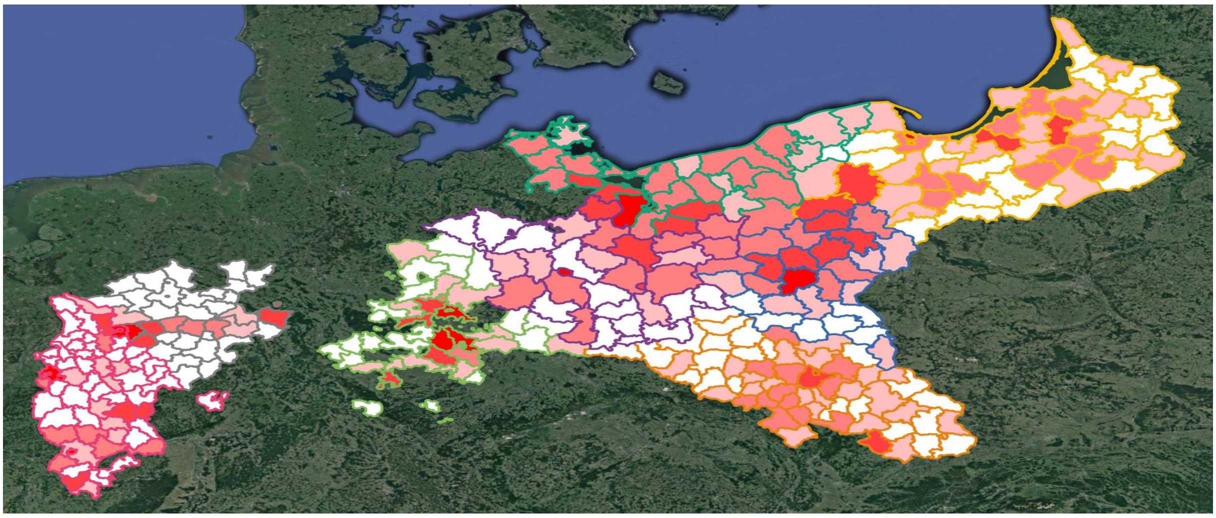


Figure 7: Prussia in 1866 - state level breaks

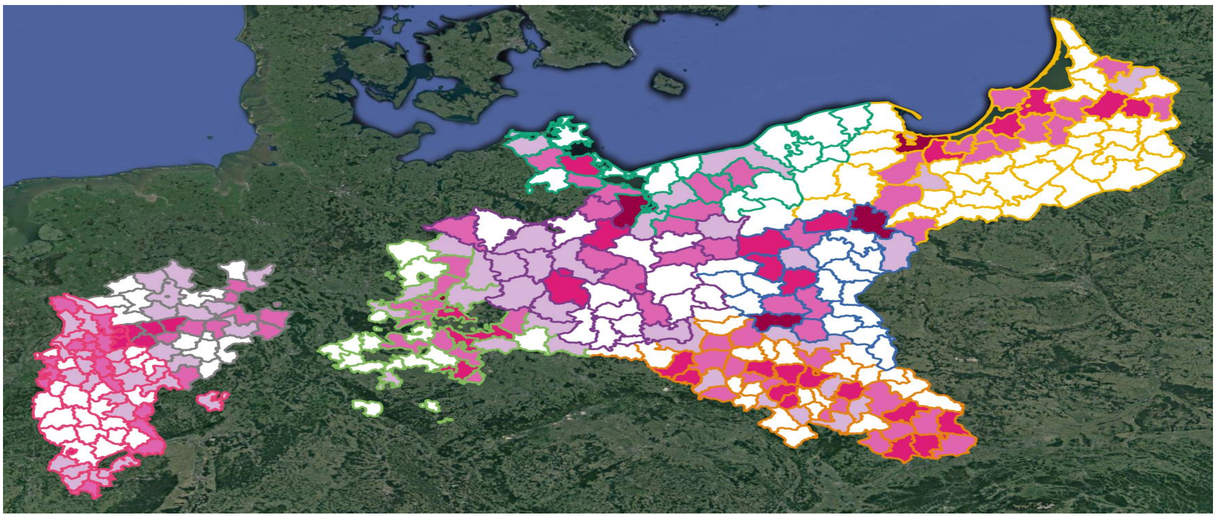
(a) Railroad Length in Prussia in 1866 - state level breaks



