Università degli studi di Trieste

Master Degree in

Data Science and Scientific Computing

Covid-19 case-study

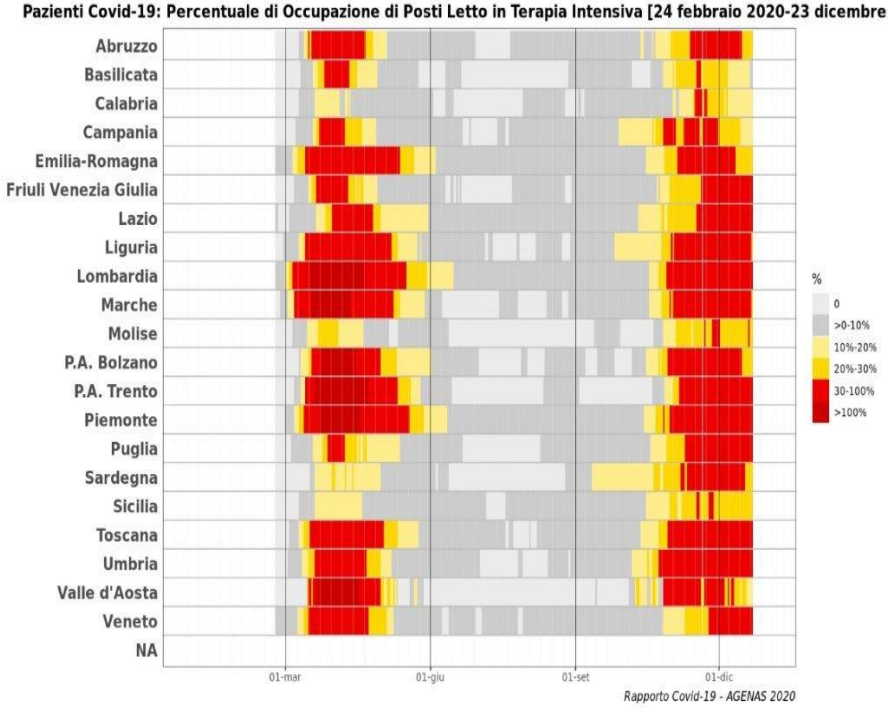
Statistical Analysis of Intensive Care in Veneto in Autumn and Winter 2020/2021

Final Project - STATISTICAL METHODS FOR DATA SCIENCE

Group A: Babaei Elham, De Santis Flavia, Doz Romina, Fodor Imola

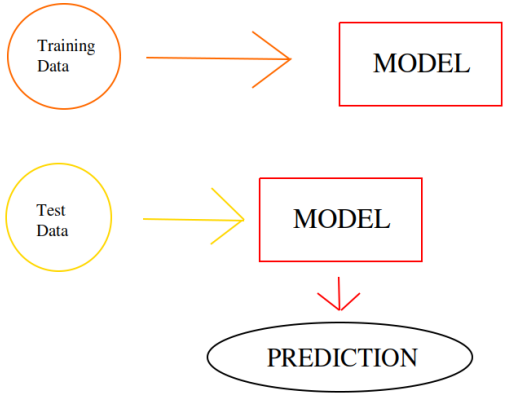
Why study intensive care

- Insufficient ICU beds to deal with covid-19 patients and also patients with other pathologies
- Increasing ICU capacity requires more equipment (in particular ventilators) and pharmaceuticals, which might be in short supply
- Increasing ICU bed numbers without increasing staff could result in increased mortality. However, doctors and nurses are not easy to find.



Why a statistical analysis

- Derive low-term predictions to get an idea of what to expect in the following weeks
- Understand which are the most relevant factors that determine the increasing of ICU patients
- Suggest possible improvements in the management of the pandemic



The dataset

- The dataset was obtained by the official website of Protezione Civile starting from 01-09- 2020 to 23-01-2021 and considering only region Veneto
- Data regarding the place of the survey (latitude, longitude, exc…) have been removed
- Data regarding variables no longer populated have been removed, while notes were considered in evaluation of the dataset but not during the modeling procedure

terapia_intensiva	Intensive Care	Intensive_care
ricoverati_con_sintomi	Hospitalised patients with symptoms	Hos_symp
data	Date of notification	Date
totale_ospedalizzati	Total hospitalised patients	Total_Hos
isolamento_domiciliare	Home confinement	Home_con
totale_positivi	Total amount of current positive cases	Total_pos
variazione_totale_positivi	Variation of current positive cases	Var_pos
nuovi_positivi	Variation of current cases	Var_cases
dimessi_guariti	Recovered	Recovered
deceduti	Death	Death
totale_casi	Total amount of cases	Total_cases
tamponi	Tests performed	Test
casi_testati	Total number of people tested	People

The dataset

- Blue variables are cumulative and therefore have been replaced by the corresponding daily changes
- Red variables have been removed because they are strongly correlated to other variables:

