Long-term time-series forecasting

Gas consumption

UNIVERSITÀ
DI PAVIA

Members:

- *Francesco Amato*
- *Filippo Maria Rognoni*
- *Francesco Minaglia*

05/2021

ROADMAP

1. Introduction

DATASET

Dataset which represent gas consumption in Italy, during two different years (in function of day of the week and day of the year).

AIMS

1. Identify a model which represent yearly trend, for long-term time-series forecast.
2. Write a function whose aim is predicting gas consumption once you put in, as parameters, the day of the week and the day of the year).

2. Process visualization

Before starting identifying the best model, it is useful to have some ideas on the process function characteristics, plotting it in three different ways:

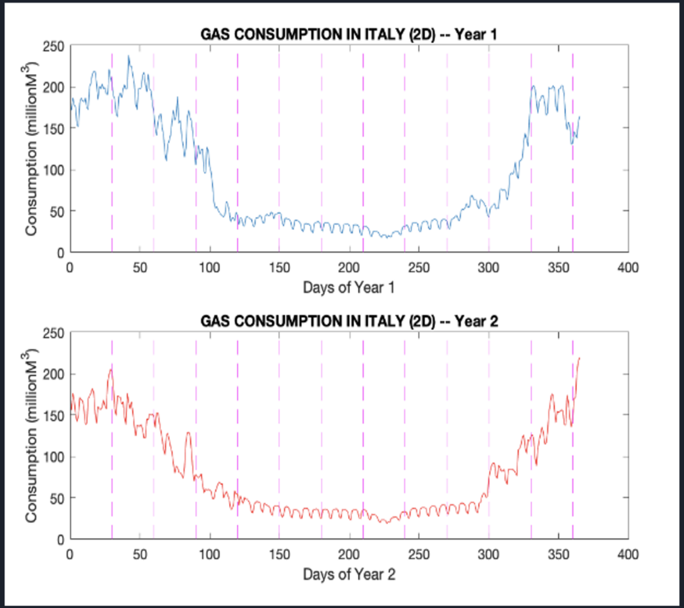
- 2D: Gas consumption in function of the day of the year;
- Spectral analysis: Allowing us to see what are the repetition that characterize our signal;
- 3D: Gas consumption in function of the day of the year, and the day of the week.

2.1 Plotting 2D

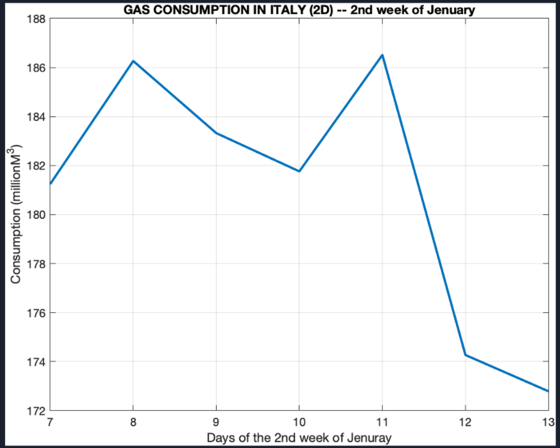
Observations:

- Gas consumption **decreases** in the **hottest months** of the year ;
- During **weekend** the gas consumption is **lower ;**
- **Stationary** process ;
- **Discrete time** process ;
- **Ergodic process**;
- **Autocovariance function** shows some **kind of relationship** between first days of the year and last days of the year ;

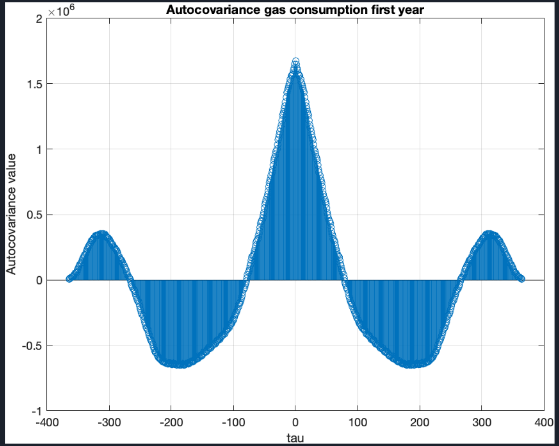
2.1 Plotting 2D



Plot 2D



2nd week of January trend



Autocovariance first year

2.2 Spectral analysis

Another process visualization that allows to notice different signal characteristics, is relative to the **frequency domain**.

