Aditya Singh

Guerry's Dataset on Crime in 1830s France

Throughout history, the increase in education is deemed as the common denominator in human progress, but human progress is the ultimate multivariate regression that is exceedingly difficult to quantify. Yet what about the hard, statistical benefits of education, such as its impact on crime rates? In this paper, I explore A.M. Guerry's seminal dataset *Moral Statistics,* the first major statistical attempt to respond to a boom in crime in France in the 1830s. Guerry collected various data points on different regions of France that he called "Moral Statistics," ranging from prostitution to military desertment rate to the amount of money spent on the lottery each year. This paper focuses on Guerry's documentation of literacy rates, to explore the preventive impact of literacy rates on lessening crime in 1830s France. Were more educated regions prone to the crime boom in the 1830s, or did the crime boom cut across different literacy rate levels, and what can that tell us about French society in the 1830s?

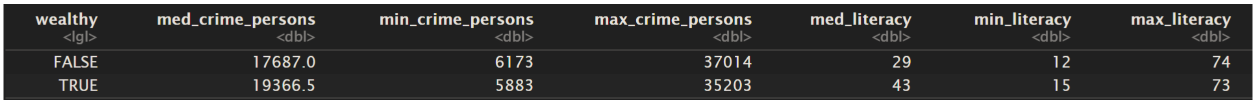
The data used to investigate this question were taken from the Guerry dataset available on GeoDa.

The first continuous variable is the literacy rate. The Guerry dataset documentation notes that literacy rate does mean the rate of the entire population, but only for military conscripts in a region. Universal military conscription was a requirement of all fit Frenchman in the 1830s, although it must be admitted that the literacy rate of military conscripts might not fully represent the literacy rate of the entire population.

The second continuous variable is the *population per crime against person*. Crime is an important indicator of quality of life in a region, making it an appropriate variable to compare to literacy rate. Note that the name can be a bit misleading, and a higher *population per crime against person* indicates a lower crime density. It should also be mentioned that the Guerry dataset provides two types of crime statistics, *population per crime against person* and *population per crime against property*, and that this paper only makes use of *population per crime against person*.

The categorical variable used in comparing literacy rate and crime rate is wealth. The Guerry dataset provides a ranked ratio from 1-89 of the wealth of each region, determined by the per capita tax on personal property, 1 being the wealthiest region. Regions ranked 1-45 are classified as wealthy regions and regions ranked 45-89 are classified as not wealthy regions.

The descriptive statistics of the dataset are shown in Figure 1. The main feature that stands out is the difference in median literacy rate between wealthy and non-wealthy regions at 43% to 29%, but all other statistics are relatively similar for wealthy and non-wealthy regions.



wealthy <lgl>	med_crime_persons <dbl>	min_crime_persons <dbl>	max_crime_persons <dbl>	med_literacy <dbl>	min_literacy <dbl>	max_literacy <dbl>
FALSE	17687.0	6173	37014	29	12	74
TRUE	19366.5	5883	35203	43	15	73

Figure 1.

We now turn to the distribution of the data. Figure 2 shows a boxplot of *population per crime against person* for all regions, and shows a very spread out distribution.

