Data and Results Visualization Project - Supplementary Material

Mattia Mancassola, Alessandro Vannoni, Paolo Roncaglioni

October 24, 2020

1 Introduction

In this document we deepen some of the results that we obtained with our analysis of the digital divide in Italy and Europe, which is available at https://digital-divide-italy.herokuapp.com/. In particular, we focused on the following three aspects:

- households with internet access at home;
- interaction with the public authorities on the internet;
- frequency in online purchases.

The deepening consists in using a new type of visualization, called **Surprise Maps**, introduced by M. Correll and J. Heer [1]. The idea is to go beyond the simple expectation, since "in the process of data analysis, one often seeks out outliers and oddities, places where the data do not match our expectations" [1]. In our case, for example, we do not want just to show the percentage of people aged 14 and over who have interacted online with public authorities for each region, but also how other factors like the wealth or the average age of a specific region impact on that type of analysis. This is important because it allows to discover hidden (and unexpected) patterns, and also to prevent the users from convincing themselves that an important pattern exists, when in fact that is not the case.

2 Theoretical Background

Surprise Maps are based on the **Bayesian Modeling**, in which a probability expresses a degree of belief in an event. This degree of belief may be based on prior knowledge about the event, such as the results of previous experiments, or on personal beliefs about the event. The way in which we can update our beliefs according to the observed data is provided by the **Bayes' Theorem**:

$$P(M|D) \propto P(D|M)P(M) \tag{1}$$

where $P(M)$ is the probability distribution modeling our expectation of random variable M and it is called *prior distribution*; $P(D|M)$ can instead be interpreted as the probability of observing the data D given the model M and it is called *likelihood*; finally, $P(M|D)$ is the *posterior distribution*, namely our updated belief in M after having observed the data D. The precise way in which the updates are computed depends on the distributions of the random variables involved and it will be shown later.

In an nutshell, what happens is that after updating the posterior, the latter will be used in the next update as the new prior, since it has become our new belief and expectation. This means that as new data is

observed, the plausibility of the models changes. Surprise Maps are so called because they quantify the distance between the prior and posterior probabilities, highlighting the unexpected events, namely those that induce large changes in our beliefs. A common way to measure such distance is the **Kullback-Leibler divergence**:

$$KL(P(\mathcal{M}|D)||P(\mathcal{M})) = \sum_{i=1}^{|\mathcal{M}|} P(M_i|D) log \frac{P(M_i|D)}{P(M_i)} \tag{2}$$

where $\mathcal{M}$ represents a selected model space.

3 Algorithm

Follows a sketch of the algorithm used to compute the "Bayesian surprise". Note that, following again the method proposed by [1], the likelihood $P(D|M)$ is computed "through a comparison of the expected data density and the posterior observed data density", indicated below as E and O respectively.

Algorithm 1: Surprise Update Algorithm

Start with equiprobable $P(M)$s
$E(x, y, t)$ is the expected data density
$O(x, y, t)$ is the posterior observed data density
for each year **do**
  Calculate $O - E$
  Estimate $P(D|M)$
  Estimate $P(M|D) = P(D|M)P(M)$
  Compute surprise as $D_{KL}(P(M|D)||P(M))$
  Normalize so that $\sum P(M)s = 1$
end for

4 Analysis

As said, starting from our previous analysis, we decided to focus on the following three datasets. Before going into the details of each dataset, it is worth to underline one fact: as said in the introduction, a Bayesian Model requires to have a *prior distribution*, namely a probability distribution that expresses our belief. That is why, as indicated in [1], the first step to build a Surprise Map is the **selection of relevant event models**. This operation usually requires domain knowledge and statistical expertise; however, Surprise Maps have an interesting advantage: they can be useful even if the selected models are relatively poor. Indeed, assuming for example a very simple model such as the uniform one, the produced surprise will be a scalar multiple of the event density, i.e. something with the same level of information of a heatmap.

For each one of the datasets, a motivation of the chosen model is given, together with the corresponding mechanism of computation. Note also that, as again indicated in [1], when you have to deal with data that have a temporal component, the frequency of the Bayesian updates has an impact of the magnitude of the surprise values. Thus, in the following examples the updates are computed by splitting the events in batches.

4.1 Interaction with the public authorities on the internet

In this section we analyze the data related to the online interactions of individuals aged 14 and over with the public authorities. Three types of interactions are taken into consideration: "obtaining information", "downloading official forms" and "submitting completed forms".

First of all, we show what would be the result of a standard per capita event rate map, to allow a comparison with the subsequent analysis and to draw some conclusions. In Figure 1 there is a choropleth map showing the event density of interactions by region. Since the data are not changed, the comments are obviously the same we did in our previous analysis, especially the ones referring to a decreasing tendency going from northern to southern regions and a clearly higher percentage of interactions for what regards the "Obtaining information" task, which is notably the simplest one among the three.

