Statistical Method for Data Science Project. The prediction of the number of Covid-19 patients in intensive care units: A case study in Campania.

Group menbers: Baurice Nafack, Cecilia Zagni and Walter Nadalin

February 23, 2022

Table of Contents

Exploratory Data Analysis

Description of the Dataset

The dataset for this study was obtained from the official website of the Protezione Civile [1] for the outbreak of Covid-19. The variables that we choose to consider to start our analysis are:

- Date
- Hospitalized.with.symptoms
- ICU
- Total.Hospitalized
- People.at.home
- Total.positives
- New.positives
- Discharged.healed
- Deceased
- Total.cases
- Covid.tests
- Cases.tested
- ~~ICU.Daily.Admissions~~

But we have to discard ICU.Daily.Admissions because more than the 45% of the data were NA.

[1] Source: https://github.com/pcm-dpc/COVID-19

Description of the Dataset

We also include:

- Color[2] -> Color of the region in a given day, according to the administrative order of November 6.
- Percentage.vaccinated -> $\frac{Total.vaccinated^{3}}{Campania.Population^{4}}$

[2]Source: https://github.com/imcatta/restrizioni$_r$egionali$_c$ovid

[3]Source: https://github.com/pcm-dpc/COVID-19

[4]Source: https://www.tuttitalia.it/campania/statistiche/popolazione-andamento-demografico/

Variables Analysis

Figure:
Hospitalized.with.symptoms, ICU and Total.Hospitalized over time

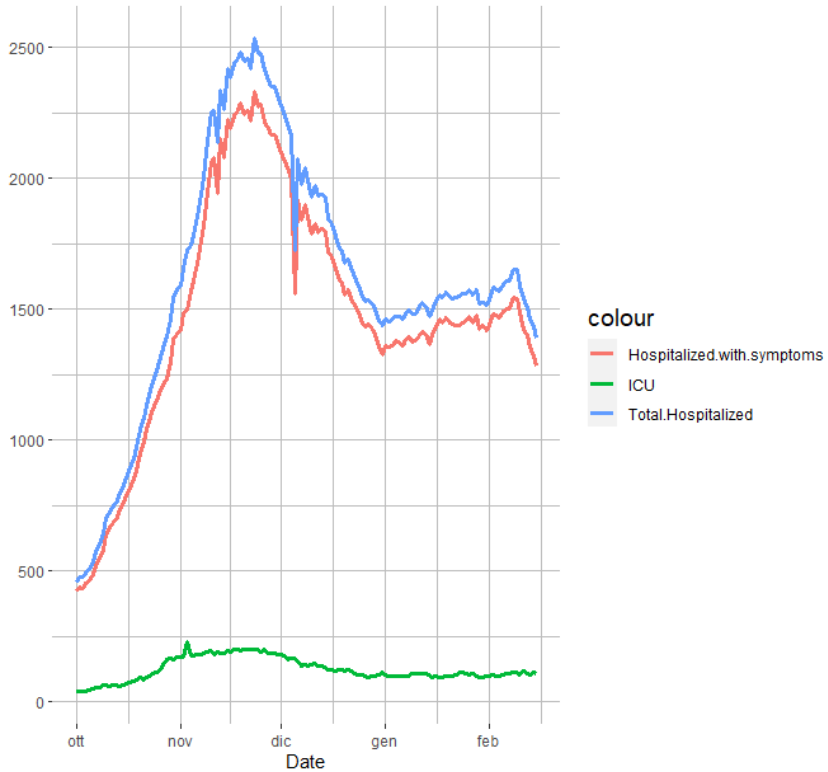
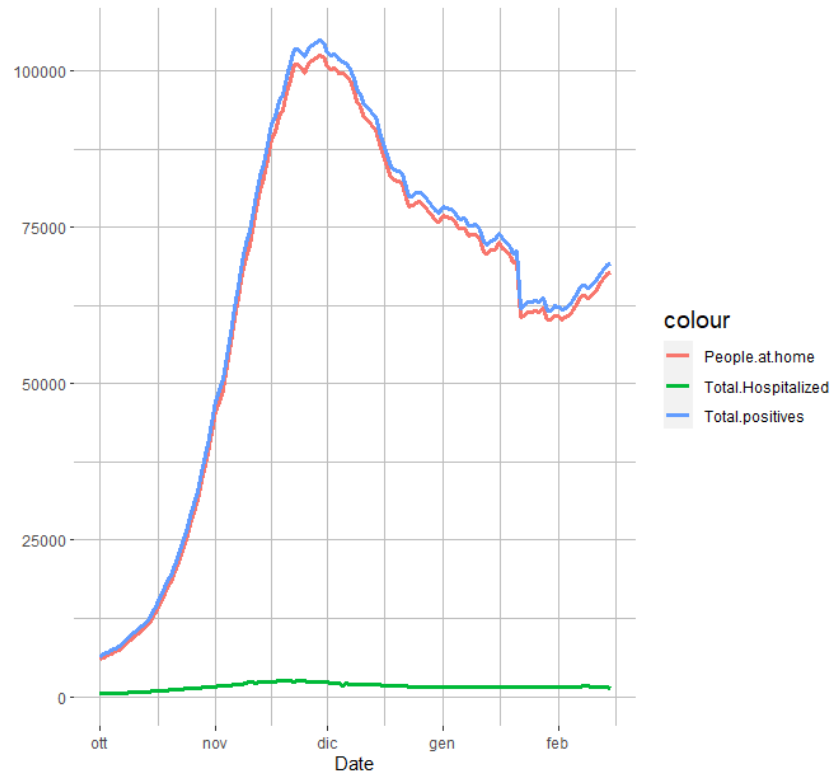


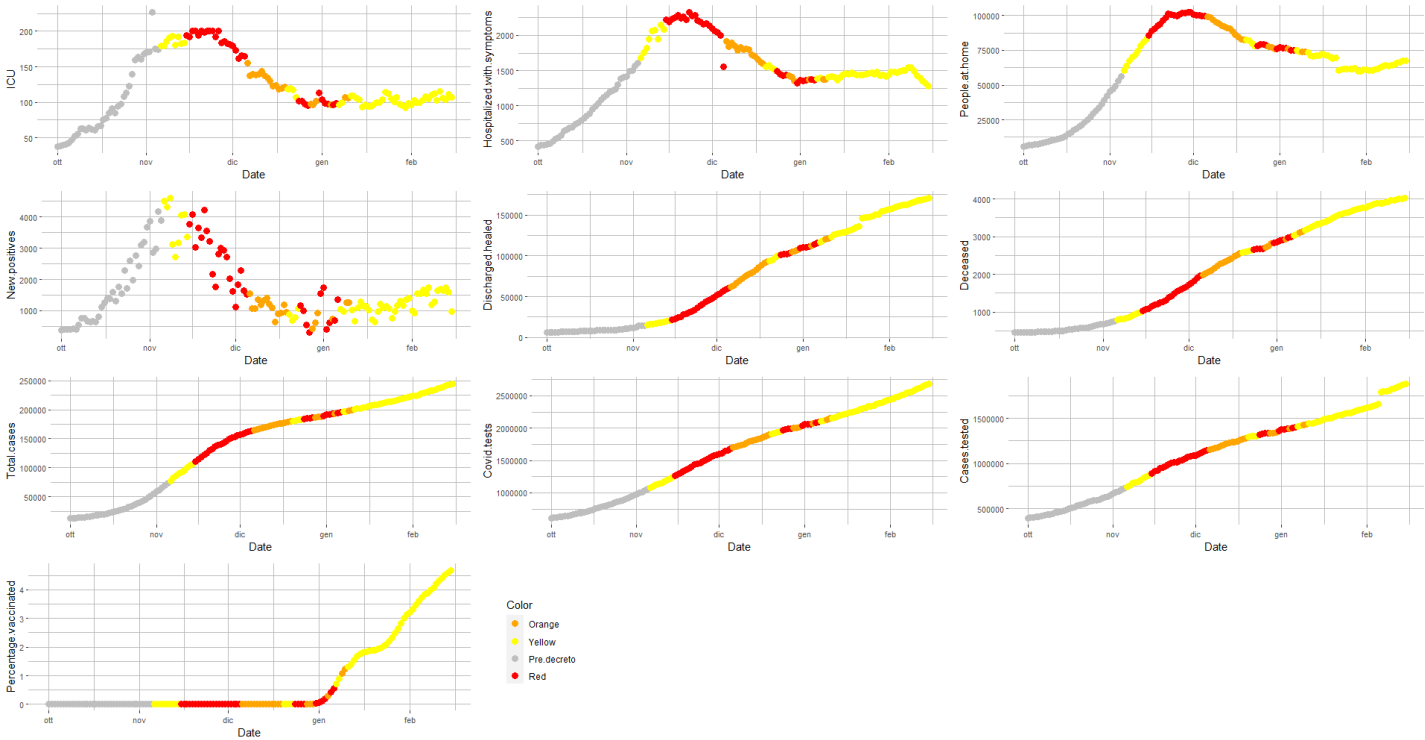
Figure: Hospitalized.with.symptoms, People.at.home and Total.positives over time



To avoid linear dependencies, we discard Total.Hospitalized and Total.positives, that are respectively the sums of Hospitalized.with.symptoms and ICU, and of Hospitalized.with.symptoms and People.at.home.

