POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Demographic winter in Italy, how have we come this far?

Nonparametric Statistics
Mathematical Engineering - Statistical Learning

Francesco Maria Mancinelli, Leonardo Marchesin, Beatrice Sisti, Alessandro Venanzi

Academic year:
2023-2024

Abstract: In recent years, Italy has experienced significant shifts in demographic trends, with fertility rates serving as a key indicator of population dynamics. Fertility rates are examined across various regions and demographic groups, in order to uncover patterns and trends in fertility behavior over the last two decades, from 2002 to 2021. Through a non-parametric analysis, the project seeks to provide insights into the factors that influence fertility decisions in Italy and their implications for population dynamics, social policies, and economic development.

https://github.com/LeoMarche24/Nonparametric_Project

Key-words: Functional Data, Splines, Permutational Test, Semiparametric Regression, Conformal Prediction

1. Introduction

The demographic winter is one of the most debated and crucial problems that affect Italy. Both domestic, foreign media and public figures talk about it and discuss possible solutions for overcoming this trend, which is causing from now on several problems on the stability of our entire system, that needs young people to sustain the social securities for the older ones. As Figure 1 shows, the decaying in the number of newborns is significant for all of the 107 provinces in the Country.
From an introductory analysis, by performing a unique permutation ANOVA test for each year[1], to assess differences in the macro-regions of Italy (namely, *North*, *Center*, and *South*), it is clear that the regional difference is significant not for all years (2002-2021), highlighting several trends and peculiarity in the geographic differences along years.
The number of newborns is the result of several factors, depending on the number of people in a reproductive age, on the changes in the culture and in the customs of people. Moreover, the total number of newborns, even if divided by the total population (as done in Section 3), is affected by the number of people living in a specific province and, since in Italy the older ages are expanding while the younger are shrinking, that statistics is not completely informative. A

[1]Note that this is not a p-value function, in fact it is a sequence of p-values evaluated on different sets of data, namely a dataset for each year.

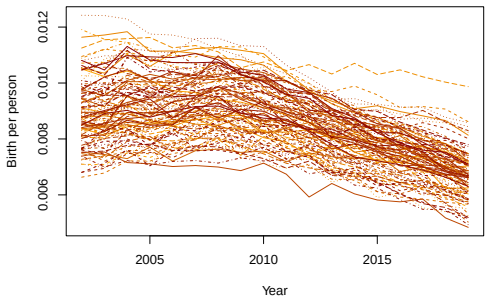

Figure 1: Total newborns per person for each province in 2002-2021

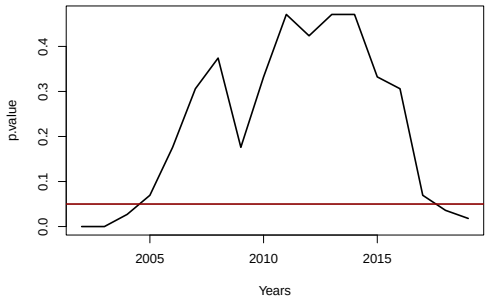

Figure 2: P-value function for the newborns, using macro-regions as factor

statistics which is surely more explanatory is the so called fertility rate, namely the number of newborns for 1000 women of the same age, and it provides much more information about the approach to the decision of having a child in the different years and geographic region. In fact, even if this informative statistics of the fertility rates would increase in future years, the number of newborns will eventually drop because of the lack of people in a fertile age. Nevertheless, this kind of data are the ones on which decisions can be made, in fact the number of people in a fertile age will increase if and only if the fertility rates increase, and moreover these rates are the data directly affected by decisions, and changes in the behavior of the population, while the number of newborns is only a consequence.

1.1. Research question

Understanding these changes and explaining them is the goal of this project, in particular it aims to study the different behaviours along the different years and geographic regions to understand where and when the fertility rates have changed, and most of all how they have changed. Nowadays the fertility rates are naturally shifting towards the right, because of the delay on making people's own family with respect to 20 years ago. The effect that should not happen is to lower these curves together with the shifting, and the goal of this work is to understand how and where this lowering effect happens and propose possible solution for overcoming this problem. A possible stakeholder for this work is the Italian Government, whose responsibility is to maintain strong our entire system, to encourage and to do whatever it takes to create a health society to enable individuals to establish a family. Indeed, understanding the behaviour in our days of the new families surely helps in targeting the incentives and the aids for the young people who are willing to have a child.

1.2. ISTAT dataset

A dataset provided by ISTAT is employed, consisting of the fertility rates, expressed as the number of newborns per 1000 women, for each year from 2002 to 2021. It provides data for each Italian province and it divides the fertility rate based on the age class of the mother, from 17 to 50 years old. Note that data from 2002 onwards include aggregate values for provinces that exist today but were not yet established in 2002. In order to explain with socioeconomic indices the phenomenon of changing in the birth rates, the data is enriched with variables concerning the education, socioeconomic status, health and population dynamics. Specifically, they include the number of university enrolls, the percentage of young individuals who have discontinued their studies, employment and unemployment rates, abortion rates, as well as immigration and emigration statistics. However, it is noteworthy that certain data are not available at the provincial level for the entire period under consideration. Consequently, regional data over a narrower time-frame will be considered.

2. Mathematical framework and methods

For each year and province, the fertility rates of the newborns for 1000 women are mathematically speaking a function

$$f : \mathbb{N}^{34} \to \mathbb{R}.$$

In fact, for each year the ratio $\frac{\text{total newborns}}{\text{total women}} \times 1000$ gives a value which is a real number, the starting space ($\mathbb{N}^{34}$) is the space of all possible ages' groups for the women (less than 17, 18, 19, ..., 49, more than 50).

Even though each record involves only discrete values, they reflect a smoothing variation in the rates that could be assessed in principle, as often as desired, and is therefore a rate function. Ramsay and Silverman [2005].

According to Jurkiewicz [2011], we assume our data to be the noisy realization of functional data

$$f_{i,j} : \mathbb{R} \to \mathbb{R}, \quad i = 2002, \dots, 2021; \ j = \text{Agrigento}, \dots, \text{Viterbo}.$$

The recording of these data are not exactly precise, as they are projection studies made by ISTAT and are therefore prone to some error or noise. Hence, the actual model is

$$n_{ijk} = f_{ij}(k) + \varepsilon_{ijk}, \quad i = 2002, \dots, 2021; \ j = \text{Agrigento}, \dots, \text{Viterbo}; \ k = 17, \dots, 50,$$

where the error parameter ε is assumed to be unbiased. Smoothing splines, Section 2.4.1, will be used.
Not only functional data will be used, but also some uni-variate and multivariate statistics will be extracted from each function, in order to study the quantities of interest of each unit. The non-parametric approach is crucial in this work, in fact the flexibility that allows to customize the test statistics and to treat all types of data without any assumptions on it - even functional data - is deeply used to understand all the variability that is explainable from a general object and in a general framework for the relationship between variables.
In principle, the observations are assumed to be all identically distributed, being said that during the analysis some of this assumption will be rejected because of an evident difference in these distribution.