  Total_hosp = Hos_symp + Intensive_care

  Total_pos = Total_cases – Recovered – Death

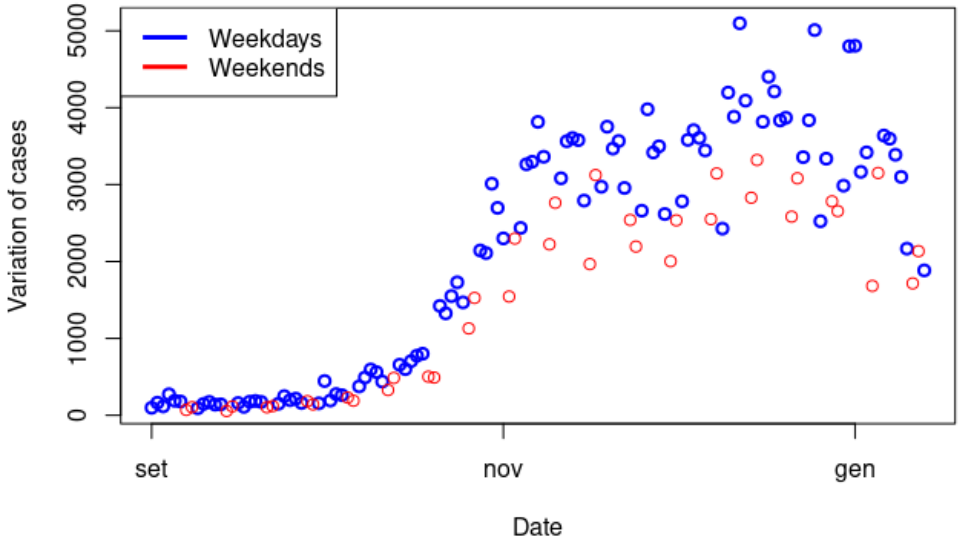
- Two new variables have been added:
  - *zone* is a categorical variable describing the “color” of the area in the current date
  - *lag_zone* is the shifted zone (7 days before the current date)
  - *season* a categorical variable (0 = autumn, 1 = winter)

Variable
Intensive_care
Hos_symp
Date
Total_hosp
Home_con
Total_pos
Var_pos
Var_cases
Recovered
Death
Total_cases
Test
People
Zone
Lag_zone
season

Variable
Intensive_care
Hos_symp
Date
Home_con
Var_cases
Recovered_today
Death_today
Test_today
People_today
Zone
Lag_zone
Season

Quality of data

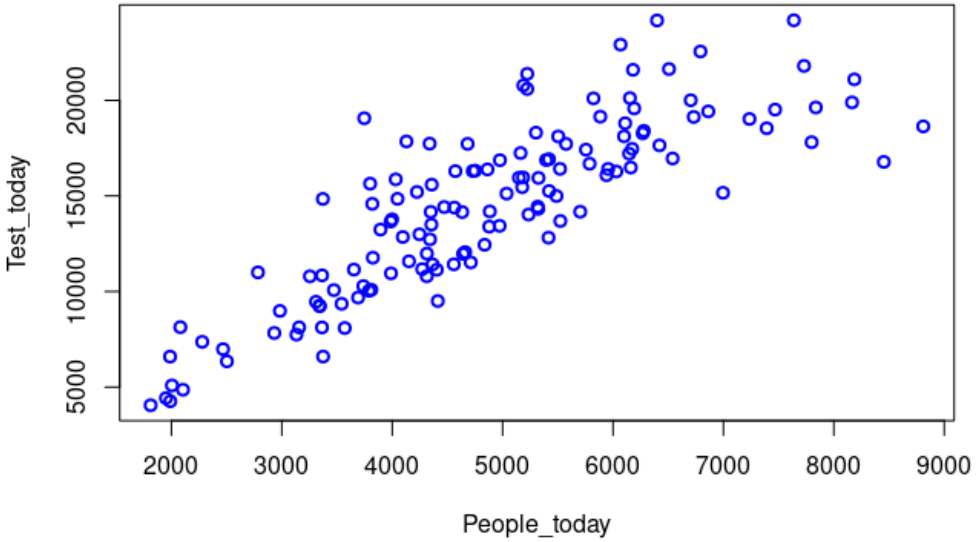
- Despite the fact that data was obtained by the official national source, the reliability depends on the procedures adopted to collect data. In this case, due to relatively frequent algorithm change and new or deleted variables, data-gathering process does not guarantee the most accurate predictions possible
- The dependent variable, intensive care, is not always the effective measured value because there are many temporal misalignments of the information flow, as reported in the notes of the dataset.
- In the weekends or holidays data collection slow down and is retrieved on subsequential weekdays.

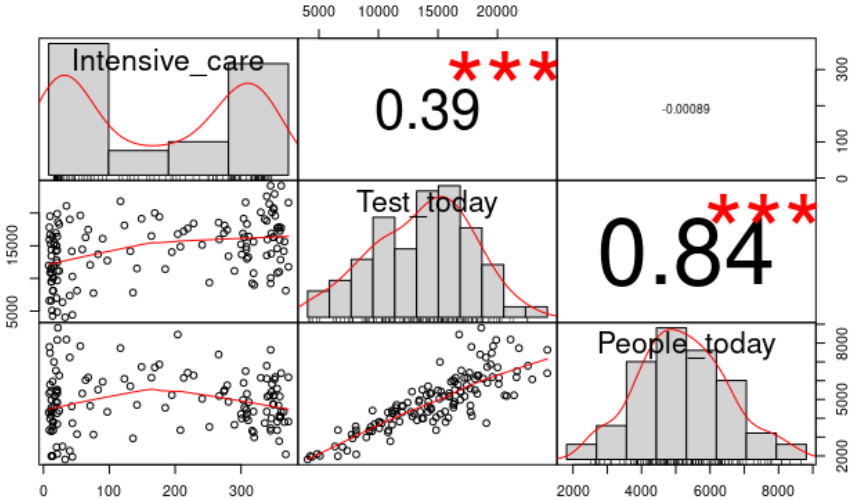


Explanatory analysis

- Before starting to create a statistical model, it is convenient to analyze the variables and their relationship with the independent variable.