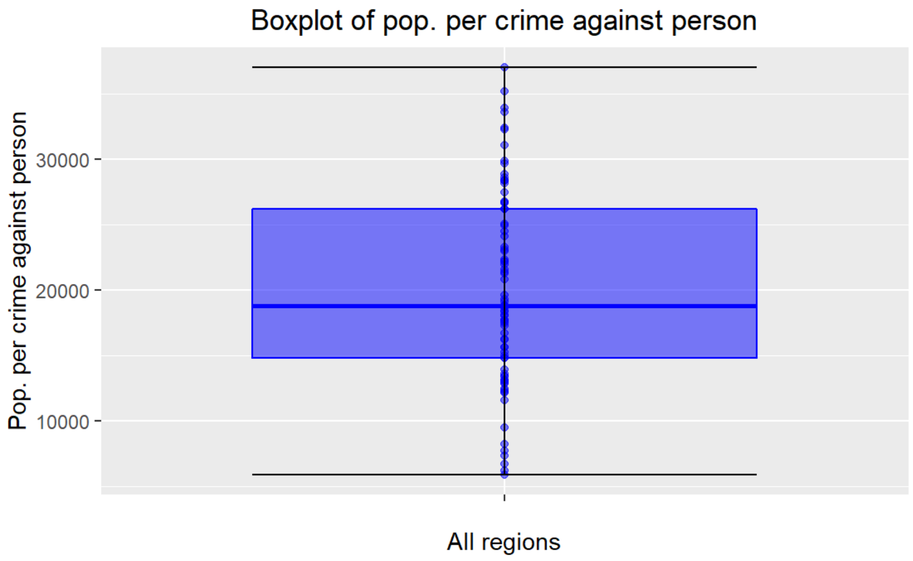


Figure 2.

There are many points both above and below the interquartile range, suggesting a rather heterogeneous spatial distribution that will be explored later on. The spread out nature of the distribution holds when *population per crime against person* is split up into the two subgroups, wealthy and non-wealthy (Figure 3).

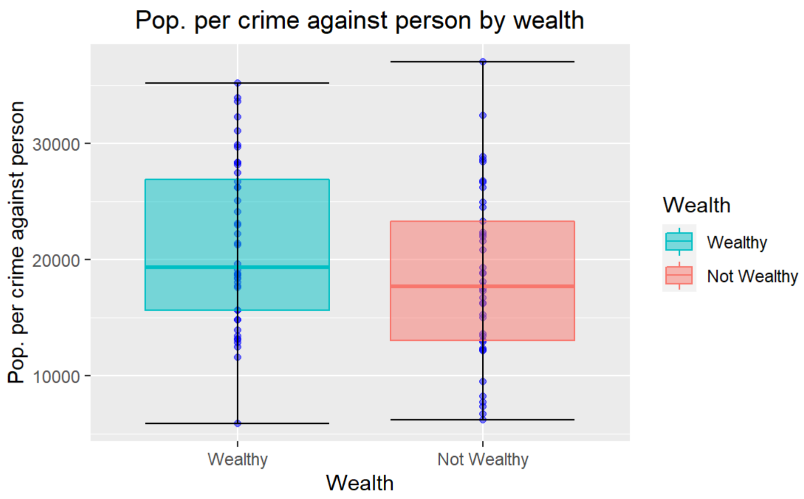


Figure 3.

The same holds for literacy rate (Figure 4) - the distribution is very spread out, suggesting a rather heterogeneous spatial distribution, which will be explored in greater detail later on. Regardless of subgroup, the distribution of literacy rate remains quite spread out (Figure 5).

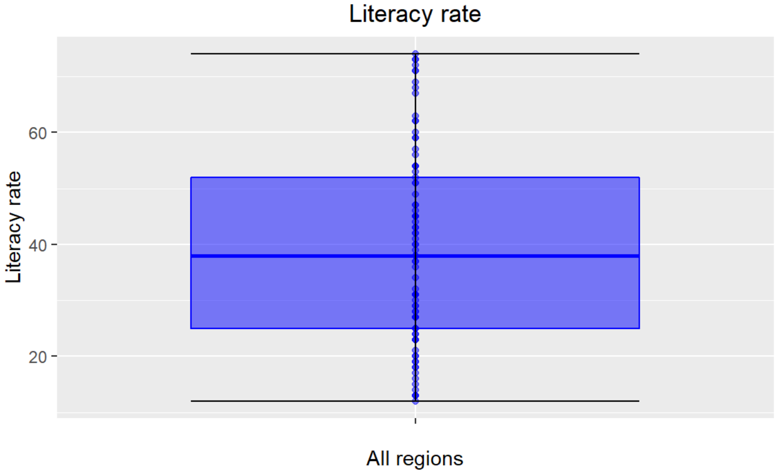


Figure 4.