Visualization of the data

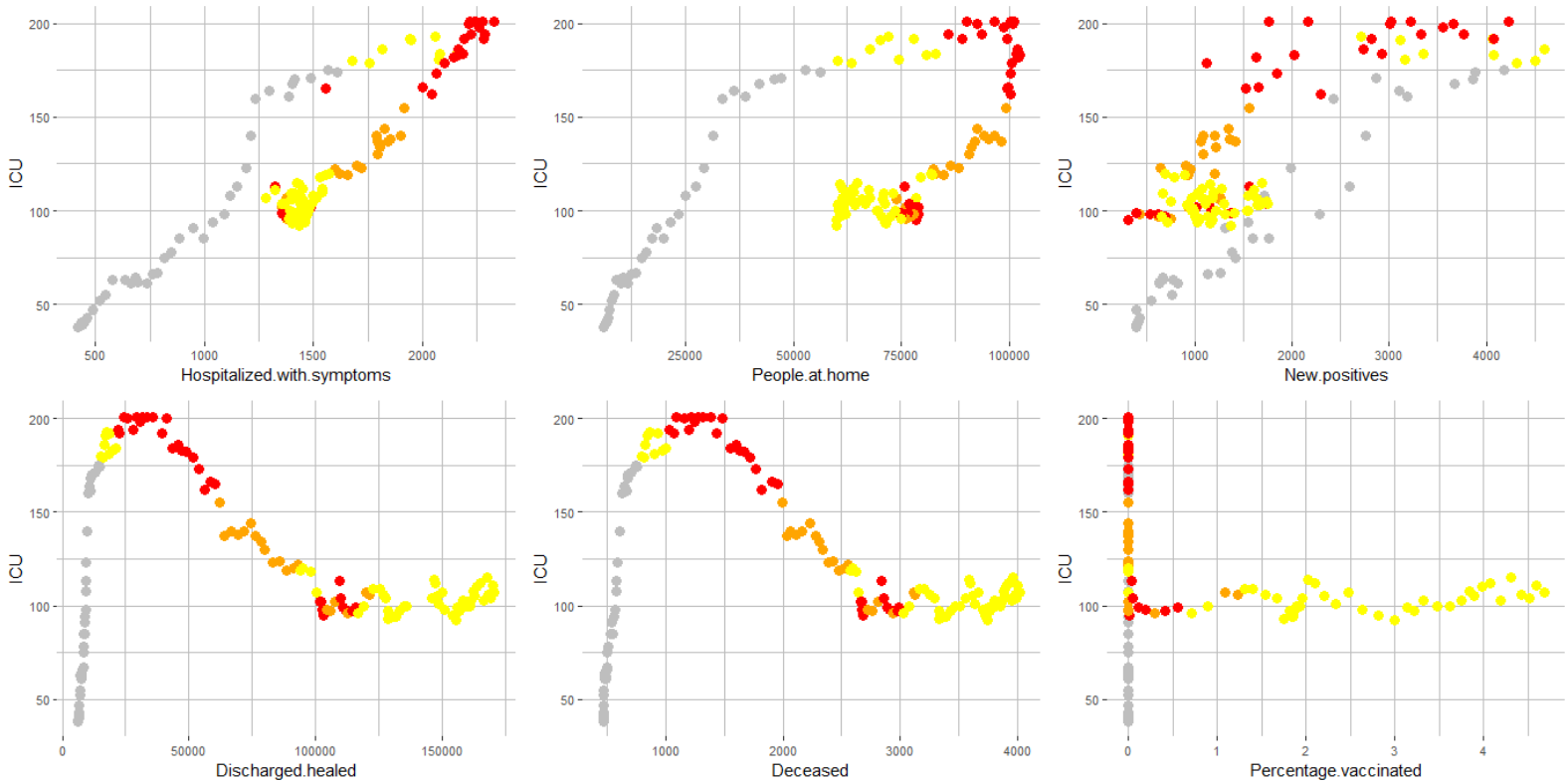
Figure: Plots of all the variables wrt Date



- Discharged.healed, Deceased, Total.cases, Covid.tests and Cases.tested are cumulative variables. Since we don't think that Total.cases, Covid.tests and Cases.tested can be relevant in modeling ICU, and to avoid linear dependencies, we choose to discard them.
- There is an outlier -> the value of ICU of November 3 -> we choose to discard it.
- There is a suspicious point -> the value of Hospitalized.with.symptoms that is clearly far from the overall trend -> for now we choose to keep it in our analysis.

Visualization of the data

Figure: Plots of ICU wrt all the other variables



- The dependency of ICU on each of the single variables doesn't seem linear.
- As we expected, since the vaccination campaign only starts on December 27, the covariate Percentage.vaccinated doesn't seem to influence the ICU -> we choose to discard it.

Attempt of data transformation

We try to apply some transformations on Discharged.healed and Deceased to see if we can obtain a more linear dependency of the ICU variable:

Figure: Logarithm and square root of Deceased

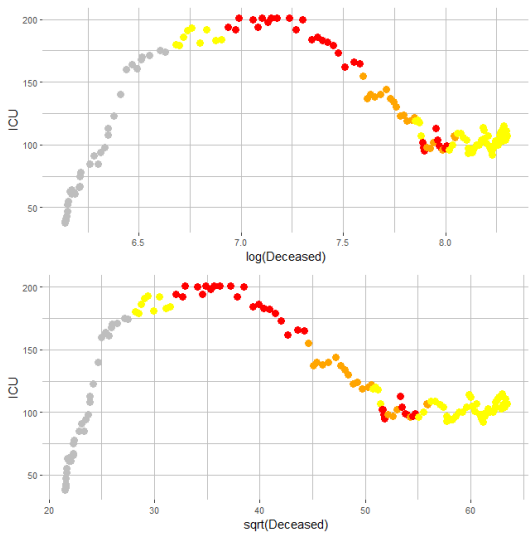
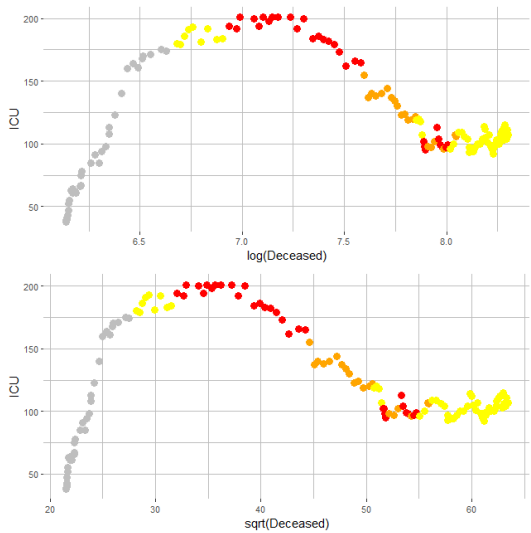


Figure: Logarithm and square root of Discharged.healed



We don't obtain any improvement, so we keep the original variables.

In conclusion, the variables that we are going to consider to build a model for the number of people in ICU are:

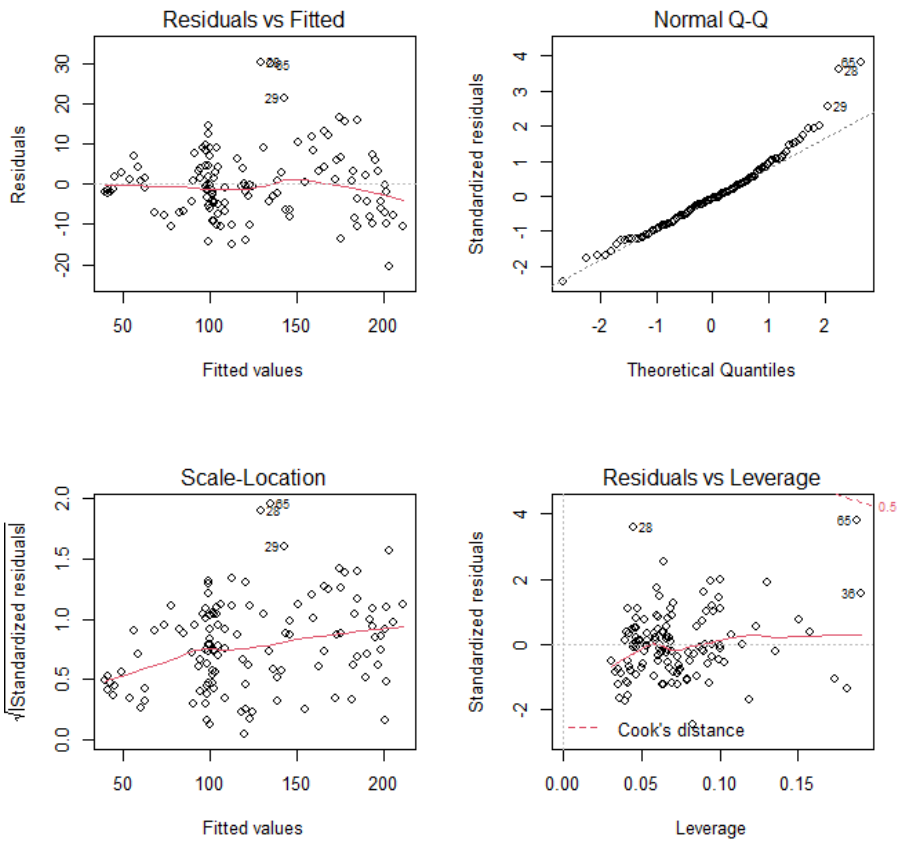
- Hospitalized.with.symptoms
- People.at.home
- New.positives
- Discharged.healed
- Deceased
- Color (considered as a factor)

Classical Approach

Linear Model

We start fitting the **complete linear model**, considering all the variables that we kept after the Exploratory Data Analysis.

Figure: Analysis of the residuals



The assumptions of **Normality** and **Homoscedasticity** look respected.