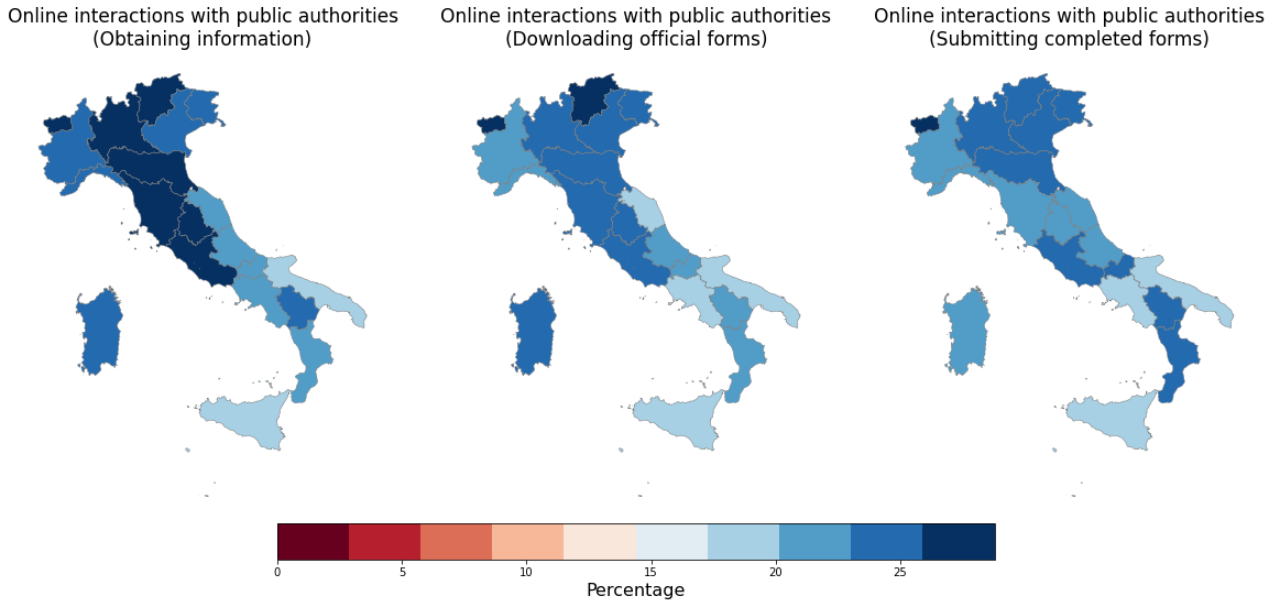

Figure 1: The **Event Density** of *interactions* by region (2018)

Let's now focus on the new analysis, the one aimed at de-biasing perceptions of event density using Bayesian methods. Two new visualizations will be provided: the first one has been built to de-bias the standard choropleth map we introduced before, while the second one will use the surprise factor to introduce a new element in the investigation and hence discover a new interesting aspect.

4.1.1 de Moivre Funnel

As clearly highlighted by [1], there are many possible biases inside a thematic map. For example, there are sometimes latent variables that affect the density of the events, causing some interesting information to be indistinguishable among the other or, vice-versa, indicating some patterns as appealing when the data are instead insignificant. One straightforward example is when we have a base rate that is not uniform, such as the population one: obviously, a very populous region will have an high number of occurrences of a particular event. The standard way in which this problem is taken into account consists in transforming the event frequencies into rates, normalizing by the population density. However, this can lead to deceptive visualizations due to the high variability resulting from the "most dangerous equation" $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ [2].

We therefore provide a new visualization of our standard event density map, following the "*de Moivre funnel*"

approach of [1]. First of all, it is useful to show the so-called *funnel plot* (Figure 2), i.e. a scatterplot where (in our case) the event rate is plotted against the sample size.