(b) Cholera Deaths in Prussia in 1866 - state level breaks



(c) Railroad Density in Prussia in 1866 - state level breaks



# 4 Empirical Research

## 4.1 Urbanization and Railroad Expansion between 1831 and 1875

The first main analysis of the paper estimates the relationship between population growth and railroad length in a county. First, I run standard OLS for different periods of the century to estimate the effect of railroad length in 1849. The census in the year 1849 provides the number of employees in the industry, service, craft, and agriculture sectors. Equation 4 shows the linear regression equation for standard OLS. $PopGrowth_{i,t_1,t_2}$ is the annualized population growth rate in county i between the census years $t_1$ and $t_2$, calculated as it is shown in Equation 2. $RKM_{i,1849}$ is the railroad length in county i in the year 1849, $Urban_i$ is a dummy variable for urban counties and $X_i$ is county level control matrix which consists area and province dummies.

$$PopGrowth_{i,t_1,t_2} = \beta_0 + \beta_1 RKM_{i,1849} + \beta_2 Urban_i + \delta_0 X_i + \epsilon_i \tag{4}$$

Table 2 shows the results of the standard OLS analysis. Railroad length in 1849 does not have a statistically significant relationship with the population growth rate between 1831-49 which indicates there is no significant difference in population growth rates before 1849 for counties with different railroad lengths in 1849. After 1849 however, railroad length in 1849 has a significant positive relationship with the population growth rate in each census period. 100 km higher railroad length is associated with 1.237% increase in annualized population growth from 1849 until 1975. The coefficient of railroad length varies between 0.715% and 2.331% in different subperiods.

Just like railroad length, being an urban county also does not have a statistically significant relationship with population growth in 1830-1849 which means there is no significant pre-1849 trend that distinguishes the urban counties and rural counties. After 1849, being an urban county associated with a higher annualized population growth of 1.491% until 1975. This result shows the rapid urbanization that Prussia experienced between 1849-1875. The interaction term between railroad length and urban dummy variable also shows the corelation between two variables and population growth rate. Model 2 at Table 2 states 100 km increase in railroad length also increases the relationship between being an urban county and annualized growth rate between 1849 and 1875 by 0.042%. This result shows how an increase in railroad also increase the post-1849 positive trend

