Big Data Project

Hertfordshire and North London Water Quality

Francesco Picciotti (matr. 854021)

Politecnico di Milano

27th August 2017

Outline

1. Data Exploration

2. Towards Time Series
   - Preprocessing
   - Modeling and Validation

3. References

Original Dataset

The original dataset contains 361101 samples (tuples) taken in the area nearby the Herfordshire and North London during a time span that goes from 2009 to 2016.
The measurements are sampled by the UK Environment Agency and other datasets are available **here**. Each tuple has the following fields:

- **_c0**: It's the row index, thus a progressive number
- **@id**: URI identifier
- **sample.samplingPoint**: The URI for making reference to a sampling point
- **sample.samplingPoint.notation**: A shorten string identifing each sampling point e.g.TH-PBRE9999
- **sample.samplingPoint.label**: The full name of the sampling point
- **sample.sampleDateTime**: The date and time when a sample was collected

Tuples' Features cont.

- **determinand.label**: A brief string identifing the determinand sampled, which is the property measured
- **determinand.definition**: A string describing the determinand meaning, its definition
- **determinand.notation**: A string which uniquely identifies the determinand
- **resultQualifier.notation**: This feature can be empty or containing "<",">" stating that is below or above the regulations
- **result**: The amount of the measured determinand
- **codedResultInterpretation.interpretation**: It is an empty column
- **determinand.unit.label**: The unit measure that expresses the **result** field

Tuples' Features cont. II

- **sample.sampledMaterialType.label**: The kind of material (strech of water, matter, ecc.) from which the determinand is sampled
- **sample.isComplianceSample**: It is a boolean to indicate whether the sample has been collected for a compliance purpose
- **sample.purpose.label**: The string describing the kind of the sampling purpose
- **sample.samplingPoint.easting**: The easting of the point on the British National Grid
- **sample.samplingPoint.northing**: The northing of the point on the British National Grid

Finding the most sampled pollutant...

In order to find the most sampled and problematic determinand (pollutant), the FPGrowth algorithm has been applied to the transaction grouped by sampling point and date.
This frequent item algorithm yields **Ammonia(N)** with a count of **17108** (which is the 60%).

...and its reason!

By analyzing the ammonia samples, it was mostly present in:

- Sewage effluents (5129 times)
- River/Running Surface Water (11080 times)

Which is the **95%** (16209/17108) of the times!
Especially the presence in the river surface water is meaningful since the Ammonia traces within water may come from:

- fertilizers
- food processing waste
- Industrial wastewater as non-conventional pollutant

The last two reasons can be both the reason since it implies the presence of industries, whereas the first cause would mean a high concentration of ammonia in the undeground water.

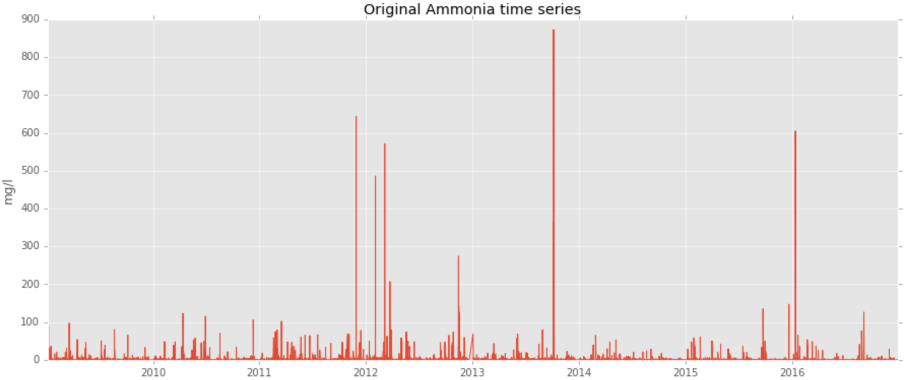
Outline

1. Data Exploration

2. Towards Time Series
   - Preprocessing
   - Modeling and Validation

3. References

Original Time series



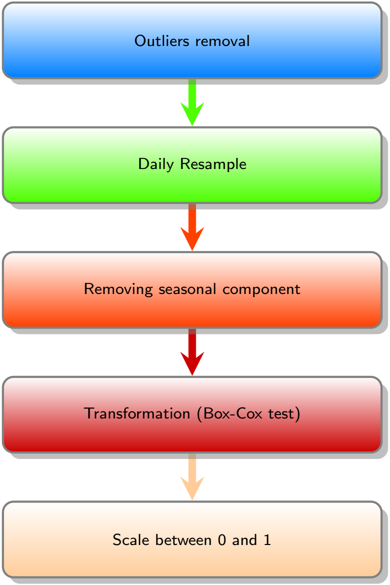
Outline

1. Data Exploration

2. Towards Time Series
   - Preprocessing
   - Modeling and Validation

3. References

Pipeline



Outliers removal

Daily Resample

Removing seasonal component

Transformation (Box-Cox test)

Scale between 0 and 1

Outline

1 Data Exploration

2 Towards Time Series

- Preprocessing
- Modeling and Validation

3 References

ARMA Modeling

The ARMA models comes from statistics background.