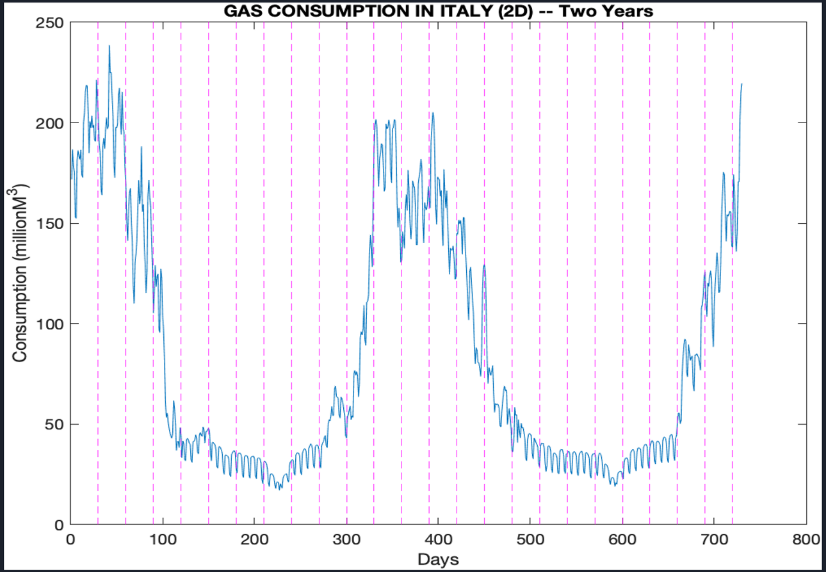
This domain can be very useful to understand and **highlight signal repetitions.**

The gas consumption of the two years was **chained**.

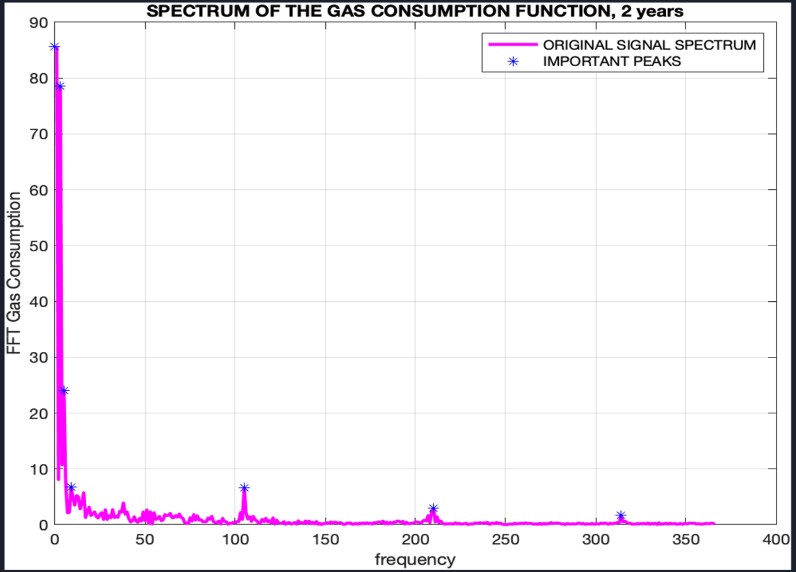
Important frequencies in signal spectrum:

- *Frequency in 0* mean gas consumption in both years;
- *Frequency in 2* yearly seasonality;
- *Frequency in 4* semestral seasonality;
- *Frequency in 104* weekly periodicity;
- *Frequency in 8* three-months seasonality;
- *Frequency in 209* four days periodicity;
- *Frequency in 313* two days periodicity.

