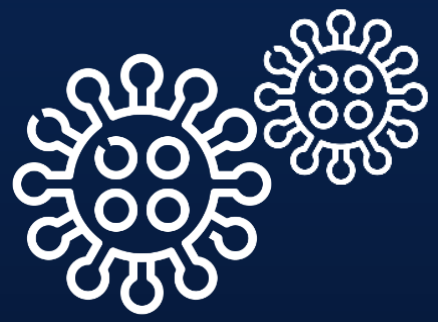
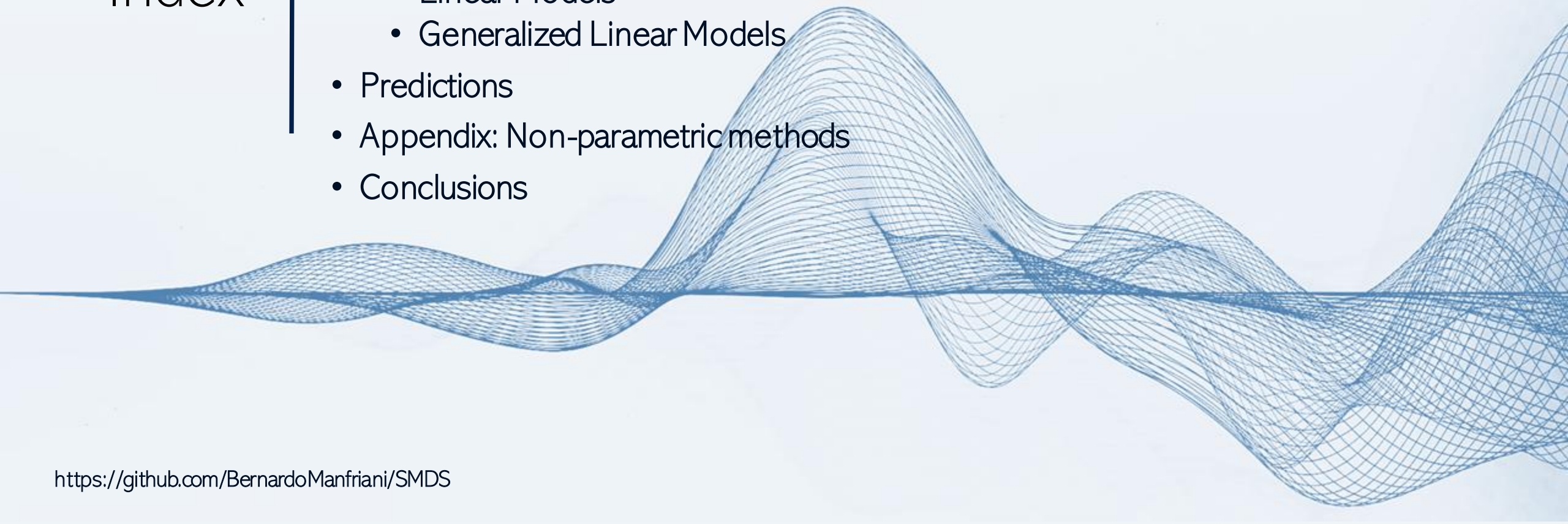
DATA SCIENCE &
SCIENTIFIC COMPUTING

Covid-19 case study



Erika Lena
Bernardo Manfriani
Francesco Ortu
Alessandro Serra

Index



https://github.com/BernardoManfriani/SMDS

Introduction

Introduction

The purpose of this project was to analyze the data for the Covid-19 spreading outbreak available on the official website of Protezione Civile and build up a statistical model for the number of new positives.

Problem specifications:

- Time period assigned: 1st October 2020 to 1st February 2021
- Region: Sicilia
- Response variable: nuovi positivi

Dataset

Dataset

- **data:** Date of notification
- **stato:** Country of reference
- **codice_regione:** Code of the Region (ISTAT 2019)
- **denominazione_regione:** Name of the Region
- **lat:** Latitude
- **long:** Longitude
- **ricoverati_con_sintomi:** Hospitalised patients with symptoms
- **terapia_intensiva:** Intensive Care
- **ingressi_terapia_intensiva:** Daily admissions to intensive care
- **totale_ospedalizzati:** Total hospitalised patients
- **isolamento_domiciliare:** Home confinement
- **totale_positivi:** Total amount of current positive cases
- **variazione_totale_positivi:** New amount of current positive cases
- **nuovi_positivi:** New amount of current positive cases (totale_casi current day - totale_casi previous day)
- **dimessi_guariti:** Recovered
- **deceduti:** Death (cumulated values)
- **totale_casi:** Total amount of positive cases
- **tamponi:** Tests performed
- **casi_testati:** Total number of people tested

variable	missing values
data	0
ricoverati con sintomi	0
terapia intensiva	0
totale ospedalizzati	0
isolamento domiciliare	0
totale positivi	0
variazione totale positivi	0
nuovi positivi	0
dimessi guariti	0
deceduti	0
totale casi	0
tamponi	0
casi testati	0
ingressi terapia intensiva	63

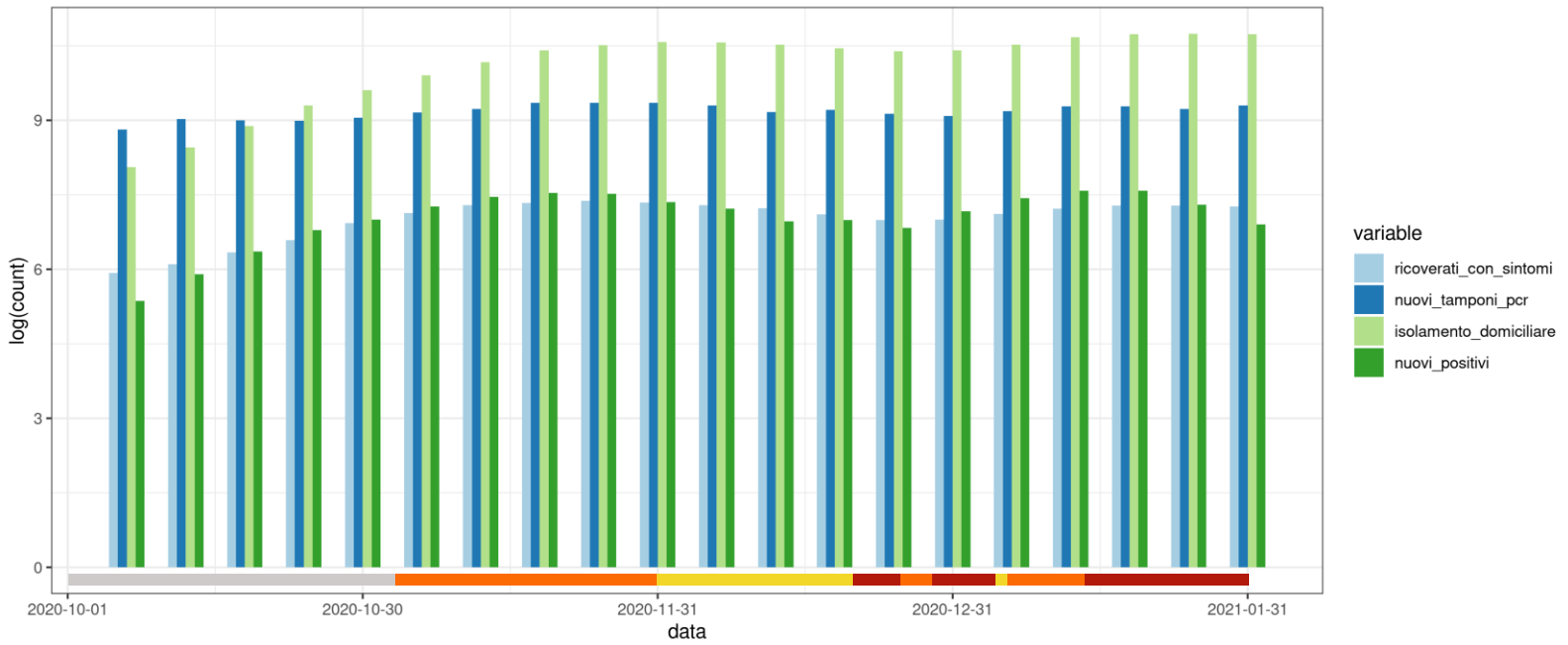
Quality of dataset

- A first phase was about cleaning the dataset to be used in model construction. For this purpose, all not useful and redundant variables have been removed.

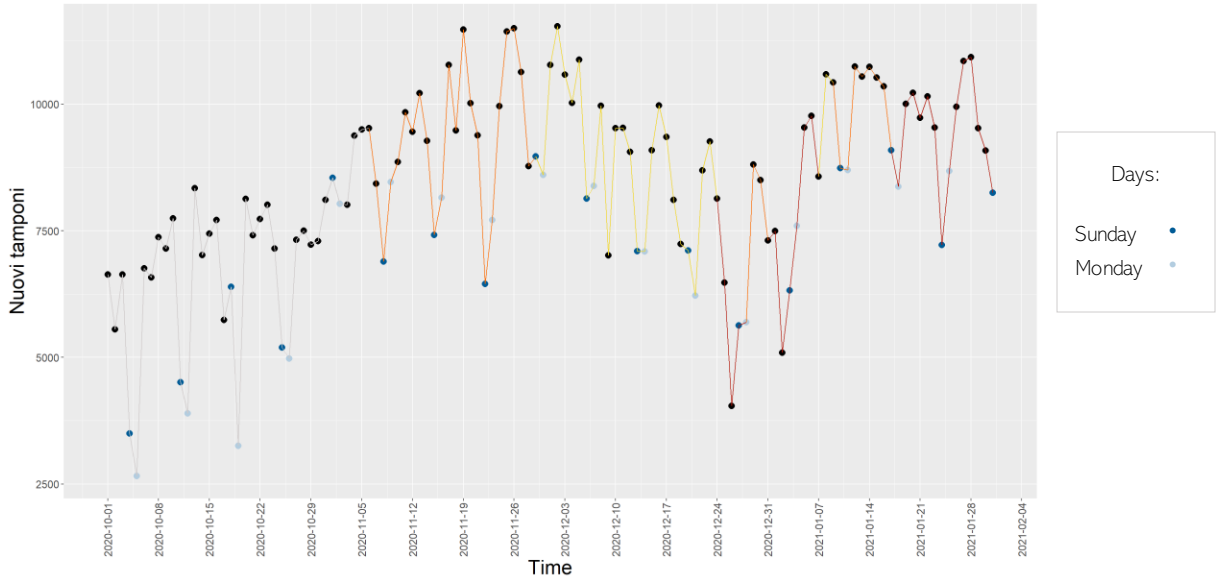
- All missing values are related to the admission to the ICU and for this reason that variable has been removed and it has been considered the number of ICU per day.

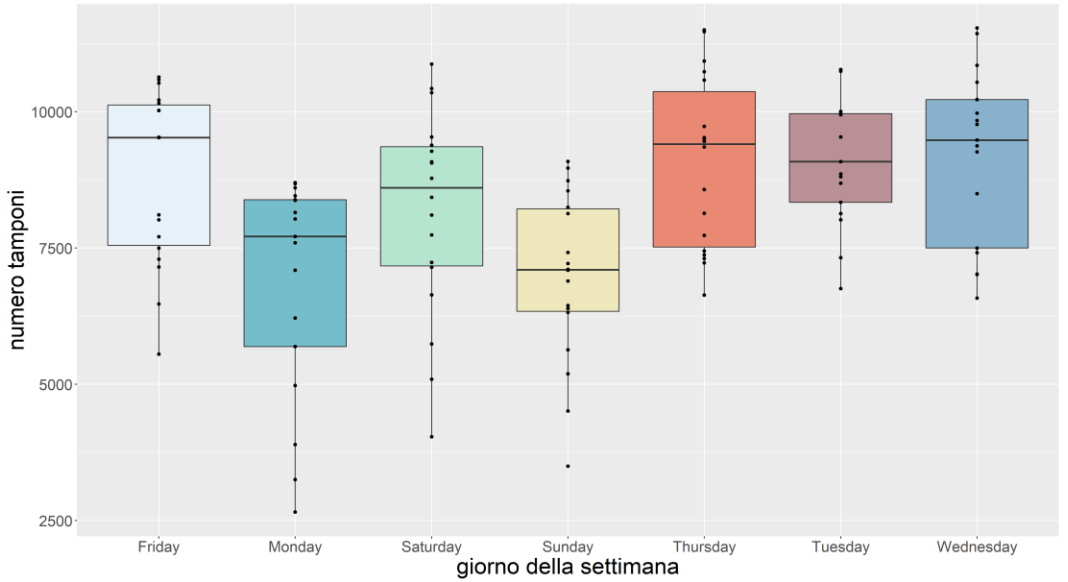
- From mid-January, Covid-19 tests are both antigenic and pcr swabs. Hence, we evaluated a new variable, `tamponi_test_molecolare`, from the subtraction of `tamponi` and `tamponi_test_antigenico_rapido`.

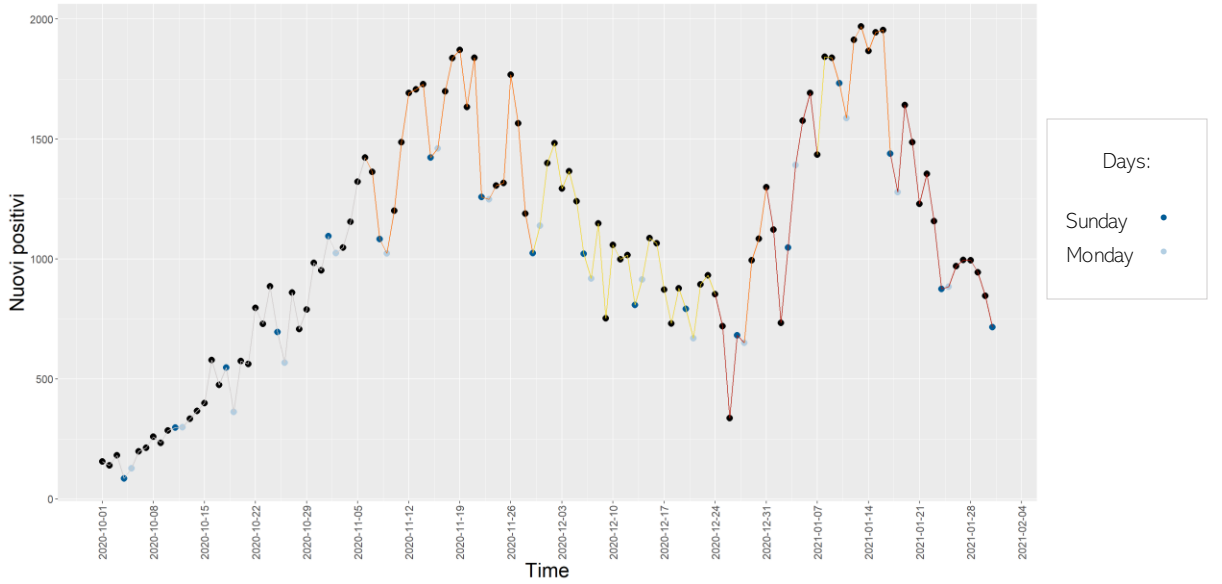
- We also added a new categorical variable, `colore`, to consider the restriction level.