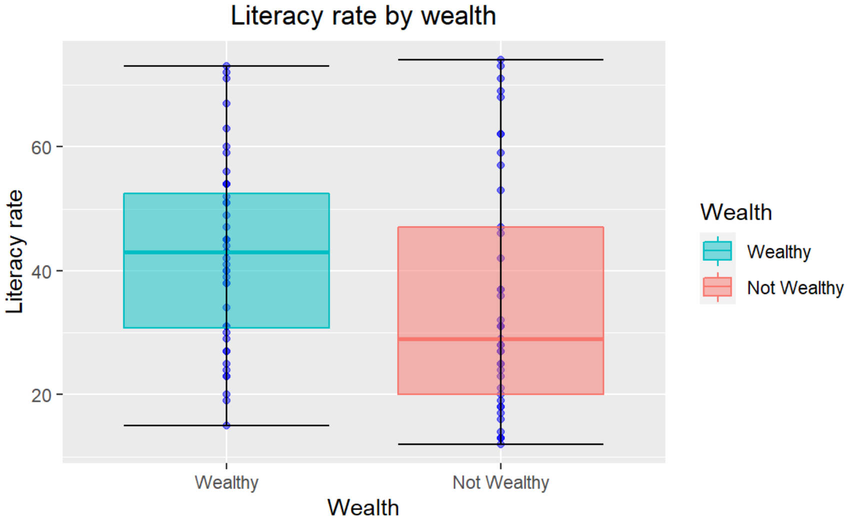


Figure 5.

To examine potential correlation between literacy rate and *population per crime against person* we use scatter plots. We first use a linear fit to plot the correlation between our two continuous variables for the subgroup of wealthy regions and the subgroup of non-wealthy regions (Figure 6).

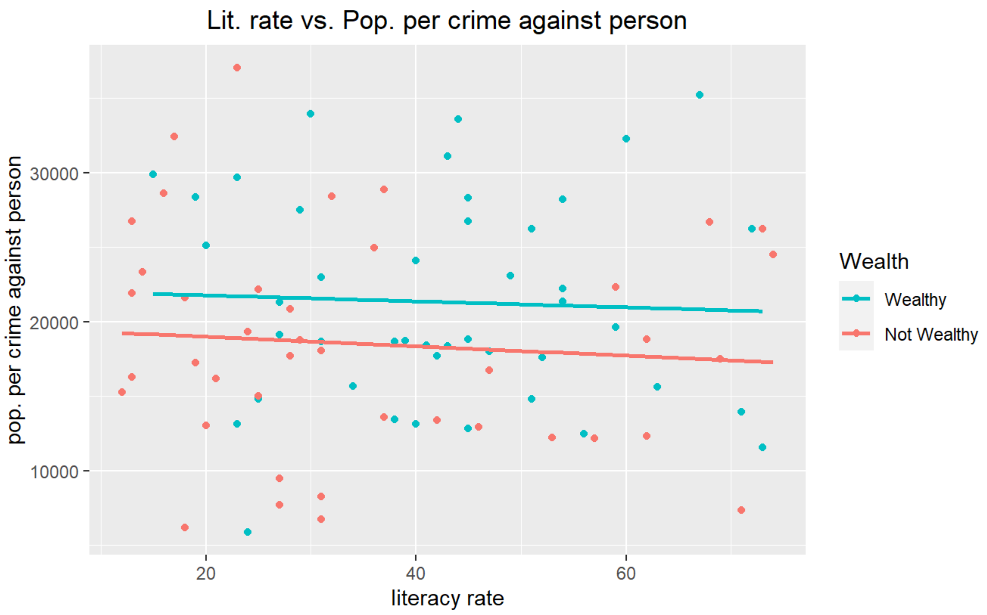


Figure 6.

As the graph shows, there is no correlation to be found in either sub-group. However, when using a LOESS fit for our scatterplot (Figure 7), we find that there is a moderate negative correlation between *population per crime against person* and literacy rate for literacy rates from 0 to 30%.