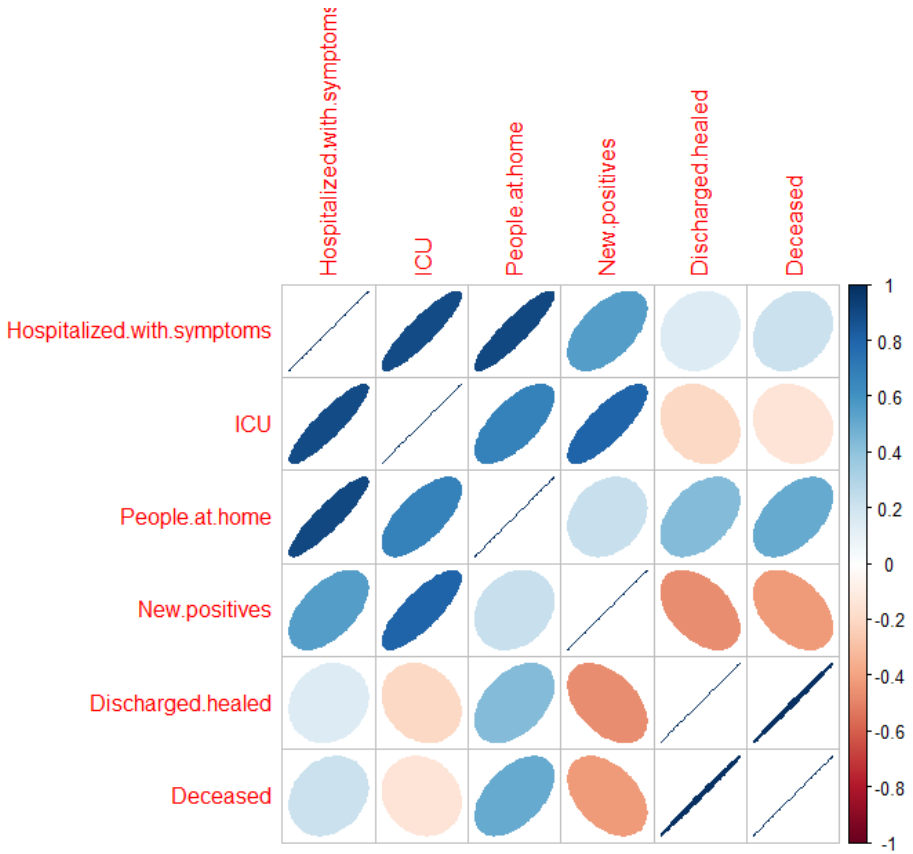
Metrics:

- AIC = 539.58
- Adj-R^2 = 0.964

However some of the covariates (Discharged.healed and Deceased) seem not to be significant (p-value $>$ 0,05).

Linear Model: Correlation Analysis

Figure: Correlation Plot



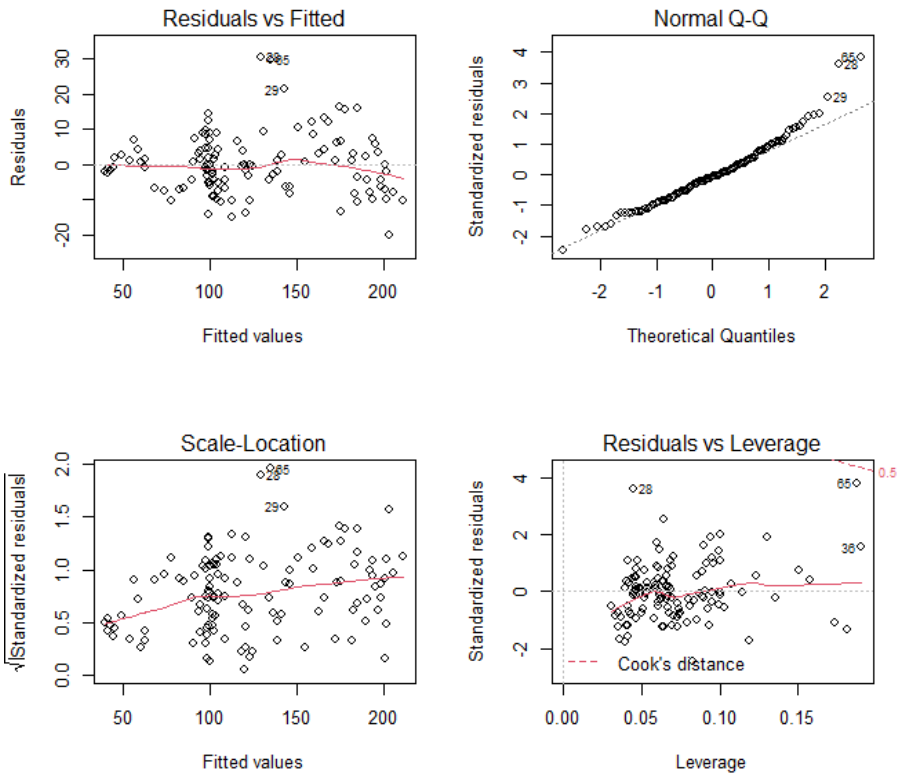
We notice that some variables are highly correlated, especially Discharged.healed and Deceased.

To overcome this problem, we try to update our model...

Linear Model: Stepwise Procedure

We update the model, discarding Discharged.healed (the covariate with the highest p-value) and we obtain:

Figure: Analysis of the residuals



The assumption are still respected.

The indexes are not significantly improved:

- AIC = 539.70
- Adj-R^2 = 0.964

For the *Occam's razor principle*, we **choose the reduced model**.

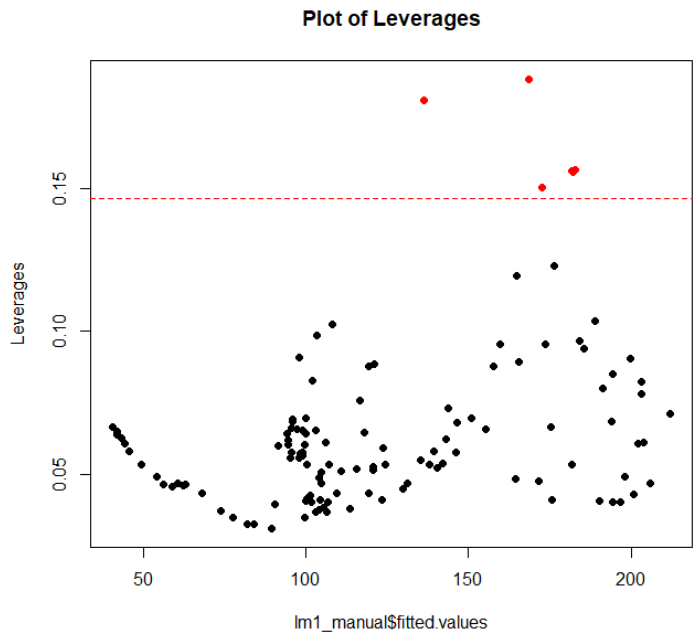
This way we can also overcome the correlation problem.

Linear Model: Detecting Influential Points

During the Exploratory Data Analysis, we detected an outlier and decided to discard it from our analysis.

Now we want to find out if there are **other points that significantly affect the estimates of our model** (i.e. Leverage).

Figure: Analysis of the residuals



Linear Model: Analysis of Influential Points

To understand how much the leverages influence our model, we fit the model without the leverages and then we investigate the relative variation of our coefficients:

Figure: Relative variation of the coefficients

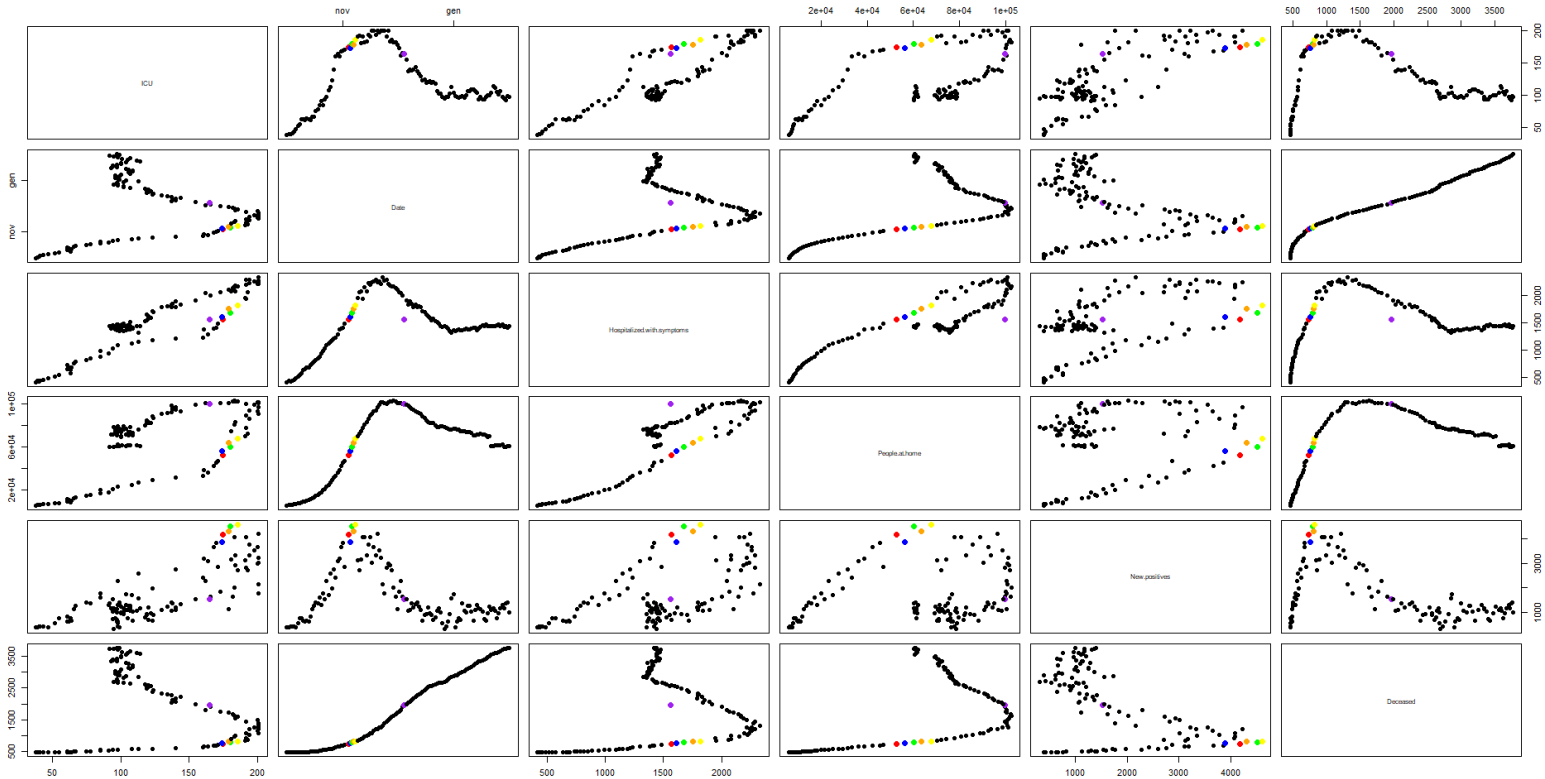
(Intercept)	Hospitalized. with. symptoms	People. at.home	New. positives	Deceased	ColorPre. decreto	ColorRed	ColorYellow
0.6595136	0.2639	0.2554	0.1645	0.5498	0.2449	0.2962	0.2312

The leverages affect the estimate heavily (there is a variation of 16% at least).