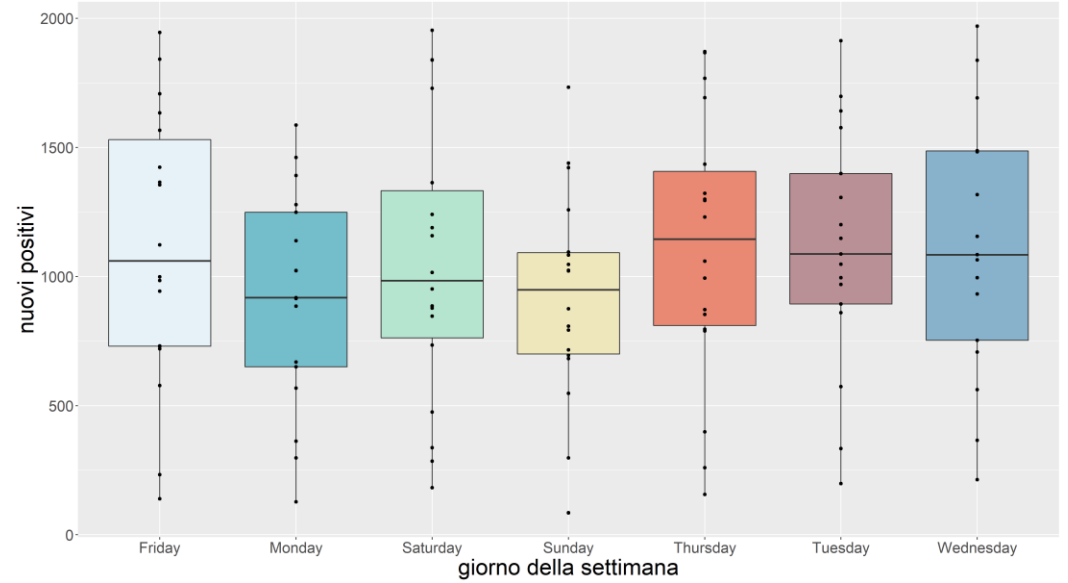


Quality of dataset



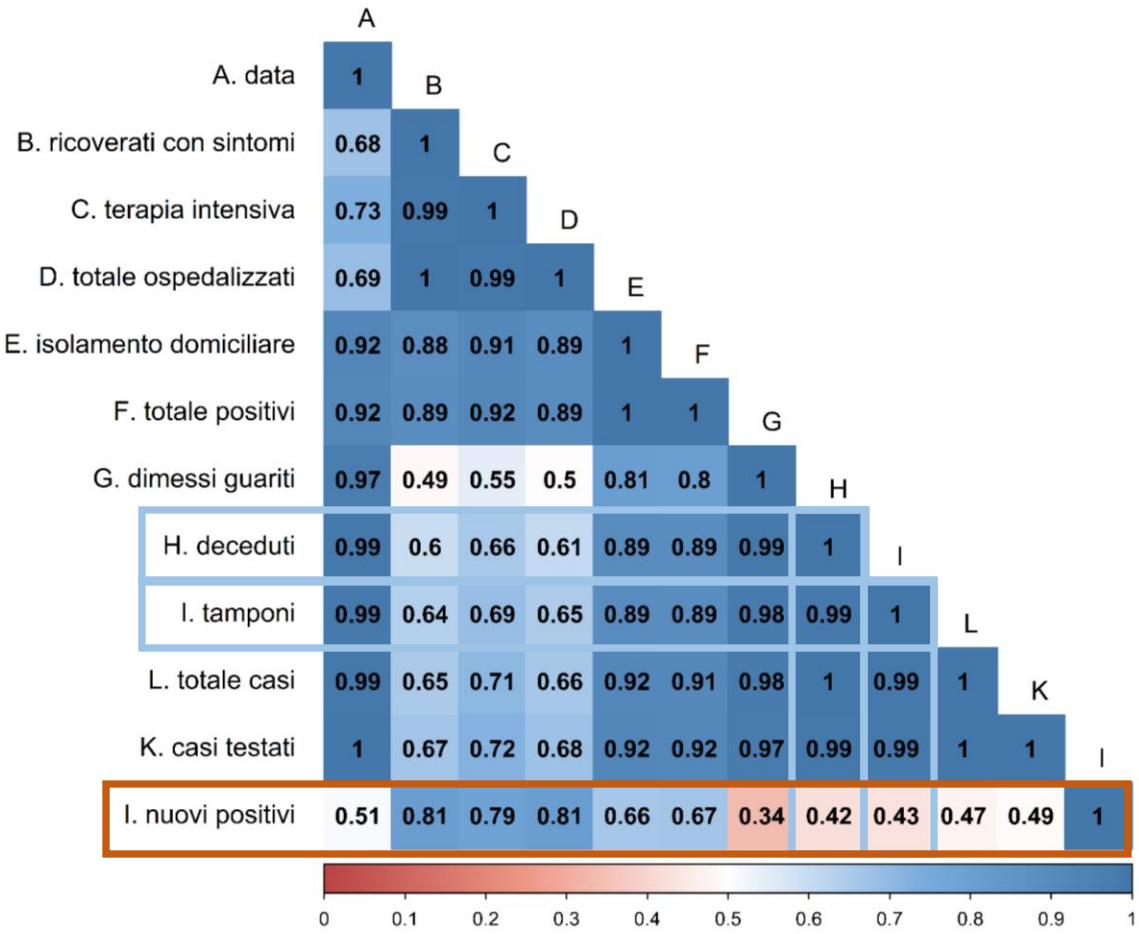






Data analysis

Data analysis



Many variables are highly related, though these relationships may be due to the dependence of the variables on each other.

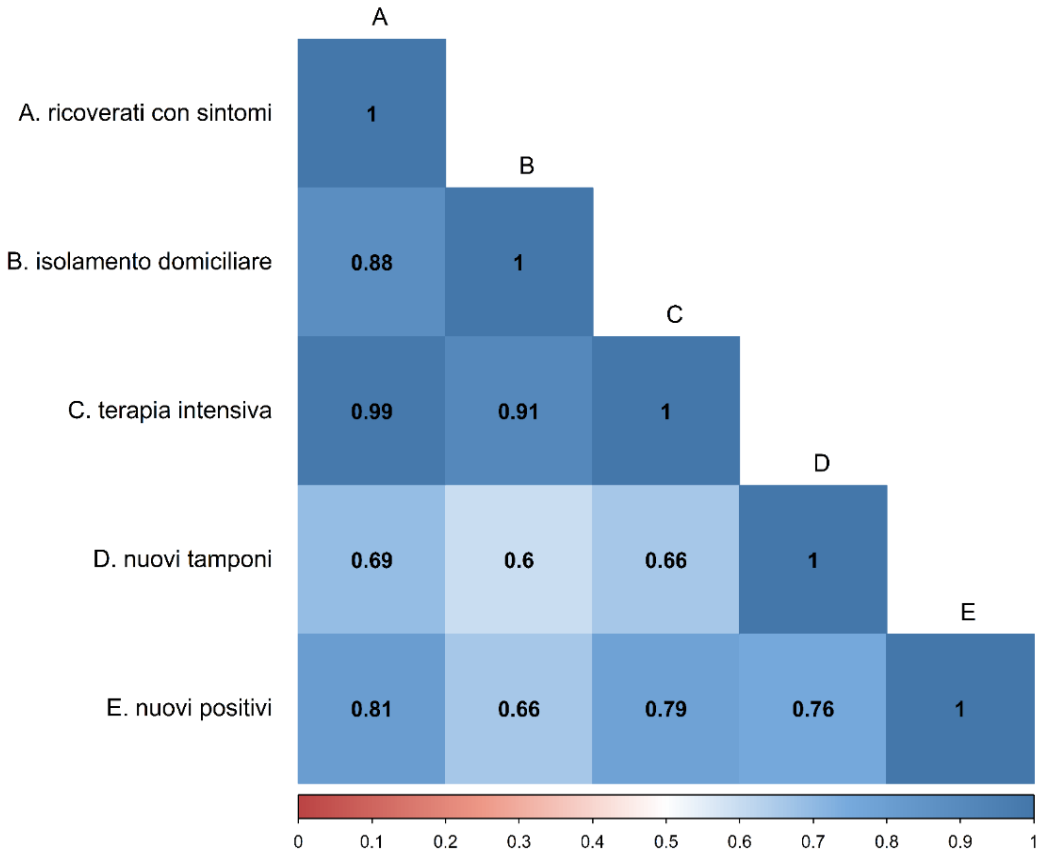
For other variables we observe a poor correlation while we would have expected something different, for example for the number of swabs carried out.

For this reason, new variables were introduced for those values which account for the total amount of cases since the beginning of the pandemic.

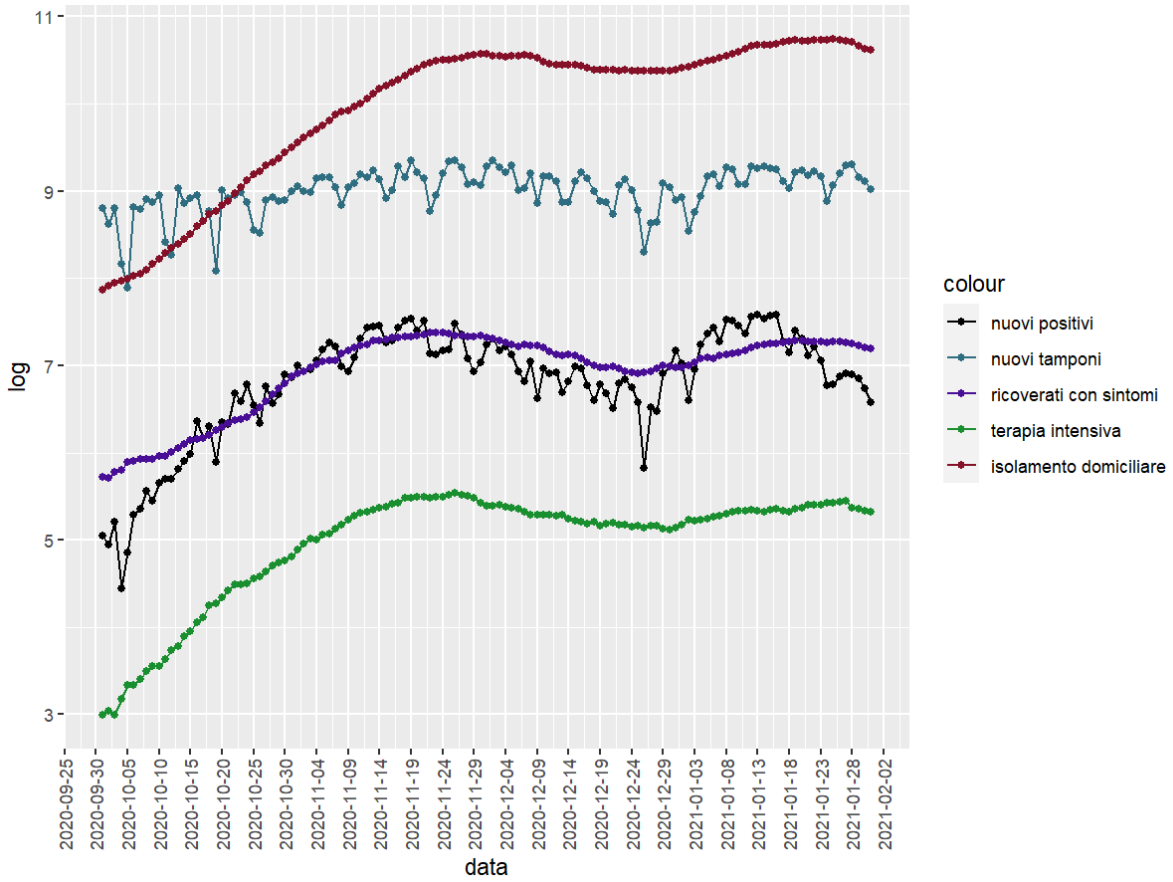
- `nuovi_tamponi_pcr`: swabs carried out daily (tamponi current day - tamponi previous day)
- `nuovi_decessi`: daily deaths (deceduti current day - deceduti previous day)