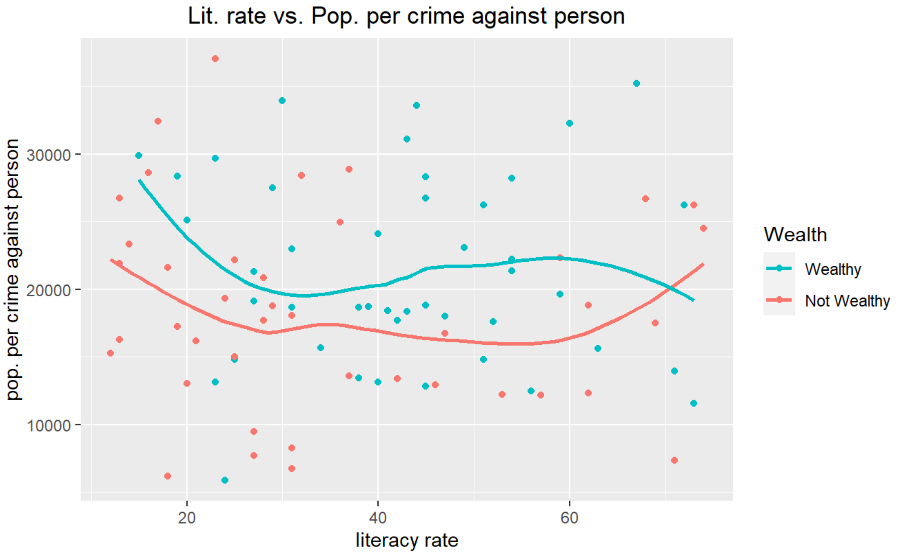


Figure 7.

This is a rather counterintuitive finding, as this implies that increases in the low ranges of literacy rates actually lead to more crime. However, it is likely that this finding is more credible for non-wealthy regions. In looking at the scatter plot for literacy rate by subgroups (Figure 5), one can see that for non-wealthy regions, there is a sizable portion of points in the low 0 to 30%

range. Thus, this downward curve is not merely a gimmick of the LOESS fit, although the correlation for non-wealthy regions is very slight. However, there are very few points in this low range for wealthy regions, and it just so happens that the few points in this low range have very high *population per crime against person* ratios. To include these few points, the LOESS fit has to include a downward curve from a literacy rate of 0 to 30%. In summary, the most credible conclusion we can draw from the scatter plots is that in non-wealthy regions, there is a very slight correlation between literacy and crime rates for the low 0 to 30% literacy rate range.

We now look at the spatial distribution of literacy rates and *population per crime against person* through quantile maps. As expected from the lack of correlation, there is little to no visual similarity between the quantile map for literacy rates (Figure 8) and the quantile map for *population per crime against person* (Figure 9).

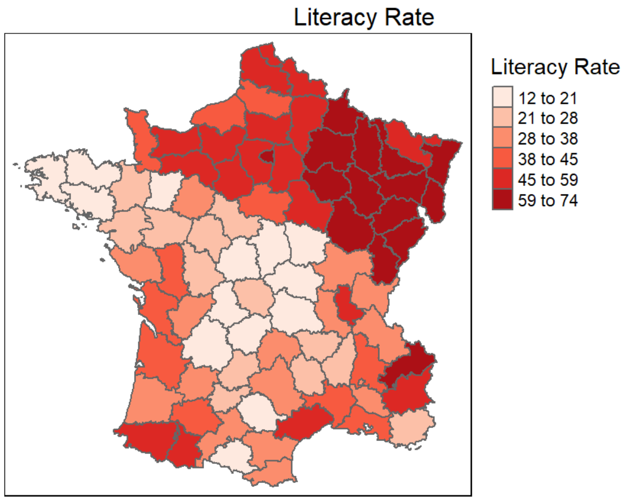


Figure 8.

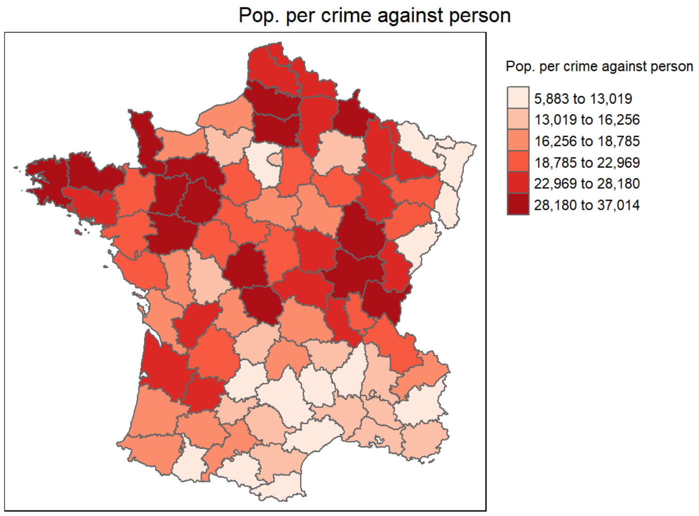


Figure 9.