Linear Model: Analysis of Influential Points

To which values do the influential points correspond?

Figure: Influential points over the dataset

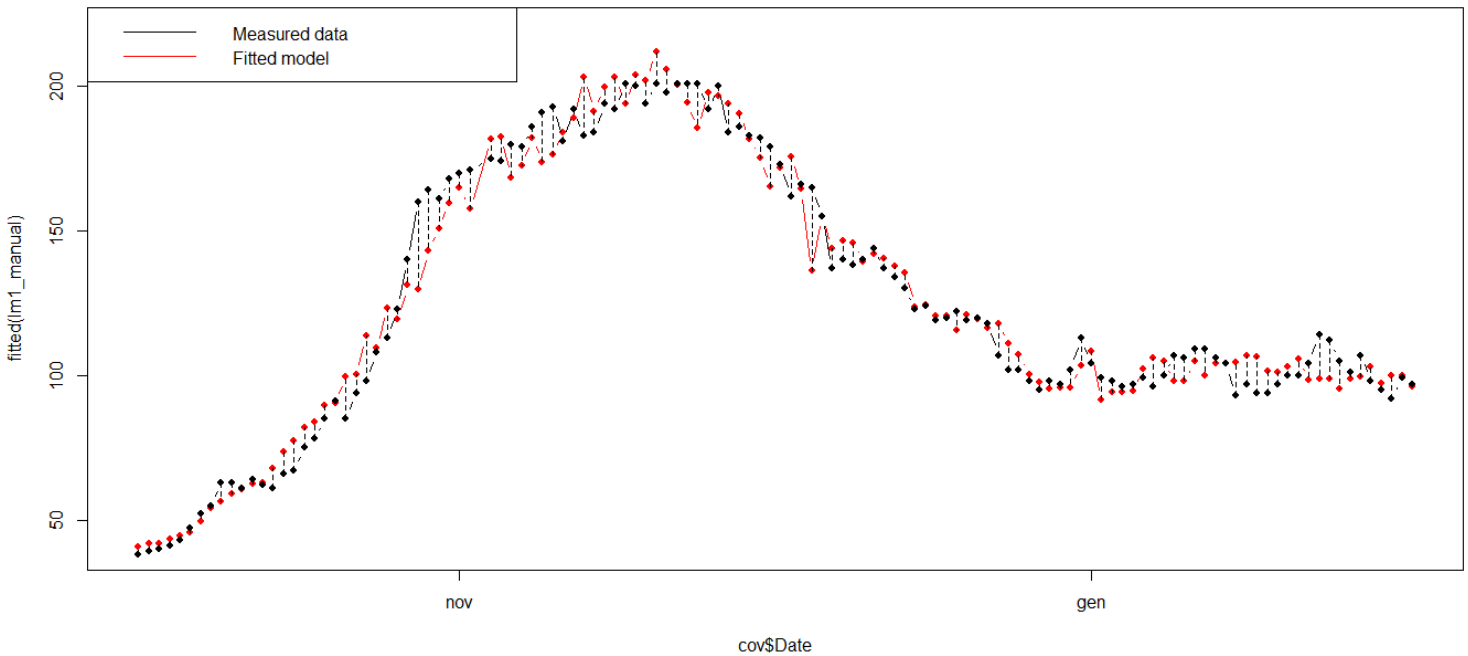


- November 4, 5, 6, 7 -> Introduction of the Region colour administrative order
- 5 December -> Suspicious value for Hospitalized.with.symptoms -> we choose to discard this value

Linear Model: visualizing the results

With the selected model we obtain the following fit:

Figure: Comparing fitted and mesured values



Metrics: AIC = 521.29 ; Adj-R^2 = 0.968

The result seems to be generally very noisy. We continue our analysis to find an improvement...

Generalized Linear Model

We fit the **complete GLM**, assuming a Poisson distribution for the response variable ICU.

Figure: GLM Summary

```
Call:
glm(formula = ICU ~ . - Date, family = poisson, data = cov_glm1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3821  -0.6366  -0.0702   0.5177   3.4733

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               3.394e+00  1.246e-01  27.233  < 2e-16 ***
Hospitalized.with.symptoms 6.134e-04  8.401e-05   7.301 2.86e-13 ***
People.at.home            7.192e-06  2.194e-06   3.278  0.00104 **
New.positives             9.396e-05  1.354e-05   6.940 3.91e-12 ***
Discharged.healed         1.403e-05  6.624e-06   2.119  0.03412 *
Deceased                 -6.322e-04  3.133e-04  -2.018  0.04359 *
ColorPre.decreto          4.173e-01  6.748e-02   6.183 6.27e-10 ***
ColorRed                 -8.452e-02  2.867e-02  -2.948  0.00320 **
ColorYellow               3.328e-02  3.603e-02   0.924  0.35555
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2119.59  on 121  degrees of freedom
Residual deviance:  112.98  on 113  degrees of freedom
AIC: 934

Number of Fisher Scoring iterations: 4
```

Good results:

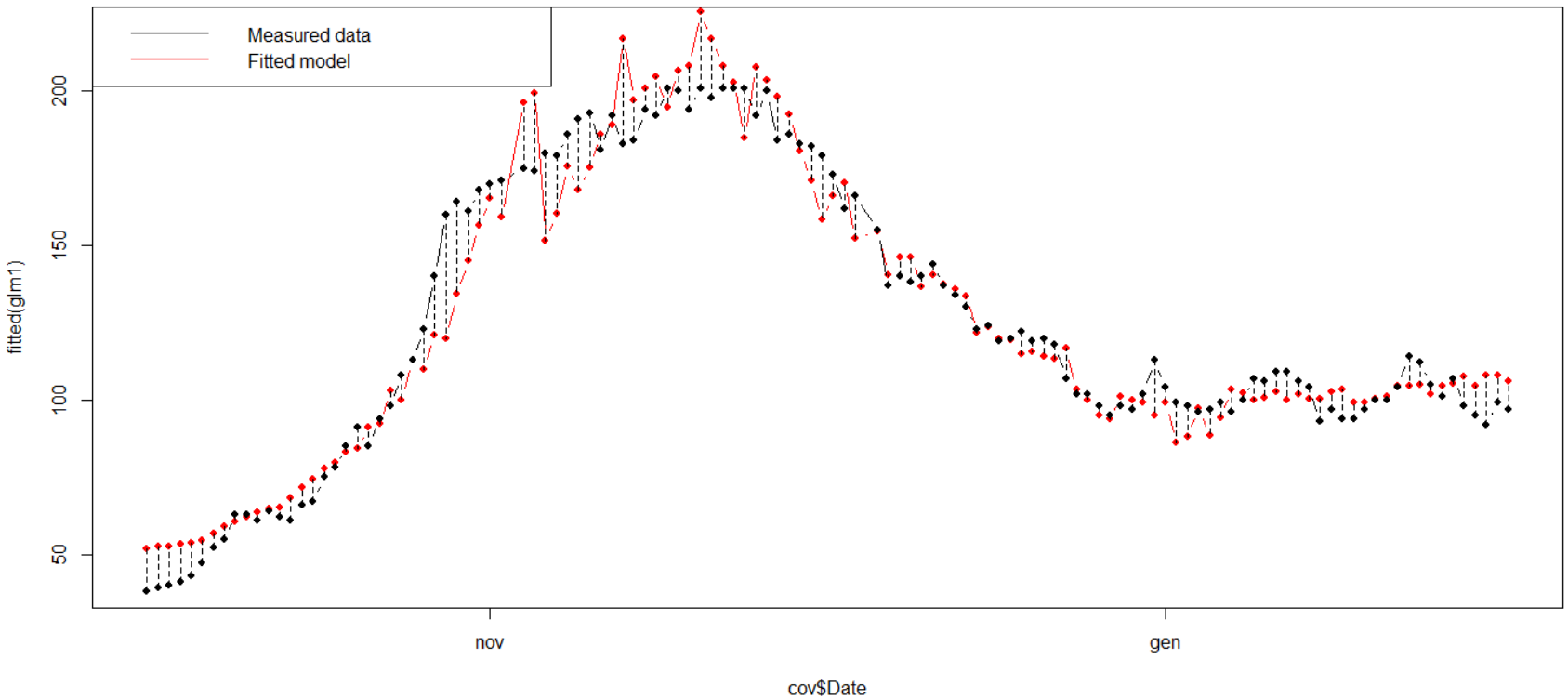
- High null deviance -> it makes sense to use more than a single parameter for fitting the model.
- The residual deviance is relatively low (and close to the number of degrees of freedom) -> the log likelihood of our model is close to the log likelihood of the saturated model.
- High p-value for the deviance goodness of fit test (0.483).
- All the covariates have a significant p-value.

But the metrics is worse than the linear model: AIC = 933.99

Generalized Linear Model: visualizing the results

With the selected model we obtain the following fit:

Figure: Comparing fitted and mesured values



As the metrics, also the plot shows that the GLM is worse than the chosen LM. Let's see if we can improve our model...

Generalized Additive Model

Once again, we start fitting the **complete GAM**, including the categorical variable Color as a linear term.