2.2 Spectral analysis



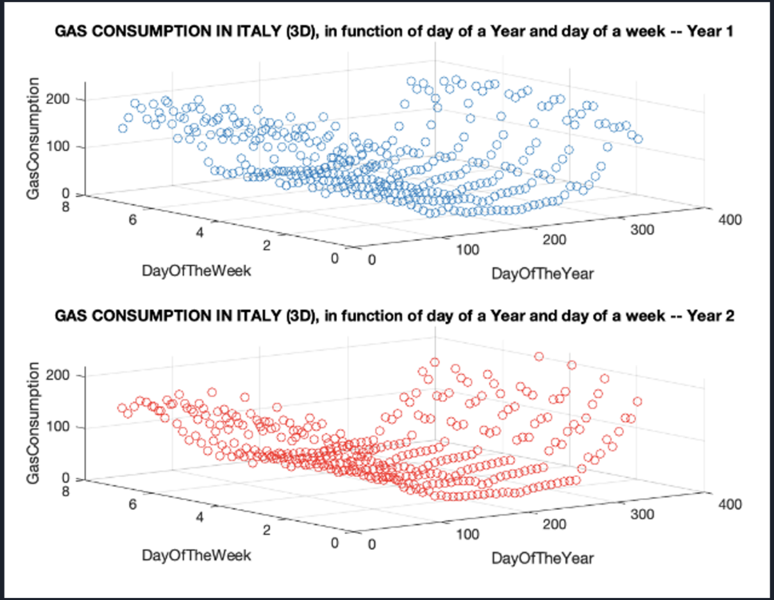
Two years 2D trend



Spectrum of our signal

2.3 Plotting 3D

For completeness we show the 3D plot of our process, depending on our two variables (day of year and day of the week)



Plot 3D

3. Model identification

Since we can identify the model through **different methodologies**, we chose some techniques to do this:

- Polynomial regression
- Neural network
- Harmonic regression

As a criterion for the selection of the method that produces the final model, we chose the one that provides **better results** than the others (3.4 section).

3.1 Polynomial regression

Creating different polynomial models and plotting them on graphs

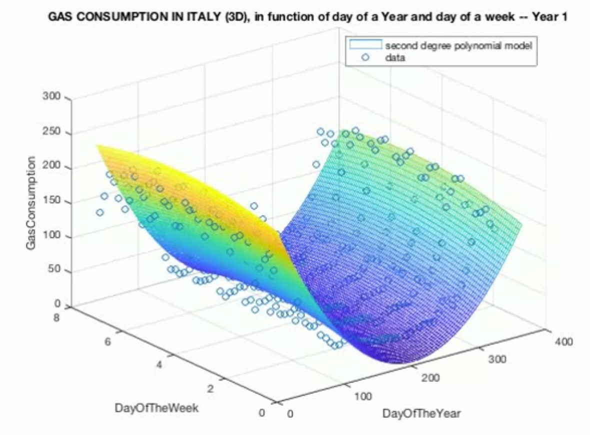
Starting from a second degree polynomial model (gas consumption trend is not constant, nor plane).

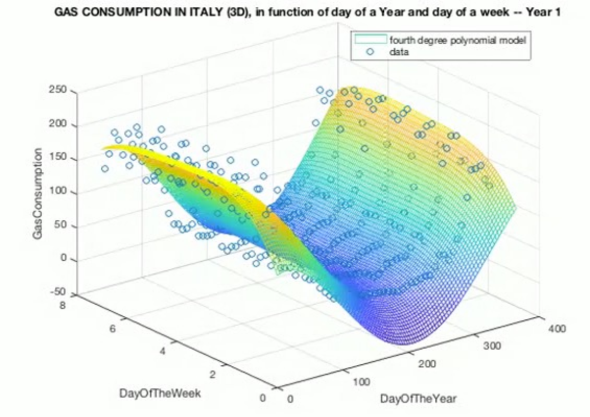
It seems like all the models are quite good, but how can we decide which is the best model?