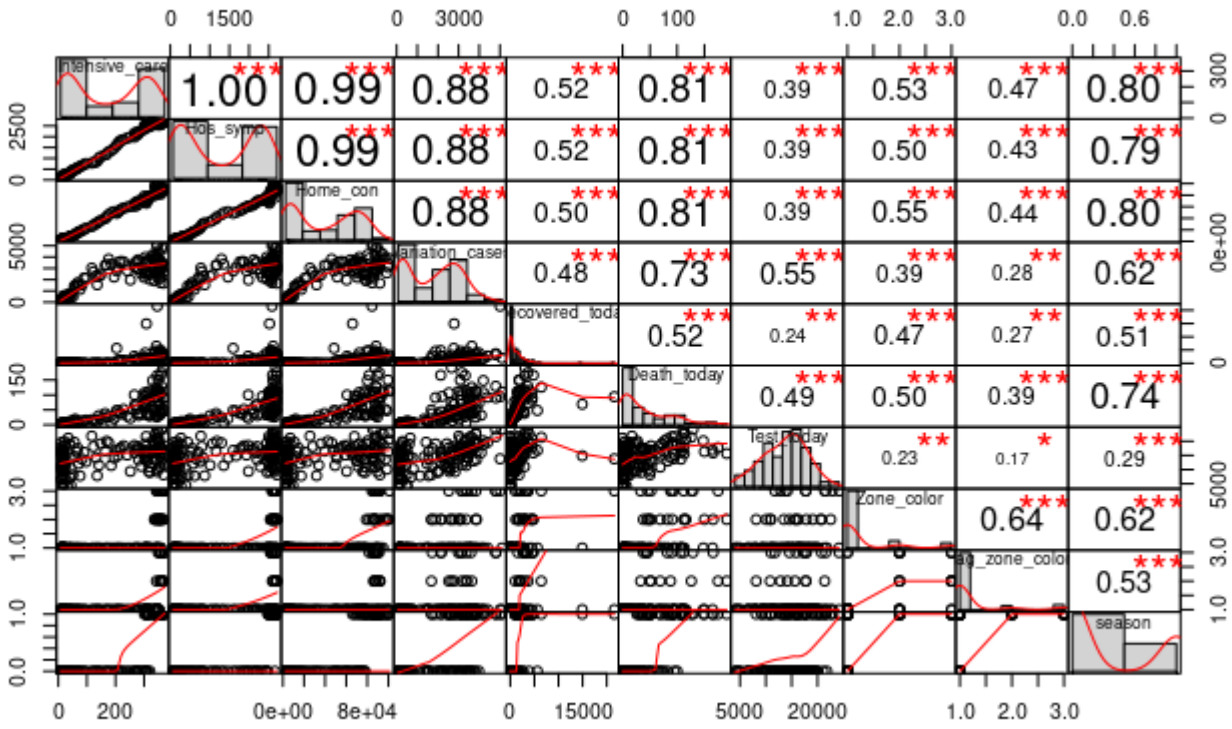
- There is a strong correlation between the people tested and the number of tests performed, so only one of them can be used in the model. The number of tests is chosen, being more correlated to the response variable.





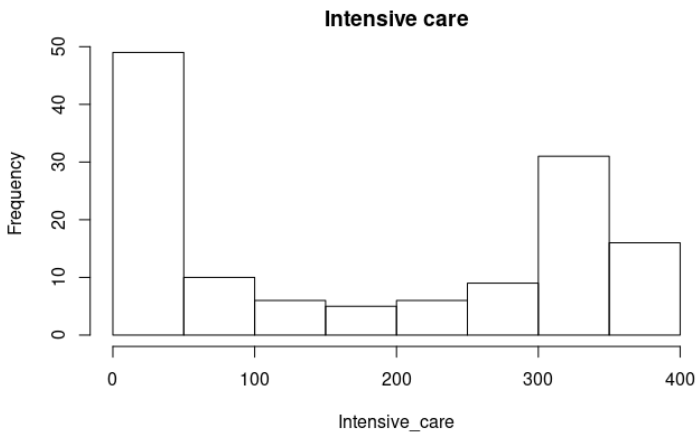
Explanatory analysis

- The other possible predictors are all correlated to the variable intensive care. However, home confinement is highly correlated with many covariates and so has been removed.



Model specification

- The aim is to find a model that describes a response variable (intensive care) using multiple predictors.
- The response variable is not normally distributed; it is discrete and non negative. So a simple linear regression model can't be used.
- At first, only the predictors of the original dataset will be used, later other covariates will be added.



Response variable
Intensive_care

Possible Predictors
Hos_symp
Date
Home_con
Var_cases
Recovered_today
Death_today
Test_today
Zone
Lag_zone
Season

Generalized Linear Model

- It is an extension of linear models, characterized by the following features:
  - linear predictor: $\lambda_i = \sum_{j=1}^{p} x_{ij} \beta_j$
  - link function: $g(E(y_i)) = \lambda_i$
  - the response variable belongs to exponential dispersion family
- The response variable of this problem (intensive care) is a count data and it is assumed to follow a poisson probability distribution in which observations are independent

Link function Linear predictor

$$\ln \lambda_i = b_0 + b_1 x_i$$

$$y_i \sim \text{Poisson}(\lambda_i)$$

Probability distribution

Generalized Linear Model

- The strategy used to select the predictors is the stepwise selection, considering the following measures:
- **AIC**: Akaike Information Criteria
- **BIC**: Bayesian Information Criteria
- **F test**: Occam's razor criteria)
- **VIF**: Variance Inflation Factor

- The link function is chosen to be the canonical link (default of GLM method)

Generalized Linear Model

- **Baseline model**

- First of all we create a baseline model, by using some of the more important and influential variables:

```
glm0 <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+Hos_symp,
            family = poisson, data=d.train)
```

AIC <dbl>	BIC <dbl>	Residual_deviance <dbl>	P_value <dbl>
1307.41	1321.94	437.29	6.807533e-35

1 row

- The P_value of F test is small so our model works better than the null model in which only intercept is included.