Table 2

Population Growth

| | Main Periods | | Subperiods | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1831-49 | 1849-75 | 1849-52 | 1852-55 | 1855-61 | 1861-64 | 1864-67 | 1867-71 | 1871-75 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Railroad Length in 1849 100 Km | 0.061 | 1.195*** | 0.788** | 1.593*** | 0.611 | 1.211** | 1.199* | 0.997* | 2.284*** |
| | (0.823) | (0.322) | (0.345) | (0.537) | (0.387) | (0.539) | (0.705) | (0.540) | (0.540) |
| | p = 0.941 | p = 0.001 | p = 0.023 | p = 0.004 | p = 0.115 | p = 0.025 | p = 0.090 | p = 0.065 | p = 0.000 |
| Urban County | 0.415 | 1.813*** | 1.718*** | 1.074** | 2.027*** | 3.199*** | 1.379** | 1.539*** | 1.793*** |
| | (0.804) | (0.314) | (0.337) | (0.525) | (0.378) | (0.527) | (0.689) | (0.528) | (0.528) |
| | p = 0.606 | p = 0.000 | p = 0.000 | p = 0.041 | p = 0.000 | p = 0.000 | p = 0.046 | p = 0.004 | p = 0.001 |
| Interaction term: RKM × Urban | 0.029 | 0.043*** | 0.043*** | 0.036* | 0.041*** | 0.037* | 0.033 | 0.059*** | 0.047** |
| | (0.030) | (0.012) | (0.012) | (0.019) | (0.014) | (0.019) | (0.025) | (0.019) | (0.019) |
| | p = 0.322 | p = 0.001 | p = 0.001 | p = 0.062 | p = 0.004 | p = 0.055 | p = 0.195 | p = 0.003 | p = 0.017 |
| Workforce in Industry in 1849 (in 1000) | −0.016 | 0.043*** | 0.038** | 0.014 | 0.065*** | 0.020 | 0.031 | 0.044* | 0.057** |
| | (0.040) | (0.016) | (0.017) | (0.026) | (0.019) | (0.026) | (0.034) | (0.026) | (0.026) |
| | p = 0.694 | p = 0.007 | p = 0.024 | p = 0.598 | p = 0.001 | p = 0.457 | p = 0.361 | p = 0.095 | p = 0.029 |
| Workforce in Service in 1849 (in 1000) | −0.008 | 0.003 | 0.000 | −0.001 | 0.004 | 0.016* | 0.007 | 0.009 | −0.008 |
| | (0.014) | (0.005) | (0.006) | (0.009) | (0.006) | (0.009) | (0.012) | (0.009) | (0.009) |
| | p = 0.573 | p = 0.523 | p = 0.953 | p = 0.925 | p = 0.556 | p = 0.081 | p = 0.549 | p = 0.341 | p = 0.384 |
| Workforce in Craft in 1849 (in 1000) | −0.033 | −0.034 | −0.059** | −0.051 | −0.047 | 0.017 | −0.009 | −0.051 | −0.026 |
| | (0.063) | (0.025) | (0.026) | (0.041) | (0.030) | (0.041) | (0.054) | (0.041) | (0.041) |
| | p = 0.594 | p = 0.163 | p = 0.025 | p = 0.215 | p = 0.112 | p = 0.675 | p = 0.875 | p = 0.218 | p = 0.529 |
| Workforce in Agriculture in 1849 (in 1000) | 9.856 | −8.083 | −6.411 | 0.738 | −19.884*** | −9.933 | −4.649 | 1.210 | −8.956 |
| | (12.893) | (5.037) | (5.401) | (8.411) | (6.064) | (8.444) | (11.048) | (8.460) | (8.458) |
| | p = 0.445 | p = 0.109 | p = 0.236 | p = 0.931 | p = 0.002 | p = 0.240 | p = 0.674 | p = 0.887 | p = 0.290 |
| Observations | 343 | 343 | 343 | 343 | 343 | 343 | 343 | 343 | 343 |
| $R^2$ | 0.039 | 0.341 | 0.279 | 0.134 | 0.242 | 0.304 | 0.105 | 0.208 | 0.270 |
| Adjusted $R^2$ | −0.005 | 0.310 | 0.246 | 0.094 | 0.207 | 0.272 | 0.064 | 0.171 | 0.236 |

*Notes:* I use the R package Stargazer (Hlavac, 2022) for the analysis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

in population growth for urban counties and validates the results of Hornung (2015) and Braun and Franke (2022).

Higher workforce participation in industry and service sectors also has positive significant relationship with annualized population growth post-1849 period. 1000 employee increase in the workforce of industry and service sectors are associated with 0.042% and 0.39% increase in annualized population growth respectively. Given the fact that participation in the industry sector is an indicator of industrialization and economic development, it is likely that similar factors led the increase in both industry sector and population growth rates that causes the positive relationship between the two variables.

The second part of the first main analysis contains fixed effect regression for the census years. I run three different fixed effect regressions. The first model regresses annualized growth rate on railroad length ($RKM_{i,t}$) while controlling for the area of a county and administrative district. $X_i$ in Equation 6 is a county-level control matrix that contains area and administrative district dummies. In Equation 6, $\psi_t$ is year fixed effects $\theta_j$ is province fixed effects $\omega_j$ and province type $\times$year fixed effects.

$$PopGrowth_{i,t} = \beta_0 + \beta_1 RKM_{i,t} + \delta_0 X_i + \theta_j + \psi_t + \omega_j + \epsilon_{i,y,j} \tag{6}$$

In addition to model 1 second model of the analysis contains the urban dummy variable and the third model contains the urban dummy variable and the interaction term between the urban dummy and railroad length variables. $M_{i,t}$ in Equation 7 and Equation 8 contains all variables from model 1.

$$PopGrowth_{i,t} = \beta_0 + \beta_2 Urban_i + \delta_0 M_{i,t} + \theta_i + \psi_t + \omega_j + \epsilon_{i,y,j} \tag{7}$$

$$PopGrowth_{i,t} = \beta_0 + \beta_2 Urban_i + \beta_3 RKM_{i,t} \times Urban_i + \delta_0 M_{i,t} + \theta_i + \psi_t + \omega_j + \epsilon_{i,y,j} \tag{8}$$