2.1. Depth measures

The concept of depth is a non-parametric tool which gives a sort of ordering (inspired to the up-down ordering on the real line) for multivariate situations where the distributions are not known. However, functional data ordering is not based on a center-outward order, but on a down-upward order. This measure is given by the Modified Epigraphic Index (MEI), which computes the amount of domain spent by the other curves of the sample above a given curve.

2.2. Permutation tests

The permutation tests aim to find a group of permutations under which, given H_0 true, all the possible outcomes of the test statistics will have the same probability to happen; comparing them to the value of the statistic obtained from the original sample gives a criterion to accept or reject the null hypothesis.
In particular, the ANOVA permutation test will be frequently used, even in its extended forms of functional and multivariate ANOVA, to test for the influence of different labels, that could be geographic region or years. The general test is the following, given $X_{i,j,k} = \overline{X}_{i,j} + \tau_k$ the quantity to test,

$$H_0 : \underline{\tau} = \underline{0} \qquad \text{vs} \qquad H_1 : \underline{\tau} \neq \underline{0},$$

where k represents different factors depending on the test.

2.3. Bootstrap

The primary task of bootstrapping is estimating the distribution of a statistic from a random sample. The bootstrap distribution is uniformly distributed over the n^n possible values of the bootstrapped sample, and the estimated quantities from the distribution are evaluated accordingly.

2.4. Non-parametric regression

The objective of non-parametric regression is to go beyond linearity, in this framework the observed relations are in some cases purely non linear. Because of that, multiple non-parametric regression techniques are used.

2.4.1. Splines

As explained in de Beer [2011] and Jurkiewicz [2011], in literature usually parametric models have been used to estimate the considered functions, but modern tools such as non-parametric splines are more able to describe all kinds of age patterns. In order to perform the smoothing of these curves, it is important to note that also their derivatives will be informative. Indeed, also the first and second derivatives will come into play, given that they represent a sort of speed and acceleration in the birth rates. Nevertheless, the model for the original curves is made with smoothing splines with penalization on the third order, because of consistency of the model with the usual smoothing procedure (command `smooth.spline`). Before proceeding with the actual smoothing process, the number of basis needs to be determined. The number of basis is selected via a generalized cross-validation approach, where the sum of the GCV obtained for each curve in the dataset is used as index. The result is 9 basis.
As for the second derivative, the smoothing with penalization on the third order would not

be effective, and in principle the curves are continuous on a compact set - $f_{ij} \in C^0[17, 50]$ -, and without loosing generality, the domain of the functions is set to be $C^5[17, 50]$. Hence, the smoothing is performed with penalization on the fifth derivative.
The different order of smoothing is needed in order to have a good prediction at the boundaries for all the original curves, and moving to the second derivative the third order would not have been enough, hence the choice to add 2 degrees. Note that for second derivative only some quantities of interest will be investigated, and far from the boundaries, resulting in a consistent analysis.

2.4.2. Semiparametric regression

Semiparametric regression combines both parametric and non-parametric components: the parametric one is used to model the relationship between the response variable and a subset of covariates, while the non-parametric component allows for more flexible modeling of the remaining covariates. The general mathematical expression is

$$Y = \beta W + f(X) + \varepsilon,$$

where $Y \in \mathbb{R}$ is the response variable, $W \in \mathbb{R}^q$ are the covariates with parametric effect, $X \in \mathbb{R}^d$ those with non-parametric effect and ε is the error term. The nonlinear part is obtained through a smoothing spline fit, building a smooth term for each covariate. The B-splines with third order penalization are used, in particular a set of knots spread evenly through the covariate's values. They are penalized by the conventional integrated square second derivative cubic spline penalty. The fitting is based on penalized likelihood, where the term to be penalized is the second derivatives of the smooths.

2.5. Conformal inference

The problem with prediction when there is no assumption on the distributions is that there is even the lack of asymptotic results. Because of this, when the model will be validated, using the last years' data, a conformal inference approach is needed to set the performance of this study.

2.6. Robust statistics

Robust methods aim at fitting the bulk of the data well, even in presence of a small proportion of outliers, giving approximately the same results as the classical methods applied to the data without outliers. To robustly estimate the location, one can discard a proportion of the largest and smallest values. A robust estimator for the location is the α-trimmed mean, defined as

$$\overline{x}_\alpha = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} x_{(i)},$$

where $\alpha \in [0, \frac{1}{2})$ and $m = n\alpha$.
Another robust statistical method is Minimum Covariance Determinant (MCD) estimator, used for estimating the mean and covariance matrix of multivariate data. Its primary goal is to find a subset of the data (a *minimum determinant* subset) for which the sample covariance matrix has the smallest determinant. Denoting the sample mean of observations in subset H with μ_H, the best clean data subset is defined as

$$H^* = \text{argmin}_H \det\left(\frac{1}{h} \sum_{x_i \in H} (x_i - \mu_H)(x_i - \mu_H)^T\right)$$

and the MCD estimator is the pair of sample mean and covariance estimates computed on this subset H^*. This approach will be used in the study in order to detect outliers in the covariates of the non-parametric regression. Note that this procedure will be used to perform a robust non-linear regression which has been inspired by the one producing the Least Trimmed Squared (LTS) model for a linear regression.

3. Introductory analysis

The first analysis is performed on these days' curves, using a trimmed point-wise mean for the last 5 years for each age, displayed in Figure 3. A result of 107 curves is obtained, one for each province.

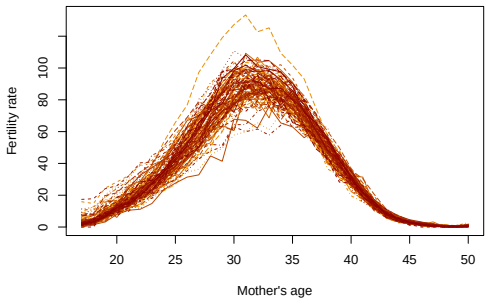

Figure 3: Curves for each province, trimmed mean of the last 5 years

To visualize the different magnitude in the Country, an exploratory plot shows the modified epigraphic index, Figure 4, for each curve. Note that a dark color means a higher curve in the fertility rates. The situation nowadays highlights a higher curve in the northern region of Italy and in some unique provinces on their own in the South, but the majority of the zones in the latter region has a very low coefficient.

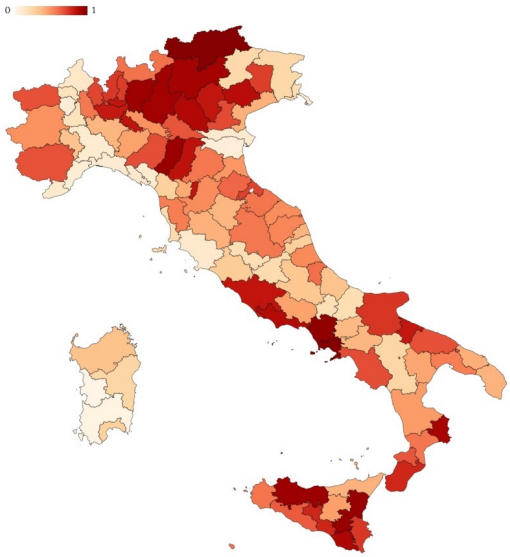

Figure 4: Modified epigraphic index of the fertility curves along Italy

3.1. Base data

The whole dataset is plotted in Figure 5 in its integrity, all the provinces and years together. The classical shape of the fertility rates is easily recognizable, with a peak for the ages that goes from 25 to 35 years old.