- the AR is the **Autoregressive** part that takes in account of the previous steps equals to the AR's degree
- the MA is the **Moving Average** part that averages the stochastic side of the process equals to the MA's degree

The ARMA(2,1) model is chosen by crossing two approaches:

- **AIC** (Akaike Information Criterion) and **BIC** (Bayesian Information Criterion) which are objective methods to select the degree that gives a good fit to the data preventing overfitting. This method computes gives the degree that enables the model to be enough complex to keep all the data's expressiveness
- **PACF** and **ACF**, the Partial AutoCorrelation Function and the AutoCorrelation Function are good indicators for finding out respectively the AR and MA degrees through a "elbow analysis".

Validation I: Whiteness test (or Anderson Test) of the error

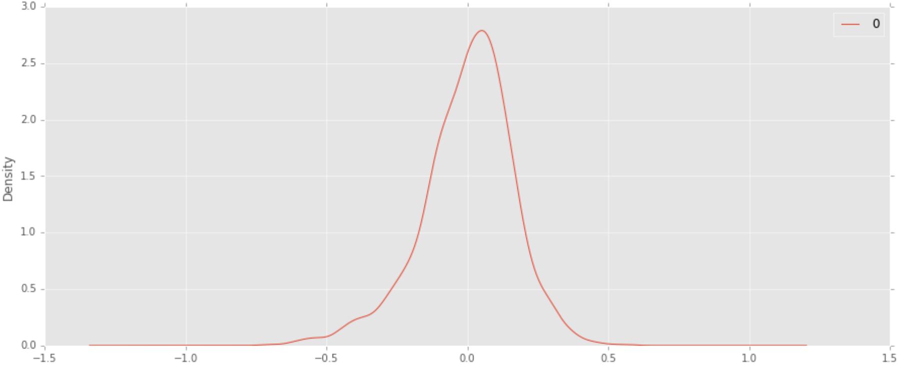


Figure: The error distribution is close to a Gaussian with mean 0 and variance 1

Validation II: Out-sample performances evaluation

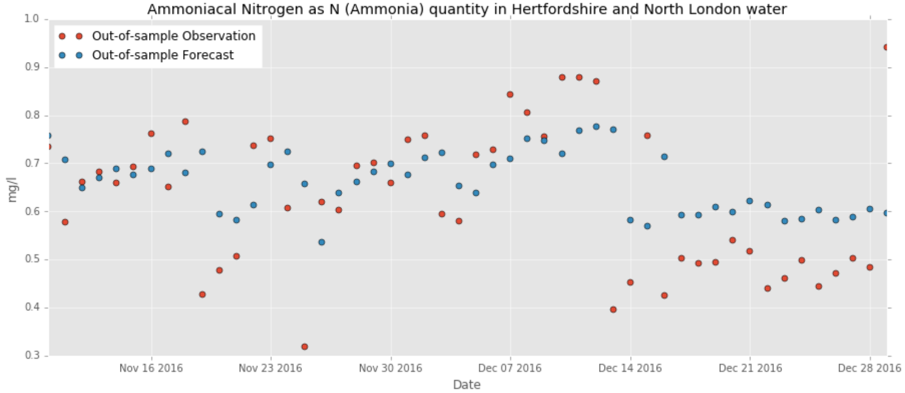


Figure: Mean Squared Error: 0.019193

Outline

1. Data Exploration

2. Towards Time Series
   - Preprocessing
   - Modeling and Validation

3. References

Guidelines

- Spark and Pyspark doc
- Spark understanding and basic knowledge from Big Data part held by prof. Ardagna during the Computing Infrustructures course
- Statsmodels python module references http://www.statsmodels.org/stable/index.html
- Notes from Model Identification and Data Analysis (MIDA/IMAD) course @ Polimi by prof. Bittanti and Savaresi
- Statistical forecasting: notes on regression and time series analysis by Robert Nau (Fuqua School of Business, Duke University) http://people.duke.edu/ rnau/411home.htm
- Cross Validated: Stack Overflow for stats guys https://stats.stackexchange.com/