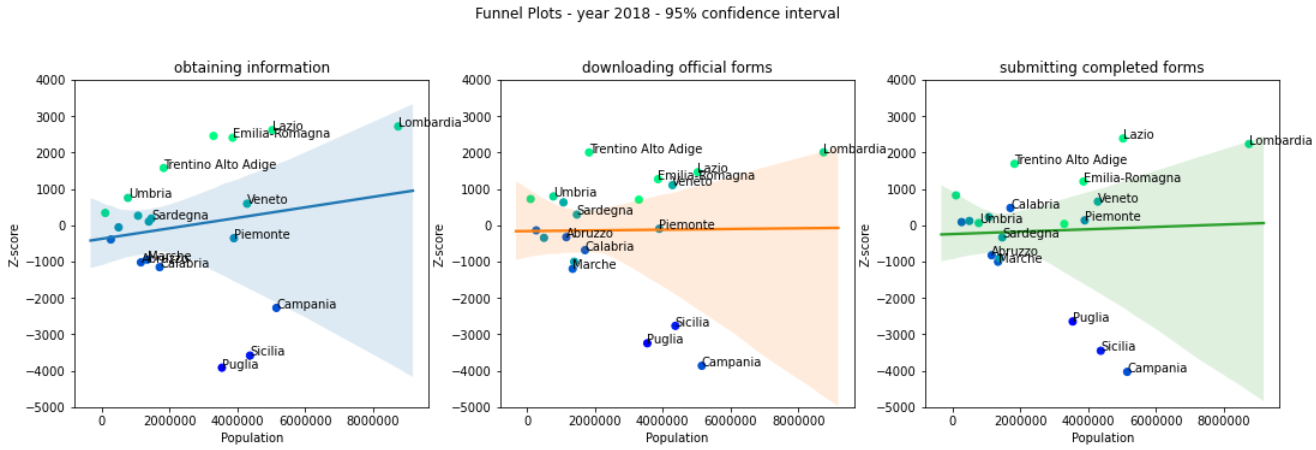

Figure 2: The funnel plots of the 2018 interactions rate with public authorities by region

Even from the above funnels we can make some interesting observations: there are regions with relatively low population which have a high interaction rate (w.r.t. the national average), such as *Trentino Alto Adige*. Other regions, instead, despite having a fairly high population, have a low interaction rate, such as *Puglia, Sicilia* and *Campania*.

Mathematically speaking, we can estimate the likelihood $P(D|M)$ in the following way: for each region, we compute the Z-score

$$Z_s = \frac{O(s) - \bar{x}}{SE_s} \tag{3}$$

where $SE_s = \frac{\sigma_s}{\sqrt{n_s}}$ is the **standard error**. Then, we can derive the probability of each event by integrating:

$$P(s|deMoivre) = 1 - \int_0^{Z_s} \phi(x)\,dx \quad \text{if } Z_s > 0 \tag{4}$$

$$P(s|deMoivre) = \left| \int_{-Z_s}^{0} \phi(x)\,dx \right| \quad \text{if } Z_s < 0 \tag{5}$$

where the two cases are due to the intention of including the sign of Z_s to differentiate between over- and under-representation, and where $\phi(x)$ represents the normal distribution.

Finally, since each event has an independent distribution, the likelihoods shall be summed and normalized by $|D|$, i.e. in our case by the number of regions:

$$P(D|deMoivre) = \frac{1}{|D|} \sum_{i=1}^{|D|} P(D_i|deMoivre) \tag{6}$$

In Figure 3 is shown the result of the above approach. As expected, the regions with a low *z-score* are shown here to have a negative surprise, meaning that their rate of interaction is low w.r.t. their population. The opposite obviously happens with the regions colored in blue. Interestingly, even if this visualization has some patterns in common with the standard density maps of Figure 1, it helps to better understand the condition of each region and also to highlight some curious situations, such as that of *Valle d'Aosta*, which, as opposed

to the previous graph, is shown here with a small surprise (especially in the "Obtaining information" task), meaning that its rate is consistent with its dimension.