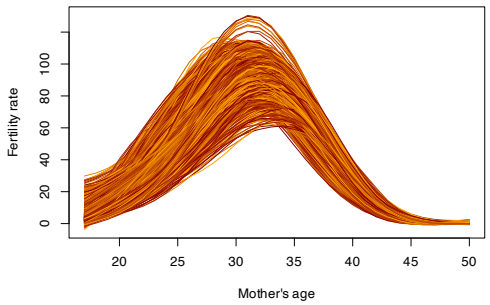

Figure 5: All curves of the dataset

In particular, it is possible to see very different behaviours around the peak of the curve, while in the right part of the domain the curves tend to flatten towards zero.
First of all, an analysis on possible outliers is performed, using tools to detect both shape and magnitude outliers. Regarding the magnitude, no outlying curves are found while for the shape there are several curves with a specific behaviour. In particular, two different types of behaviour are identified.

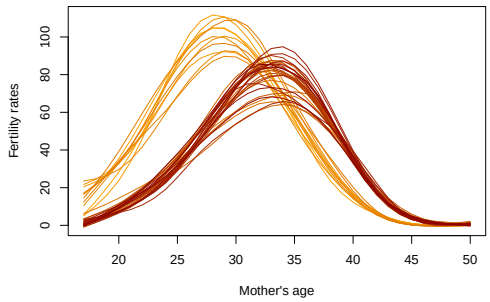

Figure 6: Outliers, colors different according to the different type

The lighter ones are those with a very early peak which is even higher, the darker ones instead have a delayed peak which results in being lower. While the second type seems to be along all Italy, the first type of outliers appears in the last years of the analysis, specifically in the southern regions.
This different feature of the two types of outliers will eventually result to be very interesting, especially the fact that the lower peak with smaller values belong to the southern regions, being that this behaviour will be the critical factor for the southern regions, deeply analyzed in the next sections.

4. Non-parametric inference

The first performed tests are made to roughly assess differences in the different years and geographic regions, in order to search for some particular features to describe the variability.

Province	Year
Agrigento	2002
Caltanissetta	2002
Caserta	2002
Enna	2002
Napoli	2002
Palermo	2002
Ragusa	2002
Siracusa	2002
Enna	2003
Caltanissetta	2004
Crotone	2004

Table 1: Outlier with an early peak

Testing for the differences in the functional data, following Pini et al. [2019] functional ANOVA permutation tests are performed to determine the significance of various factors. The first to be testes is the geographic factor for the curves, which results in a highly rejected hypothesis, with even the p-value function which is a flat line in zero, namely hypothesis rejected in all the points. The L^2 norms of the medians for the three groups are reported in Figure 7 to detect which of the regions has the highest curve. Clearly the Northern region has the highest curves, followed by the Central region and lastly the Southern. The same approach is performed looking at the differences in the years, and again the hypothesis of non-significance of the factor *year* is rejected, and moreover rejected in all the domain. In Figure 8 there are the L^2 norms for each year.

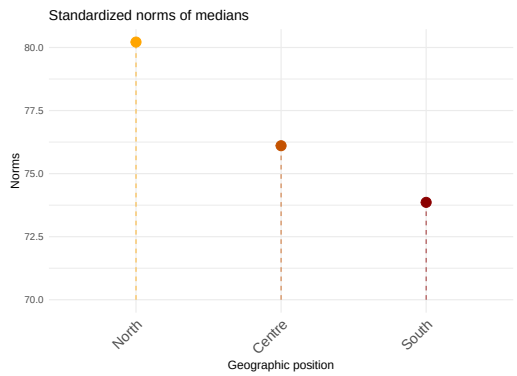

Figure 7: L^2 norm of the medians grouped for geographic region

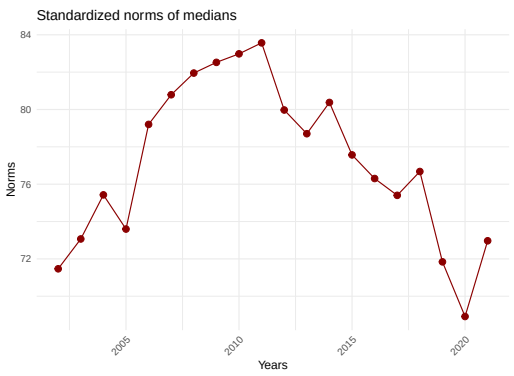

Figure 8: L^2 norm of the medians grouped for year

Here a first interesting result is shown: the progression in the norms does not follow the trend in the total newborns. As it was said in Section 3, the total number of newborns is not so informative as a statistics, while thanks to the fertility rates it is evident that the trend is not constantly decreasing, instead the higher value of this statistics was reached in 2011, and from there it plummeted significantly, a fact that could be explained by the economic crisis with its extended effects. Searching for the main measures made by the politics with respect to the family legislation, it is possible to divide this temporal interval in three: while in the first part a big amount of resources was invested in facilities for the new families, in the years from 2007 to 2013 no action was taken, obviously due to the economic crisis. This reflects the observed curve, in fact after a period where many aids were given to young families, the curves of the fertility rates rocketed, and after a few years without any investments on them the curves immediately fell.

Because of this argument, and to make the analysis more *easy-to-read*, a division for the years is made according to the following three periods[2]:

- From 2002 to 2005;
- From 2006 to 2013;
- From 2014 to 2018.

4.1. Quantities of interest

To address all the variability of the curves through functional analysis is very complex both in terms of tests and interpretation, and because of this a *quantity of interest* research is made, pointing out some of the main features of the functional data and analyzing them in a multivariate way. The most important quantities of interest in these curves are obviously their maximum abscissa, being the most common year for women to have a child, and their maximum ordinate, giving an overall measure of how much high is the curve and of the shifting toward the right already mentioned in Section 1.1. A new multivariate dataset is created with both the ordinate and the abscissa of the maximum of each province for each year. For consistency with the above results, the first step is to perform some multivariate ANOVA that result in rejecting the null hypothesis, both for the difference in the geographic region and for the difference in the years. A comment should be done for the previous division in three groups of years, in fact despite being rejected for the multivariate dataset, regarding the value of the maximum the distribution inside the group of years do not depend on the years for almost all the regions. As a result, the maxima of these curves can be taken to be identically distributed inside the groups, making this analysis even more significant for this quantity. Moreover, testing for differences of the three groups of years of the analysis, the null hypothesis is rejected, indicating differences in the maxima for the three different periods.
The first thing to note is that these two quantities are overall negatively correlated, namely the more early is the maximum in the curve the more it is high, which is an intuitive fact, the younger is the age with the most number of pregnant women, the more these women will be. In order to quantify this negative correlation, a linear regression is performed and the coefficient for the regression is taken as test statistics.