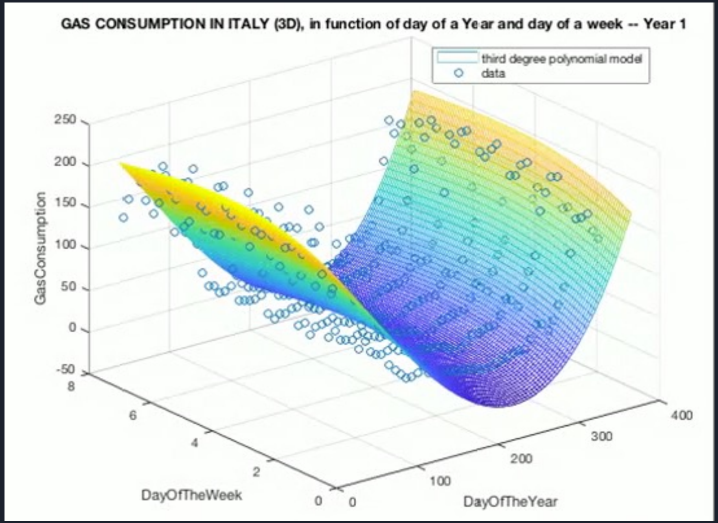
1. SSR (not a good idea) More complex model always wins
2. **Figures of merit:**
    a. **Subjective criterion** (*Test F*) 4th wins
    b. **Objective criteria** (*AIC, MDL, FPE*) 4th wins
3. Cross-validation (first year gas consumption data to identify the model, second year gas consumption data for cross-validation).

In order to choose the best polynomial model we will use cross-validation.

3.1 Polynomial regression







3.1 Polynomial regression

Polynomial models results, identification

Polynomial model degree:	SSR_Identification	TEST F	AIC	FPE	MDL
2°	$2.6444e^5$	$1.673e^3$	12.52	$2.733e^5$	12.60
3°	$2.1417e^5$	83.09	12.33	$2.262e^5$	12.44
4°	**$1.8057e^5$**	**64.84**	**12.18**	**$1.960e^5$**	**12.35**

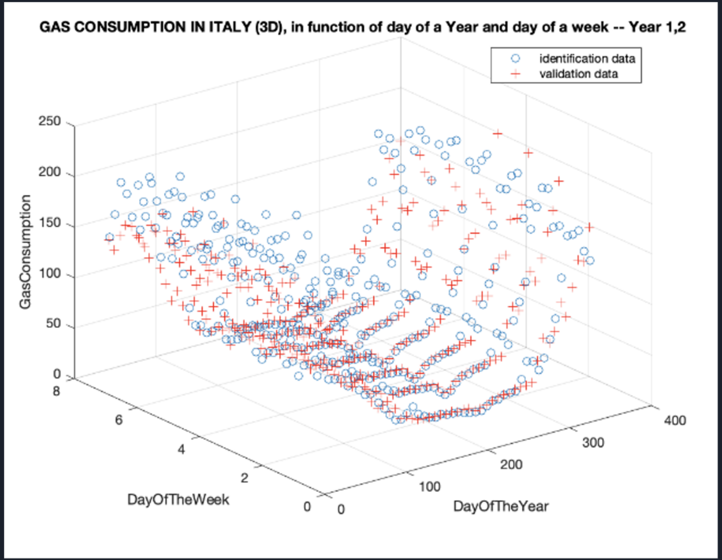
Cross-validation:

We considered the various polynomial models previously shown and saw which was the one that minimized the SSR (hence also the standard deviation) between validation data and expected data.