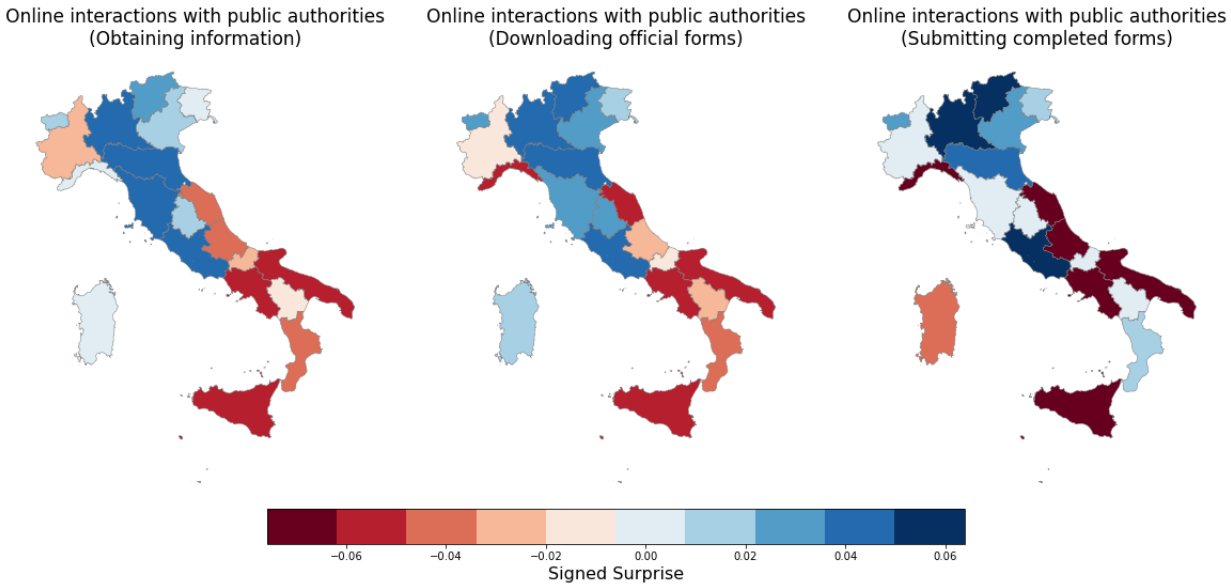

Figure 3: Signed Surprise Maps - de Moivre Funnel

4.1.2 Base Rate - GDP per capita

In this second part we will again use a surprise map, but this time to introduce a new element in the investigation: the **Gross Domestic Product (GDP)**. The idea is to use this indicator to formulate an hypothesis about our data, hence specifying a prior distribution. More precisely, assuming that the regions with higher GDP per capita have higher rates of interaction, we will use a one-to-one function, considering as our expectation the rate among the observed values which corresponds to the region with the nth highest GDP per capita. For example, if the *Lombardia* region as the highest GDP per capita, our expectation is that it will have the highest rate of interaction with the public authorities.

This model is called **Base Rate**, since it assumes a per-region rate. Considering S as our domain of regions, we can compute the likelihood as follows:

$$P(D|BaseRate) = 1 - \frac{1}{2}\sum_{i=1}^{|S|}|O(i) - E(i)| \tag{7}$$

In Figure 4 is shown the result of this approach. Also in this case the result is pretty meaningful, since we can draw conclusions that before we were not able to draw. The *Lombardia* region, for example, is clearly penalized by the fact of having the highest GDP per capita, but, especially in the last two tasks, clearly not the highest interaction rate. This type of considerations can be carried out for each region. Of course, it must be considered the simplicity of the model, which can for sure be improved by mixing it with other models.

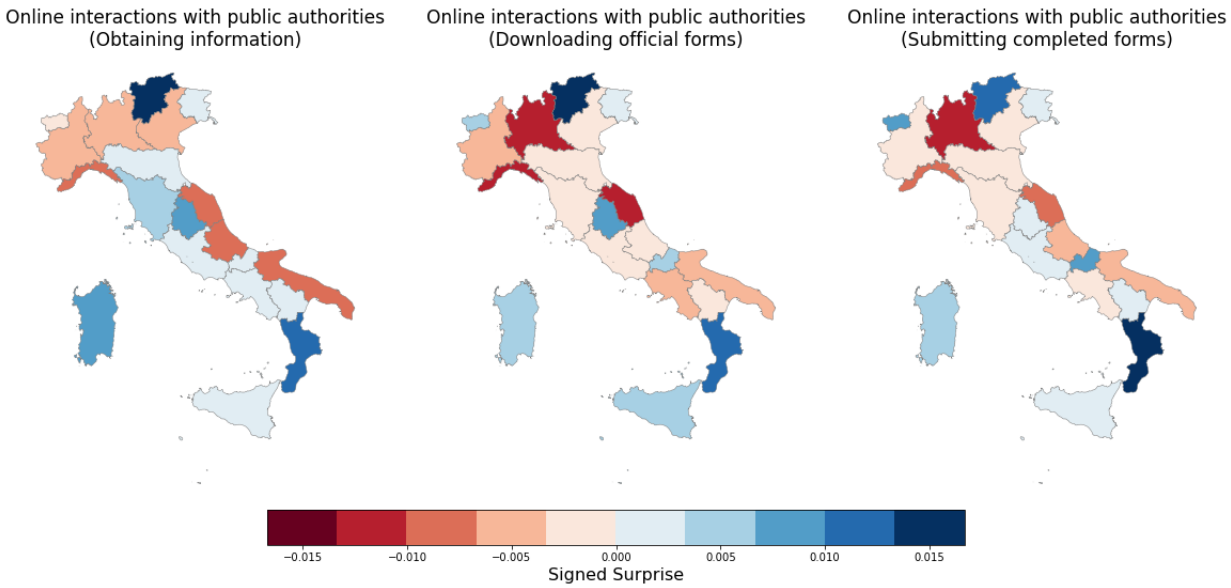

Figure 4: Signed Surprise Maps - Base Rate (GDP per capita)

4.2 Internet usage over the territory

Let's proceed with the same approach, considering here the *internet use of individuals over 6 years of age*. The procedure is the same: showing an initial standard density map (of the type used in the main analysis) and then highlighting some interesting aspects by adopting the Surprise Map approach. The goal is again to see if and how some factors have an impact on the topic under investigation, i.e. in this case the internet usage at a regional level. In particular, two factors will be considered: the *graduates rate*, namely the percentage of individuals who have a certain degree (of any level), and the *average age*.