Figure: GAM Summary

```
Family: poisson
Link function: log

Formula:
ICU ~ +s(Hospitalized.with.symptoms) + s(People.at.home) + s(New.positives) +
    s(Discharged.healed) + s(Deceased) + Color

Parametric coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.73158    0.03830 123.530   <2e-16 ***
ColorPre.decreto  0.03331    0.09570   0.348    0.728
ColorRed         -0.00303    0.03339  -0.091    0.928
ColorYellow       0.01604    0.04065   0.395    0.693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                edf Ref.df Chi.sq  p-value
s(Hospitalized.with.symptoms) 1.000  1.000 16.959 3.87e-05 ***
s(People.at.home)             4.950  6.066 81.514  < 2e-16 ***
s(New.positives)              1.215  1.391  0.959    0.376
s(Discharged.healed)          1.000  1.000  0.057    0.811
s(Deceased)                   1.000  1.000  0.023    0.880
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.984   Deviance explained = 98.4%
UBRE = -0.50101  Scale est. = 1         n = 122
```

Metrics:

- Adj-$R^2 = 0.984$
- AIC $= 863.89$

We immediately notice that a lot of covariates are not significant. Let's try to simplify the model...

Generalized Additive Model: stepwise procedure

We perform a backward step-wise procedure, to discard all the non-significant covariates, and we obtain:

Figure: GAM Summary

```
Family: poisson
Link function: log

Formula:
ICU ~ +s(Hospitalized.with.symptoms) + s(People.at.home) + s(Discharged.healed)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.74570    0.00879   539.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                edf Ref.df Chi.sq p-value
s(Hospitalized.with.symptoms) 1.000  1.000  42.13  <2e-16 ***
s(People.at.home)             4.883  5.949 228.30  <2e-16 ***
s(Discharged.healed)          1.000  1.000  70.89  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.983   Deviance explained = 98.3%
UBRE = -0.5697  Scale est. = 1         n = 122
```

Improvement in the metrics:

- Adj-$R^2 = 0.983$
- AIC $= 855.51$

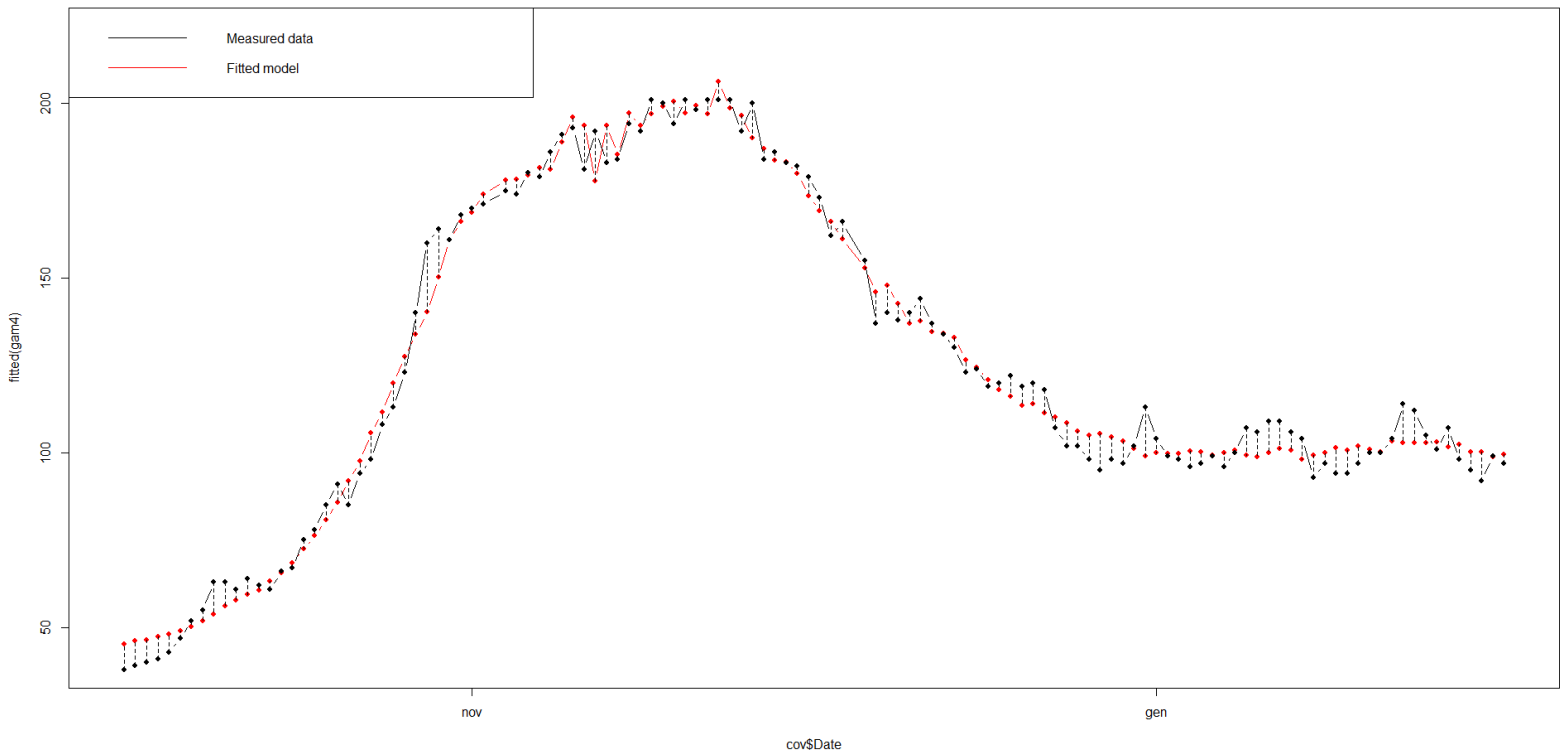
High p-value of the anova test (0.8679).

We choose to consider the **reduced model**.

Generalized Additive Model: visualizing the results

With the selected model we obtain the following fit:

Figure: Comparing fitted and mesured values



Even if the indexes are worse than the chosen LM, **the plot seems slightly** (less noise in the period December/Genuary). Can we still improve our model?

Moving windows approach

But what about the days before?

One could argue that the amount of people in the ICU in a day is not only given by the people who entered that day: **a person could stay in ICU for many days!**[5]

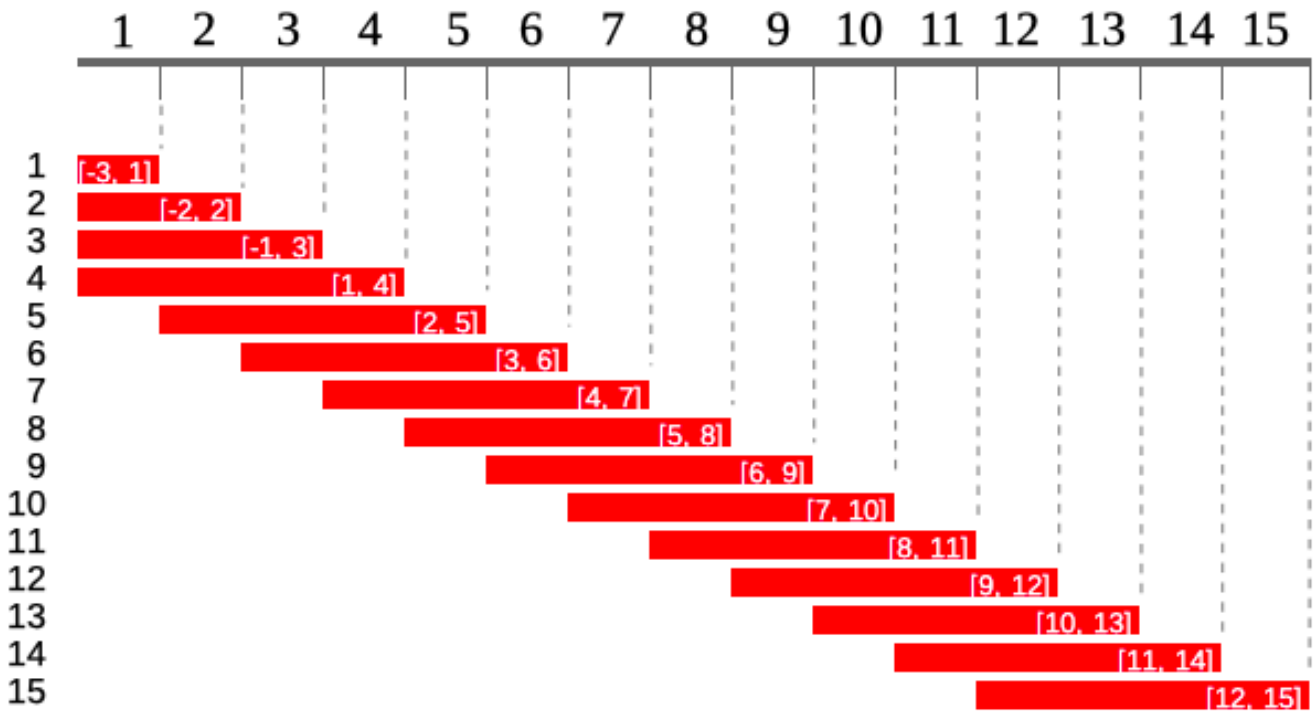
How can we take into account this fact?
We try to do it with *moving windows*

[5] *Rees E.M., Nightingale E.S. and Jafari Y* **COVID-19 length of hospital stay: a systematic review and data synthesis.** BMC Med 18, 270 (2020).

What are moving windows?