Table 3 shows the positive relationship between railroad length in a county and annualized population growth. The railroad length coefficient is significant and consistent across all three models. The associated annualized population growth increase with a 100 km increase in railroad length varies between 1.114% and 1.231%. Similar to Table 2, fixed effect regression also shows urban counties experienced higher population growth during the period. According to the results of model 2, being an urban county is associated with an additional 1.278% increase compared to rural counties during 1831-1875, and the same

Table 3

Annualized Population Growth

| | Baseline Model | With Urban Dummy | With Interaction Term |
|---|---|---|---|
| | (1) | (2) | (3) |
| Railroad Length in 100 Km | 1.114*** | 1.231*** | 1.232*** |
| | (0.167) | (0.169) | (0.169) |
| | p = 0.000 | p = 0.000 | p = 0.000 |
| Urban County | | 1.278*** | 1.021*** |
| | | (0.222) | (0.335) |
| | | p = 0.000 | p = 0.003 |
| Interaction term | | | 5.044 |
| | | | (3.318) |
| | | | p = 0.129 |
| Year Fixed Effects | Yes | Yes | Yes |
| Province Fixed Effects | Yes | Yes | Yes |
| Province × Year Fixed Effects | Yes | Yes | Yes |
| Observations | 4,116 | 4,116 | 4,116 |
| $R^2$ | 0.112 | 0.122 | 0.122 |
| Adjusted $R^2$ | 0.082 | 0.092 | 0.092 |
| *Notes:* | I use the R package Stargazer (Hlavac, 2022) for the analysis. | | |
| | | | *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ |

value is 1.021% for model 3. Finally, the interaction term also has positive relationship with annualized population growth which indicates higher railroad length increases the relationship between being a urban county and annualized population growth. However, it is statistically insignificant as its confidence interval is just below the 90% threshold.

Results from fixed effect regression are in line with the results of the first analysis. With different specifications, both analyses show having access to more railroad is associated with more annualized population growth between 1831-1875. The magnitude of urbanization is also shown by the analyses as there is a significant positive trend in population growth for urban counties compared to rural counties throughout the period and the trend is increasing with railroad length in all models.

## 4.2 Relationship Between Railroad Length and Cholera-Caused Deaths

In the second analysis of the paper, to obtain the relation between railroad length in a county and cholera deaths, I run three different fixed effect regressions with different specifications. In the first model, I use all years where there is at least one cholera-caused death occurred in Prussia, which are the years highlighted in Figure 3. The first model regresses cholera deaths ($CD_{i,t}$) on railroad length ($RKM_{i,t}$), population density ($Dens_{i,t}$), and population ($Pop_{i,t}$) with county fixed effects ($\theta_i$), year fixed effects ($\psi_t$), and county type (*Landkreis* or *Stadtkreis*)×year fixed effect ($\omega_j$).

$$CD_{i,t} = \beta_0 + \beta_1 RKM_{i,t} + \beta_2 Dens_{i,t} + \beta_3 Pop_{i,t} + \theta_i + \psi_t + \omega_j + \epsilon_{i,y,j} \tag{5}$$

In addition to the first model, the second model contains weighted neighbor cholera deaths ($WAvg_{i,t}$) and the third model contains both $WAvg_{i,t}$ and an interaction term between $WAvg_{i,t}$ and $RKM_{i,t}$. $Z_{i,t}$ is a matrix that contains the variables from the first model.

$$CD_{i,t} = \beta_0 + \delta_0 Z_{i,t} + \beta_4 WAvg_{i,t} + \theta_i + \psi_t + \omega_j + \epsilon_{i,y,j} \tag{7}$$

$$CD_{i,t} = \beta_0 + \delta_0 Z_{i,t} + \beta_4 WAvg_{i,t} + \beta_5 RKM_{i,t} \times WAvg_{i,t} + \theta_i + \psi_t + \omega_j + \epsilon_{i,y,j} \tag{8}$$

Table 4 shows the results from the three fixed effect regression models. The railroad length has a positive relationship with cholera deaths, with more than 95% confidence level and a coefficient of 59.268. However, while adding the weighted average of neighbors' cholera death numbers reduce both the coefficient and the significance level, adding an interaction term between the weighted average and railroad length in addition to the weighted average completely diminishes the relationship between railroad length and cholera deaths, as the relationship decreases to only 9.367 more deaths associated with 100 km increase in railroad length and the p value increases to 0.751. Given the fact that the covariation between railroad length and cholera deaths is mostly explained by the mean of other counties' weighted deaths, one way to interpret these findings is the relationship is mostly about how cholera transitions across counties, which is in line with the literature.