As said, we first consider the standard visualization of the index taken into consideration, which can be observed in Figure 5. It represents the percentage of individuals (by region) who make use of Internet. Even though the percentages are not so different, there is a clear distinction between northern and southern regions.

4.2.1 Graduates rate

In this section we want to use the percentage of graduates of each region to formulate an hypothesis on what could be the different levels of internet usage. In particular, as described in the previous example concerning the analysis with the GDP indicator (section 4.1.2), we will associate a higher percentage of graduates with a higher internet activity.

The result of this analysis is shown in Figure 6. We can still see the same distinction between the northern and southern regions, however there are some interesting exceptions, like *Puglia*, *Basilicata* and *Calabria*, which manage to position themselves slightly higher than the expectations. This means that the level of use of Internet is high w.r.t. their rates of graduates. *Campania*, on the other hand, has a far worse rank than expected. Other minor "surprising" outcomes can be seen in the other regions, both in positive and negative sense.

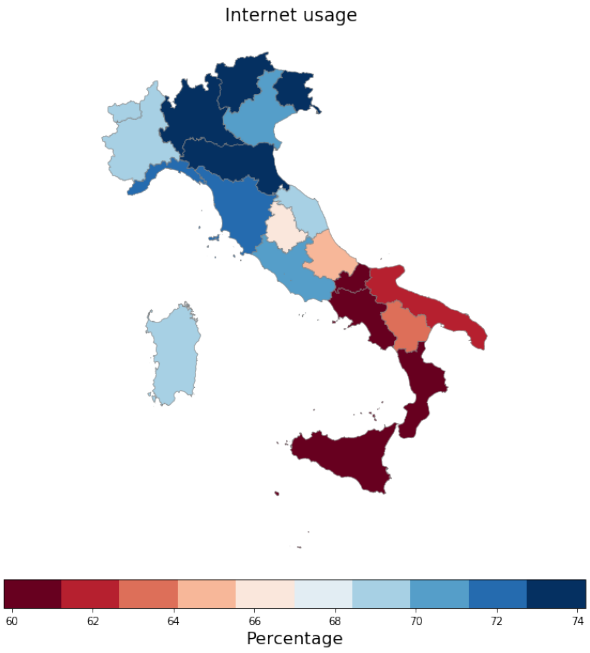

Figure 5: The **Event Density** of *internet usage* by region (2018)

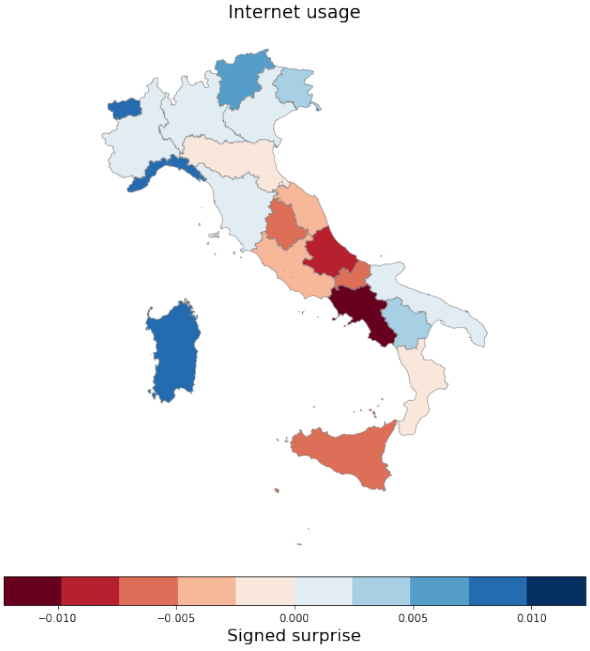

Figure 6: Signed Surprise Maps - Base Rate (graduates rate)

4.2.2 Mean age rate

Next, we propose a surprise map using instead the *average age* of the population by regions. As usual, we are likely to assume that a lower average age in the population is associated with higher internet activity of individuals. The result is shown in Figure 7.

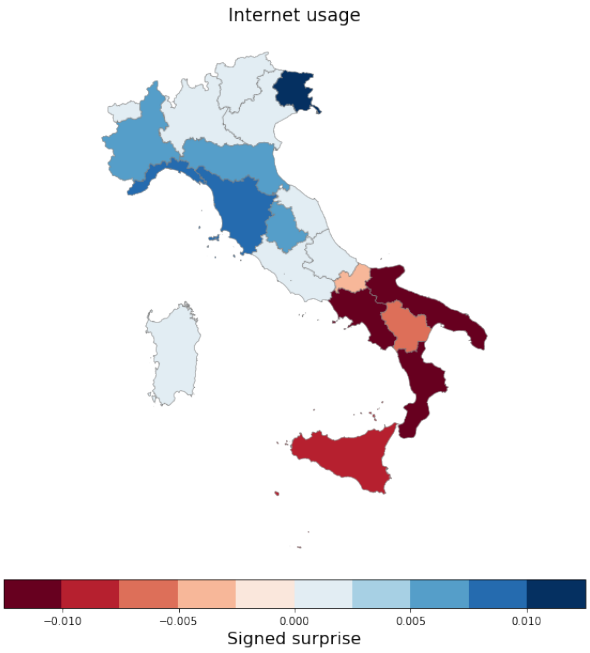

Figure 7: Signed Surprise Maps - Base Rate (mean age)