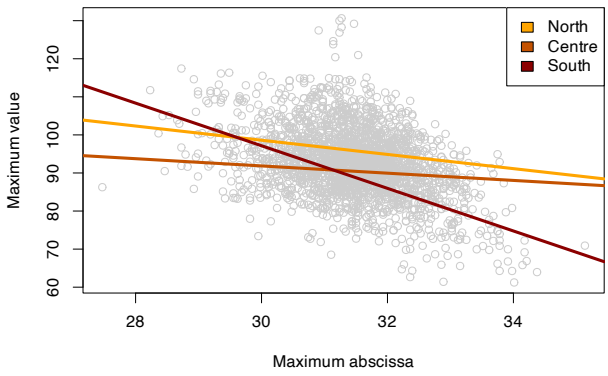

Figure 9: Maxima and the regression lines for area

[2] A sort of elbow's analysis in a qualitative way is used together with the different legislative periods, to highlight the difference in the norms.

In Figure 9, there is a graphical representation of the dataset containing the maxima and the three different lines of the models considering the three different geographical macro-regions. The overall lines are all with negative slope.
Moreover, this coefficient is evaluated for the three different macro-regions and for the three different periods of the analysis, and a confidence interval using the reversed percentile interval (95% confidence interval) of the bootstrapped[3] distribution is evaluated.

Zone	Lower	Estimate	Upper
North 2002-2005	-5.8557821	-3.88224796	-1.7763718
North 2006-2011	-3.1574974	-1.93358540	-0.6816319
North 2012-2018	-3.8004178	-2.42257909	-1.1835017
Center 2002-2005	-1.4380307	0.34327881	2.1009551
Center 2006-2011	-0.7211677	0.65987097	2.0863432
Center 2012-2018	-1.1953645	0.02356706	1.2435242
South 2002-2005	-8.2791845	-7.09528693	-5.9710253
South 2002-2005	-6.8479150	-6.15205058	-5.4461859
South 2002-2005	-7.5586217	-6.67412364	-5.7289206

Table 2: Table of the 95% confidence intervals

There is clear evidence to state that almost all these confidence intervals are below zero, and that the tendency for the maximum ordinate decreases going on with age. In particular, the Southern part of Italy shows this behaviour in a much more pronounced way with respect to the other two areas, indicating that in older ages it is much more difficult for people in South Italy to have a baby. The Central part of Italy has this particular behaviour where the three coefficients are not significantly different from zero, despite the overall one is negative because of the different intercept.

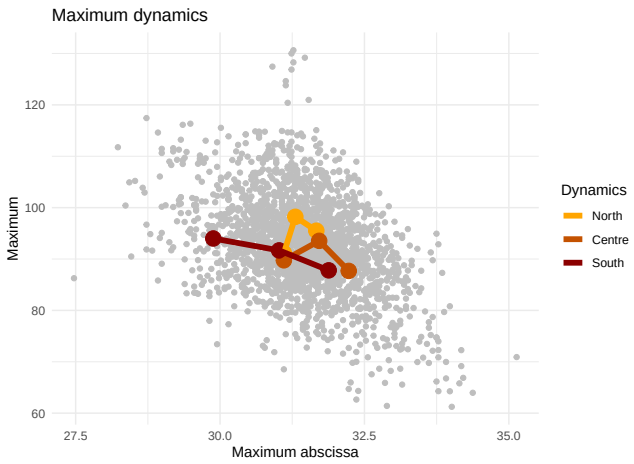

Figure 10: Dynamics of maxima for area and period

As a final remark on these value of the maxima, the dynamics in the three different periods for North, Centre and South Italy are reported in Figure 10. It is clear that the South area has a completely different dynamics, which is constantly decreasing, while both Centre and North are firstly jumping up and then declining. Moreover, while in these last two dynamics it is possible to see that the advancing in the abscissa of the maximum does not imply a huge decreasing of the ordinate of this maximum, as highlighted in Table 2. Indeed, considering only these macro-periods the value of the maximum increases with the abscissa of it. It means that in the South having a family at a older age is much more difficult, while in the other parts of Italy

[3]The resampling scheme is to permute with replacement the residuals of the regression and re-fit the model.

it is easier to make a family even one or two years after, with respect to 20 years ago. This dynamics is very interesting and meaningfully, indeed understanding where is the maximum of the fertility curve and how much is this maximum gives a direct interpretation of the behaviour of the population, that differs in the different part of the Country. The increasing behaviour in the maximum abscissa is similar in all the regions, despite having differences in the numbers, and reflects the need for a longer and longer time to construct a family and having a baby, which is typical of our days, as highlighted in many other social studies. The main difference, and the most important is the magnitude of this maximum, in fact while in North and Centre Italy for the last period it is the same as the first period, in south Italy the level is way below, which is why the rates for the south are the lowest, as shown in Figure 7, pointing out the fact that going on with age in the South of Italy having a family becomes more and more complicated.
Recalling table 1 it is possible to compare this dynamics and the table, in fact the curves which result as outlier with a very early peak are all in the southern regions, and the plot of the dynamics highlights this fact, being the South in 2002-2005 the point to the extreme left in the abscissa of the maxima.

4.2. First derivative

The first issue with this kind of analysis is how to compute such a quantity, in fact two are the main drivers for evaluating the derivative of a functional datum: the first is to differentiate the spline basis on which the raw data are fitted, the second is to make subsequent differences in the raw data to have the approximation of the derivative. Despite the latter implies less manipulation on the data, the aim will be to evaluate the maxima of these derivatives, and having a very spiky curve could lead to poor estimates. As reported in the Figure 11, the one evaluated with the derivatives on the basis leads to a smoother curve which was chosen for the following analysis.

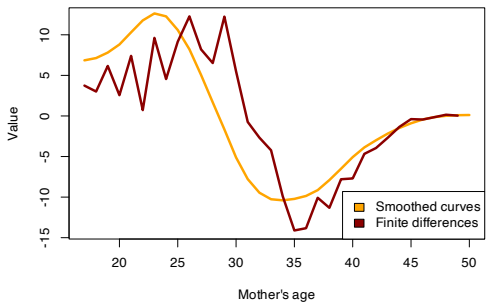

Figure 11: Two types of derivatives

As a first approach, the study on the outliers is developed, both for shape and magnitude. No magnitude outliers are present, while two types of shape outlier are found (Figure 12), in this case involving only Southern regions with two different dynamics: low peak delayed and high peak early in the domain, that will again be the most interesting dynamics' feature in the study. Again, even in this context the two types of outliers reflect what the conclusions on the dynamics will be. It is important to note what the shift towards the right or the left of these curves means, in fact the first derivative of the fertility rates represent a sort of velocity on the rate's curve. In less mathematical terms, the maximum represents the starting point of the trend inversion from an accelerating curve to a decelerating curve, in other words the maximum slope point of the curve, and the magnitude of the maximum is the value of this slope in that

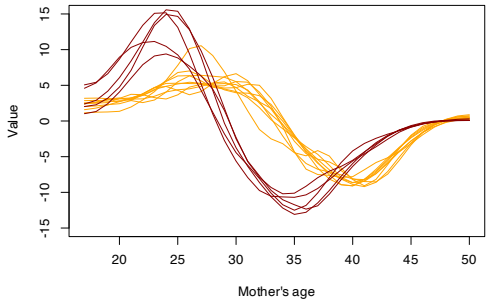

Figure 12: Shape outliers in the first derivative

point. The influence on these curves of the two factors geographic macro-region and years is again tested with the permutation ANOVA, and rejected in both cases. The sequence of the norms both divided by geographic region and years is not so different from the one in Section 4.1. Again with these curve the analysis on the maxima is performed, pointing out the main characteristics of the first derivatives, which reflect the latter dynamic in the original curve. Moreover, this point indicates the maximum slope that is present on the raw data, indicating how fast women are starting to having children going on with age.