Generalized Linear Model

- **Adding the variable Death_today**

```
model.glm <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+Hos_symp+
                 poly(Death_today,3), family = poisson, data=d.train)
```

AIC <dbl>	BIC <dbl>	Residual_deviance <dbl>
1282.38	1305.62	406.26

1 row

- The above table shows that AIC and BIC are smaller compared to baseline model. So we add *Death_today* to the model.

Generalized Linear Model

- **Adding the variable Recovered_today**

```
model.glm <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+Hos_symp+
                 poly(Death_today,3)+Recovered_today, family = poisson, data=d.train)
```

AIC <dbl>	BIC <dbl>	Residual_deviance <dbl>
1281.09	1307.24	402.97

1 row

- The above table shows there is a negligible improvement in AIC and BIC. As the simpler model is prefered, we <u>do not add</u> *Recovered_today*.

Generalized Linear Model

- **Adding the variable Test_today**

```
model.glm <- glm(Intensive_care ~ Date+poly(Variation_cases,2)+Hos_symp+
                 poly(Death_today,3)+Test_today, family = poisson, data=d.train)
```

AIC <dbl>	BIC <dbl>	Residual_deviance <dbl>
1273.46	1299.61	395.34

1 row

```
vif(model.glm)
```

```
                    Date poly(Variation_cases, 2)1 poly(Variation_cases, 2)2
                  4.9052                    5.1471                    2.6281
                Hos_symp     poly(Death_today, 3)1     poly(Death_today, 3)2
                 10.4290                    9.3075                    5.7127
poly(Death_today, 3)3                 Test_today
                  2.8018                    2.3227
```

- Although AIC and BIC are improved, we <u>do not add</u> *Test_today* because it causes to VIF greater than 10 for some variable.

PREDICTION

EXTRA: Predicting on shifted data

- The current time (t) and future times (t+1, t+n) are forecast times and past observations (t-1, t-n) are used to make forecasts
- We could frame our forecast problem with an input sequence of 7 past observations to forecast 7 future observations and use the data as follows:

TRAIN

date_only <date>	terapia_intensiva <int>	deceduti <int>	lag_deceduti <int>
2020-09-01	9	2122	2107
2020-09-02	9	2123	2116
2020-09-03	12	2123	2117
2020-09-04	10	2126	2119
2020-09-05	9	2130	2120
2020-09-06	12	2130	2120
2020-09-07	13	[illegible]	2120
2020-09-08	12	[illegible]	2122
2020-09-09	12	2135	2123
2020-09-10	12	2135	2123

Previous 1 2 3 4 5 6 ... 14 Next

TEST

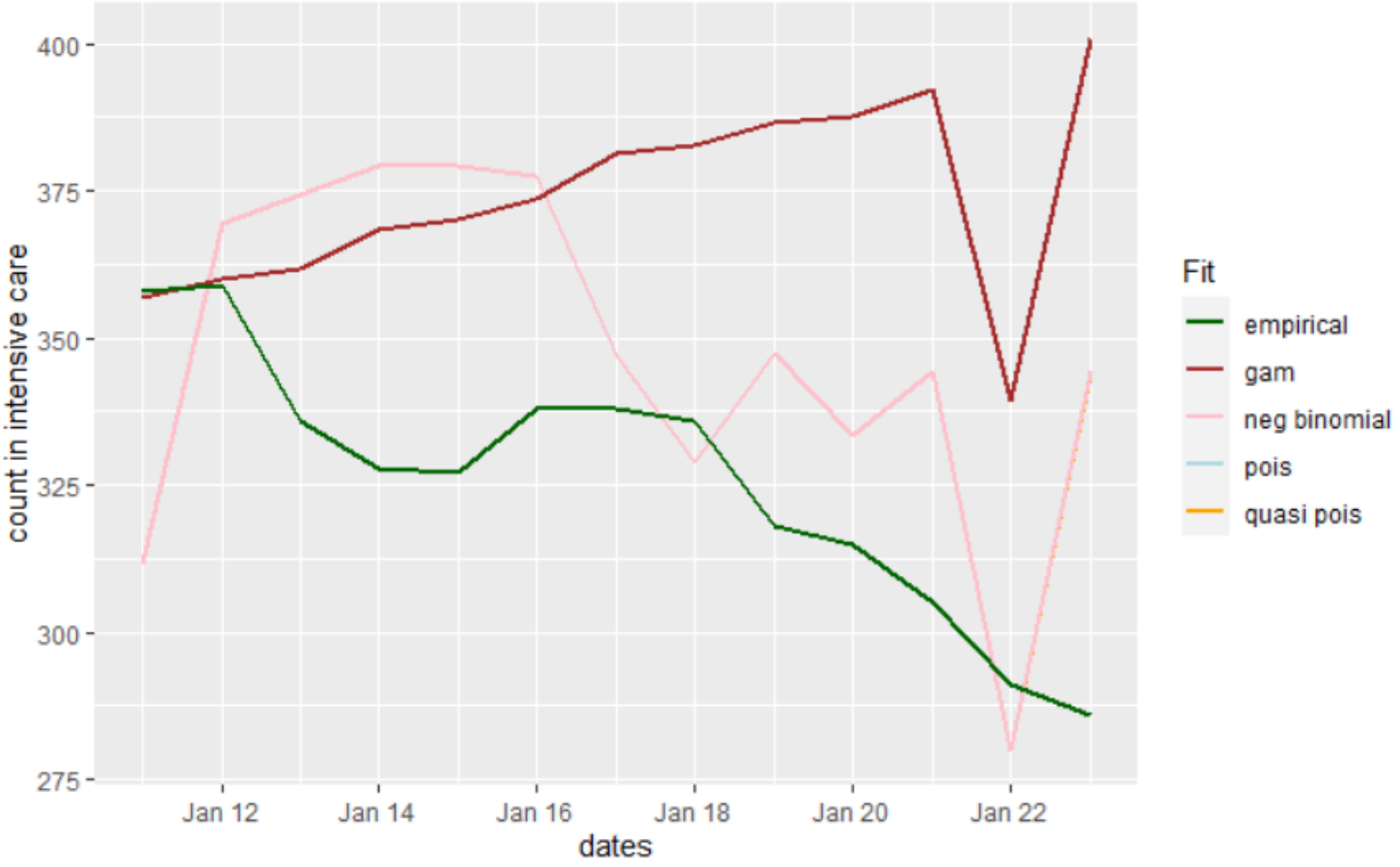
date_only <date>	terapia_intensiva <int>	lag_deceduti <int>
2021-01-11	358	6813
2021-01-12	359	6988
2021-01-13	336	7114
2021-01-14	328	7157
2021-01-15	327	7263
2021-01-16	338	7345
2021-01-17	338	7389
2021-01-18	336	7427
2021-01-19	318	7593
2021-01-20	315	7684