There is a clear demarcation line in the center, below which we have negative surprises values, leading to a result consistent with the *event density* map in Figure 5. Indeed, regions like *Campania*, *Puglia* and *Calabria* have a relatively young population, but we have already seen how they have a low (if compared to other regions) level of Internet usage. *Friuli-Venezia Giulia*, instead, stands out for the great positive surprise value, showing an high level of Internet usage despite having an old population (w.r.t. the national average). *Trentino Alto Adige*, if compared with the *event density* map (and also with the previous surprise map), does not seem to lead to a surprise in any sense, meaning that it has a level of Internet usage consistent with the quite young population (w.r.t. the national average). Other similar observations can be carried out for the other regions.

4.3 Frequency in online purchases

This section focuses on the frequency of online purchases by region and aims to exploit surprise maps to explore correlations between this activity and indicators such as GDP per capita, household internet access and level of education.

In all the sub-sections, the analysis will be carried out by looking at two different perspectives: the one of frequent online buyers (people having made at least one online purchase in the past three months), and the one of online non-buyers (people having never made an online purchase). Those two categories are analyzed annually to see their evolution and how much surprise they generate with regard to different priors. Those two extremes in the frequency of purchase have been chosen, rather than intermediate ones, to obtain observable differences in the data.

We start by showing the standard event map provided directly from the data for the year 2018. Figure 8 highlights the usual north-south divide: as already noted in our previous analysis, it appears that the inhabitants of the northern regions are generally confident in online shopping and do it very frequently, while online purchases become more sporadic in the southern regions.

We now try to find correlations among these data and the aforementioned indicators, to reduce possible biases and uncover unexpected patterns.

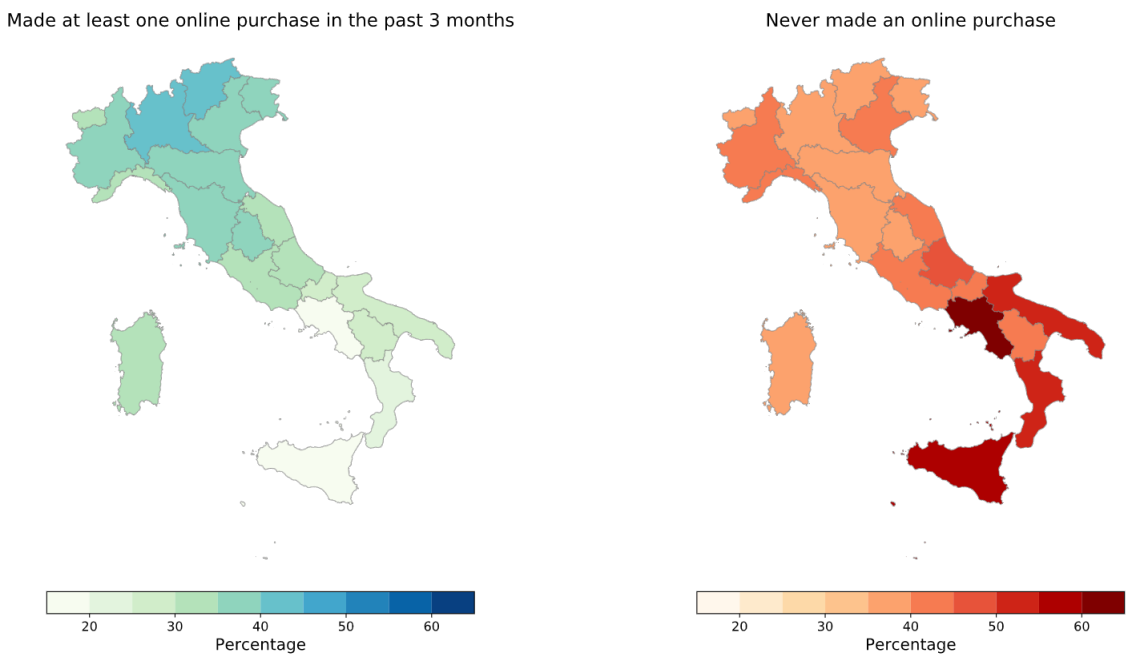

Figure 8: The **Event Density** of *online purchases* by region (2018)

4.3.1 GDP per capita

The surprise maps of Figure 9 are produced by using the **GDP per capita** by region as a Base Rate model for the frequency of online purchases and, vice versa, in the non-use of the internet for shopping. We once more assume that an higher GDP per capita translates to a more frequent online shopping activity; this model allows us to draw some new conclusions.