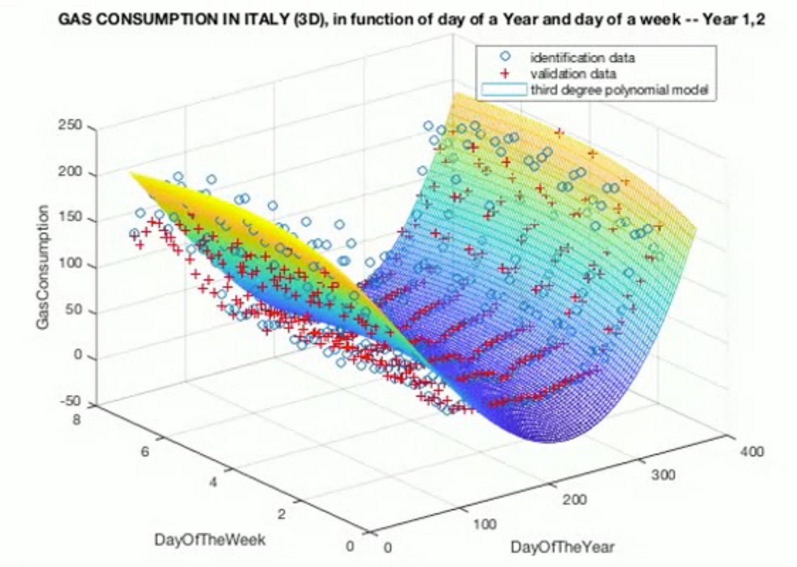
Polynomial models results, cross-validation:

Polynomial model degree:	SSR_Validation	SD
2°	$2.1648e^5$	24.38
3°	**$1.9883e^5$**	**23.37**
4°	$2.2411e^5$	24.81

3.1 Polynomial regression



Identification and Validation data



Best model for cross-validation

3.2 Neural network

Multi-layer perceptron neural network characteristics:

- **2 layer** MLP (1 hidden layer, 1 output layer)
- **Data division**: 1st year identification, 2nd year validation.
- **Activation function:** tanh(z)
- **Training algorithm:** Bayesian-Regularization
- **Performance:** MSE minimization, consequently also SSR e SD

Here's some results, in identification (left) and cross-validation (right), obtained **modifying** the **neuron's number** in the hidden layer.

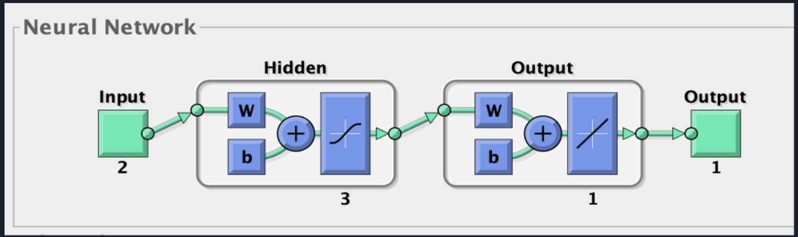
Number of Neurons	SSR	SD
3	$9.2184e^{4}$	15.89
5	$6.5197e^{4}$	13.36
8	$5.4082e^{4}$	12.17
10	$3.1955e^{4}$	9.35
12	**$2.9674e^{4}$**	**9.10**

Identification results

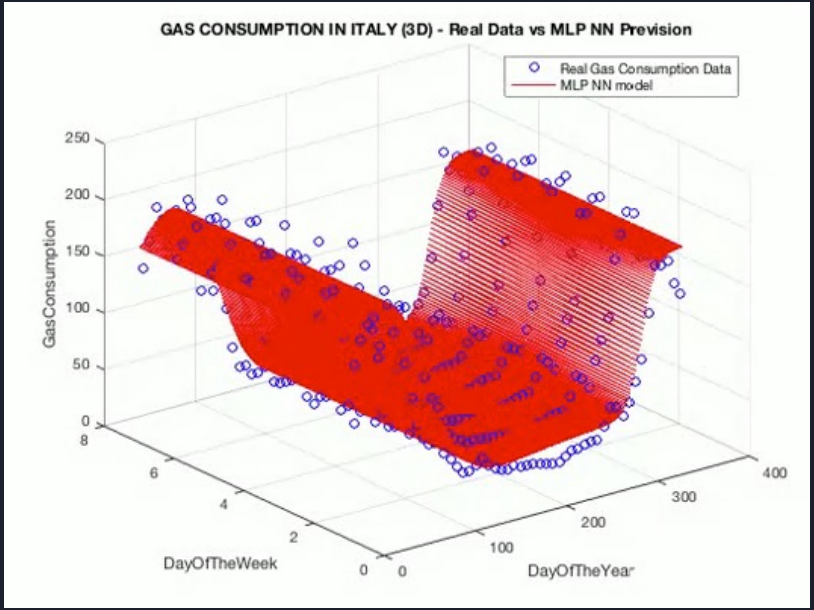
Number of Neurons	SSR	SD
3	**$2.4047e^{5}$**	**25.66**
5	$2.5286e^{5}$	26.32
8	$2.6414e^{5}$	26.90
10	$2.7289e^{5}$	27.34
12	$2.6331e^{5}$	26.86