Data analysis

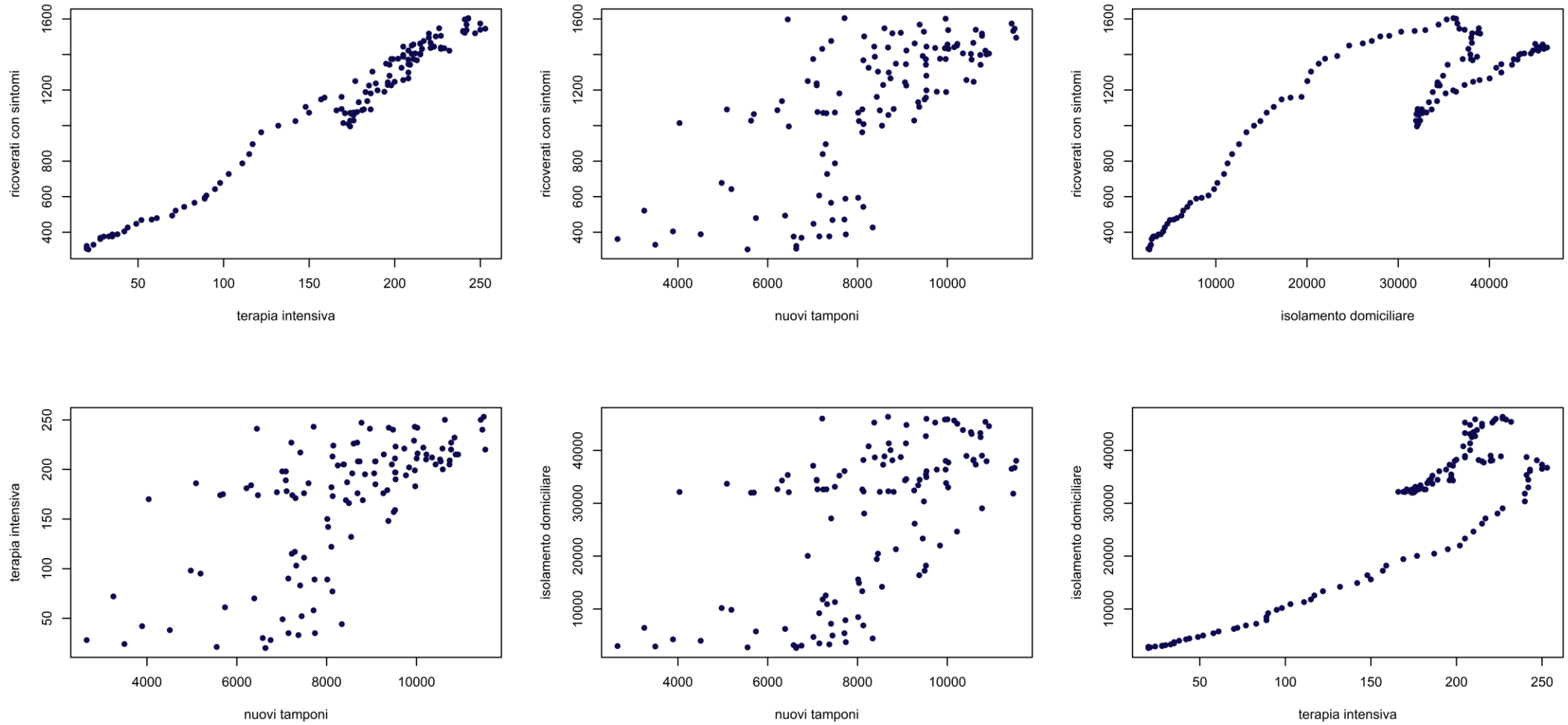
Correlations between relevant variables



Trend of relevant variables over time

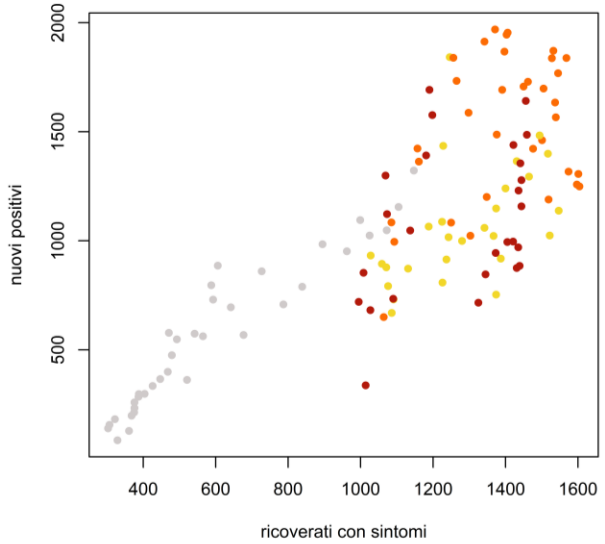


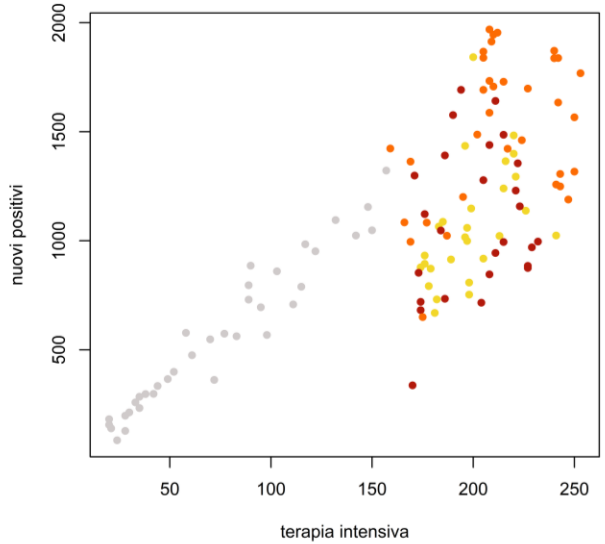
Data analysis

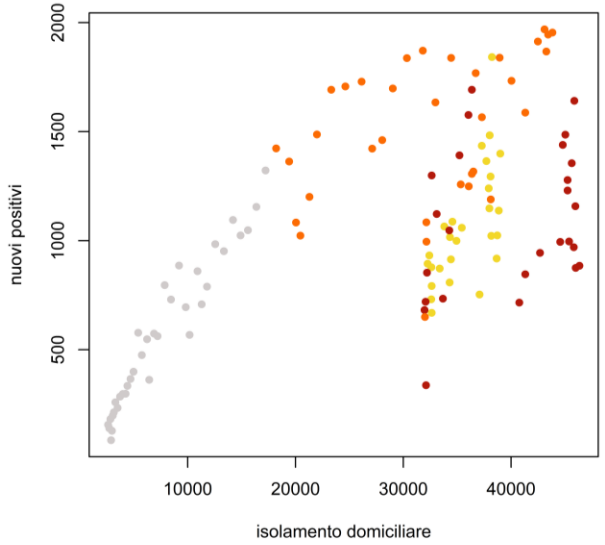


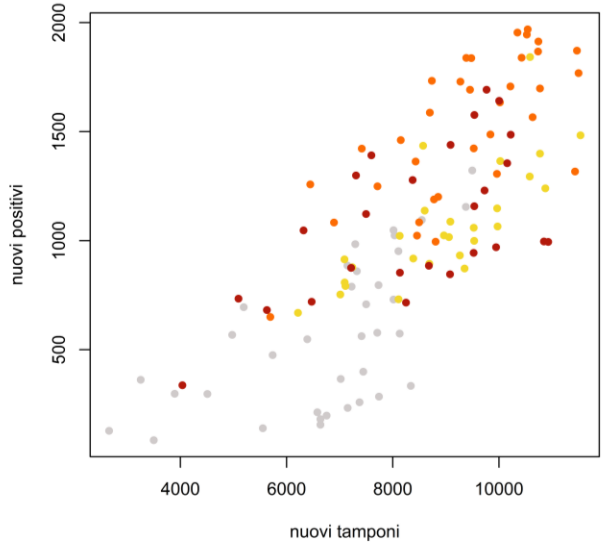
Correlation between pairs of covariates

Analysis of covariates









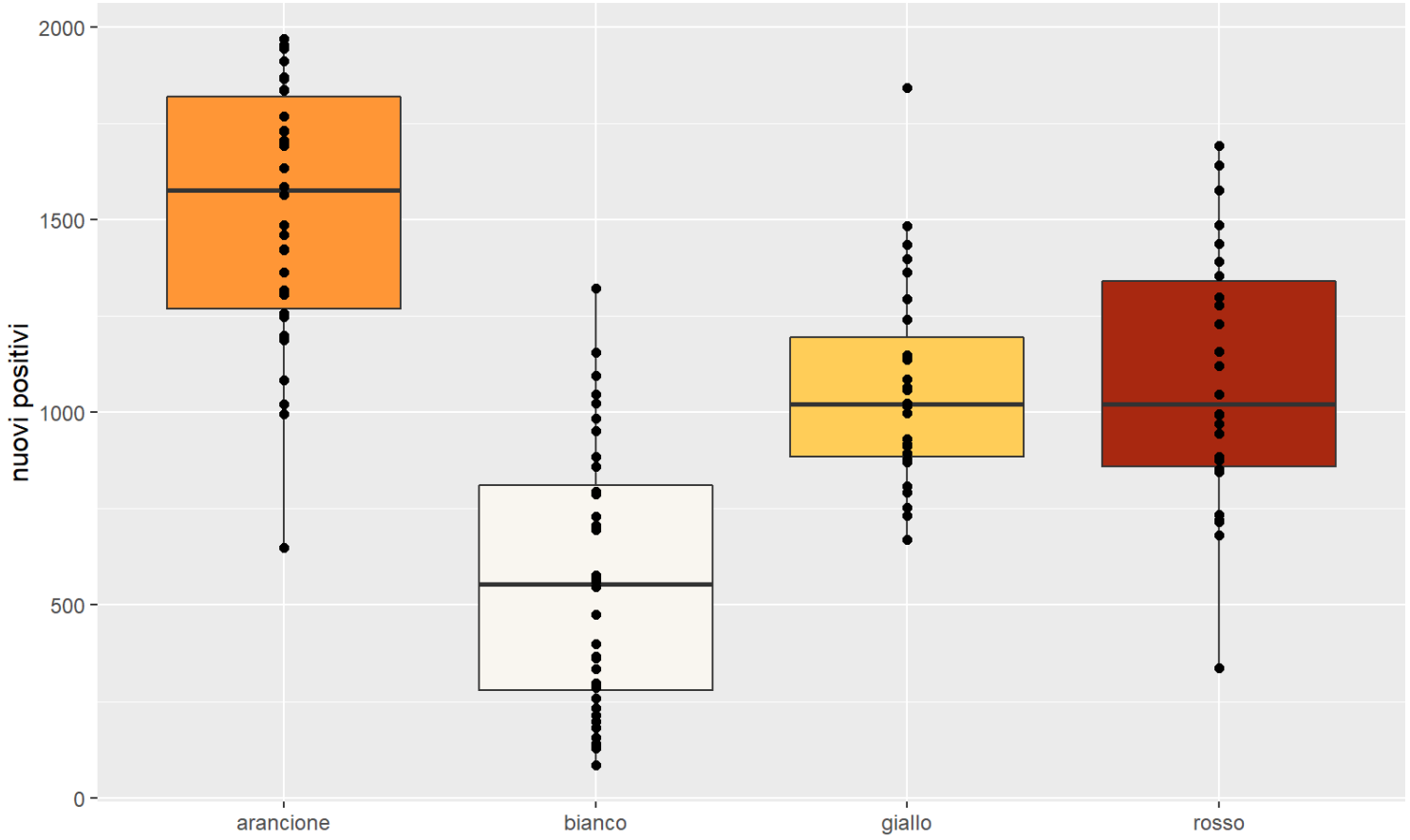
Once the covariates have been chosen, their relationships with the response variable are further investigated.

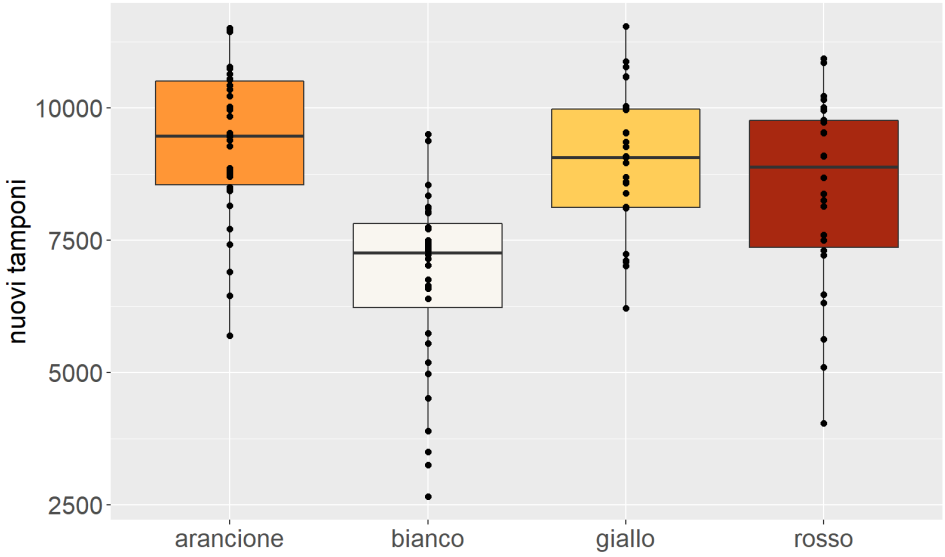
Each point was associated with the color of the region at the time of observation.

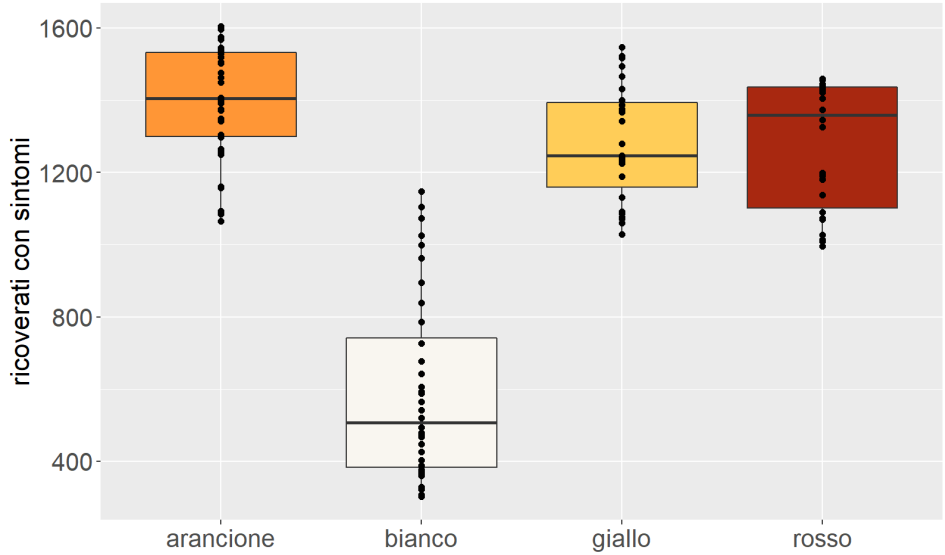
Regional color is related to the restrictions to which the region was subjected in the period considered.
We had 4 different levels of restrictions, which were relevant in determining the number of people infected.

Analysis of covariates

Distribution of the response variable and relevant covariates with respect to different regional colors.







Models

Linear Models

A first attempt was made trying to model our response variable using multiple linear regression. The variable of interest is model through the following functional relationship:

$$y = \beta_0 + \sum_{i=1}^{p-1} \beta_i \cdot x_i$$

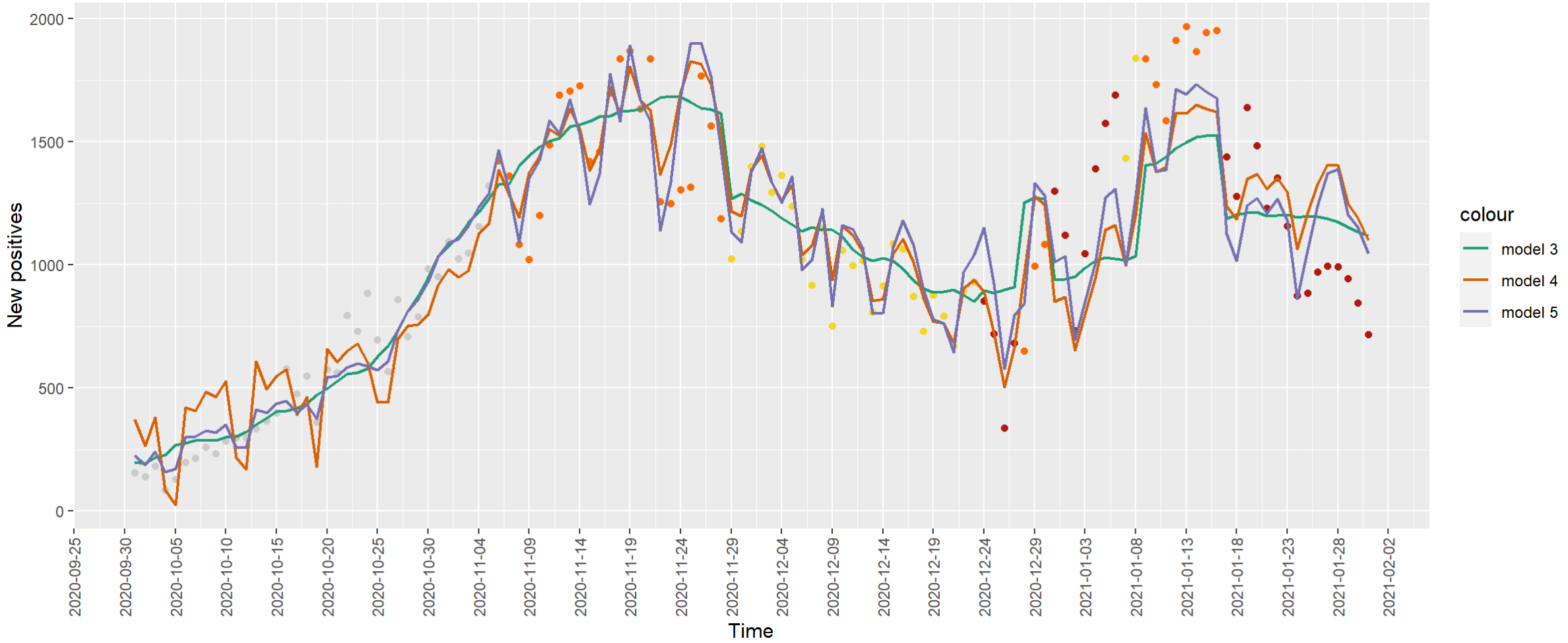
The models were built starting with the most correlated covariate and incrementally adding other covariates, which could improve the model providing information not yet considered.

Linear Models

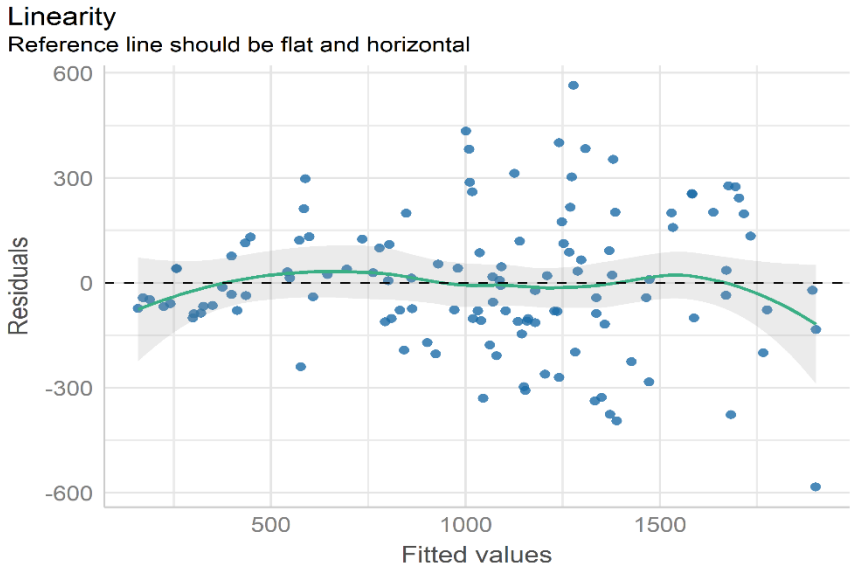
Comparing performances between linear models

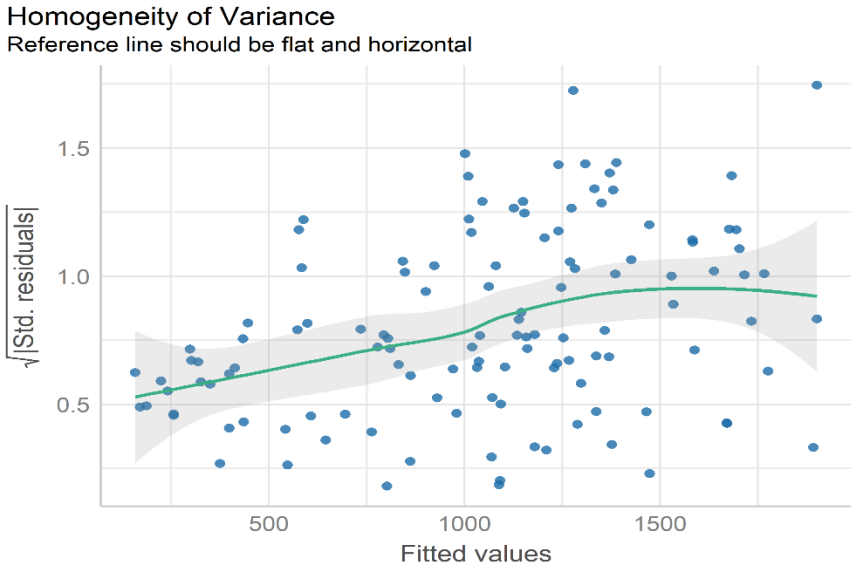
	Formula	R2	R2_adjusted	RMSE	Sigma	AIC wt	BIC wt	Performance Score	AIC
Model 1	nuovi_positivi ~ ricoverati_con_sintomi	0.6504469	0.648	284.980	287.326	0	0	0.00%	1745.554
Model 2	nuovi_positivi ~ ricoverati_con_sintomi + color	0.7302628	0.721	216.700	255.588	0	0	26.44%	1719.672
Model 3	nuovi_positivi ~ ricoverati_con_sintomi * color	0.7400996	0.724	245.732	254.136	0	0	28.72%	1721.102
Model 4	nuovi_positivi ~ ricoverati_con_sintomi + nuovi_tamponi_pcr + color	0.7978830	0.789	216.700	222.187	0	0.485	67.16%	1686.174
Model 5	nuovi_positivi ~ (ricoverati_con_sintomi + nuovi_tamponi_pcr) * color	0.8403247	0.825	192.609	202.753	1.000	0.515	100.00%	1669.182