The *Umbria*, *Calabria* and *Sardegna* regions exhibit in both cases an higher inclination towards online shopping than their GDP per capita would suggest: this is especially notable for *Calabria*, which is among the lowest ranking in frequency of purchase. On the other hand, *Lazio* and *Campania* show the opposite tendency, but while the *Campania* region was already in the lowest spot, *Lazio* seems more surprising. The negative trend of those two regions is consistent even in the following analyses.

4.3.2 Internet access

Since it is a necessary prerequisite, the possibility of accessing the internet can be used as an appropriate predictor of frequency of online purchases: Figure 10 shows the two surprise maps with household internet access as the Base Rate.

In this case, the surprise factor for most of the regions is limited, meaning that the correlation is stronger and internet access rate can effectively predict online shopping activity. The notable exceptions lie in the *Molise*, *Basilicata* and *Calabria* regions, which in this case exhibit a positive surprise, while *Sardegna* and the aforementioned *Lazio* and *Campania* stand out as negative surprise regions.

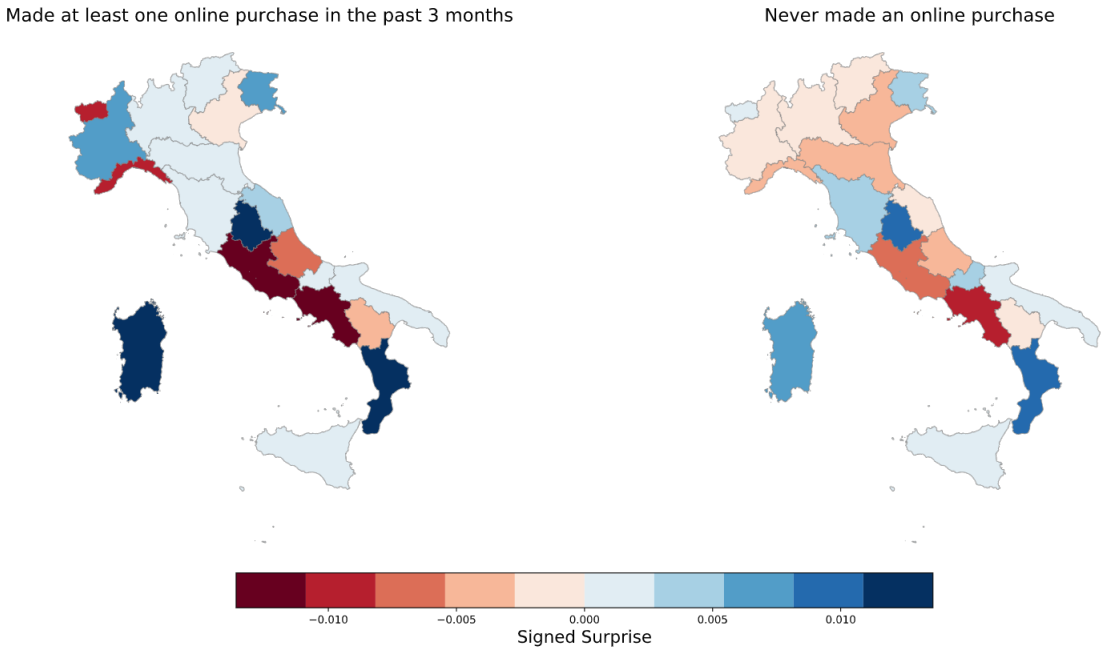

Figure 9: Signed Surprise Maps - Base Rate (GDP per capita)