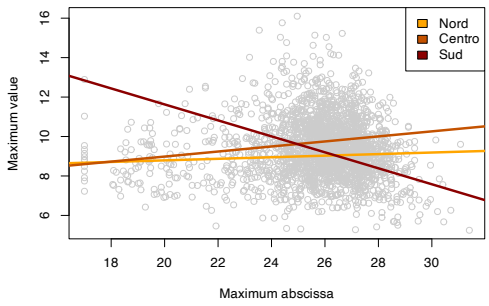

Figure 13: Maxima of the first derivative and regression lines for area

In this case the abscissa and the value of the maximum are not so clearly negative correlated, in fact in Figure 13 it is clear that the slope for the Southern region is negative while the others are not significant. Analyzing again the differences in the *class of years* defined in Section 4 and the difference between these are significant. Within these groups the permutation ANOVA test with years as factor results in acceptance of the null hypothesis for almost all the combination, namely even when talking about the first derivative the observation in these groups can be considered as identically distributed, while the difference between these groups remain significant.
As done in Section 4.1 a bootstrapped confidence interval for the coefficient of the linear regression is evaluated and reported in Table 3.

In the South regions having a delayed peak in the first derivative means having a maximum inclination of the fertility curve which is lower, while in the other regions there is not a clear pattern. In almost all the intervals the value 0 is contained, meaning that there is not clear evidence to state if it is positive or negative. The dynamics in the group of years points out the same conclusions already stated.

Zone	Lower	Estimate	Upper
North 2002-2005	0.07	0.17	0.27
North 2006-2013	0.03	0.08	0.13
North 2014-2018	-0.03	0.09	0.21
Center 2002-2005	-0.10	0.06	0.23
Center 2006-2013	0.00	0.07	0.15
Center 2014-2018	-0.05	0.04	0.14
South 2002-2005	-0.93	-0.67	-0.42
South 2006-2013	-0.50	-0.36	-0.20
South 2014-2018	-0.22	-0.06	0.08

Table 3: Table of the 95% Confidence intervals

4.2.1. Minima

A similar kind of reasoning can be made for the minimum of the first derivative, indeed the minimum represents the point of maximum decreasing of the fertility curves, pointing out how much the curve decrease when it is on its maximum decreasing point. In this case all the intervals are significantly above zero, indicating that the later in time the maximum decreasing is, the more it will be steep, as one can imagine. In particular, the dynamics of the medians of the minima is displayed in Figure 14 and clearly it is possible to see that this dynamics is common to all of our observation. Nowadays the minima tend to be later in time, as seen in many previous sections this is a common trait of many characteristics in these rates. In particular for the minima they have first a decreasing trend and then an increasing one, meaning that the steepness increases as the age increases, as one might expect. In this feature, in the

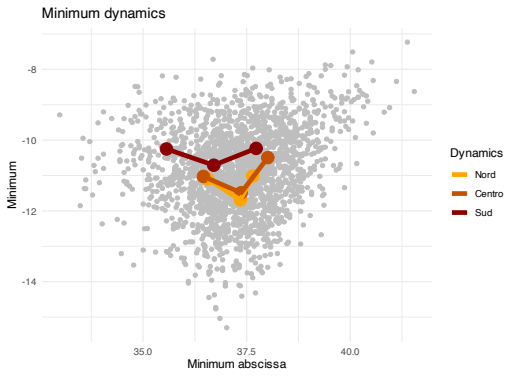

Figure 14: Dynamics of minima of the second derivative for area and period

last period, the central region is near the southern region while the northern one is away, a dynamic in counter-tendency with the others presented until now. This means that nowadays having a child in later ages is much more easier in northern Italy, letting the decrease be less steep, and it is in this characteristics that the northern region gain the advantage that let their curves being higher than the others.

4.3. Second derivative

For the second derivative, as anticipated in Section 2.4.1, the smoothing with 5 degree splines is needed in order to restore the value of the maximum of these curves, and for consistency the whole curves have not been analyzed and the focus is in studying the maximum of the curves. The interpretation for the peak of the second derivative can be easily developed thinking about it as the maximum of the acceleration of the fertility curves, namely the years for which women

start significantly to have children, which indeed is a very important feature of the curve and of the project in general. Starting and having children is a very thoughtful decision, hence understanding how these maxima have changed during time can be highly informative on how the behaviour and custom of young people have changed in these years.

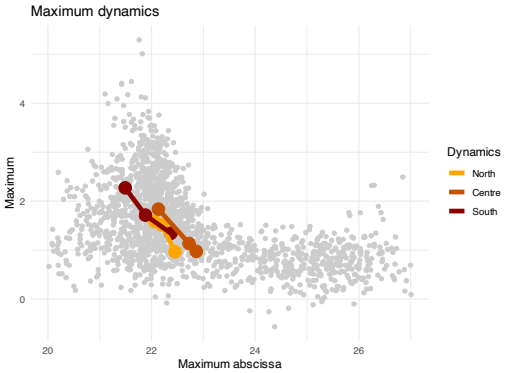

Figure 15: Dynamics of maxima of the second derivative for area and period

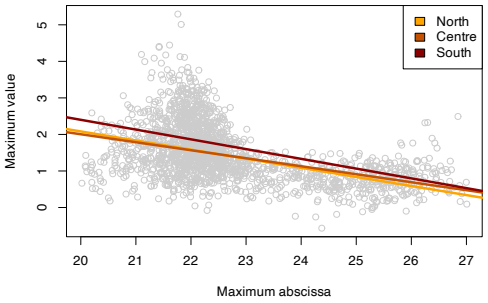

Figure 16: Maxima of the second derivative and regression lines for area

North, Centre and South have a similar trend, where higher values of the abscissa are associated with lower values of the maxima. Figure 16 shows that the slopes of the three lines are all negative and similar to each other, in particular North and South lines seem to be almost parallel. It means that as age in which women start to have children increases, the magnitude of this growth decreases and it happens in the same way for North and South Italy, even though the intercept of the last one is higher. In Figure 15, the dynamics of the maxima of the second derivative for the three different periods considered is evidently decreasing and moving forward higher value of the abscissa. It represents how over the three periods of years, namely 2002-2005, 2006-2011 and 2012-2018, the moment in which women start to have a child shifts toward higher ages meanwhile acceleration decreases. This is an expected behaviour, in fact for young people making their own family is a more and more difficult decision, and this kind of decision is delayed, moreover the intensity of this growth is naturally decreasing as the percentage of women making their own family as soon as it is possible is lowering in the years.