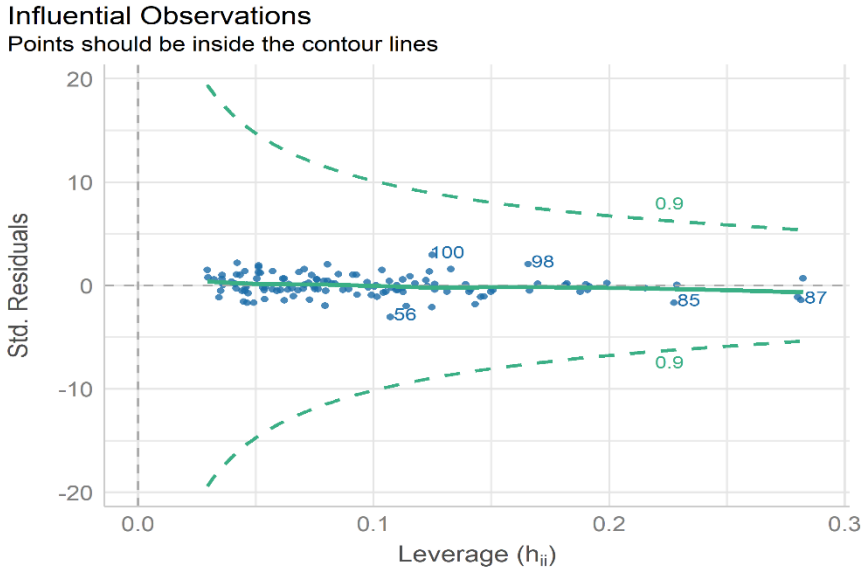
Linear Models

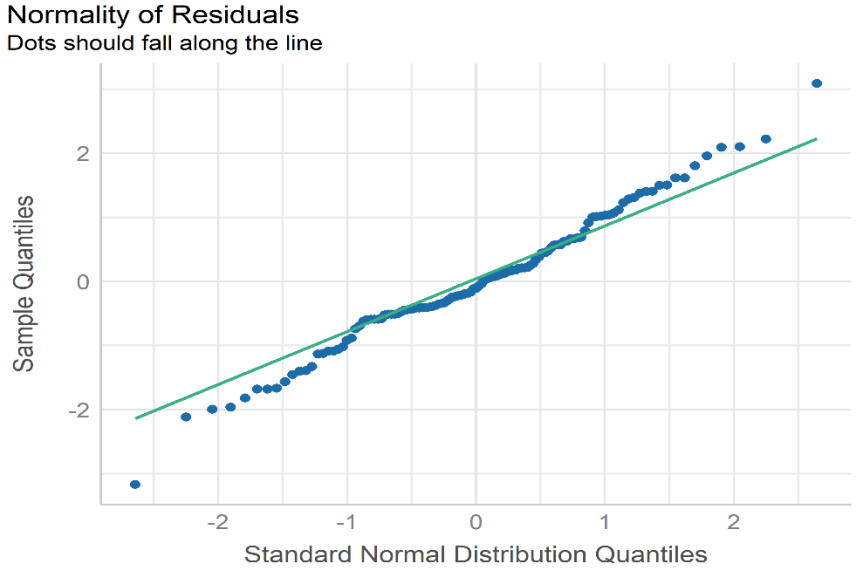


Linear Models



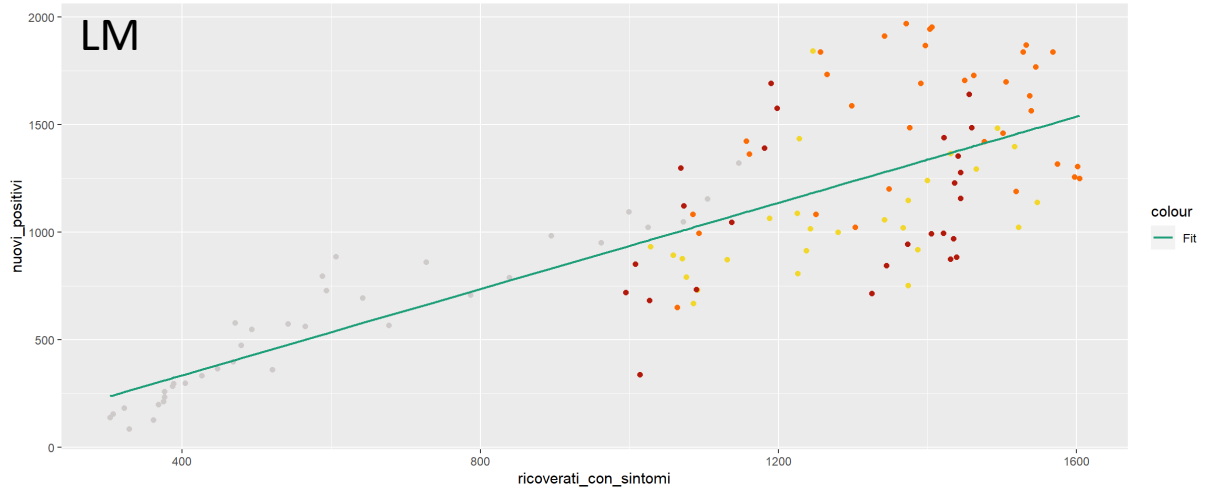




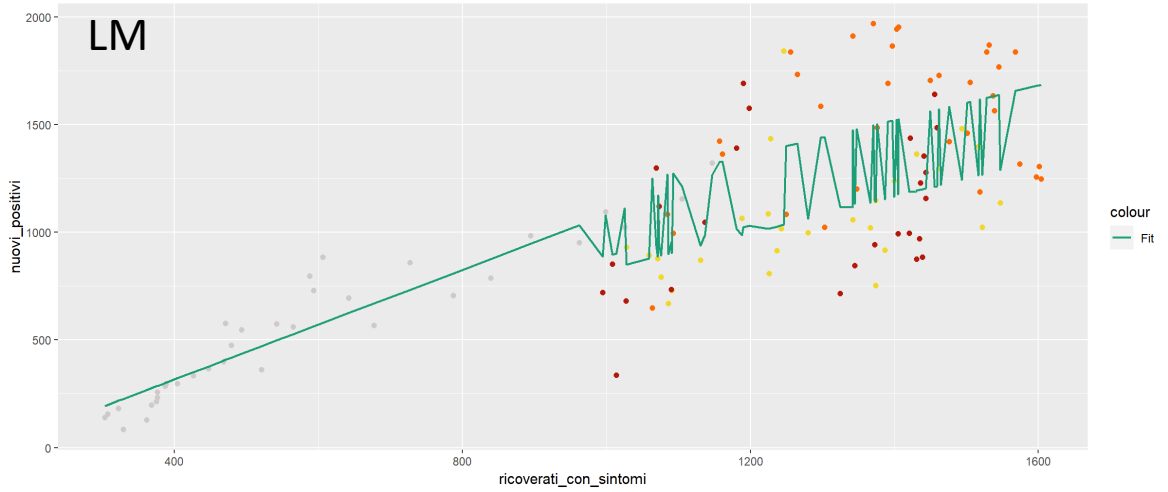


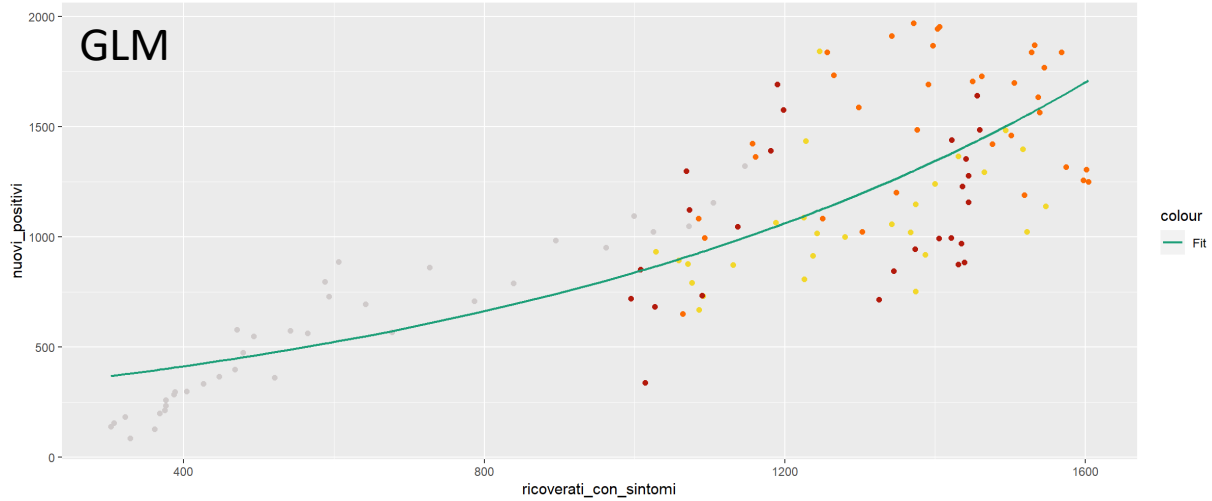
Linear Models

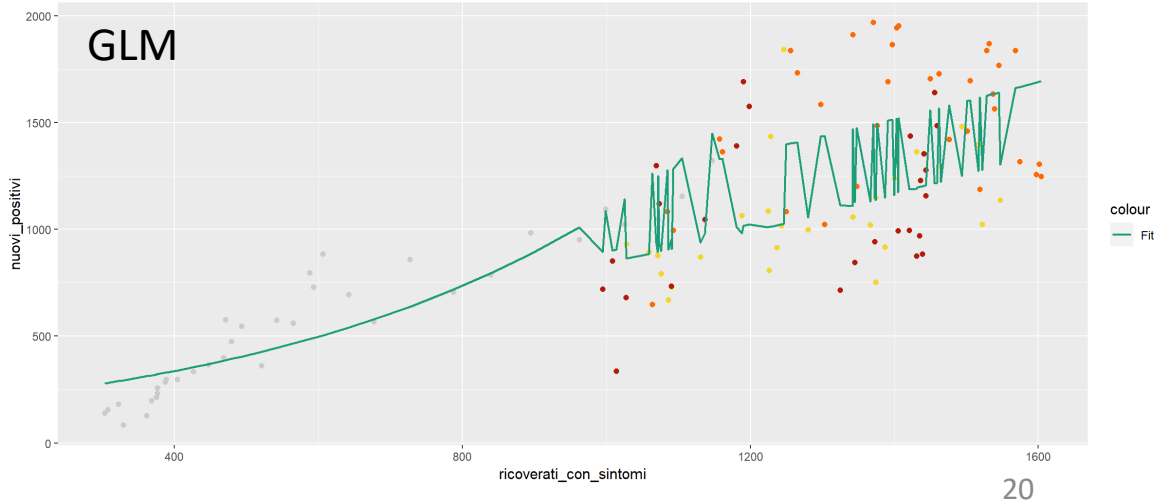
nuovi_positivi ~ ricoverati_con_sintomi



*nuovi_positivi ~ ricoverati_con_sintomi*colore*







Generalized Linear Models

Generalized Linear Model is a generalitazion of linear regression in which the expected value of the response variable Y is related to a linear combination of the predictors through a *link* function. Besides it allows higher flexibility regarding the variance which is computed as a function of the predicted value.

The general structure is: $$g\big(\mathbb{E}(Y)\big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

The GLM assume the Y to follow a specific distribution, the ones most commonly used for modelling counting variables are *Poisson, QuasiPoisson, NegativeBinomial.*

Generalized Linear Models – Poisson Regression

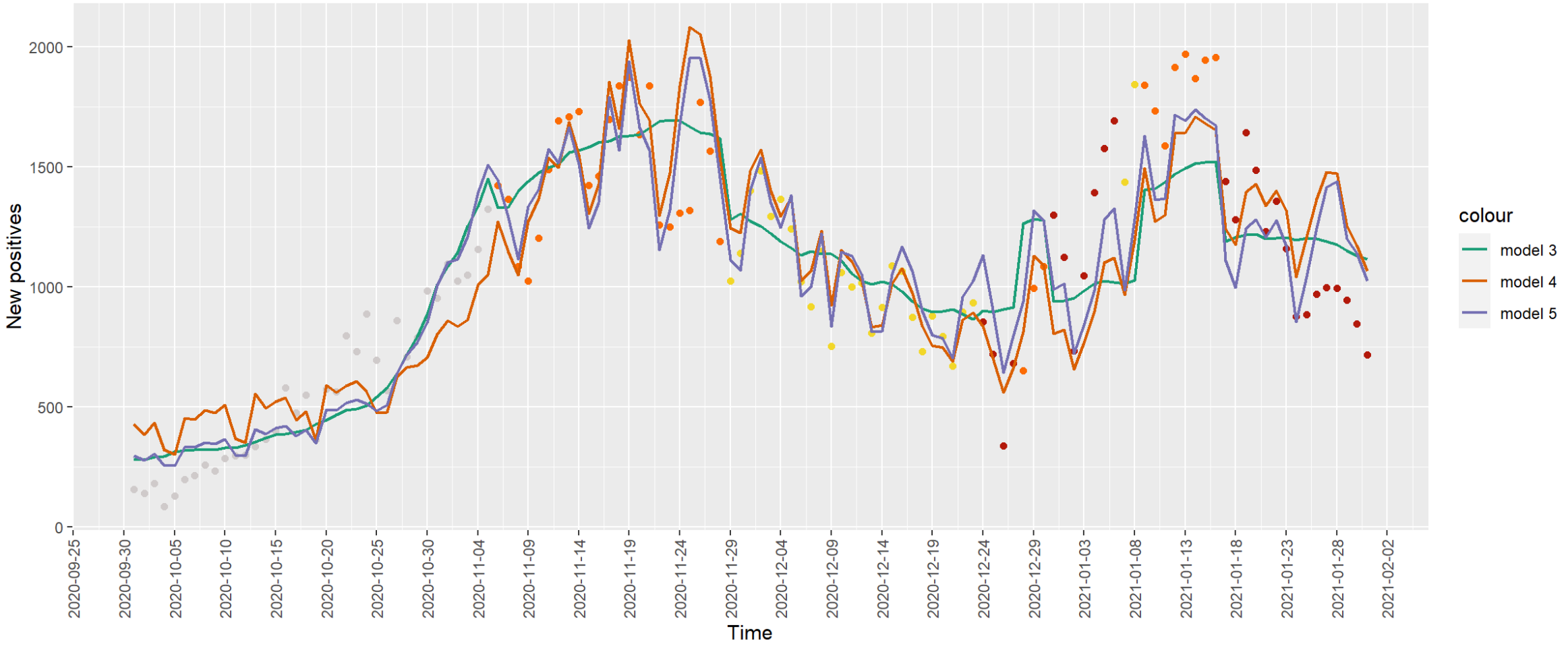
Poisson regression operates under the assumptions that the *response variable* Y follows a Poisson distribution and the logarithm of its expected value can be modelled by a linear combination of paramaters which have to be estimated.

$$x \in \mathbb{R}^n,\ \log\big(E(Y|x)\big) = \alpha + \beta' x$$

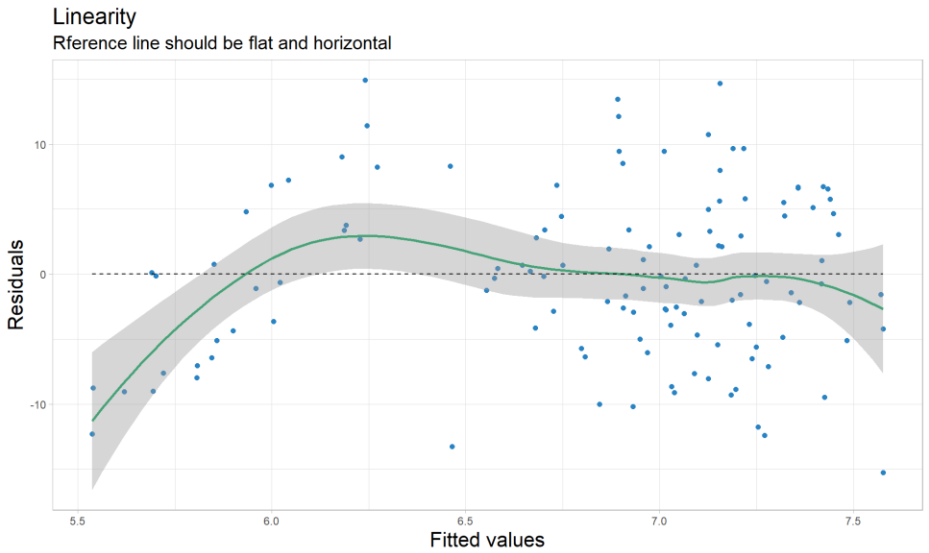
The model assumes the mean being equal to the variance, which can lead to a poor fit in case of overdispersion.