Previous 1 2 Next

EXTRA: Predicting on shifted data

Prediction

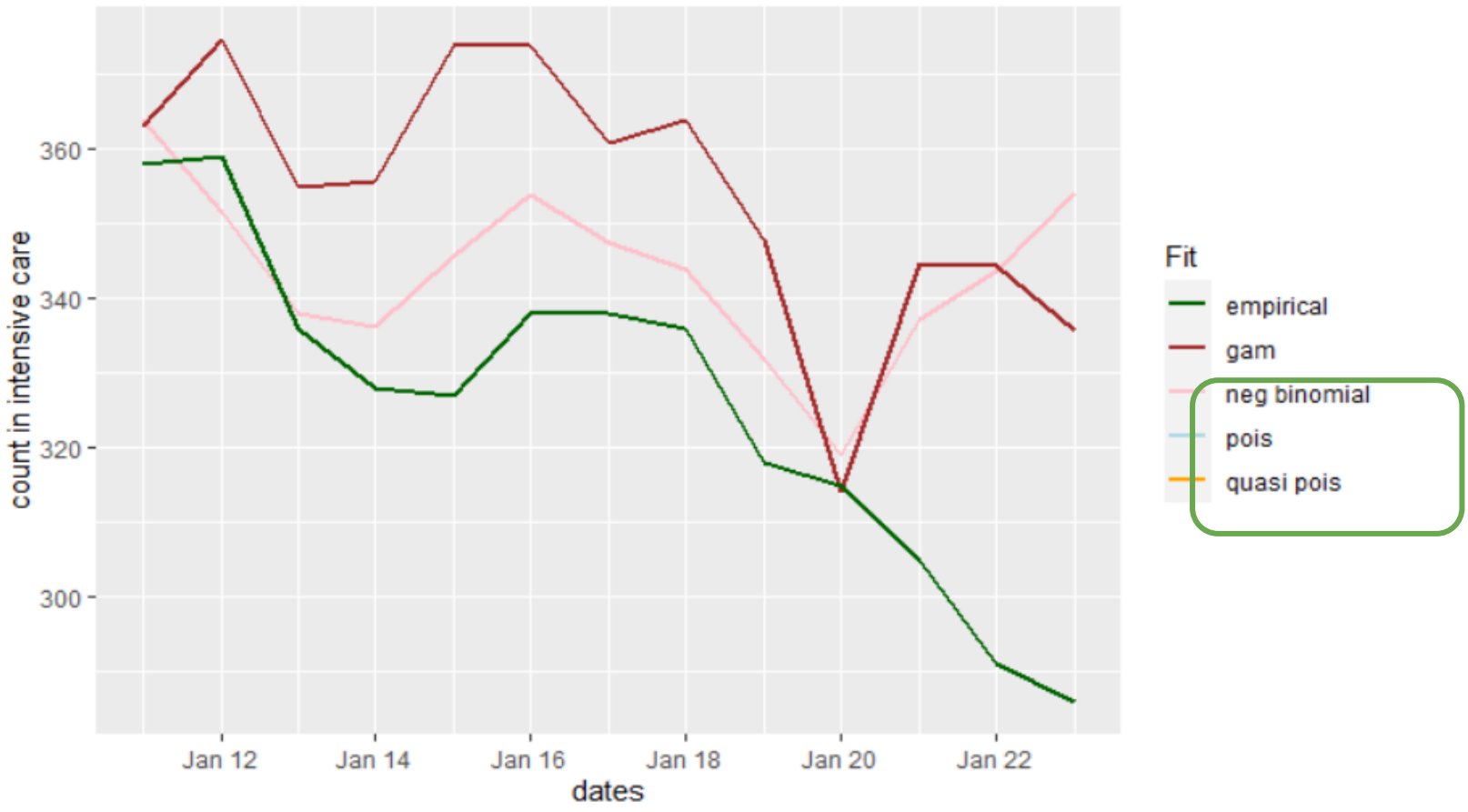
- Shifted data + proved well-suited model
- terapia_intensiva ~ date_unix + poly(lag_nuovi_positivi,2) + as.numeric(type.convert(variazione_guariti_dimessi)) + poly(lag_deceduti,3) + lag_nuovi_positivi