5. Non-parametric regression

Having observed that the quantities of the maximum ordinate and abscissa of the fertility rate functions and their derivatives yielded interesting results in Section 4, non-parametric regression will focus on these quantities, in order to understand which variables actually affect them. From now on in this analysis, the maximum abscissa of the rates will be related to as `MAxDomain`, while the maximum ordinate as `Max`. Moreover, due to the fact that only few data for the covariates are available for each province, the regression will be performed considering the regional quantities of interest, obtained by computing the trimmed mean with $\alpha = 0.1$ of the provincial values. For each region and year, the following features are considered:

- `Emigrations`: proportion of emigrants from that region over the resident population;
- `Immigrations`: proportion of immigrants in that region over the resident population;
- `Dropouts`: percentage of young people between 18 and 24 years old who dropped out of education;
- `Women.enrolled`: proportion of women enrolled in university over the total number of enrolled individuals;
- `Employment.rate`: measure of the incidence of employed people between 15 and 64 years old;

- `Unemployment.rate`: measure of the incidence of unemployed people between 15 and 64 years old;
- `Abortions.2529`: proportion of women aged 25-29 who have had an abortion out of the total number of women who have had an abortion;
- `Abortions.3034`: proportion of women aged 30-34 who have had an abortion out of the total number of women who have had an abortion.

Note that some of these variables do not cover the entire period of 2002-2021, for instance `Women.enrolled` is available from 2008, `Dropouts` from 2004 to 2020, `Abortions.2529` and `Abortions.3034` from 2010 to 2021. Hence, different models have been defined based on the availability of data. The initial model is a generalized additive model with cubic spline smoothing, with the following mathematical expression

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon_i, \quad i = 1, \dots, n,$$

where Y_i is the response variable, X_j is the j^{th} covariate and ε_i is the error term.
The procedure used to obtain the best models is as follows: at first, all covariates have been considered and those resulting not significant have been sequentially discarded. Considering that the pairs of variables `Emigrations`,`Immigrations` and `Employment.rate`, `Unemployment.rate` appear very correlated, as displayed in (Figures 17, 18), hence, even when both highly significant, only the best one has been kept.

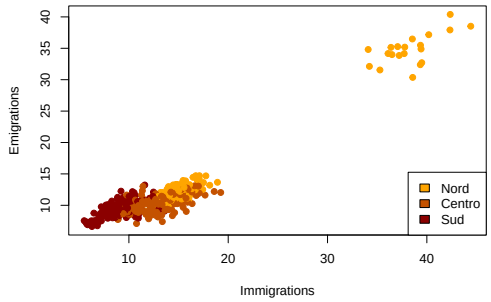

Figure 17: Scatterplot of Immigrations and Emigrations

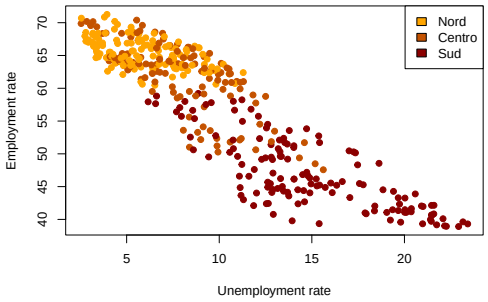

Figure 18: Scatterplot of Unemployment.rate and Employment.rate

5.1. Base data

For both the models with `MaxDomain` and `Max` as response variables, the most valuable models consider as covariates:

- `Immigrations`;
- `Employment.rate`;
- `Women.enrolled`.

It's noteworthy that both variables associated to abortion are not significant, hence they do not influence neither the maximum value of the fertility rate nor the women's age at which this maximum is reached, proving the fact that limiting the right of abortion is not an effective method to increase the fertility rate in its maximum that is a general measure of highness of the curves.

5.1.1. Max Domain model

The model with the covariates reported above achieves an adjusted R-squared of 0.586. Indeed, by linearizing the `Immigration` covariate, the model reaches an $R^2_{adj} = 0.573$ and has the following form

$$MaxD = \beta_0 + \beta_1 * Immigrations + f_1(Employment.rate) + f_2(Women.enrolled) + \varepsilon. \quad (1)$$

The reduced model is chosen, due to the better interpretability of the linear coefficient. By a visual exploratory analysis of these variables, some outliers related to `Immigration` are evident in Figure 19. Through a Minimum Covariance Determinant approach with $\alpha = 0.95$, those outliers are found and they are all related to the observations of *Valle d'Aosta* region. In fact, in the last 20 years, this region has experienced a massive immigration phenomenon as it is reported in VDA [2022].

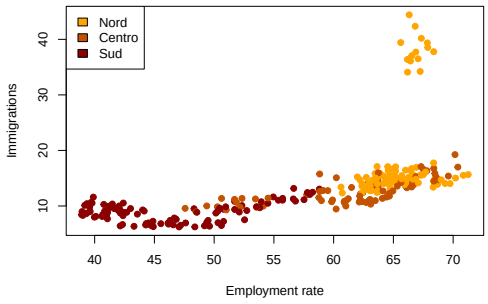

Figure 19: Scatterplot of Employment.rate and Immigrations

These values would have affected the fit of the model resulting in a poor estimates not really looking at the bulk of the data. With the spirit of the least trimmed of squares, the fit was performed only looking at this core of the data, namely on the data that the MCD fit indicates as not outliers, and this model reaches an $R^2_{adj} = 0.556$.
After a check on the distribution of the residual, as done in Figure 20, that validate the Gaussian hypothesis, the p-values of the covariates are reported in Table 4, resulting all significant.

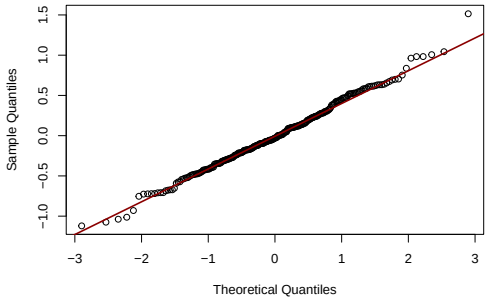

Figure 20: QQ-norm of the residuals of MaxDomain model

The coefficient for the linear part is $\beta_1 = -0.09436$: the lower the immigration within a region, the higher is the age for which the maximum of the fertility rate is observed. Regarding the effects of the other two covariates, plots are made isolating the effects of each covariate by fixing the others to their median value. Figure 21 shows that with high e low values of employment

Covariate	P-value
Immigrations	1.4e-05
s(Employment.rate)	$<$ 2e-16
s(Women.enrolled)	0.00176

Table 4: Table of the p-values of MaxDomain model

rates the ages of maximum fertility rates are lower, while for mid values of employment rates the ages are higher. Instead, Figure 22 explains that there is a general increasing trend, with the age of the maximum of the fertility rates increasing as the proportion of women enrolled at university is higher, which may seem a counter-intuitive result, but it is deeply related with the fact that when there is no limitation in the academic studies for women a more inclusive and better environment is built, which leads to an increasing fertility rate.