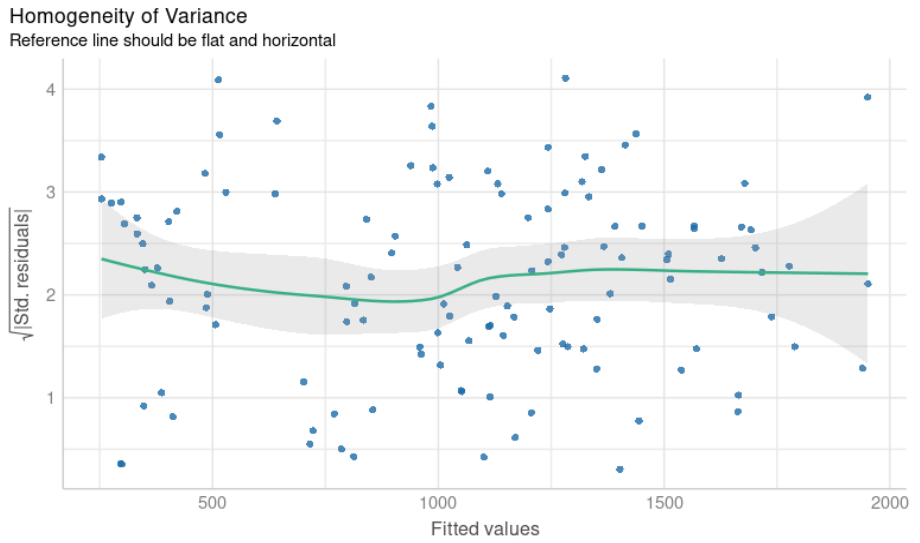
	Formula	AIC	RMSE	Sigma	AIC wt	BIC wt	Performance Score
Model 1	nuovi_positivi ~ ricoverati_con_sintomi	11485.561	300.153	9.280	0	0	0.00%
Model 2	nuovi_positivi ~ ricoverati_con_sintomi + color	10264.652	277.068	8.826	0	0	18.67%
Model 3	nuovi_positivi ~ ricoverati_con_sintomi * color	8549.686	253.153	8.060	0	0	34.02%
Model 4	nuovi_positivi ~ ricoverati_con_sintomi + nuovi_tamponi_pcr + color	8298.537	239.295	7.858	0	0	44.05%
Model 5	nuovi_positivi ~ (ricoverati_con_sintomi + nuovi_tamponi_pcr) * color	6149.989	205.439	6.754	1.00	1.00	100.00%

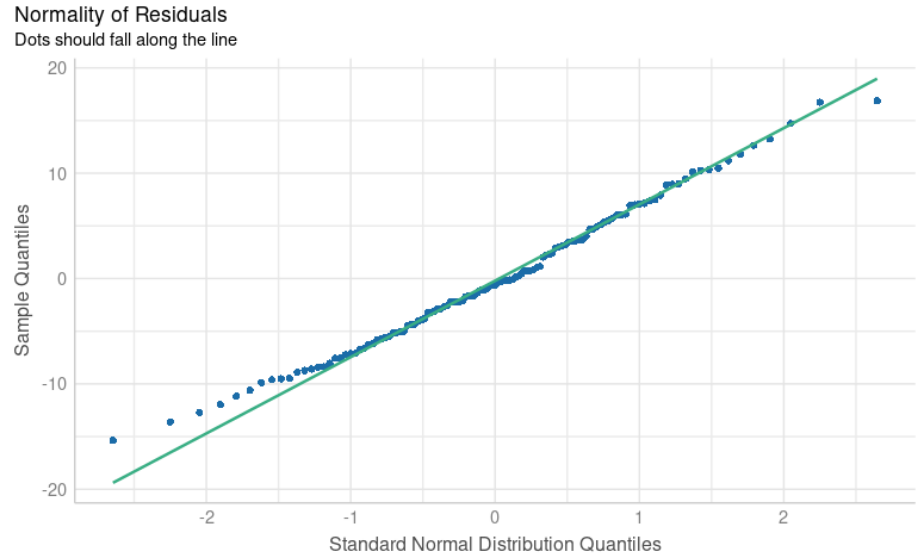
Generalized Linear Models – Poisson Regression

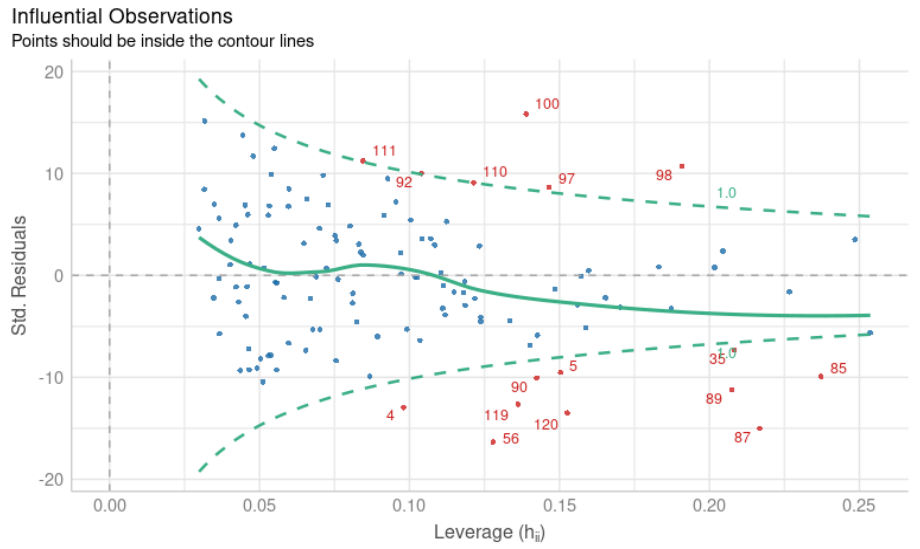


Generalized Linear Models – Poisson Regression



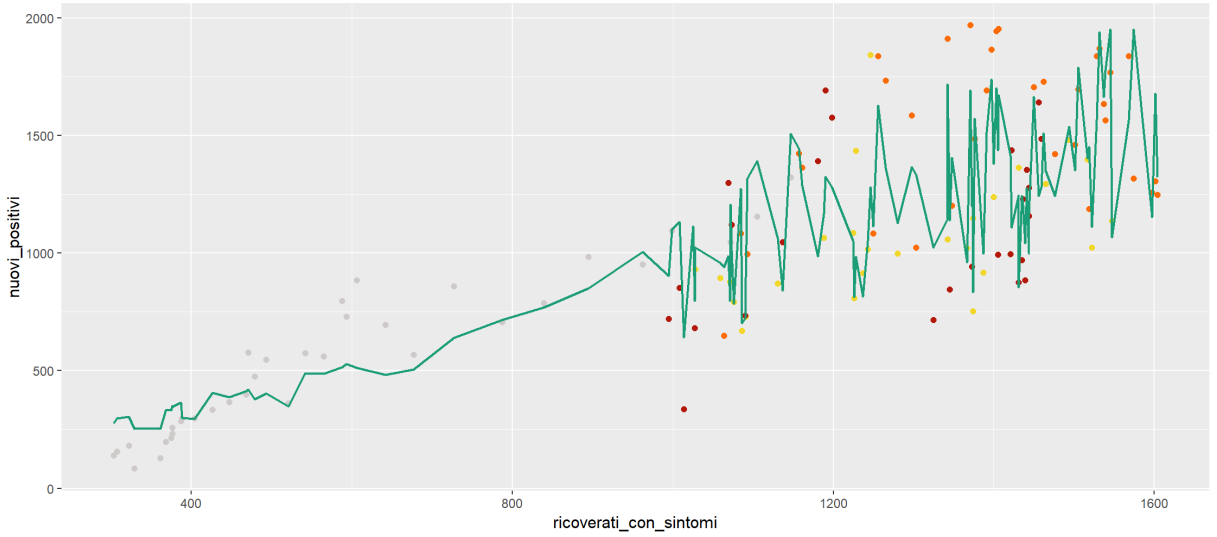


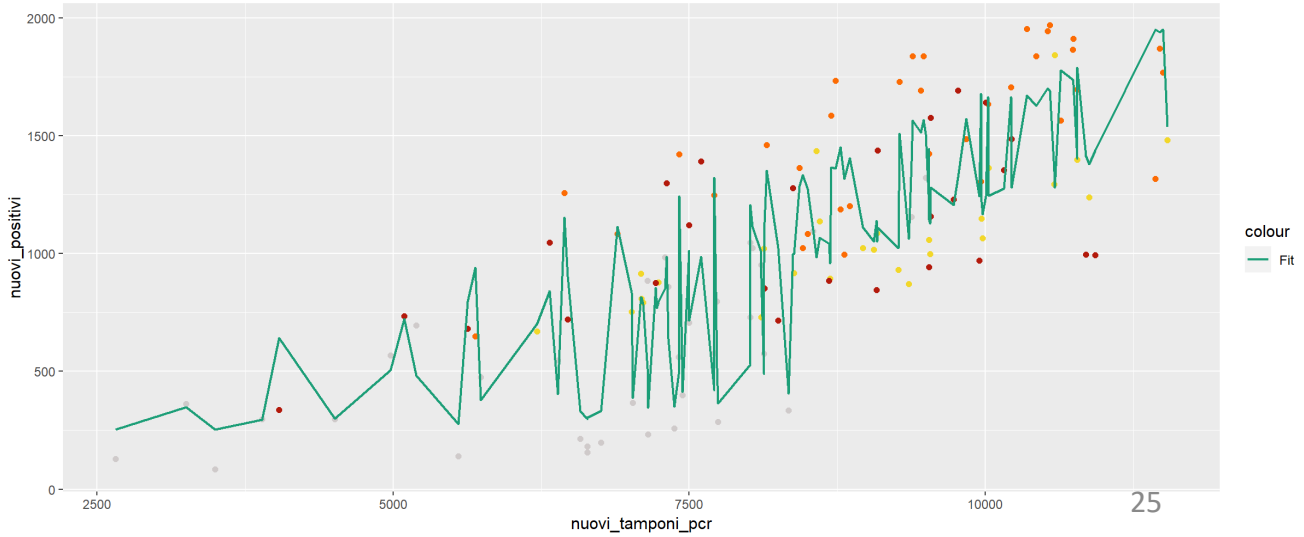




Generalized Linear Models – Poisson Regression

$$y \sim (ricoverati\_con\_sintomi + nuovi\_tamponi\_pcr) * colore$$





Generalized Linear Models – Poisson Regression

We have run an Overdispersiontest over the model previously analyzedand we got the following output.

```
>> check_overdispersion(glm_mod5)

# Overdispersion test

       dispersion ratio =   45.398
  Pearson's Chi-Squared = 5039.228
                p-value =  < 0.001

Overdispersion detected.
```

We can test for overdispersion in classical Poisson regression by computing the sum of squares of the n standardized residuals: $\sum_{i=1}^{n} z_i^2$, and comparing this to the χ^2_{n-k} distribution, which is what we would expect under the model.

We thus need to switch to other more flexible models in which the dispersion parameter is not fixed to 1.

Generalized Linear Models - Negative Binomial

Negative binomial regression is a generalization of Poisson regression where the condition of the mean and the variance is flexible. The assumptions are that Y follows a negative-binomial distribution and the link function is the logarithm. This model is commonly used instead of the Poissonian one, when the dataset shows signs of overdispersion.

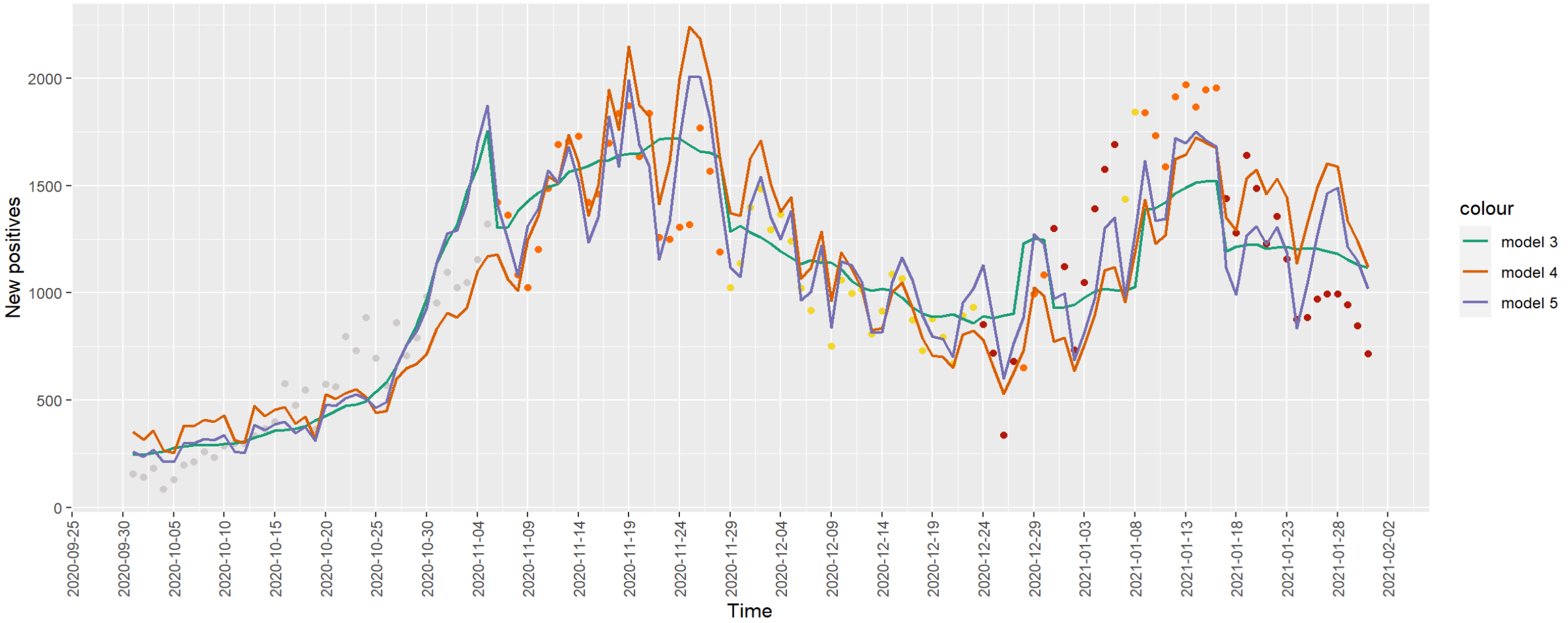
Probability mass function of Negative Binomial: $\binom{k+r-1}{k} \cdot (1-p)^r p^k$

Generalized Linear Models - Negative Binomial

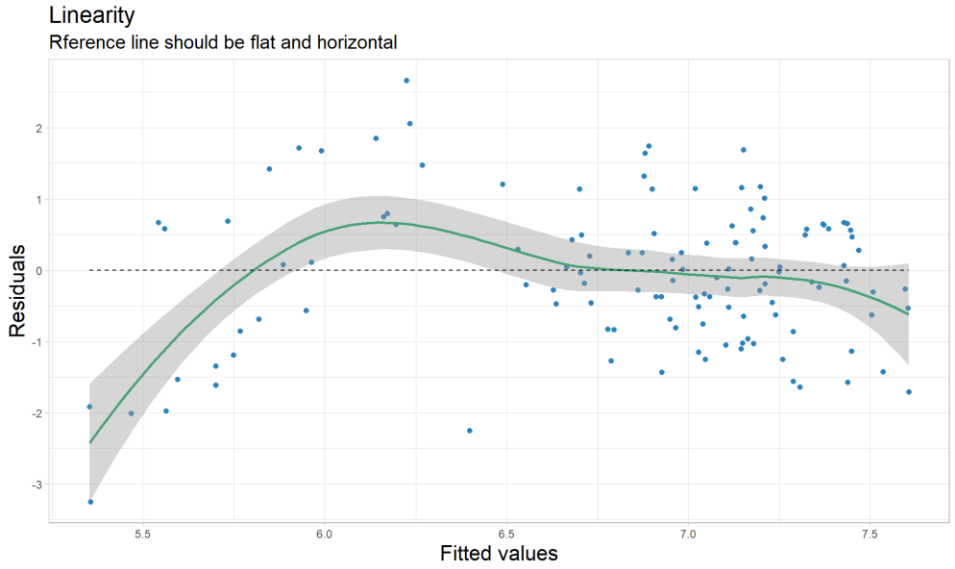
Comparing performance between generalized linear models with negative binomial distribution.