Thus, the only reasonable conclusion is that there is very little correlation between literacy rates and *population per crime against person.* However, the spatial distribution of literacy rates illustrates a very clear North-South divide, and in particular the northeastern part of France is comprised of extremely high literacy rates. We discuss this finding in the following section.

We now discuss how our finding that there is little correlation between literacy rate and *population per crime against person* would inform a proper path of action in 1830s France to combat crime. The philosophy that education will prevail over crime is certainly not backed by our analysis of literacy rates and crime, literacy rates being one of the "fruits" of education. For example, despite the Northeastern part of France being committed to high literacy rates, it still has many regions with a lot of crime.

The other major alternative to combating crime is a much more hard-nosed, preventive method. Guerry's dataset includes a ranked ratio of desertion rates for every region. Thus, a logical way to explore whether the preventive method would have more efficacy is to compare quantile maps for desertion rates (Figure 10) and *population per crime against person* (Figure 9).

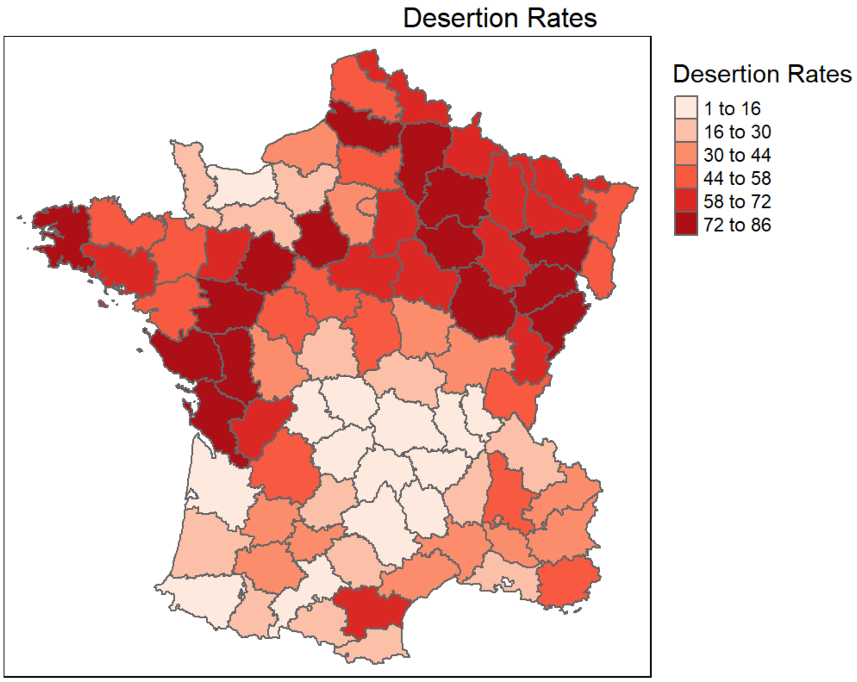


Figure 10.

The main feature of the quantile map for desertion rates is a swath of regions in Southern France with very high desertion rates (they are ranked in the 1-16 range on the index, 1 being the maximum desertion rate). Similarly, in the quantile map for *population per crime against person,* there is a lot of crime in many regions in Southern France. Although this visual comparison is a bit speculative, it implies that there may be a correlation between high desertion rates and high crime rates. Unfortunately, as Guerry only provides desertion rates as a ranked ratio, we cannot use a scatter plot analysis to precisely examine the correlation.

In conclusion, it was very difficult to isolate a single variable that strongly correlated with *population per crime against person* for regions in 1830s France. Literacy rates certainly had very little correlation with crime, and adding in the extremely heterogeneous spatial distribution of *population per crime against person* rates, the evidence strongly suggests that crime rates in 1830s France were likely a complicated result of many different variables, and a multivariate analysis would likely be a better tool.