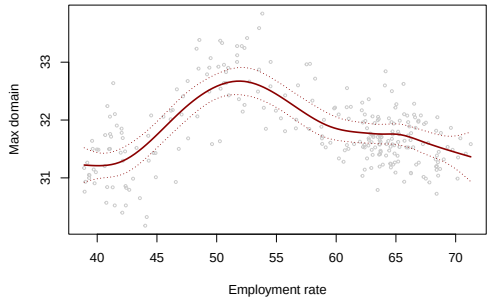

Figure 21: Effect of Employment.rate in MaxDomain model

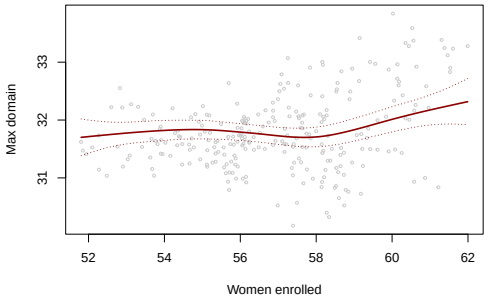

Figure 22: Effect of Women.enrolled in MaxDomain model

5.1.2. Max model

As it was said in Section 5.1, the same covariates as in the previous model turn out to be significant, with resulting adjusted R-squared of the model equal to 0.62. Also in this case, we consider the effect of the `Immigrations` covariate to be linear, reaching an $R^2_{adj} = 0.582$. The model has the following expression

$$Max = \beta_0 + \beta_1 * Immigrations + f_1(Employment.rate) + f_2(Women.enrolled) + \varepsilon. \quad (2)$$

For the same argument as in section 5.1.1, the reduced model is selected. The same outliers found before are removed following the same approach. The model above reaches an adjusted R-squared of $R^2_{adj} = 0.56$. After the check on the residuals in figure 23, table 5 contains the p-values for each covariate, resulting all highly significant.

Covariate	P-value
Immigrations	0.0225
s(Employment.rate)	$<$ 2e-16
s(Women.enrolled)	2.07e-06

Table 5: Table of the p-values of Max model

In this case, the coefficient related to `Immigrations` is $\beta_1 = 0.6791$, meaning that the higher immigration in a region, the higher is the value of the maximum fertility rate, which is the same conclusion obtained in the section 4, namely the earlier are the women making children

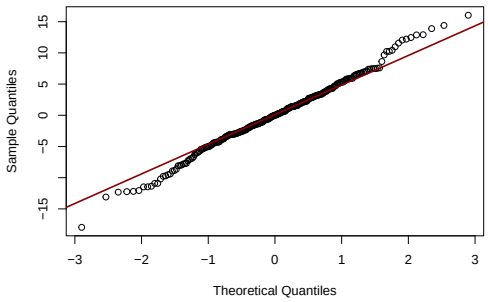

Figure 23: QQ-norm of the residuals of Max model

the more this peak will be high. As to the effects of the nonlinear part, the plots are obtained with the same techniques of fixing the other covariates. This analogy with the non-parametric inference is repeated, in fact in figure 24 it is clear how the minimum of `Employment.rate` match the same values in which in the previous model (figure 21) the maximum for age is achieved. Instead, for lower and higher values of `Employment.rate`, the maximum value of fertility rate is on the same range and correspond to lower numbers in terms of the ages. So, overall it is possible to say that the employment rate is a factor that pushes the curves to the right until a certain rate, after which the `MaxDomain` returns towards the left. Roughly speaking, surpassing this middle range of employment rate, the fertility rate increases as if women were less employed, highlighting again the fact that a prosperous society is the perfect environment for making a family. figure 25 shows an overall effect of `Women.enrolled` of the same type as the one in the previous model, with a more pronounced decreasing in the maximum as the number increases.

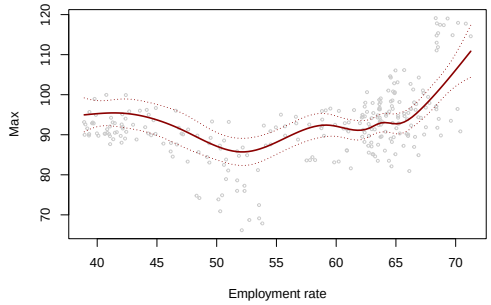

Figure 24: Effect of Employment.rate in Max model

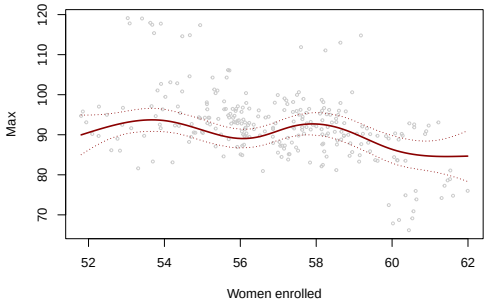

Figure 25: Effect of Women.enrolled in Max model

5.2. Regression on second derivatives for Max

When considering the second derivatives of data the most valuable model with `Max` as response variables considers as covariates:

- `Immigrations`;
- `Employment.rate`;
- `Women.enrolled`.

These are the same as the one selected for the models of the original data. Observations related to Valle d'Aosta are not taken into account as those are still outliers for such covariates, following the same reasoning above. The `Immigrations` covariate has been set to have a linear effect, in order to enhance the interpretation of its effect. Such model achieves an adjusted R-squared of 0.479, with the following expression:

$$Max = \beta_0 + \beta_1 * Immigrations + f_1(Employment.rate) + f_2(Women.enrolled) + \varepsilon. \quad (3)$$

After the check on the residual with figure 26, the p-values of the covariates are reported in table 6, resulting all highly significant.

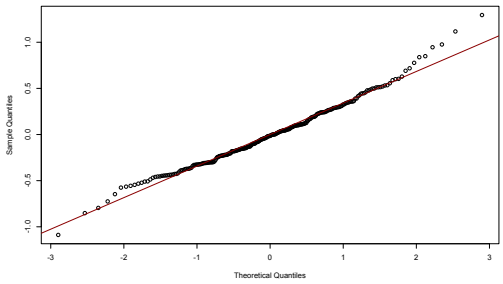

Figure 26: QQ-norm of the residuals of Max model for second derivatives

Covariate	P-value
Immigrations	0.0148
s(Employment.rate)	$<$ 2e-16
s(Women.enrolled)	$<$ 2e-16

Table 6: Table of the p-values of Max model for second derivatives

The coefficient for the linear part is $\beta_1 = -0.04389$: the lower the immigration within a region, the higher the maximum of the second derivatives of fertility rate is observed, which is in contrast with what observed for the maximum of the curves and for the abscissa in which the maxmum is reached, indeed immigration increases the starting of the ascent of the curve and decreases the maximum of it. As to the effects of the other two covariates, Figure 27 shows a decreasing trend with a peak at the right boundary; however, it may lack informativeness near the boundaries so no inference is made on that part of the graph. As the `Employment Rate` increases, the maximum tends to decrease. Instead, Figure 28 illustrates a minimum around 56 and a maximum around 58. The increasing trend between these values is notable, being significant as it is far from the boundaries showing how the higher is the value `Women.enrolled` the higher the `Max`, again coherently with the previous analysis.