Death numbers in neighbors have a clear relationship with cholera deaths in a county.

The weighted average of neighbor cholera deaths has a coefficient higher than 500 for both the second and third models with a confidence level of more than 99%. It is hard to interpret the coefficient, but it can be interpreted as additional 3420 death in a neighboring county with a 100 km distance is associated with roughly 50 more cholera deaths. This relationship can explain the spacial aspect of cholera outbreaks and also the clusters in Figure 6b and Figure 7b.

Model 3 in Table 4 also shows the relationship between the weighted average of neighbor counties' deaths and railroad length. The interaction term can be interpreted as the change in the relationship between the weighted average and cholera death in a county with a 100 km increase in railroad length. According to model 3, a 100 km increase in railroads will increase the value of cholera deaths associated with the weighted average by 145.392 with a more than 99% confidence level. This outcome strengthens the interpretation of the previous paragraph as well.

Population in all models has a near 0 coefficient with high p values. Given the fact that the main way for cholera to spread in small distances (inner county) is contaminated water sources and not human-to-human contact, these results can be seen as not surprising. On the other hand, density has a much higher coefficient, varying from 0.011 to 0.006 which means 1 person in $km^2$ is associated with 0.011 to 0.06 more cholera deaths on average. However, the results for density are statistically insignificant for all models.

The first analysis takes into consideration every year where there is a cholera-caused death. The underlying assumption in this specification is all years where there is a cholera death are comparable with respect to cholera spread dynamics and have the same attributes that impact cholera death numbers. However, different stages of an epidemic might have different transmission dynamics. For example, the final year of an epidemic might have different attributes than peak years and the spread dynamics might be fundamentally different in the final and the peak years. 1868 and 1833, for example, might be seen as good examples as both are the last year of their cholera spread and one can argue that it is not necessarily a spread but more of an end of an epidemic. In addition to that, one also can argue that there might be a scale effect of the disease on its spread dynamics. A higher magnitude of the spread of a disease might make the effects of some aspects irrelevant and might make the other ones much more important than before. Therefore, if the underlying assumption does not hold, that would mean that the years in the first analysis are not comparable and the results of the models would be incorrect.

Table 4

Cholera Deaths

| | Baseline Model | With Weighted Avg | With Weighted Avg + Interaction Term |
|---|---|---|---|
| | (1) | (2) | (3) |
| Railroad Length in 100 Km | 59.268** | 46.125 | 9.367 |
| | (29.721) | (28.190) | (29.409) |
| | p = 0.047 | p = 0.102 | p = 0.751 |
| Density | 0.011 | 0.006 | 0.007 |
| | (0.021) | (0.018) | (0.018) |
| | p = 0.593 | p = 0.729 | p = 0.717 |
| Population | 0.000 | 0.000 | 0.000 |
| | (0.001) | (0.001) | (0.001) |
| | p = 0.902 | p = 0.899 | p = 0.890 |
| Weighted Avg | | 531.128*** | 500.584*** |
| | | (31.927) | (32.177) |
| | | p = 0.000 | p = 0.000 |
| Weighted Avg x Railroad Length in 100 Km | | | 145.392*** |
| | | | (52.340) |
| | | | p = 0.006 |
| Year Fixed Effects | Yes | Yes | Yes |
| County Fixed Effects | Yes | Yes | Yes |
| County type × Year Fixed Effects | Yes | Yes | Yes |
| Observations | 7,546 | 7,546 | 7,546 |
| $R^2$ | 0.350 | 0.491 | 0.494 |
| Adjusted $R^2$ | 0.312 | 0.462 | 0.465 |

*Notes:* I use the R package Stargazer (Hlavac, 2022) for the analysis. *** p < 0.01, ** p < 0.05, * p < 0.1

As a robustness check whether the stage effect or scale effect has an impact on relations between the model's dependent and independent variables, I run the same analysis for two different sets of years. The first robustness check contains years where there are more than 1000 cholera-caused deaths throughout Prussia, which eliminates the years that cholera epidemics start to vanish. The second robustness check contains years where there are more than 5000 cholera deaths which also eliminates the years with relatively low cholera death numbers. Table 5 shows the results of both analyses. Columns 1, 3, and 5 show the results of the first analysis, and columns 2, 4, and 6 show the results of the second analysis.