We use a **window around a sample** to calculate a local mean **The window is then slid forward** by one sample to process the next data point and so on[6]

Figure: running windows of length 4, each one of the 15 windows considers 4 elements



[6] We use the package runner to implement sliding windows in R

A first attempt: linear model

We add the **mean of the number of people occupying the ICU the 7 days before** the current and we fit a LM
Then we apply a step procedure and drop some features, we keep:

- **hospitalized with symptoms**
- **new positives**
- **colors**
- **added feature**

A first attempt: goodness of fit

Table: indices of goodness of fit before and after the step procedure

	Before	After
AIC	507.06	502.00
R^2	0.973	0.973

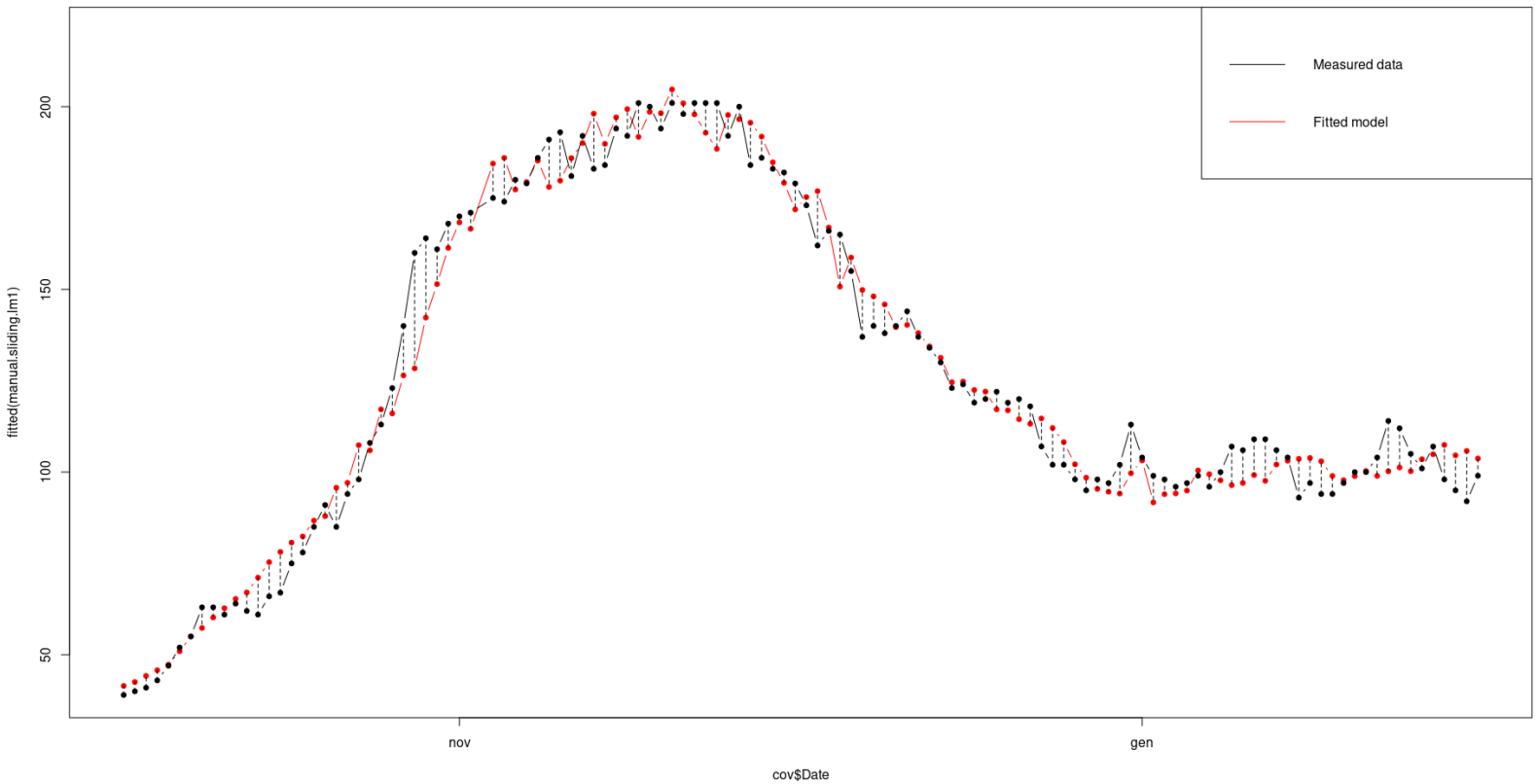
We choose the reduced one since it seems slightly better and we follow the *Occam's razor principle*

Moreover the ANOVA test using a χ^2 gives a p-value of 0.83 suggesting to reject the model with all the features

A first attempt: visualizing the results

With the selected model we obtain the following fit

Figure: comparing fitted and measured values



The result seems to be **too sensitive** during the peak of november

A first improvement: sliding window for the new positives

This results seems **better than the previous ones**, but we aren't satisfied

How can we *improve*?
Maybe the number of people in ICU today is not directly dependent on the number of **new positives** measured today, but for sure it depends on the one of the **previous days**

A first improvement: goodness of fit

Therefore we remove the new positives and add a variable that consider their mean in the 7 previous days and we keep:

- **hospitalized with symptoms**
- **people in home quarantine**
- **colors**
- **two added features**

Table: indices of goodness of fit before and after the step procedure

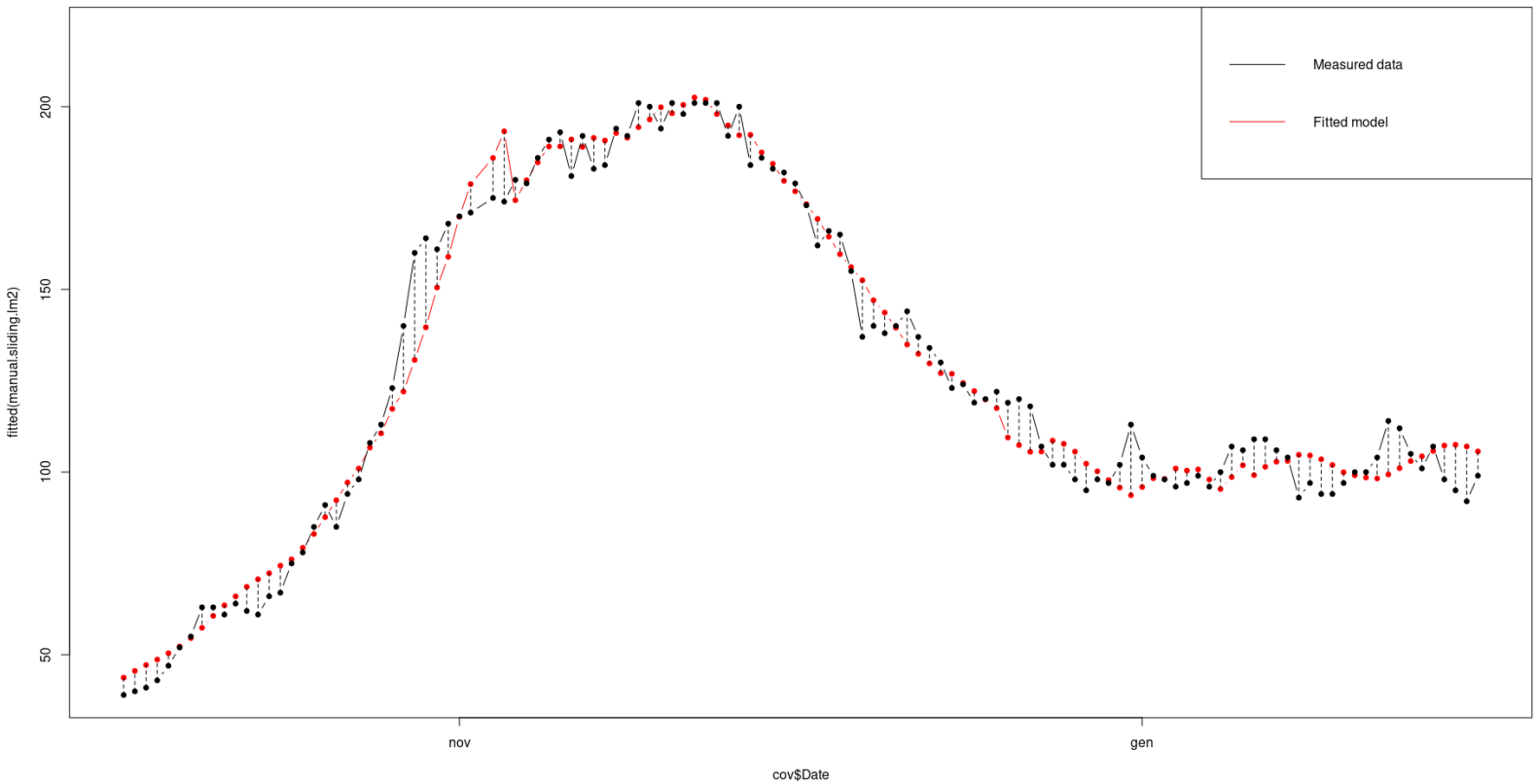
	Before	After
AIC	502.28	501.61
R^2	0.974	0.973