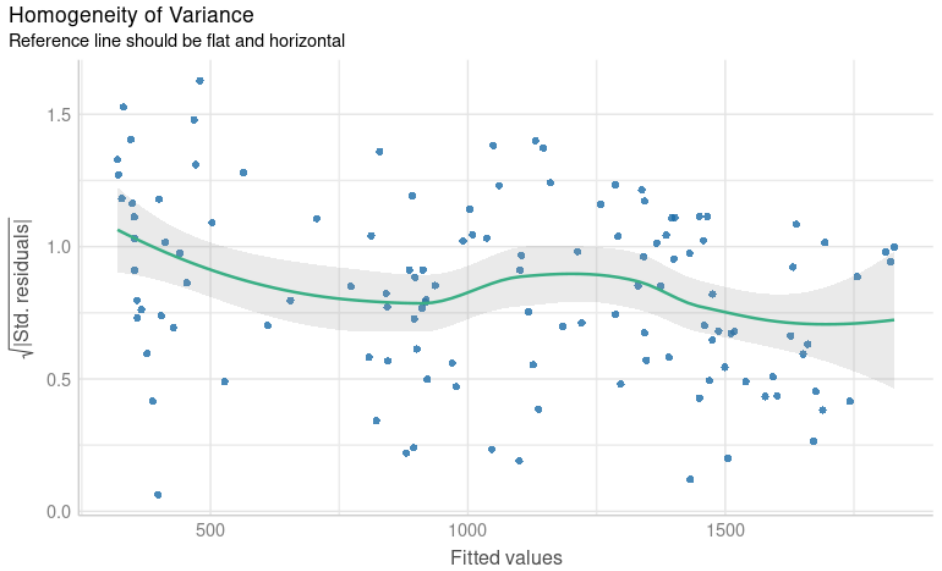
Formula		AIC	RMSE	Sigma	AIC wt	BIC wt	Performance Score
Model 1	nuovi_positivi ~ ricoverati_con_sintomi	1754.076	310.756	1.020	0	0	26.87%
Model 2	nuovi_positivi ~ ricoverati_con_sintomi + color	1749.925	302.622	1.033	0	0	31.68%
Model 3	nuovi_positivi ~ ricoverati_con_sintomi * color	1724.282	262.545	1.045	0	0	34.29%
Model 4	nuovi_positivi ~ ricoverati_con_sintomi + nuovi_tamponi_pcr + color	1731.239	267.698	1.037	0	0	43.22%
Model 5	nuovi_positivi ~ (ricoverati_con_sintomi + nuovi_tamponi_pcr) * color	1692.763	221.425	1.065	1.00	1.00	71.43%

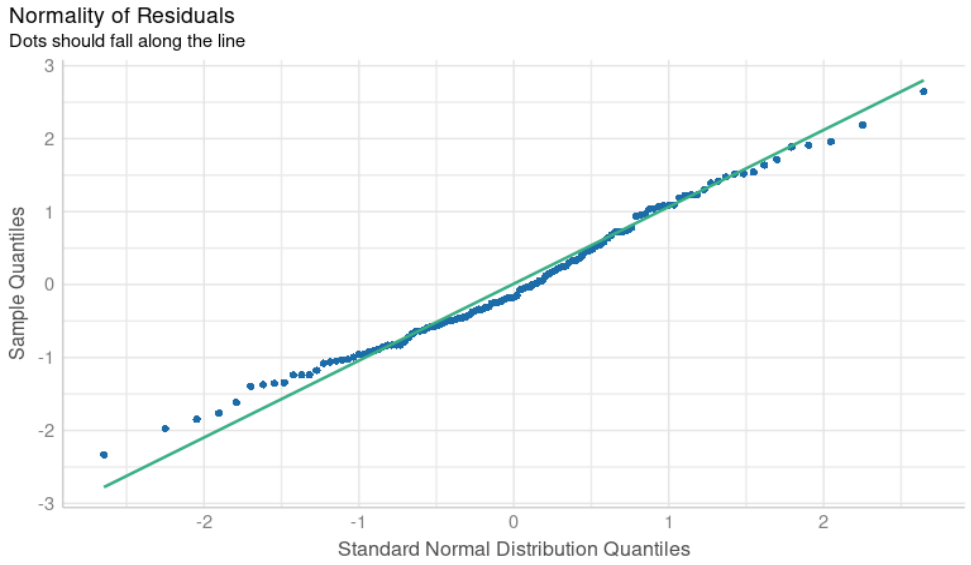
Generalized Linear Models - Negative Binomial

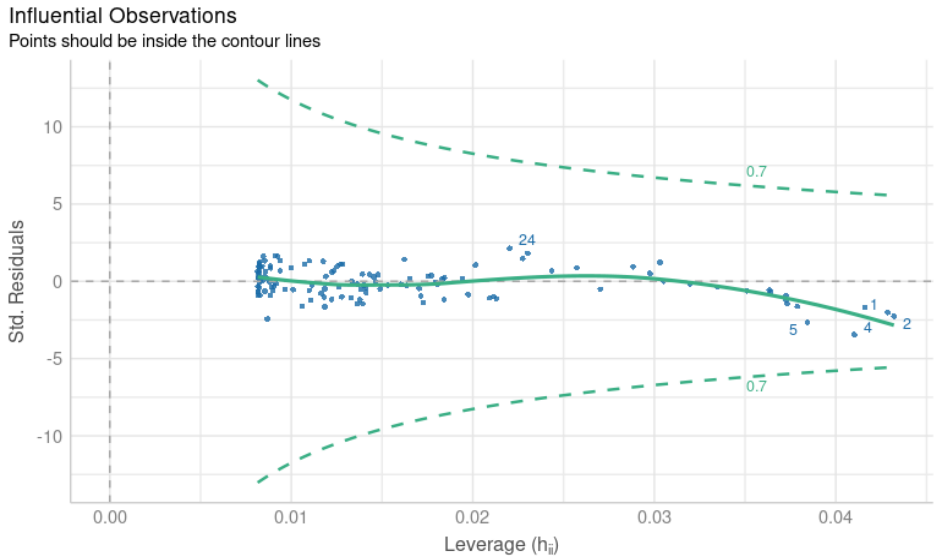


Generalized Linear Models - Negative Binomial









Predictions

Aim of predictions

Provide 15 days-forward predictions and check their accuracy

We start using the best LM and GLM models

Increase complexity in order to improve the prediction accuracy

Evaluation of predictions

To evaluate the accuracy of the observed preditctions the Symmetric Mean Absolute Percentage Error (SMAPE) has been used.

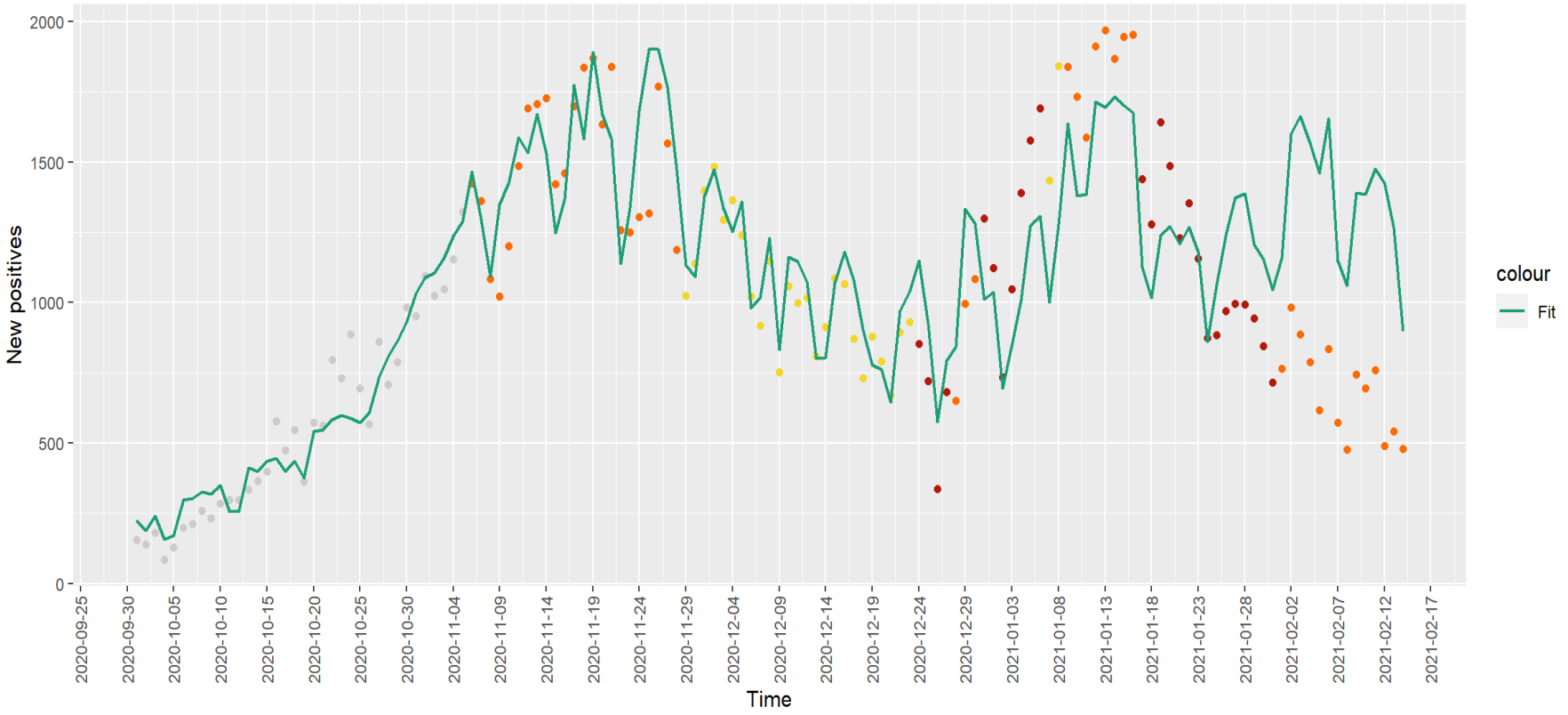
SMAPE is defined as: $$SMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|A_t| + |F_t|)},$$

and it has been used to compare the predicted values with the actual values for the first 15 days of February.

	Formula	ACCURACY
LM	`nuovi_positivi ~ (ricoverati_con_sintomi + nuovi_tamponi_pcr) * colore`	0.666282
GLM	`nuovi_positivi ~ (ricoverati_con_sintomi + nuovi_tamponi_pcr) * colore`	0.6654311
NGLM	`nuovi_positivi ~ (ricoverati_con_sintomi + nuovi_tamponi_pcr) * colore`	0.6768924

Predictions

*glm(nuovi_positivi ~ (ricoverati_con_sintomi + nuovi_tamponi_pcr) * colore)*

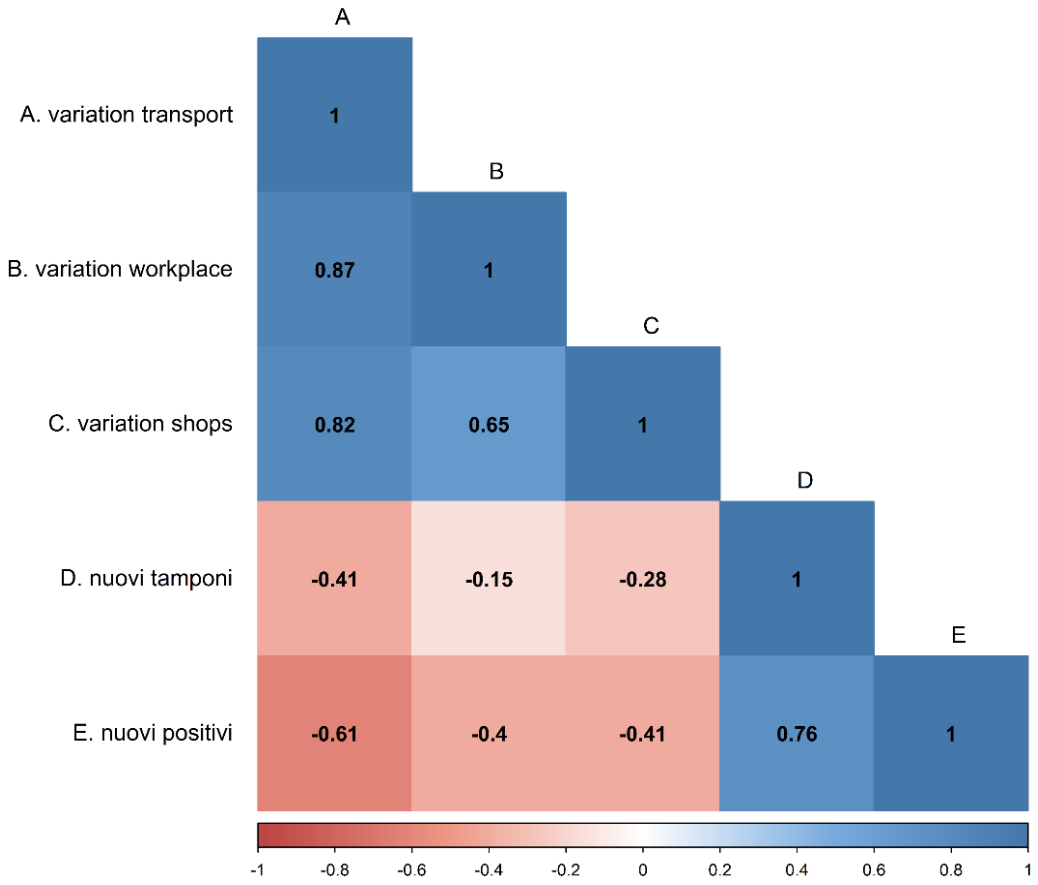


All the models previously considered overestimate the number of new positives on the first two weeks of February.

This behavior might be due to the gap between the variable color and the actual level of restrictions.

To handle this issue, we firstlytried to use the data about movements of people; at a later time,we introduced a new numerical variable (livello_di_restrizioni).

Prediction: Google data

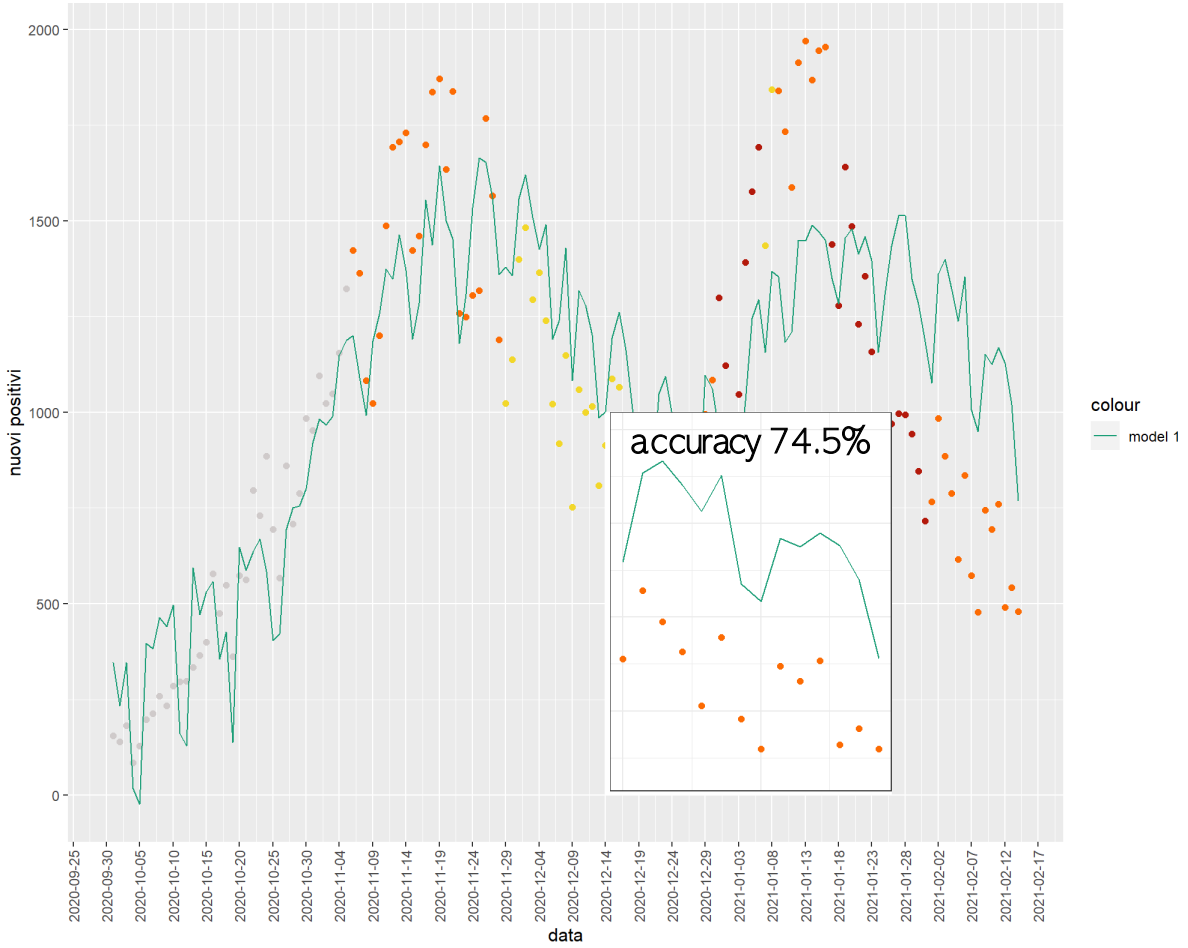


- To further improve the prediction accuracy, we add a variable which considers people movements with respect to a given baseline.
- We used the data released by Google on the "Covid-19 mobility report" [1].