Table 5

Cholera Deaths in High Epidemic Years

| | Baseline Model | | With Weighted Avg | | With Weighted Avg Interaction Term | |
|---|---|---|---|---|---|---|
| Cholera Deaths higher than | > 1000 | > 5000 | > 1000 | > 5000 | > 1000 | > 5000 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Railroad Length in 100 Km | 81.497** | 116.508** | 62.758* | 77.011 | 15.953 | 2.687 |
| | (36.259) | (52.303) | (34.328) | (49.377) | (37.057) | (54.972) |
| | p = 0.025 | p = 0.026 | p = 0.068 | p = 0.119 | p = 0.667 | p = 0.962 |
| Population Density | 0.013 | 0.022 | 0.007 | 0.014 | 0.007 | 0.015 |
| | (0.025) | (0.039) | (0.022) | (0.033) | (0.022) | (0.033) |
| | p = 0.612 | p = 0.569 | p = 0.740 | p = 0.662 | p = 0.732 | p = 0.651 |
| Population | 0.000 | 0.001 | 0.000 | 0.001 | 0.000 | 0.001 |
| | (0.001) | (0.002) | (0.001) | (0.002) | (0.001) | (0.002) |
| | p = 0.991 | p = 0.741 | p = 0.994 | p = 0.745 | p = 0.976 | p = 0.739 |
| Weighted Avg | | | 537.667*** | 547.069*** | 507.155*** | 516.032*** |
| | | | (32.979) | (35.009) | (33.358) | (35.757) |
| | | | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 |
| Weighted Avg x Railroad Length in Km | | | | | 143.636*** | 146.431** |
| | | | | | (54.407) | (59.174) |
| | | | | | p = 0.009 | p = 0.014 |
| Year Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| County Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Year × Province Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 5,488 | 4,116 | 5,488 | 4,116 | 5,488 | 4,116 |
| $R^2$ | 0.382 | 0.428 | 0.516 | 0.551 | 0.519 | 0.554 |
| Adjusted $R^2$ | 0.335 | 0.370 | 0.479 | 0.506 | 0.482 | 0.508 |

*Notes:* I use the R package Stargazer (Hlavac, 2022) for the analysis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

The results of both robustness check analyses are very similar to the results in Table 4. The coefficient of railroad length is significant at the baseline model while losing its significance and decreases in value when the weighted average and the interaction term are added to the model. The weighted cholera deaths average of the neighbors has a positive relation with cholera deaths in a county and increasing railroad length by 100 km strengthens the relationship between weighted average and cholera deaths by 143.636 in the first robustness check and by 146.431 in the second robustness check with more than 95% confidence interval for both checks. The population has a near 0 coefficient while the model's coefficient for population density is statistically insignificant. These results increase the reliability of the underlying assumption of model 1, thus increase the reliability of the model's results as well.

# 5 Conclusion

Fremdling (1977), Braun and Franke (2022), and Hornung (2015) similarly stated that industrialization and urbanization in Prussia during the 19th century were highly linked on many levels. By using both simple OLS and fixed effect regressions, Table 1 and Table 2 shows the positive relationship between railroad length and population growth in 19th-century Prussia.

Even though the data lack fundamental measures of economic development, urban population, and urban population density growth over the period can be used as a proxy for economic development in general. (Acemoglu et al., 2002). As both analyses in Table 1 and Table 2 show there is a positive trend in urban county growth compared to rural county growth, which shows the urbanization during the time and the link between railroad length and urbanization as the positive trend in urban county growth is increasing with the railroad increase, the results supports the evidence in the literature and shows the relationship between railroad expansion throughout the century and the economic development in Prussia.

The results from the second analysis of the paper (Table 3 and Table 4) show that the relationship between railroad length and cholera is almost exclusively via weighted neighbor cholera deaths. This result is in line with (Mari et al., 2012) and recent literature which states that human mobility and transportation are good predictors of cross-region spread of cholera disease. The results are also partially validated by Eskişehir/Ottoman Anatolian railroad example in 1892 stated by Schweig (2022) and the inner parts of Nigeria example in 1918 stated by Hogbin (1985). In both examples, the diseases spread across regions by railroad.

Another important result from section 4 is the relationship between the weighted average of neighbor cholera deaths and cholera deaths in a county. It shows there is a strong positive relationship that can partially explain the clusters in the maps in Figure 6b and Figure 7b. Table 4 and Table 5 also show the relationship between railroad length and weighted average. Results from all models that include the interaction term of the two independent variables show there is a highly significant positive relation.