EXTRA: Predicting on shifted data

Prediction

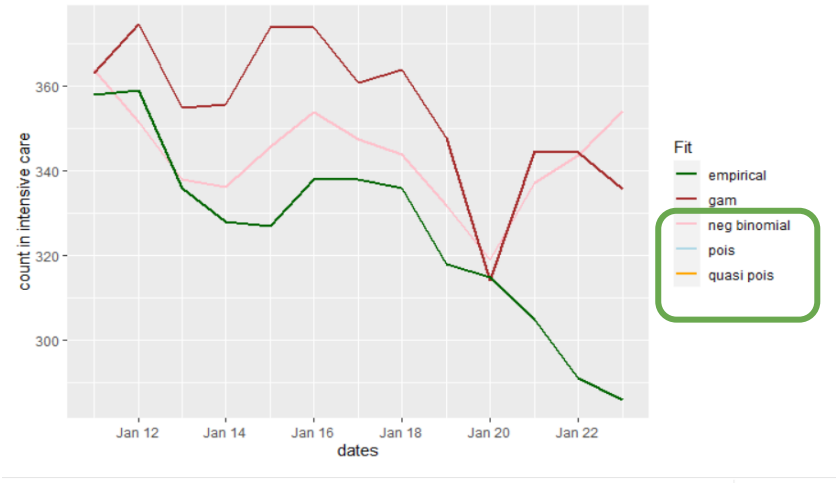
- Shifted by 14 data + new model
- terapia_intensiva ~ date_unix + ma_decessi + ma_tot_osp + as.numeric(type.convert(variazione_tot_osp)) + perc_esaurito_ti



EXTRA: Predicting on shifted data

Details of the flow

- terapia_intensiva ~ date_unix + ma_decessi + ma_tot_osp + as.numeric(type.convert(variazione_tot_osp)) + perc_esaurito_ti
- NOTE: with the ultimate goal to have a good prediction; the data was not scaled



Model comparison

- **GLM**
  - **Poisson**
  - **Quasi-Poisson**
  - **Negative binomial**
- **GAM**
- **Random Forest**

	df <dbl>	AIC <dbl>
model.glm	8.00000	1282.379
model.glm.quasi	8.00000	NA
model.glm.nb	9.00000	1152.315
model.gam	12.54147	1036.664

4 rows

- The above table shows AIC improves considerably from the first GLM Poisson model to the Negative Binomial one, and gets even better for the GAM model

Model comparison

- **Quasi-Poisson** GLM is used we have overdispersion, i.e. the variance of Y is greater than its theoretical value
- In general, the quasi-likelihood approach allows to deal with overdispersion problems: it is possible to specify var(Yi) so that there is more variability with respect to the exponential family.

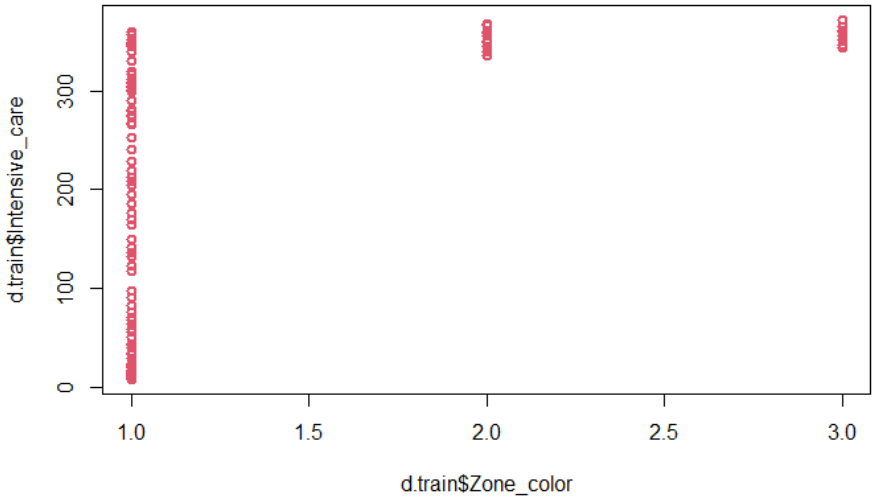
- **Negative Binomial** is an alternative model that can be considered when data exhibits overdispersion.
- Interpretation: probability to observe z failures until the pre-specified number of successes k is observed.
- Compared with Poisson: it has an extra parameter; it proves to be more flexible; mean is larger than variance and then it accommodates overdispersion; Poisson is a limiting case of negative binomial (if $p \rightarrow 1$ and $k \rightarrow 0$ then $kp \rightarrow \lambda$).
- Recall that negative binomial emerges as a mixture of Poisson when each unit Y is Poisson with mean λ and λ are drown from a Gamma distribution.