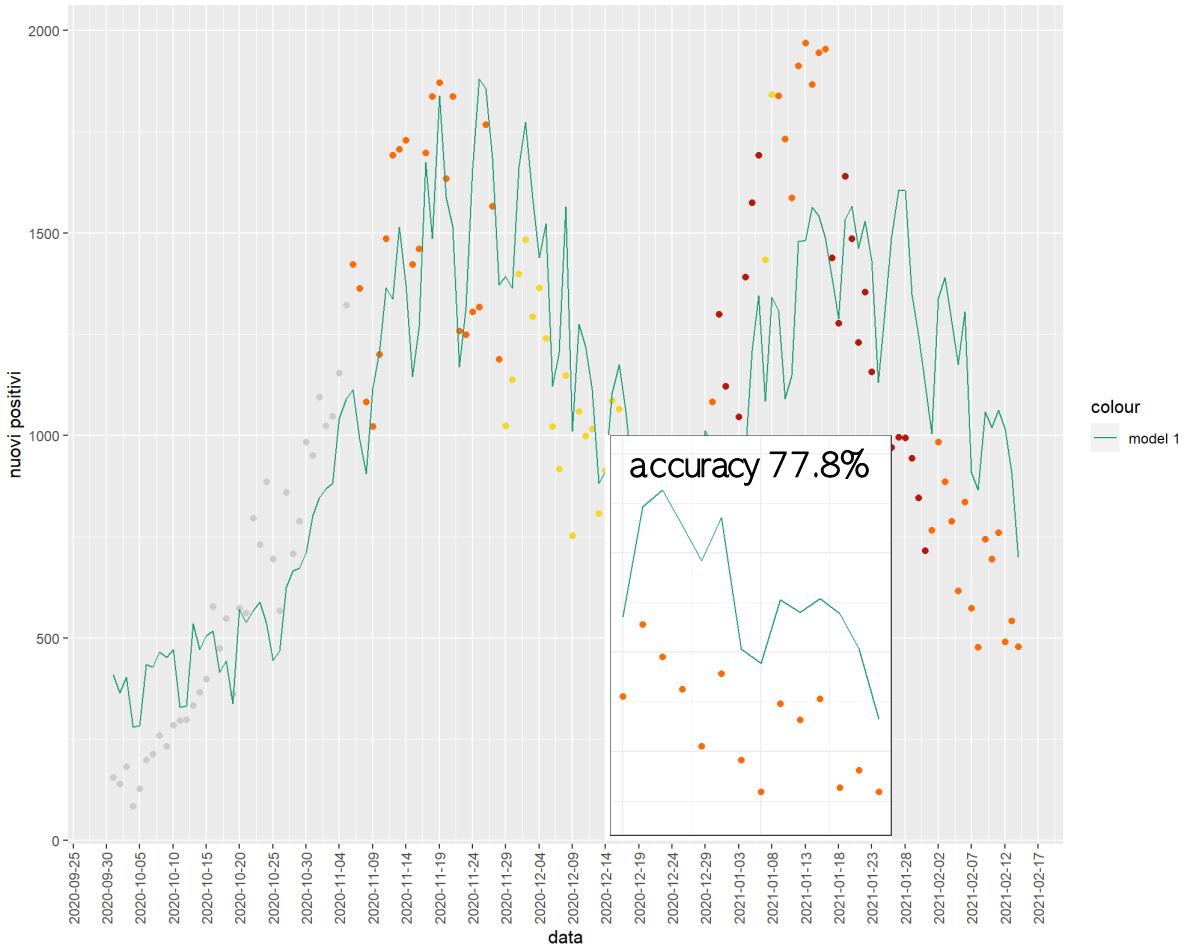
[1] https://data.humdata.org/dataset/google-mobility-report

Prediction

$$y \sim ricoverati\_con\_sintomi + nuovi\_tamponi + var\_station$$

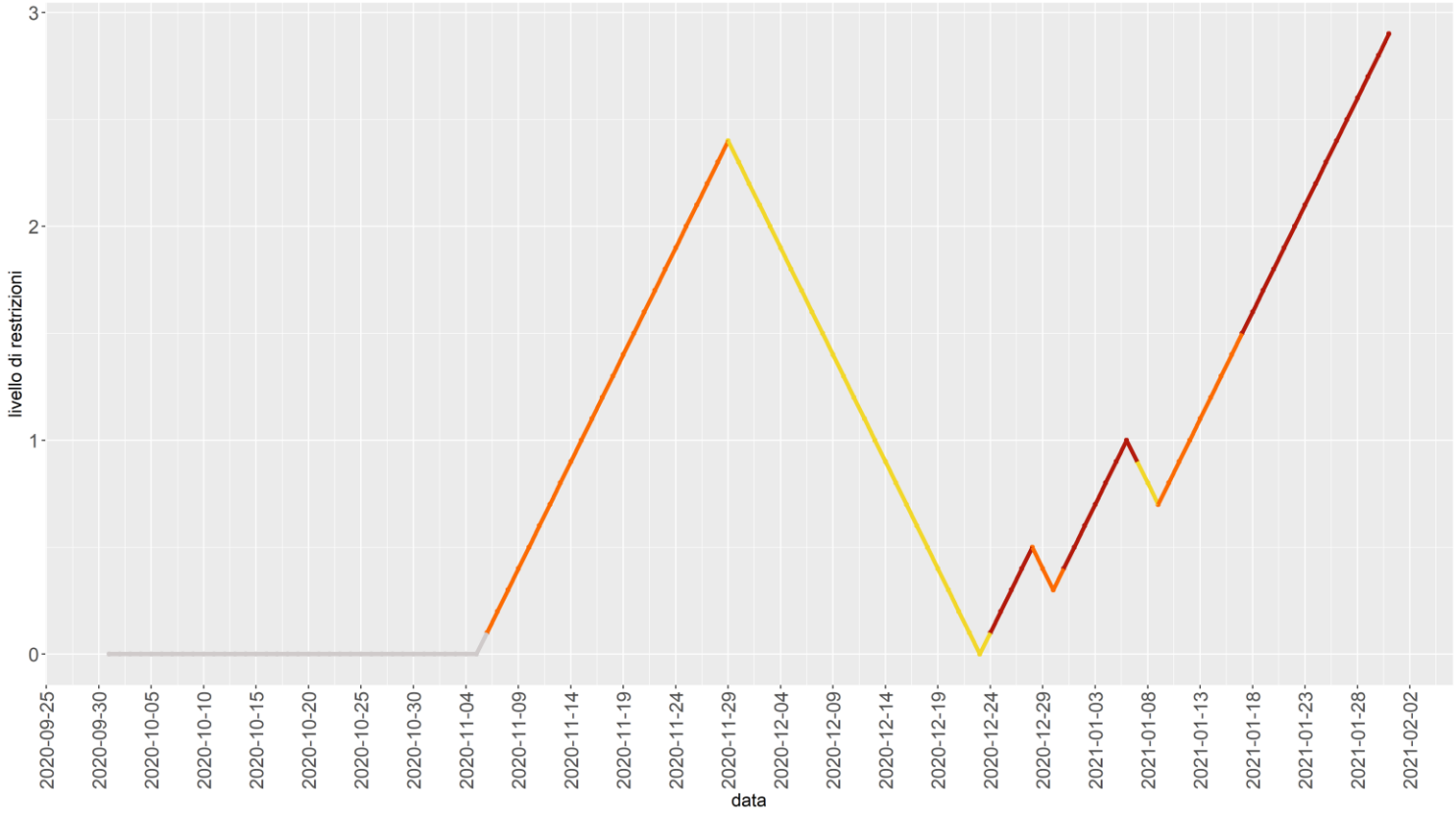


LM



GLM

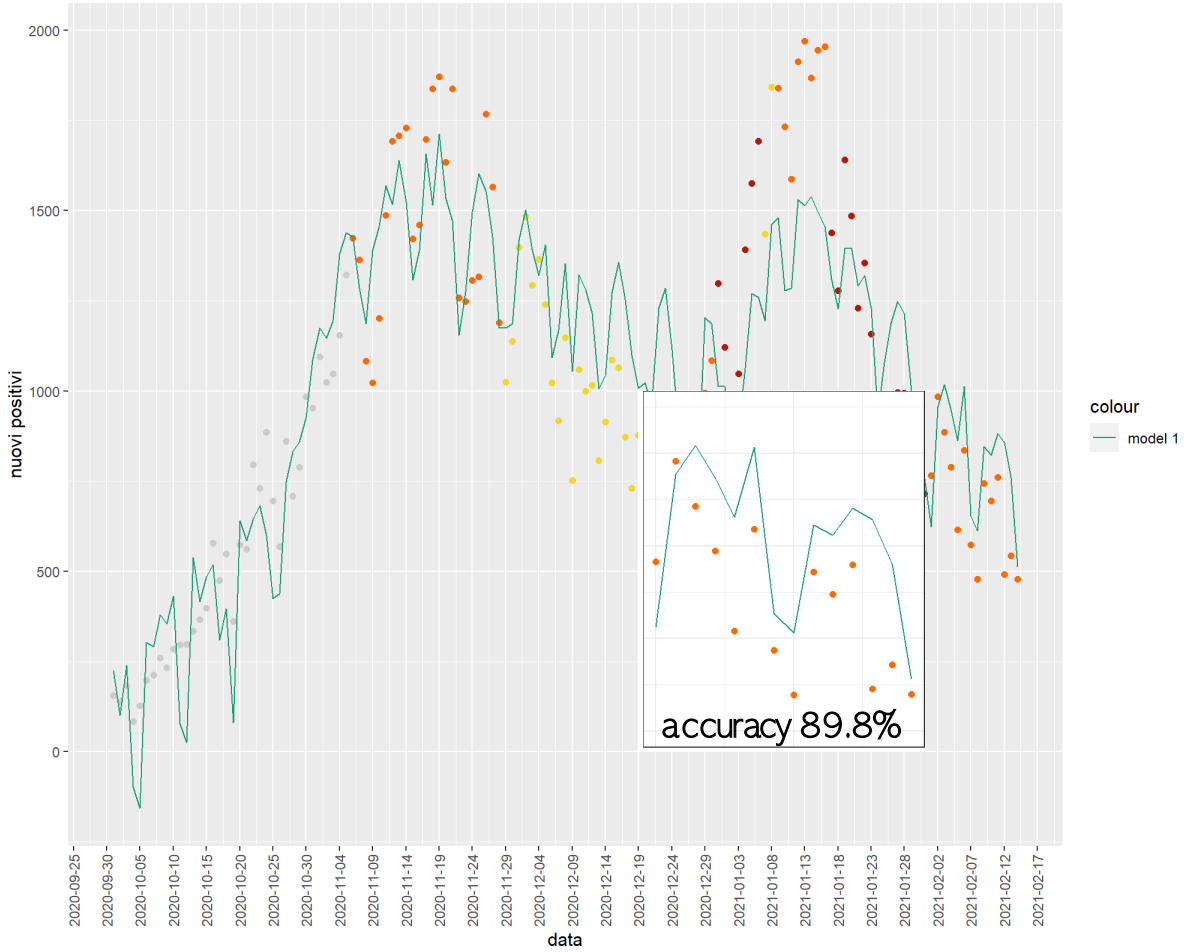
Prediction: model of restrictions



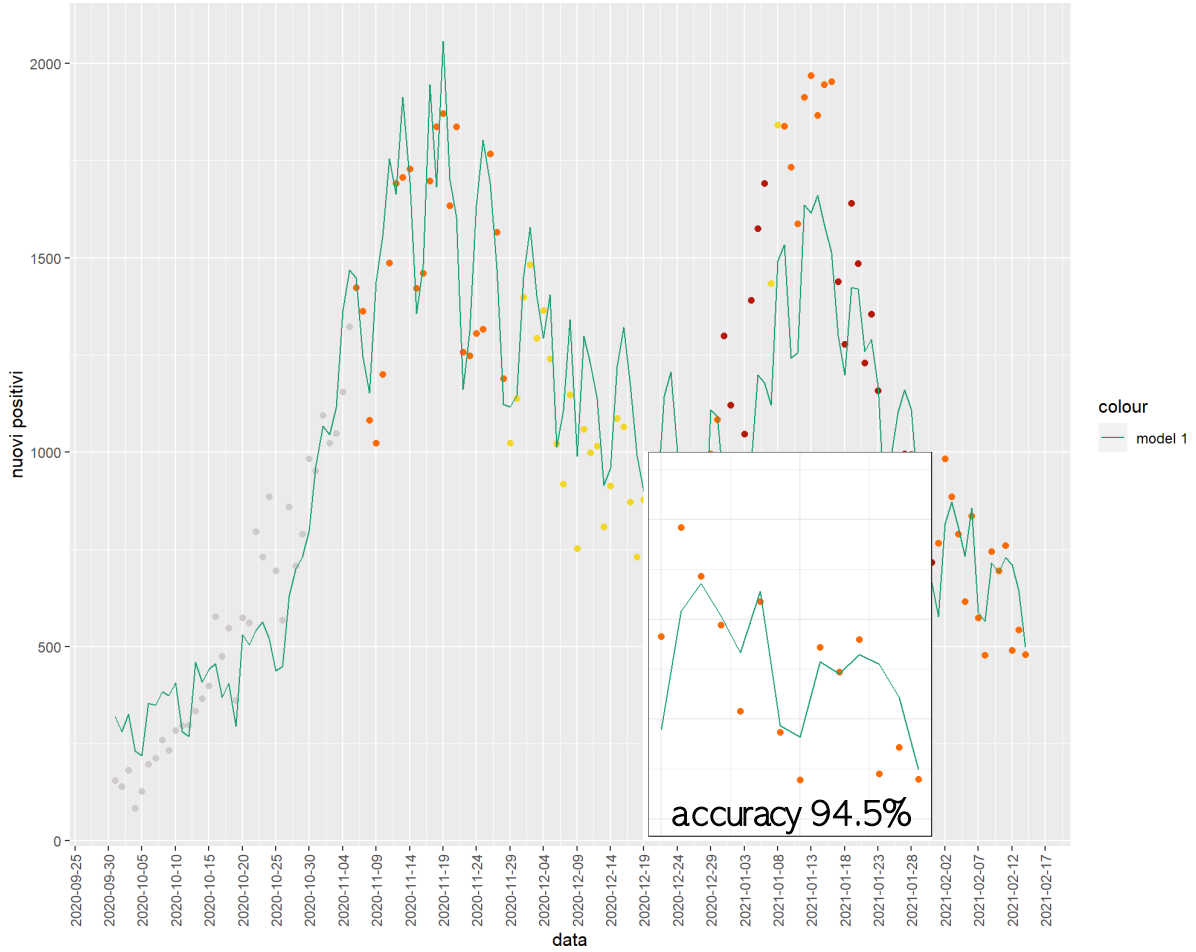
- Instead of a categorical value for the region color, a numerical variable has been used to account for the level of restrictions.
- The value depends on the previous regional color as well as the current one and the number of days elapsed from the last color change.

Prediction

$$y \sim ricoverati\_con\_sintomi + nuovi\_tamponi + livello\_di\_restrizioni$$



LM



GLM

Appendix: Non-parametric methods

Non-parametric methods

- LOESS regression is a nonparametric technique that uses local weighted regression to obtain a smooth curve. For $i = 1 \ldots \mathrm{n}$ the measurement y_i and the corresponding x_i of the vector x of p predictors are related by: $y_i = g(x_i) + \epsilon_i$, where g is the regression function and ϵ_i is a random error.

- SPLINE methods model the relationship between the response and a continuous covariates, employing a linear combination of smooth function (B-splines) of the predictor.

  The general structure is: $\mathbb{E}(Y \mid X = x) = m(x)$ when m is a deterministic smooth function.

- We tried these two models but we noticed that fitted models result similar to LM and GLM models. It is hard to assess the quality of the fit due to the fact that common performance metrics used in LM and GLM are less informative in the non-parametric framework. Moreover predictions were not so good and for this reason we did not include them here.

Non-parametric methods

- We tried also using MARS, MARS is a semi-parametric method that recursively adapt a curve to the regression surface.