All of the results from both main analyses are strictly limited to correlation and do not show any evidence of a causal effect. Another limitation of the models is there is no control for any economic growth and development indicator. The data that had been used in the analysis does not give the chance for controlling for fundamental measures of economic development and economic growth during the time that also has a relationship with railroad constructions. The only economic development indicator other than the urban population that the data contains is the size of labor working in the crafts, industrial, service, and agriculture sectors only in 1848 which is not suitable for a panel study.

For the second analysis of the paper, if the economic growth of a city is correlated with railroad length or the weighted average of cholera deaths, and the number of cholera deaths in a county during an epidemic, it will lead to biased estimators. As it is stated before, the rapid industrialization in Prussia during the period is highly linked with railroad constructions. Both Fremdling (1977), Braun and Franke (2022), and Hornung (2015) highlight that railroad expansion was the leading sector that drove industrialization during the period by both creating demand for iron and coal and increasing the market for German coal. In addition, according to Hornung (2015), until the mid-1860s German railroad contractions are mostly aimed to connect big cities. This means more developed cities benefited more railroad construction compared to undeveloped regions. This implies railroad length is not independent from the economic development of a city, and it might mean the coefficients of the models are biased if it also correlated with the number of cholera deaths because no economic development control was added to the models.

Another problem with the cholera deaths models might be due to the distribution of industries across counties. Some industries might create more suitable environments for cholera to spread. For example, given that 19th-century factories and mines are notorious for their unsanitary conditions, it is very possible that more industrialized cities and counties with a higher share of the workforce employed in the mining sector were more susceptible to cholera outbreaks than the rest of the country. Considering the relationship between these attributes and the economic development of a county, thus railroad length in a county, leaving these variables out might bias the results of the models.

# References


Acemoglu, D., S. Johnson, and J. A. Robinson (2002): "Reversal of fortune: Geography and institutions in the making of the modern world income distribution," *The Quarterly journal of economics*, 117, 1231–1294.

Braun, S. T. and R. Franke (2022): "Railways, growth, and industrialization in a developing German economy, 1829–1910," *The Journal of Economic History*, 82, 1183–1221.

Dunlavy, C. A. (1991): "Mirror images: Political structure and early railroad policy in the United States and Prussia," *Studies in American Political Development*, 5, 1–35.

Fremdling, R. (1977): "Railroads and German economic growth: a leading sector analysis with a comparison to the United States and Great Britain," *The Journal of Economic History*, 37, 583–604.

Hlavac, M. (2022): *stargazer: Well-Formatted Regression and Summary Statistics Tables*, Social Policy Institute, Bratislava, Slovakia, r package version 5.2.3.

Hogbin, V. (1985): "Railways, disease and health in South Africa," *Social Science & Medicine*, 20, 933–938.

Hornung, E. (2015): "Railroads and growth in Prussia," *Journal of the European Economic Association*, 13, 699–736.

Mari, L., E. Bertuzzo, L. Righetto, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo (2012): "Modelling cholera epidemics: the role of waterways, human mobility and sanitation," *Journal of the royal society interface*, 9, 376–388.

Ohadike, D. C. (1981): "The influenza pandemic of 1918–19 and the spread of cassava cultivation on the lower Niger: a study in historical linkages," *The Journal of African History*, 22, 379–391.

Pang, J., Y. He, and S. Shen (2023): "High-Speed railways and the spread of Covid-19," *Travel Behaviour and Society*, 30, 1–10.

Ross III, R. S. (2015): *Contagion in Prussia, 1831: The Cholera Epidemic and the Threat of the Polish Uprising*, McFarland.

Schweig, A. (2022): "Progressing into disaster: The railroad and the spread of cholera in a provincial Ottoman town," *History of Science*, 00732753221113151.

Sirkeci, I. and M. Yüceşahin (2020): "Coronavirus and migration: analysis of human mobility and the spread of Covid-19," *Migration Letters*, 17.

Snow, J. (1856): "On the mode of communication of cholera," *Edinburgh medical journal*, 1, 668.

Tilly, R. (1966): "The political economy of public finance and the industrialization of Prussia, 1815–1866," *The Journal of Economic History*, 26, 484–497.