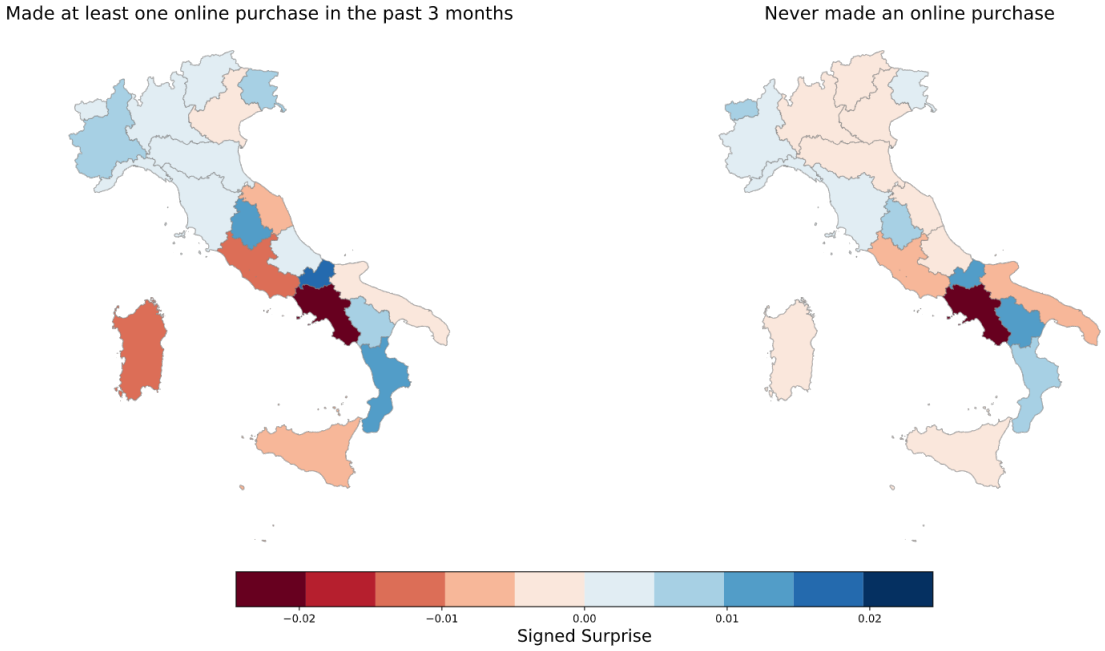

Figure 10: Signed Surprise Maps - Base Rate (household internet access)

4.3.3 Level of education

The maps of Figure 11 are produced by using the level of education by region as prior: specifically, the data taken into consideration to approximate the level of education, corresponds to the percentage of people with at least an upper secondary school certificate.

Here the data seems to yield more surprise in all regions, both positive and negative, than previous cases. This could indicate that the level of education is less effective than other indicators to predict online purchases. Along with the usual negative surprise for the *Campania* region, there are also positive surprise regions such as *Sardegna*, *Puglia* and the northern regions (especially *Trentino Alto Adige*).

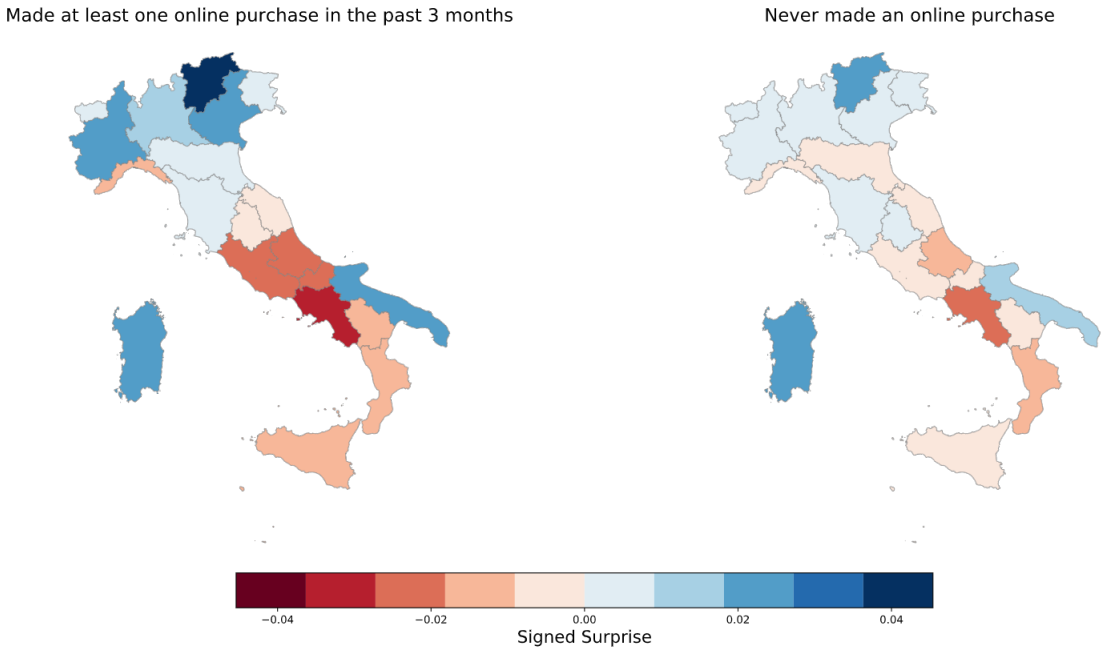

Figure 11: Signed Surprise Maps - Base Rate (Level of education)

References

[1] M. Correll and J. Heer. "Surprise! Bayesian Weighting for De-Biasing Thematic Maps". In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 651–660. DOI: 10.1109/TVCG.2016.2598618.

[2] Howard Wainer. "The Most Dangerous Equation". In: *American Scientist - AMER SCI* 95 (May 2007). DOI: 10.1511/2007.65.249.