Cross-validation results

3.2 Neural network



Neural network architecture



Real gas consumption VS nn model

3.3 Harmonic regression

This identification method follows the same methodology of the polynomial regression (using harmonics instead of polynomials, according to **Fourier series theory**).

The **division of the data and the initial observations are the same** we discussed in polynomial regression (first year data identification, second year cross-validation).

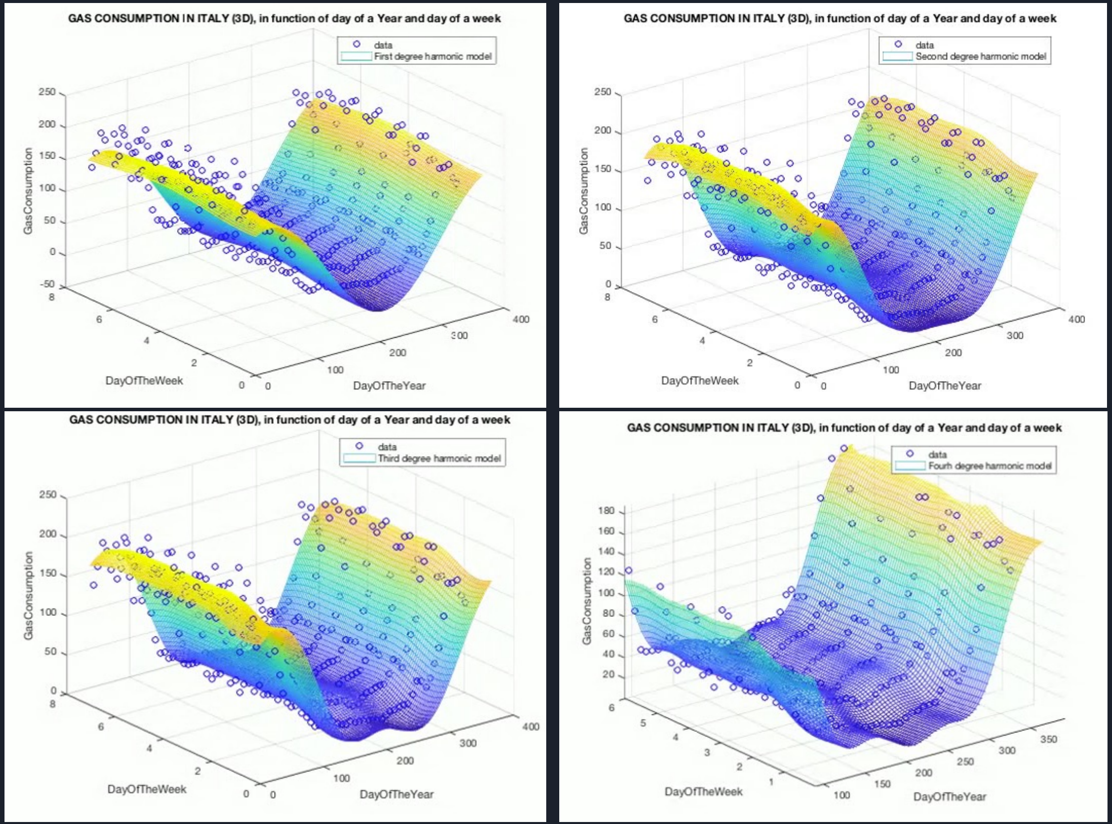
We'll use another time:

- **SSR, objective/subjective criteria**, to compare different identified models;
- **Cross-validation,** to understand which is the best model.