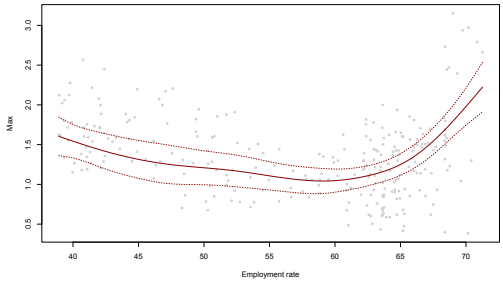

Figure 27: Effect of Employment.rate in Max model for second derivatives

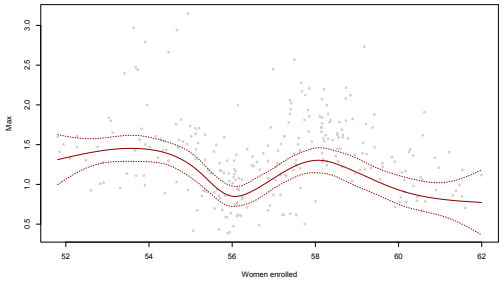

Figure 28: Effect of Women.enrolled in Max model for second derivatives

6. Conformal prediction

Assuming exchangeability for the covariates the approach of conformal prediction is employed to assess the reliability of models in Sections 1, 2 and 3. This method offers a more comprehensive evaluation of model predictions by providing confidence estimates for each prediction made. In order to do so, a full conformal approach is used, considering as Non Conformity Measure (NCM) the residuals of the models. The conformal prediction intervals are computed at 95% confidence level. Given that in Section 4 the inference is conducted up to 2018, the interest is focused on the prediction for the maximum value of the fertility rate curves and the corresponding women age in 2019.

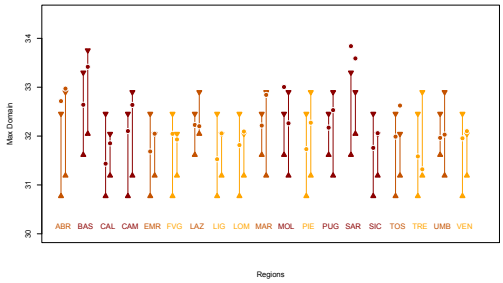

Figure 29: CPI and true value for Max-Domain in 2019 and 2021

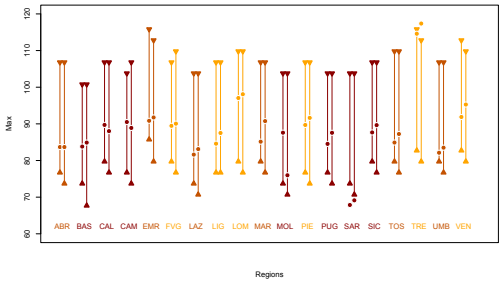

Figure 30: CPI and true value for Max in 2019 and 2021

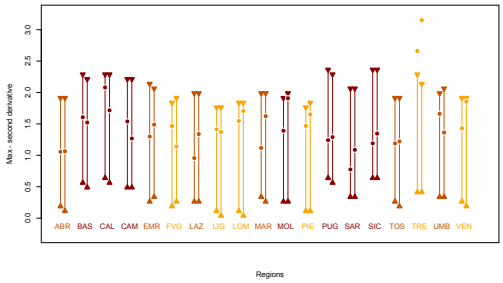

Figure 31: CPI and true value for Max in 2019 and 2021

By computing the conformal prediction intervals for each region for the ages of the mother corresponding to the maximum of the fertility rates, the model is able to correctly predict values for almost all regions, specifically the true value of 16 out of 19 regions are contained in the intervals. Remind that Valle d'Aosta is not included in the model as it is an outlier for

the covariates. The values that fall out are those of Abruzzo, Molise and Sardegna regions, for which the true values are higher than the predicted. In the case of the maximum fertility rate, all regional values are contained in the prediction intervals, apart from the one of Sardegna, that is lower that the predicted. Model (2) is capable of predicting 94.7% of values for 2019. Finally, the model for the second derivative consistently generates confidence intervals that encompass the actual data for each region. Also in this case, 18 regional values are correctly predicted as they are contained in the prediction intervals. The only values missed is the one of Trentino Alto Adige region, whose true value is higher than the one predicted. The results can be visualized in the left bar of each abscissa on the Figures 29, 30 and 31. Note that the colors are associated with the area each region belongs to, with yellow associated to North, orange to Centre and red to South.
Moreover, since data are available up to 2021 and considering that this is the first year in which the effects of Covid-19 can be seen, a comparison between the predictions of 2021 and 2019 is presented in Figures 29, 30 and 31. The aim is to qualitatively check whether models presented in Sections 1, 2 and 3 are able to capture the behaviour of the maximum values also in 2021. The predictions are pretty good for all three models and the one for the `Max` value of the fertility rates actually performs better. It can be concluded that the coviarates selected in Section 2.4 are able to well explain the phenomenon.

7. Conclusions and possible developments

As said in section 1.1 the fertility rates are naturally shifting toward the older ages, as one might expect. In fact, nowadays families are typically created later in life as some results on section 4 highlight. Young people dedicate significantly more time studying and working, particularly in Italy, in order to attain a situation where creating a family is feasible. Considering this reality, it is not an inevitable outcome for birth rates to decrease, as observed in certain regions, notably during the period 2006-2013. Addressing this trend should be the primary objective of any policy regarding birth rates. The models presented in this work attempt to explain the anticipated changes and quantify the effects of debated topics such as abortion, employment rate, women's university enrollment rates, alongside the dynamics of emigration and immigration. Italy has already experienced conditions to maintain stable birth rates despite delayed fertility. The covariates elucidate these conditions: firstly, the employment rate must surpass the presented threshold of 53%; secondly, a highly debated topic like abortion is deemed inconsequential as it lacks significance when up to this context. All things considered, the findings of this work suggest that fostering a healthy and prosperous society leads to higher fertility rates. This should be the primary objective of policymakers.

References

Mur. URL https://www.mur.gov.it/it.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 1995. URL https://www.jstor.org/stable/2346101.

Joop de Beer. A new relational method for smoothing and projecting age-specific fertility rates: TOPALS. *Demographic Research*, 24(18):409–454, 2011. doi: 10.4054/DemRes.2011.24.18. URL https://www.demographic-research.org/volumes/vol24/18/.

Istituto Nazionale di Statistica. www.istat.it, 2023. URL https://www.istat.it/.

Tomasz Jurkiewicz. Multidimensional smoothing in tables of fertility rates. 2011. URL https://bibliotekanauki.pl/articles/658372.

Alessia Pini, Lorenzo Spreafico, Simone Vantini, and Alessandro Vietti. Multi-aspect local inference for functional data: Analysis of ultrasound tongue profiles. *Journal of Multivariate Analysis*, 170:162–185, March 2019. ISSN 0047-259X. doi: 10.1016/j.jmva.2018.11.006. URL http://dx.doi.org/10.1016/j.jmva.2018.11.006.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer New York, 2005. ISBN 9780387227511. doi: 10.1007/b98888. URL http://dx.doi.org/10.1007/b98888.

Regione VDA. Immigrazione in valle d'aosta. 2022. URL https://www.regione.vda.it/statistica/statistiche_per_argomento/immigrazione/default_i.aspx#:~:text=Gli%20ultimi%20dati%20di%20tipo,riferimento%20al%2031%2F12.