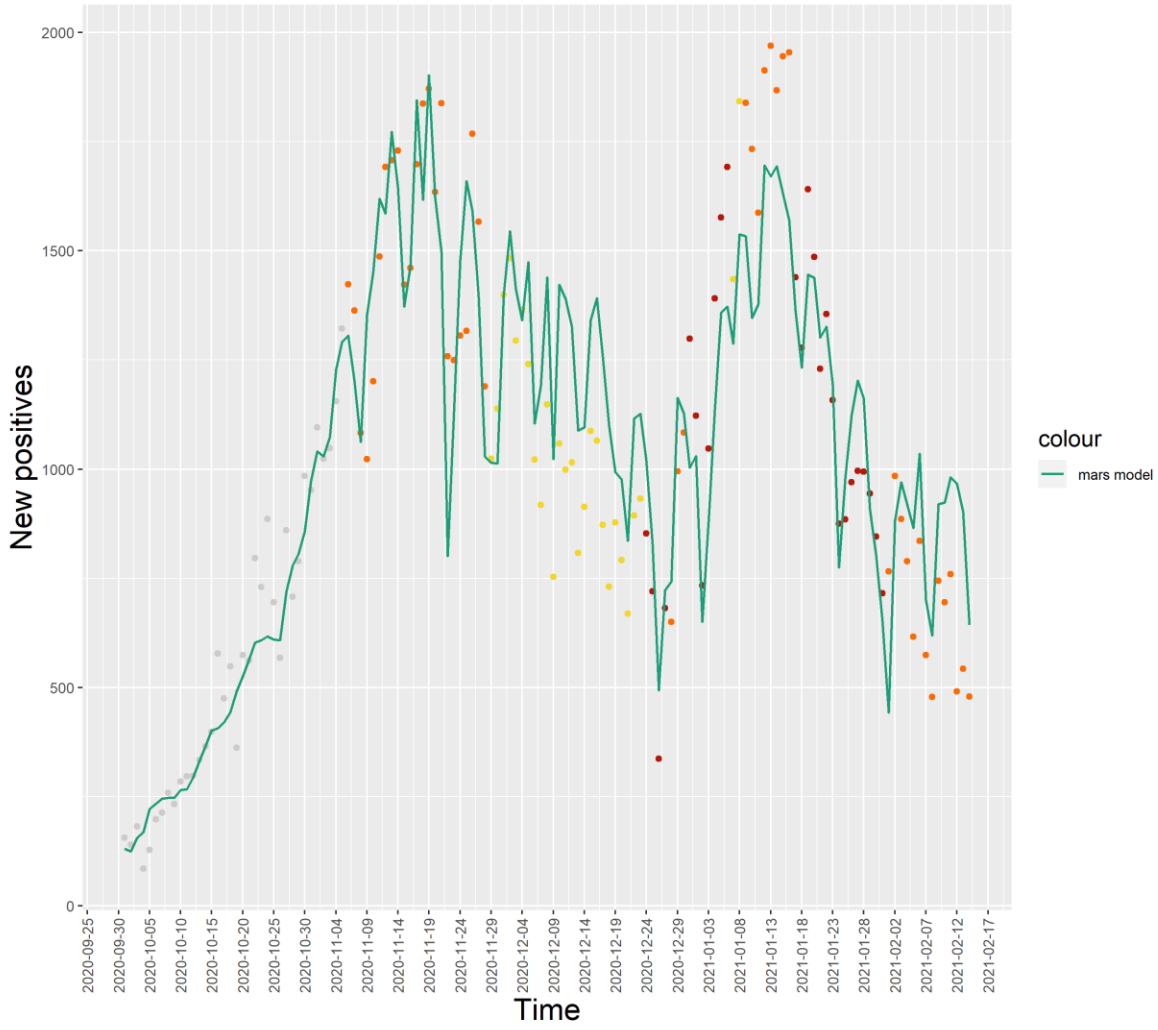
The general structure is: $f(x) = \sum_{i=1}^{k} c_i \, B_i(x_i)$

In fact, MARS is an adaptive procedure for regression, which adds model terms handling automatically the intrinsic nonlinearity in our model. These models yield very accured fits and quite good predictions, employing automatic routines and it is hard to get information about the effective behavior of the model, as for now we decided not to proceed any further.

Non-parametric methods

y ~ ricoverati_con_sintomi + nuovi_tamponi + livello_di_restrizioni

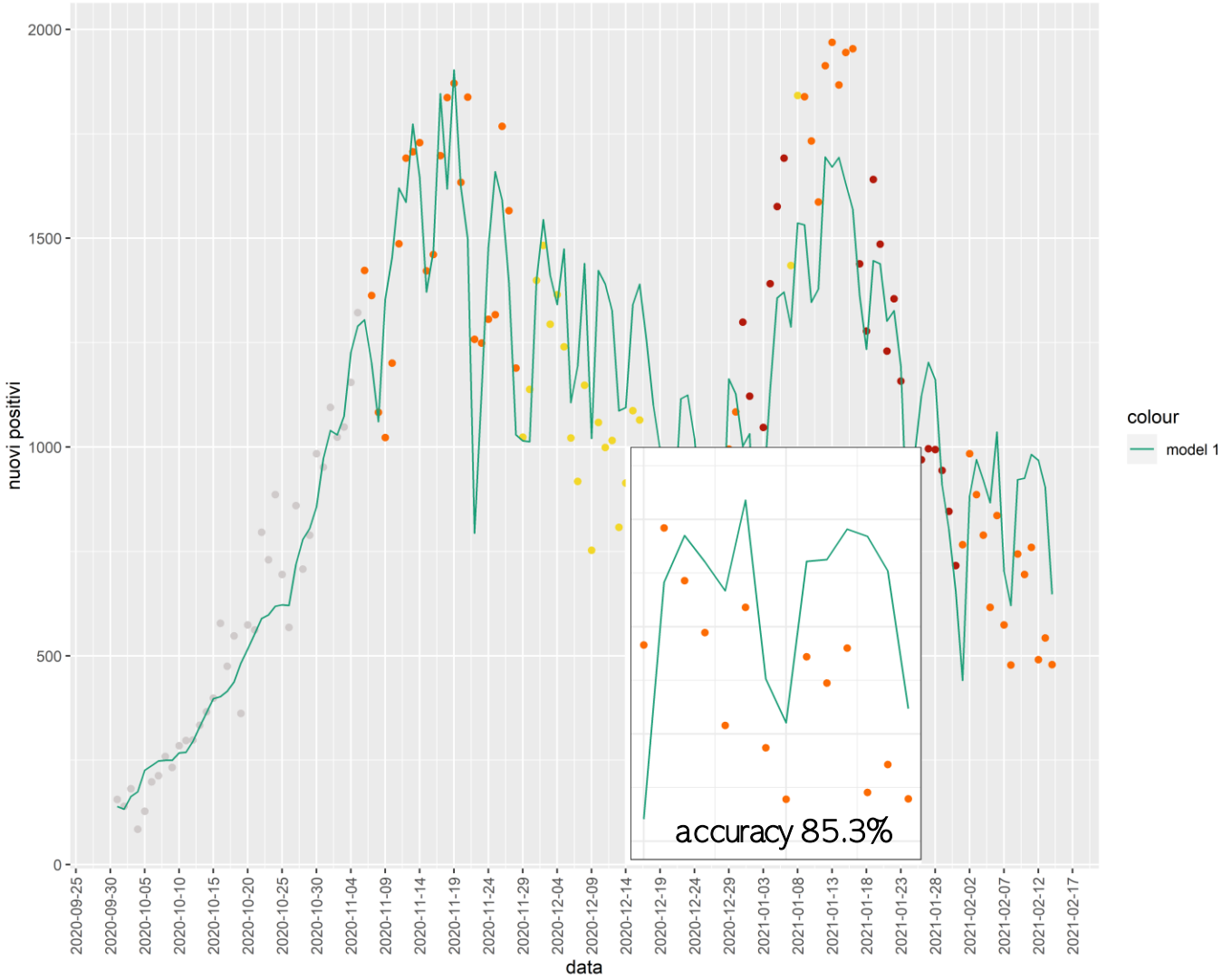
GLM

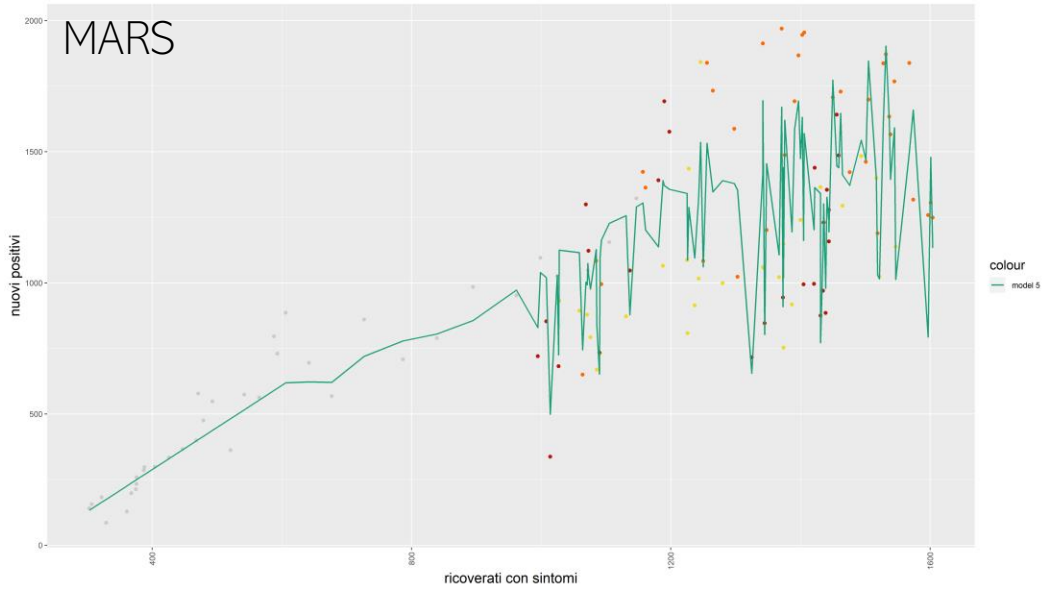


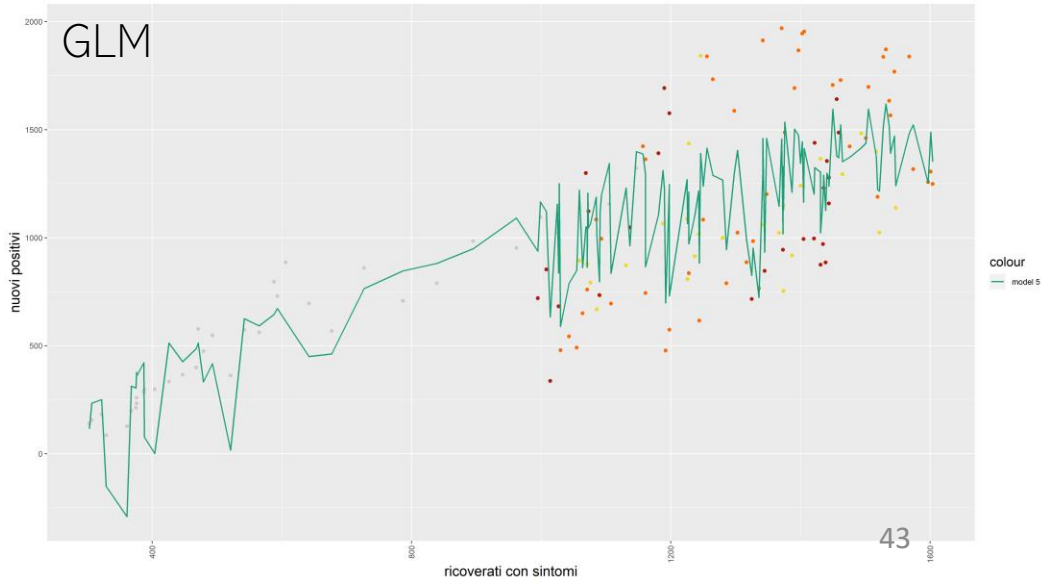
MARS

Non-parametric methods

y ~ ricoverati_con_sintomi + nuovi_tamponi + livello_di_restrizioni



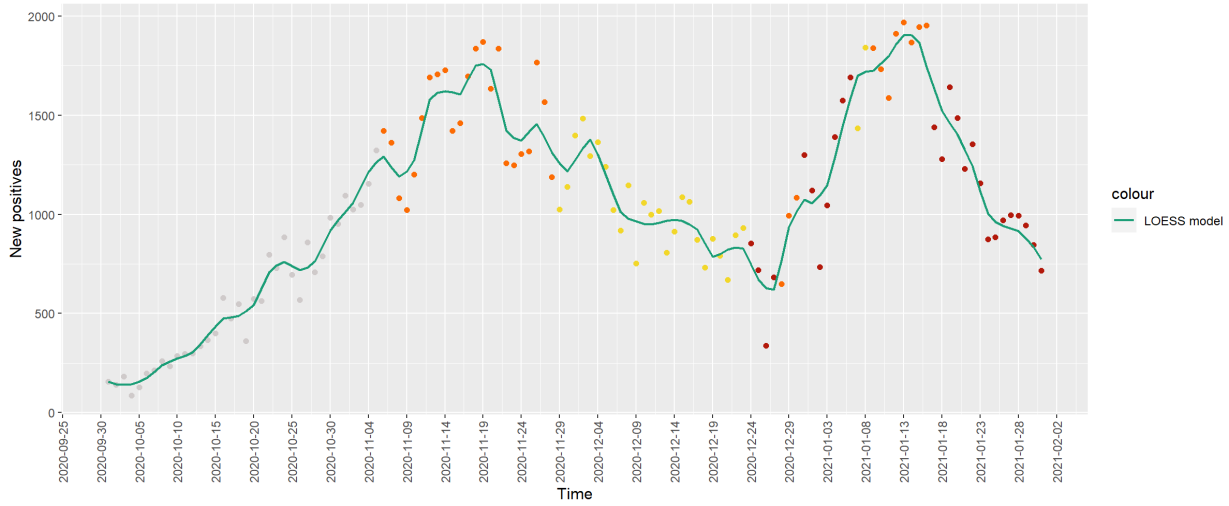




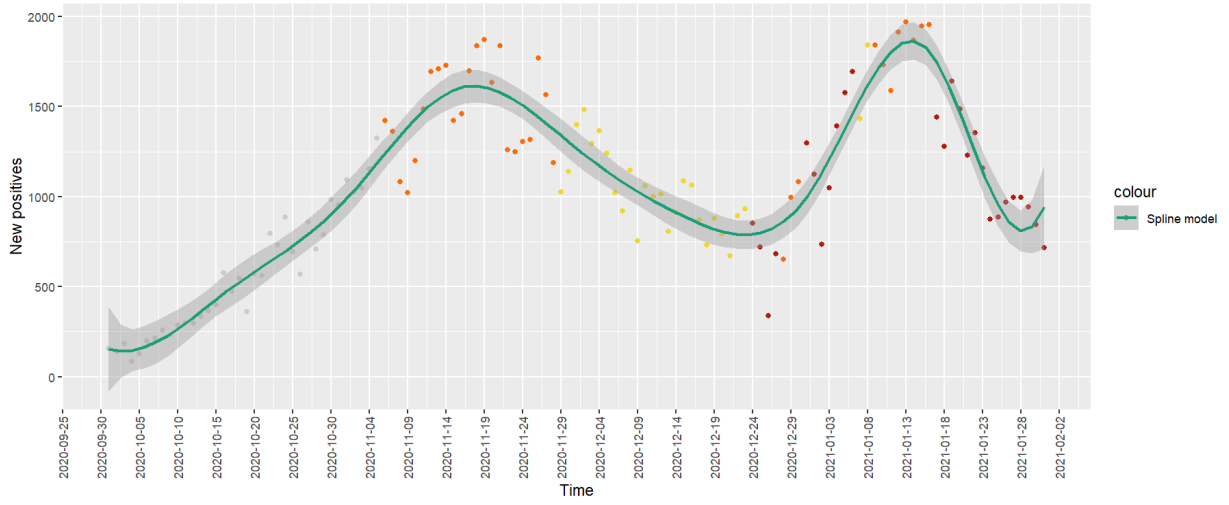
Non-parametric methods

- We also tried to fit the model using as covariate the variable `data`, even if there is not a real cause-effect relationship between this variable and our response variable.

- Anyway, it can be seen how both LOESS and SPLINE methods can catch the trend of our response variable over time.



Local Polynomial Regression



Spline

Conclusions

Conclusion

- LM and GLM models fits well, however there seems to be an apparent overfitting. This issue may be caused by the behavior of `nuovi_tamponi_pcr`. This could be due to a very strong day-of-week effect that dominates the subtler patterns.
- LM and GLM are quite good even if there is still space for improvement.
  - LM have a higher variance and a non-normal residual distribution. Running the Shapiro-Wilks test we assessed that the hypotesis of a gaussian distribution for our response variable is not respected.
  - GLM are worse with respect to the linearity of residuals, and they present numerous outliers, even if this issue is mitigated by employing negative binomials.

- Predictions are quite good on average, and they improve a lot when new covariate `livello_di_restrizioni` is introduced.
- Non-parametric regression shows a better understanding of the irregularity of the set but checking the accuracy and compare models is more difficult. Several metrics employed in assess the quality of the fit for LM and GLM cannot be used in a non-parametric framework.

Sitography - R packages used

- Data Analysis Using Regression And Multilevel/Hierarchical Models,Jennifer Lynn Hill,Andrew Gelman, Cambridge University Press

- Lüdecke D, Ben-Shachar M, Patil I, Waggoner P, Makowski D (2021). "performance: An R Package for Assessment, Comparison and Testing of Statistical Models." *Journal of Open Source Software*, **6**(60), 3139. doi: 10.21105/joss.03139.

- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

- Wang W, Yan J (2021). *splines2: Regression Spline Functions and Classes*. R package version 0.4.5, https://CRAN.R-project.org/package=splines2.

Thank you for
the attention