The ANOVA test results in a p-value of 0.21 so the reduced model is preferred

A first improvement: visualizing the results

With the best model we get the following fit

Figure: comparing fitted and measured values

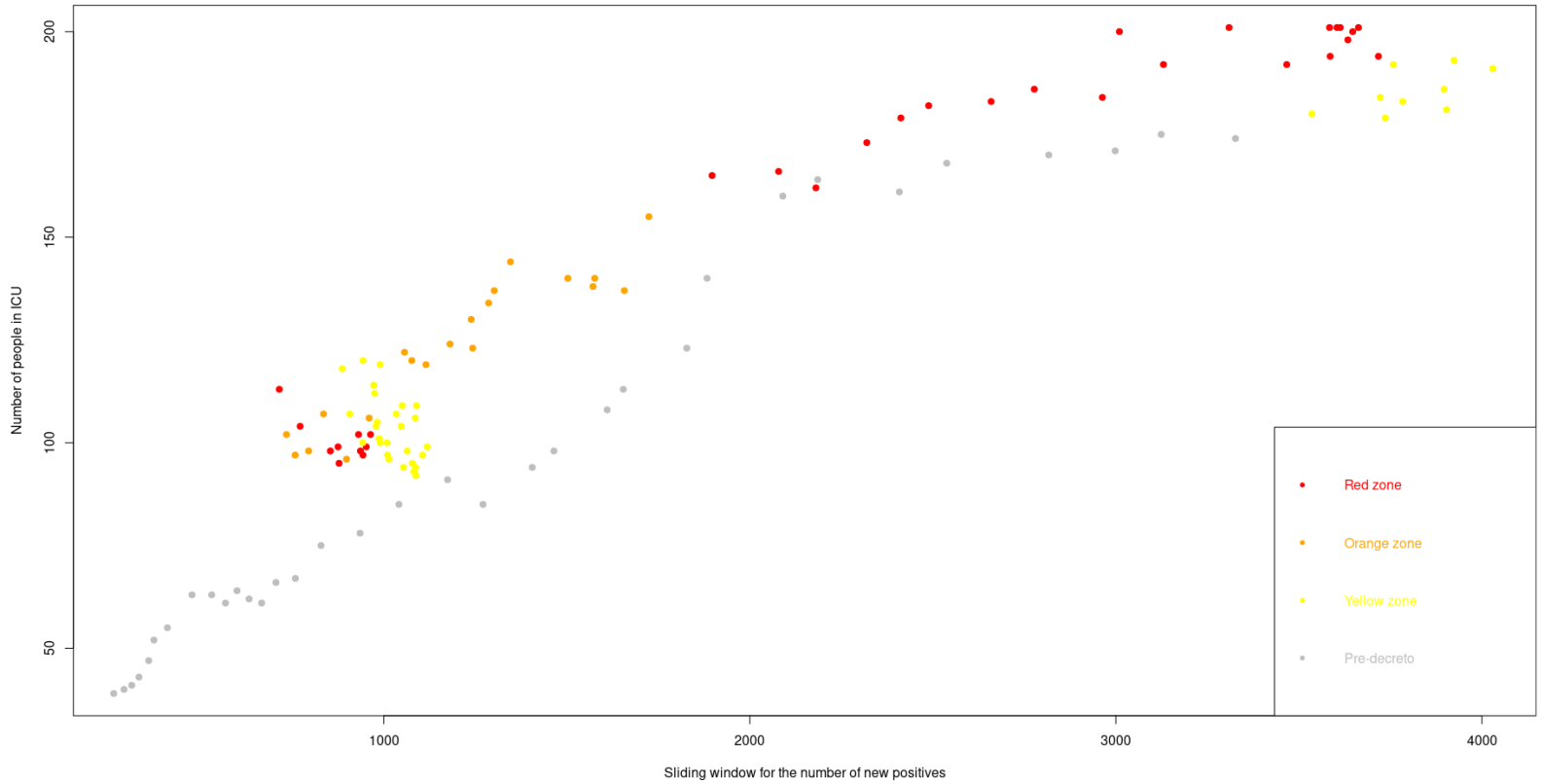


The result seems to be **better**

A final improvement: introducing a non-linear relation

But.. Let's take a look a the following plot

Figure: number of people in ICU with respect to the sliding window for the new positive



There's seem to be a **non-linear relation**

A final improvement: choosing the square root

Trying different relations we discover that a **square root** relation fits well the data

We consider the covariate with this transformation, and we keep:

- **people in home quarantine**
- **discharged healed**
- **colors**
- **two added features**

Table: indices of goodness of fit before and after the step procedure

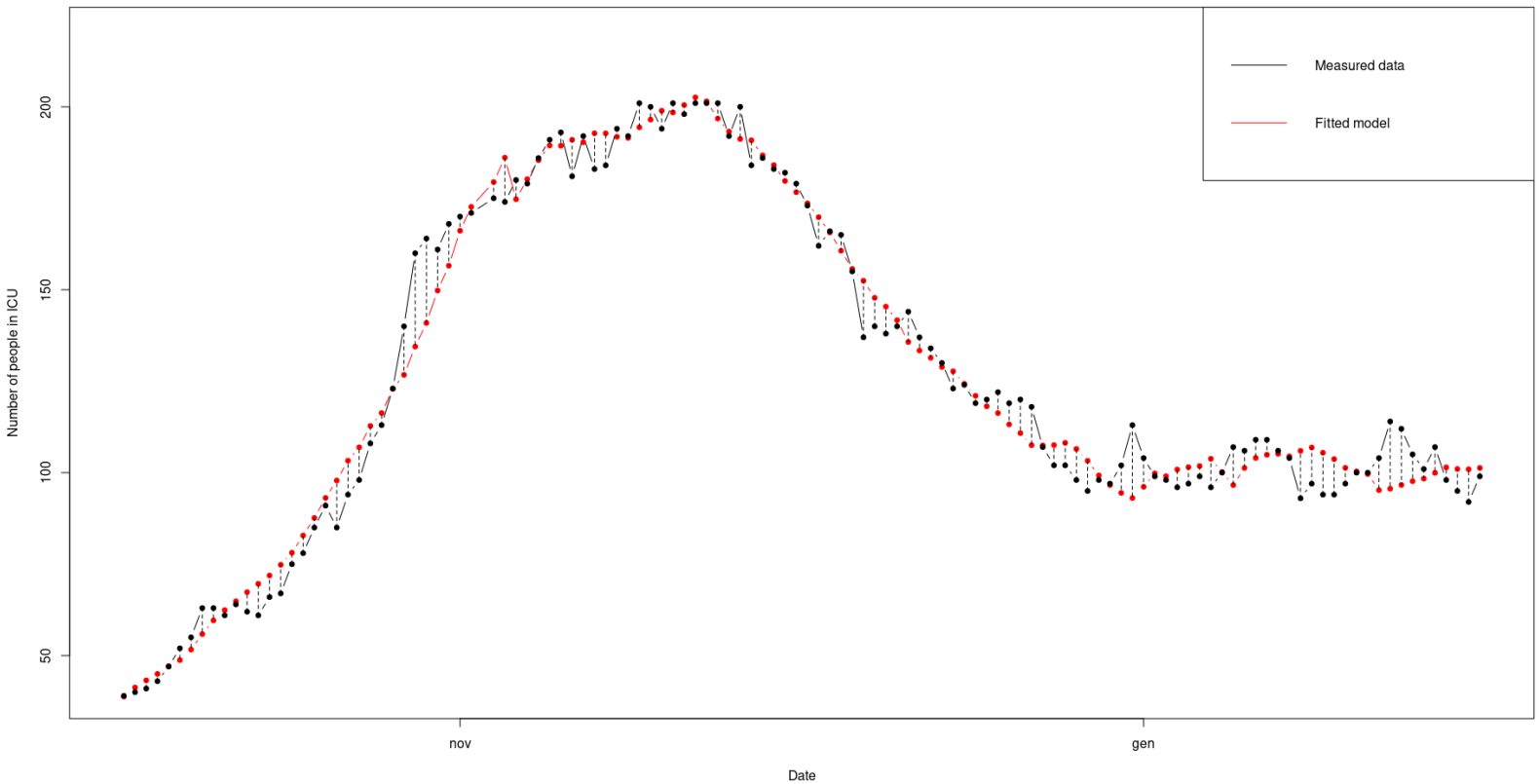
	Before	After
AIC	496.00	493.66
R^2	0.976	0.975

The ANOVA test results in a p-value of 0.47 so the reduced model is preferred

A final improvement: visualizing the results

With the best model we get the following fit

Figure: comparing fitted and measured values



The result seems to be **the best obtain so far!** It doesn't seem to get well the rapid changes of February though

Trying other models: GLM and GAM

We also consider **GLM** and a **GAM** with the same features of the best LM model and apply a step procedure to discard some features

In both case the indices and the ANOVA test suggest to keep the reduced model

Table: indices of goodness of fit of the various models

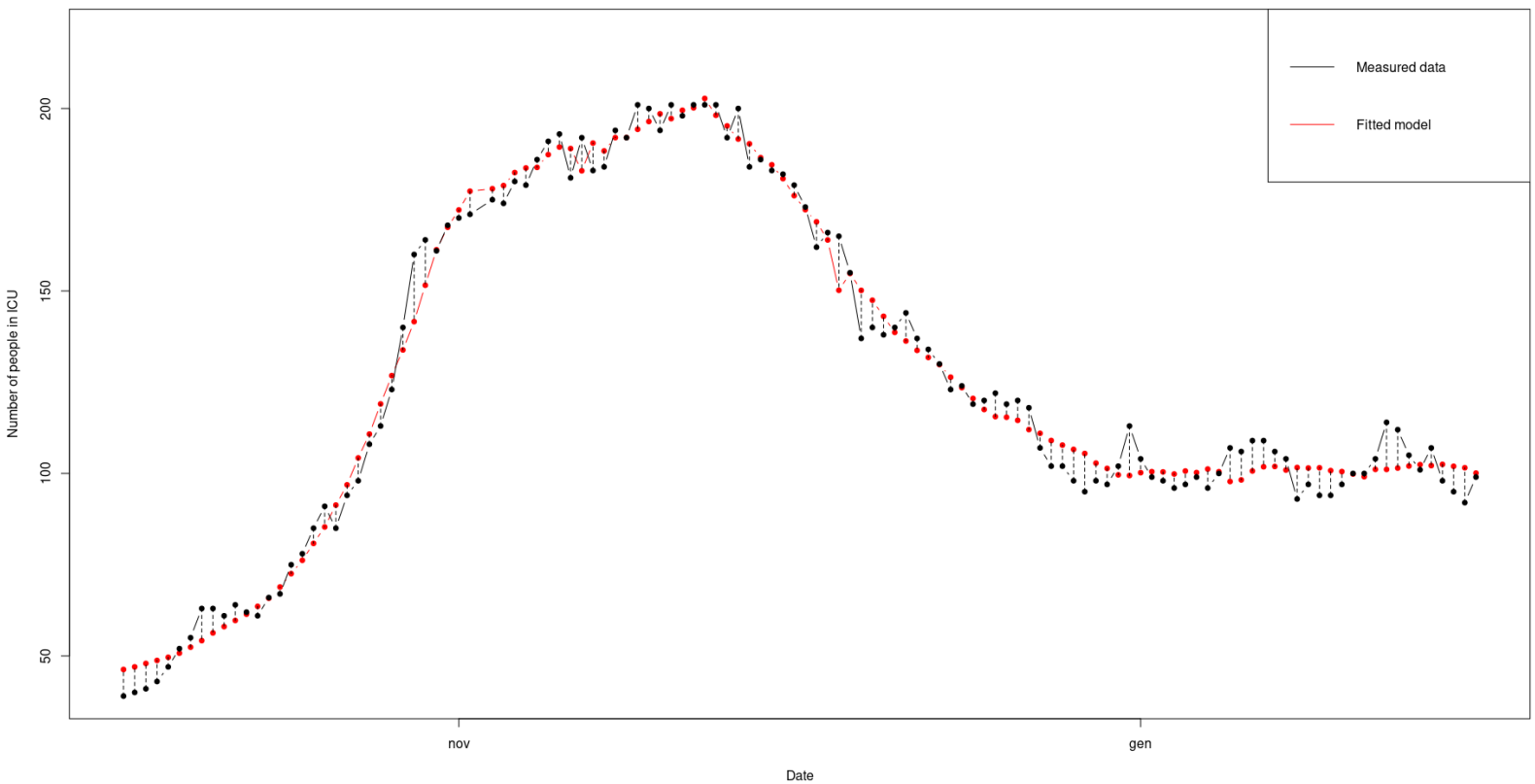
	GLM	GAM	LM
AIC	905.87	852.15	492.65
R^2	0.955	0.983	0.975

Comparing the AIC the LM seems to be best however the R^2 for the GAM is pretty good! What's up with that?