We won't start using constant model, but first degree harmonic, because gas consumption **trend isn't constant.**

3.3 Harmonic regression

Plot (Different harmonic models)



3.3 Harmonic regression

Harmonic models results, identification

Harmonic model degree:	SSR_Identification	TEST F	AIC	FPE	MDL
1°	$2.3475e^5$	/	12.39	$2.4127e^5$	12.45
2°	$1.9969e^5$	344.5	12.25	2.0979^5	12.35
3°	$1.0416e^5$	24.64	11.62	$1.1185e^5$	11.76
4°	**$8.8673e^4$**	**87.96**	**11.49**	**$9.7736e^4$**	**11.67**

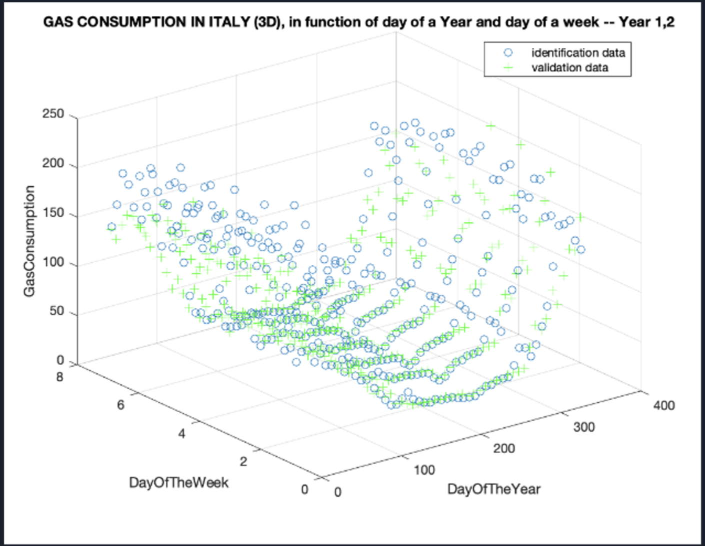
Cross-validation:

We considered the various polynomial models previously shown and saw which was the one that minimized the SSR (hence also the standard deviation) between validation data and expected data.

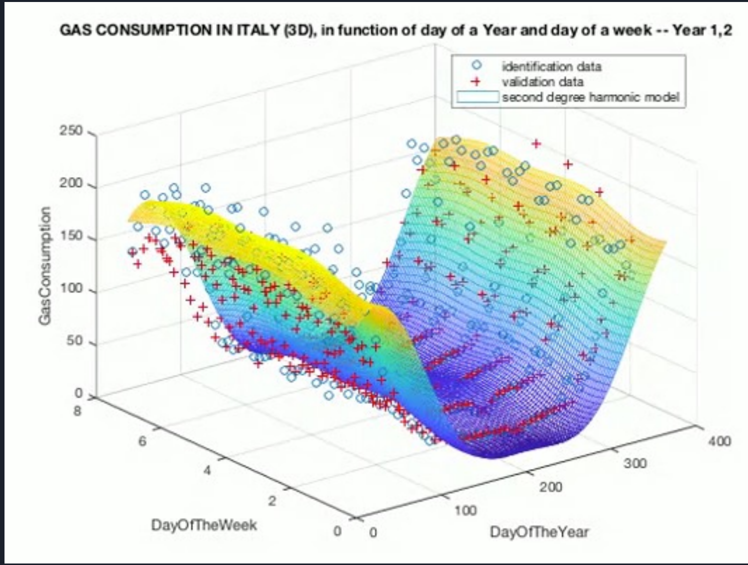
Harmonic models results, cross-validation:

Harmonic model degree:	SSR_Validation	SD
1°	$2.8801e^5$	28.09
2°	**$1.9169e^5$**	**22.91**
3°	$2.0297e^5$	23.58
4°	$2.1635e^5$	24.34

3.3 Harmonic regression



Identification and Validation data



Best model for cross-validation

3.4 Compare different models

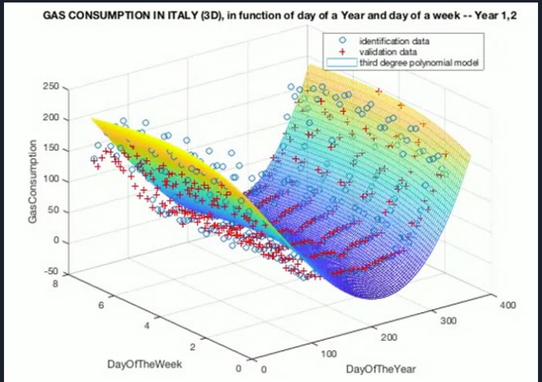
In order to decide which model is better we compare them using two figures of merit: SSR and standard deviation.