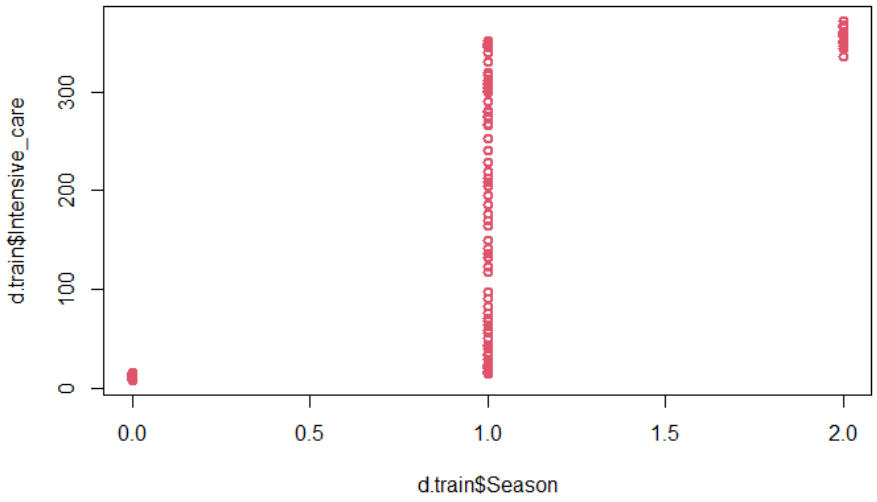
Model comparison

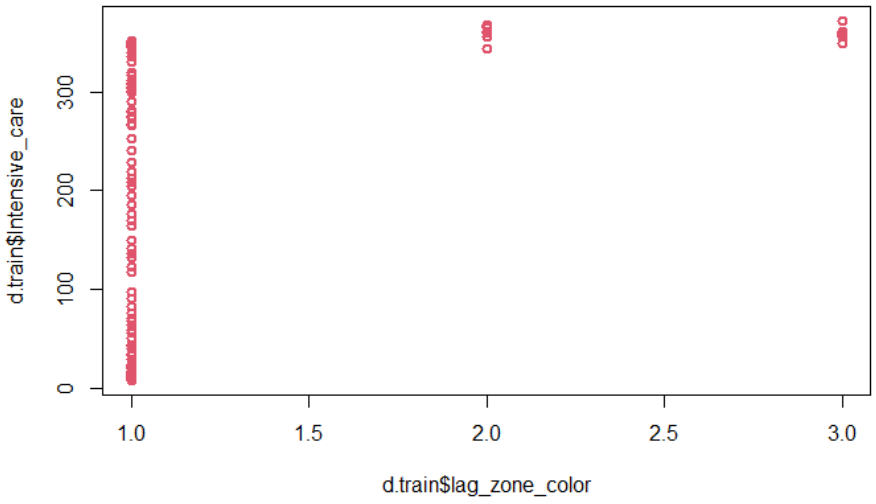
Covariates

- **Zone_color**
  - **lag_zone_color**

- **Season**
  - **0** for Summer
  - **1** for Autumn
  - **2** for Winter







Covariates

Covariates

- After adding **Zone_color**

aic <dbl>	bic <dbl>	RD <dbl>
1282.403	1302.739	408.2799

1 row

- After adding **lag_zone_color**

aic <dbl>	bic <dbl>	RD <dbl>
1277.932	1301.175	401.8098

1 row

- After adding **Season**

aic <dbl>	bic <dbl>	RD <dbl>
1269.217	1295.365	391.0946

1 row

Predictive information criteria on model comparison

- **MSE**
- **RMSE**
- **NRMSE**

Predictive information criteria on model comparison

Prediction with covariates

- Model comparison without covariates

models <chr>	MSE <dbl>	RMSE <dbl>	NRMSE <dbl>
glm	14826.364	121.76356	0.3826636
glm.quasi	14826.364	121.76356	0.3826636
glm.nb	13616.123	116.68815	0.3667132
gam	17729.581	133.15247	0.4184553
rf	3185.755	56.44249	0.1773806

5 rows

- Model comparison <u>with</u> covariates

models <chr>	MSE <dbl>	RMSE <dbl>	NRMSE <dbl>
glm	13562.0959	116.45641	0.36598495
glm.quasi	13562.0959	116.45641	0.36598495
glm.nb	11197.6030	105.81873	0.33255414
gam	18710.8662	136.78767	0.42987954
rf	404.3742	20.10906	0.06319628

5 rows

Prediction with covariates Season

- Model comparison without *Season*

	df <dbl>	AIC <dbl>
model.glm1	8.0000	1277.932
model.glm.quasi1	8.0000	NA
model.glm.nb1	9.0000	1142.087
model.gam1	14.5545	1039.844