Trying other models: a more detailed look to GAM

Indeed if we look at the fitted values we see that the **GAM is less sensitive** to the rapid changes during February: this may be an hint that it could be better than the LM

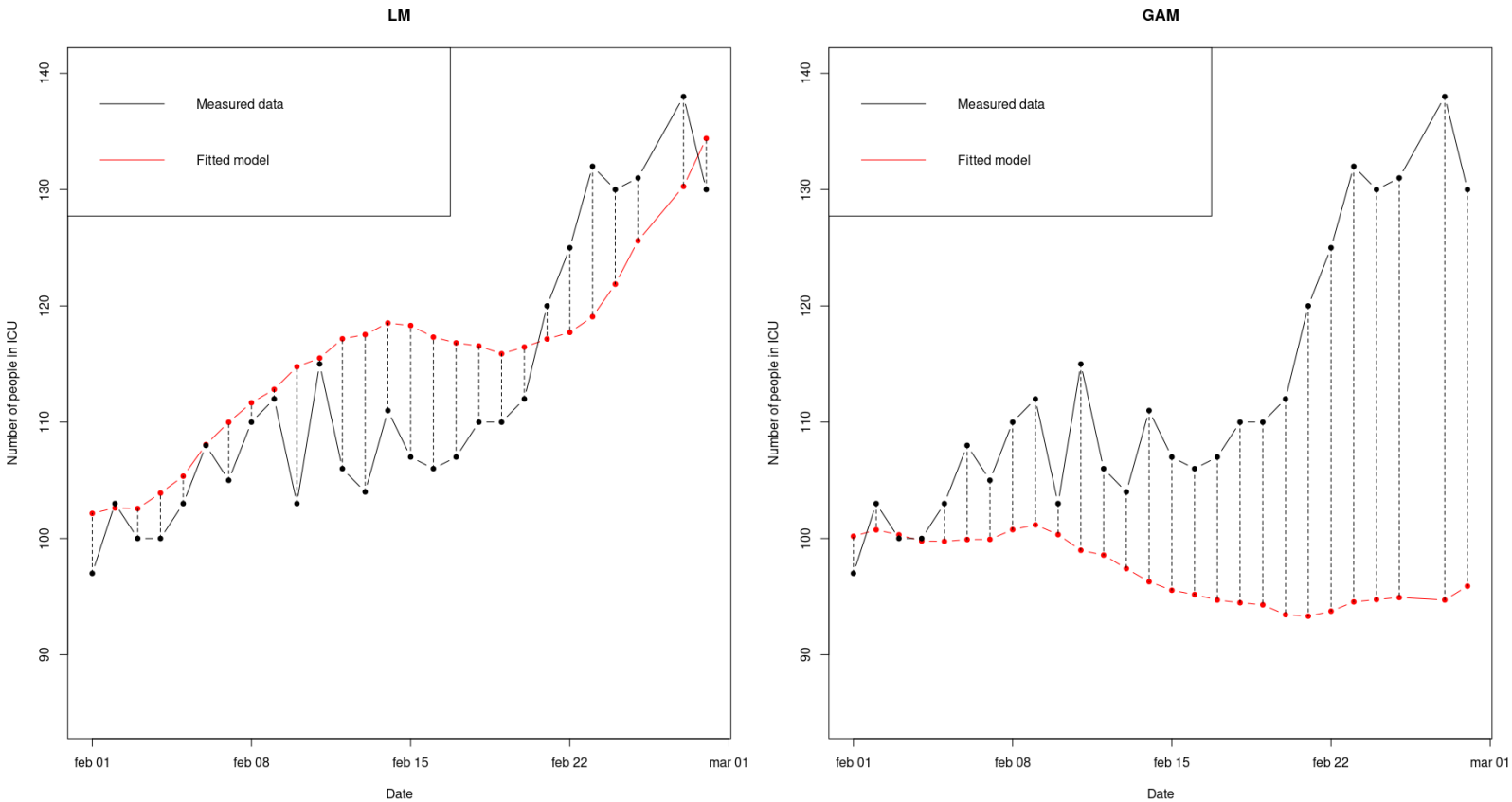
Figure: comparing fitted and measured values



Trying other models: comparing predictions of LM and GAM

In our case we can *cheat* in a certain sense and take a look at which model will **perform best on unseen data**

Figure: comparing predictions between LM and GAM for the whole month of February



Conclusions

Model chosen

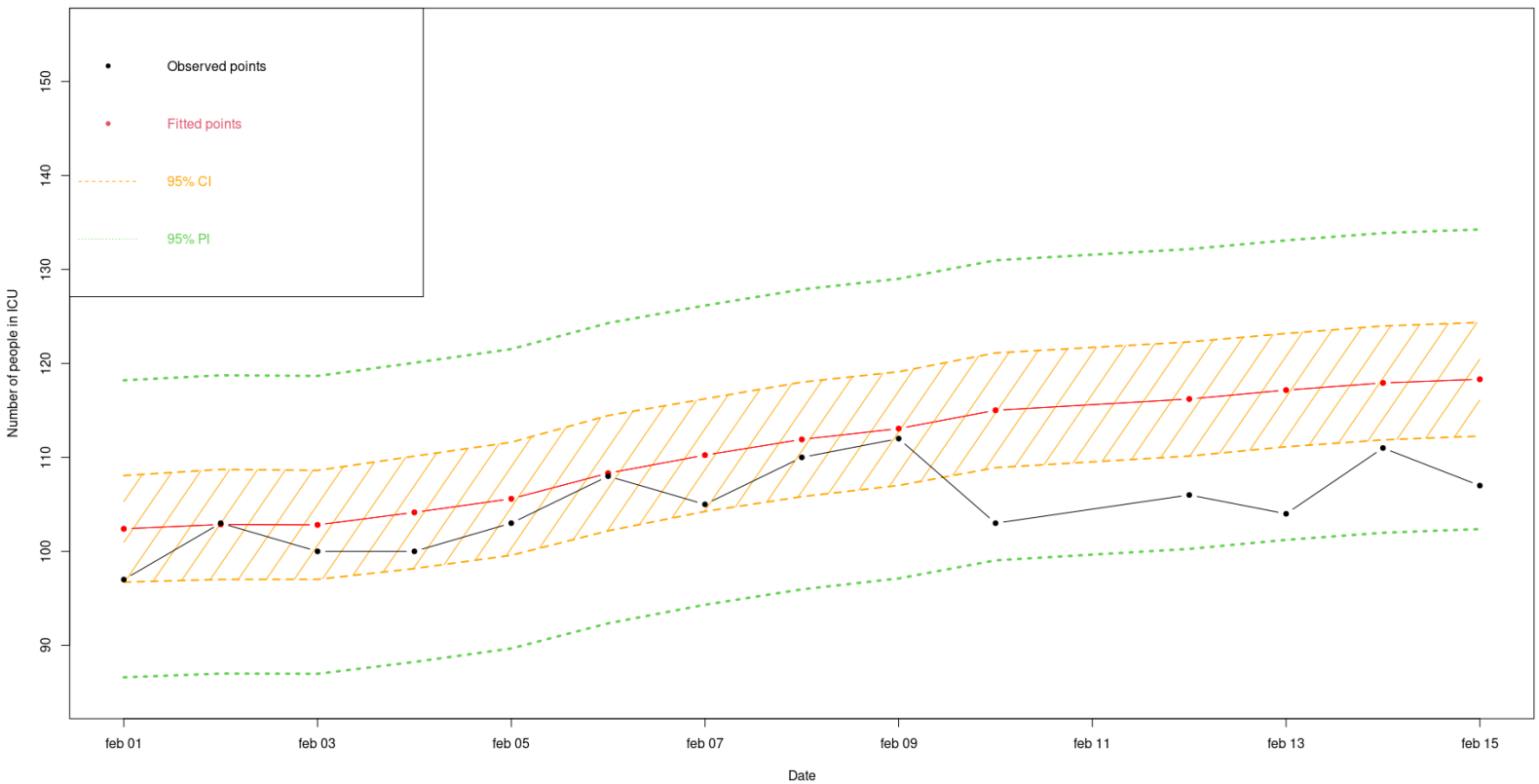
As a final outcome of everything discussed, we decided to use a LM using moving windows:

- The simplest model might give a poor description of the real system that produced the data.
- A simple explanation sometimes works best if you manage to capture well the essence of the whole system: perhaps sliding windows are a good way to simplify the description.

A detailed prediction with LM

Here we report a prediction done using the final model along with the 95% Confidence and Prediction Intervals

Figure: predictions for the first to weeks of February



As we can see all the observation fall inside the PIs estimated

Final notes and comments

The **LM seems to give fairly good results**

We observe that considering the **data relative to the past** help us to give a more accurate model: this result seems quite reasonable

The data on region **color turned out to be significant** while we decide **not to consider the data on vaccines**: probably this information becomes important if we consider a later period

Thank you for your attention!