Identification Method	SSR	SD
Polynomial model	$1.9883e^{5}$	23.37
Neural network	$2.4047e^{5}$	25.66
Harmonic regression	**$1.9169e^{5}$**	**22.91**

Comparison between different models, in validation

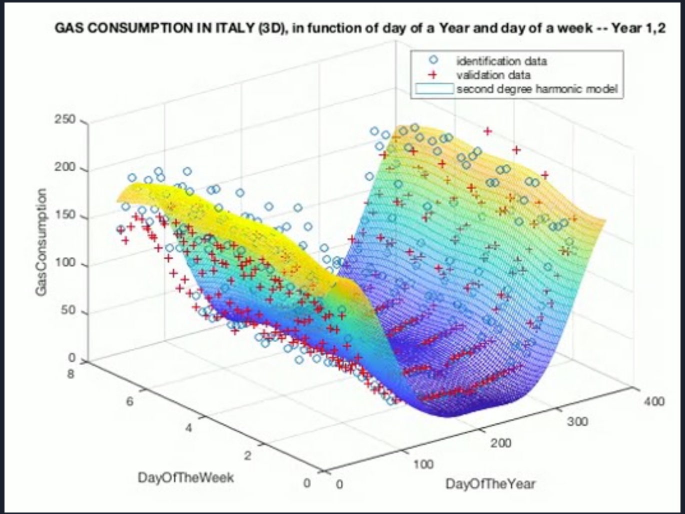
3.4 Compare different models

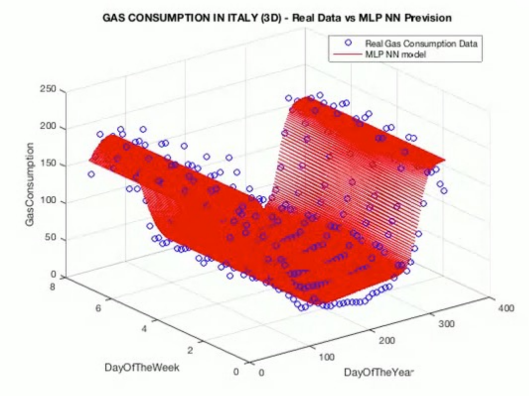
Different models graphs



Best polynomial model

Best harmonic model





Neural network model

4. Prediction function

Once identified best model for our dataset, in our case the 2nd degree harmonic model, a prediction function which **forecasts gas consumption depending on two parameters** (days of the week and day of the year) has been written.

You can also **change manually**, via numerical slider, the **value of those parameters** in order to see how the gas consumption prediction changes.

```
% Insert parameters value
% Remember that in DayOfTheWeek notation 1 is Sunday, 7
% is Saturday
DayOfTheWeek= 3
DayOfTheYear= 342

% Call the prediction function, and passing the parameters useful
% to perform prediction
prediz=fun2;
s_hat = prediz(ThetaLS2(1),ThetaLS2(2),ThetaLS2(3),ThetaLS2(4),ThetaLS2(5
```

```
DayOfTheWeek = 3
DayOfTheYear = 342

s_hat = 160.9136
```

5. Link to our repository project (GitHub)

Here's our **full project repository**: LINK

If you want to see only the **live script**, here's the link: LINK

6. Conclusion

Note that the best model was built on the basis of the data in our possession. If you want to predict the gas consumption in an “**abnormal situations**”, you have to consider that the forecast might be very different from what really happens.

The identification of models is a really **useful “tool”**, which allows you to make **predictions built on data** in possession of any process that you are considering.

Myriad of other applications of this theory.

Important component, among other things, in the **"decision making"** process.

Thanks for your time !