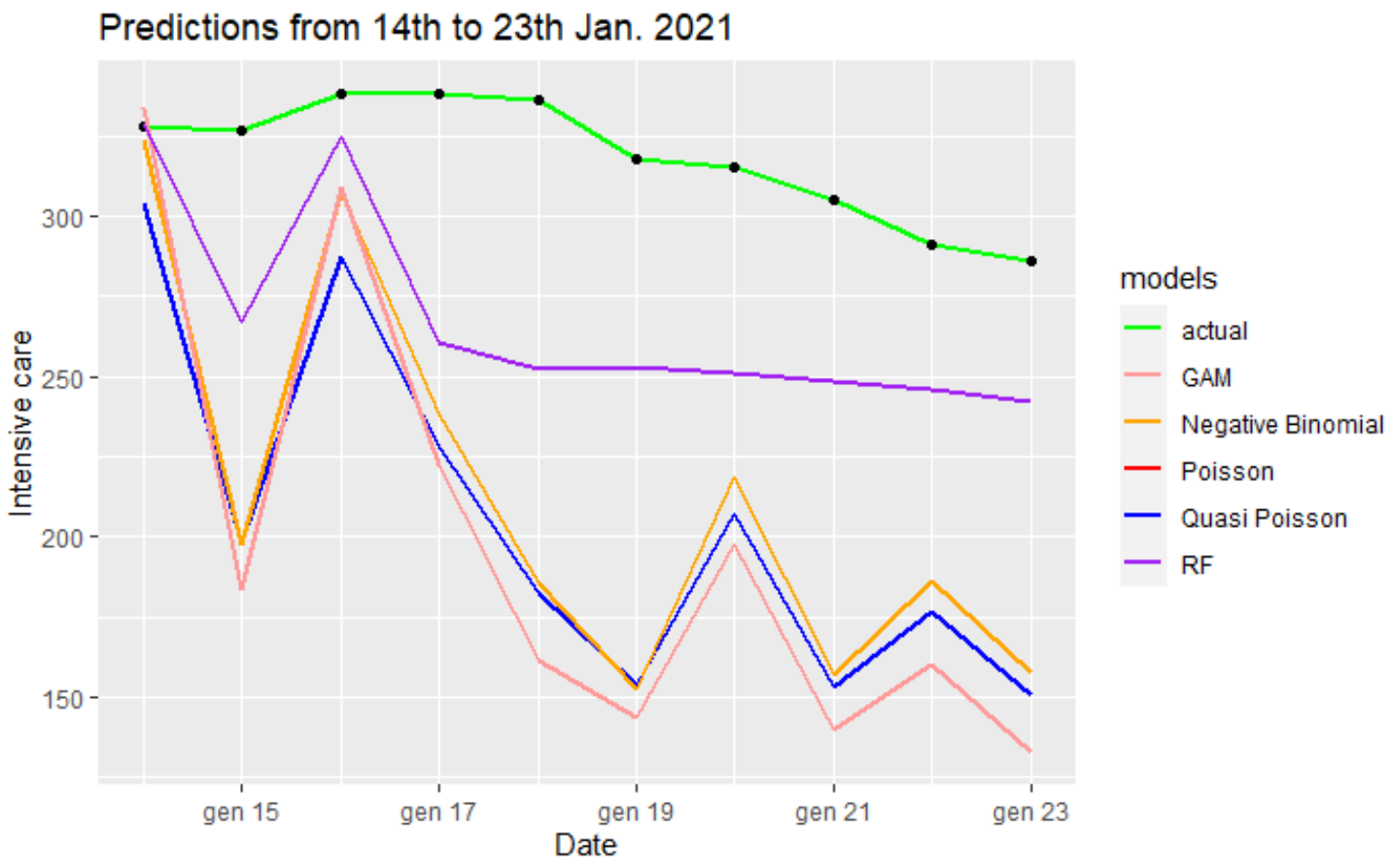
4 rows

- Model comparison <u>with</u> *Season*

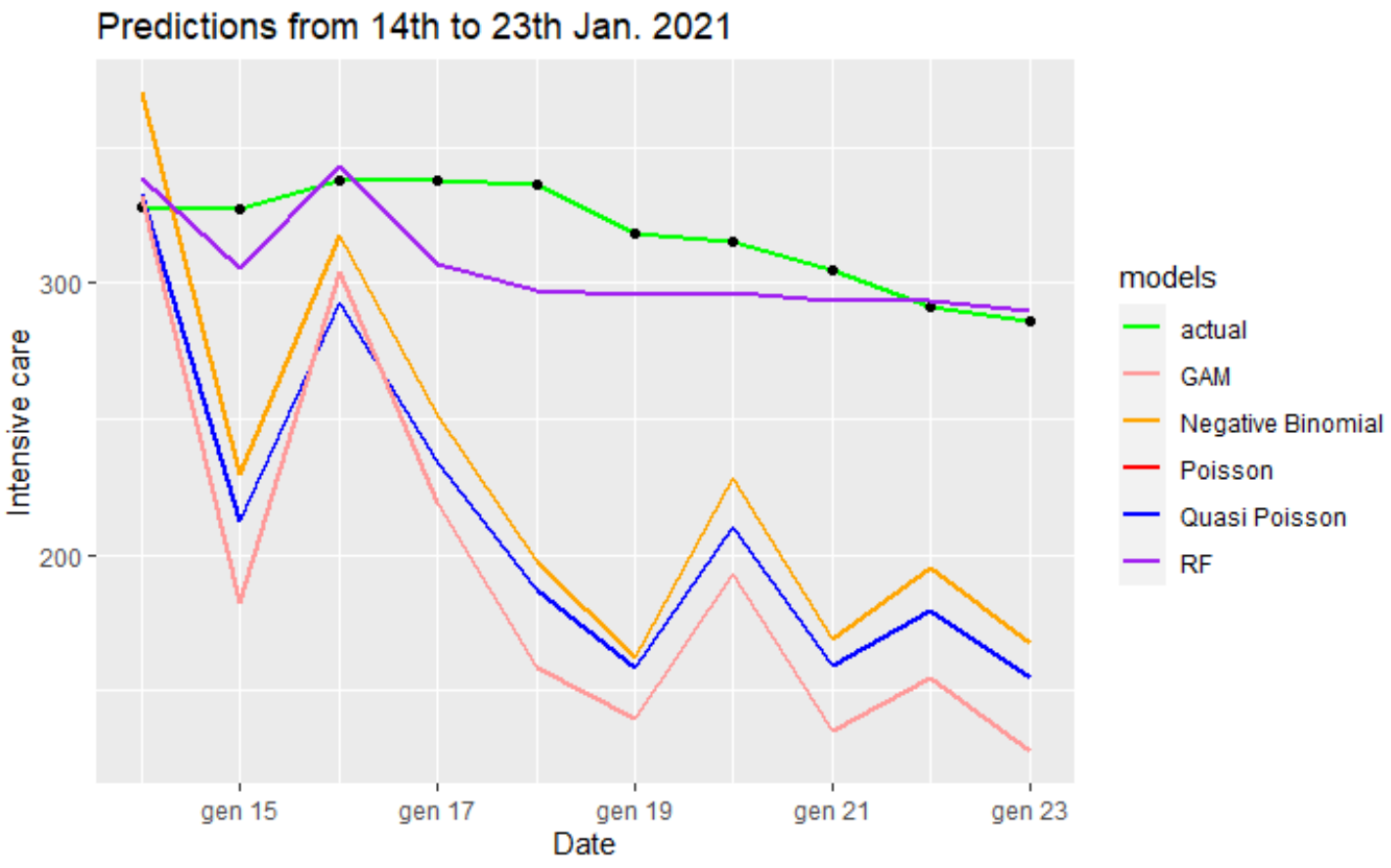
	df <dbl>	AIC <dbl>
model.glm1	9.00000	1269.217
model.glm.quasi1	9.00000	NA
model.glm.nb1	10.00000	1137.186
model.gam1	15.44354	1036.889

4 rows

Prediction without covariates



Prediction with covariates



References

- https://www.agenas.gov.it/covid19/web/index.php?r=site%2Fheatmap
- https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(20)30161-2/fulltext
- https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab