The European Commission's science and knowledge service

Joint Research Centre

From big data to smart data

Álvaro Gómez Losada and Néstor Duch Brown
Joint Research Centre (Seville, Spain)

NAEC Conference, 16 April 2019

European Commission

Background – what do we already know?

- Knowledge Discovery in Data Bases (KDD) – Fayyad et al. (1996):
- -> **Domain knowledge** + application of empirical discovery algorithms
- Volumen of data is not equivalent to knowledge
- Expensive resources to store and analyze big amount of data
- Scientific production related to new machine learning algorithms and artificial intelligence is vast (and growing)

Three empirical use cases: machine learning for small data sets

- Three examples in which the application of machine learning is possible in small data sets:

  (1) a price recommendation system.

  (2) a forecasting system for ozone in cities.

  (3) a forecasting system for demographic population.

Application 1: a price recommendation system

- Source of data: Amazon market place
- Objective: forecasting product prices
- Size of data: **~ 5 MB**
- Methodology: collaborative filtering for recommendation systems
- Technology: *python language*
- Characteristics: unequally spaced time series (*events*)
- Dissemination: *Monograph in Business information Systems* (Springer, 2019) – to appear

Application 1: a price recommendation system

Transforming a TS in the (user, item, frequency) triple emulating a Recommendation scheme

Table 1. Some equivalences used in this study to build the forecasting recommender system (RS) from a given time series (TS).

Concept	Symbol	RS equivalence
Set of distinct values in TS	$\mathcal{U}$	Set of users with cardinality m
Set of distinct values in TS shifted	$\mathcal{I}$	Set of items with cardinality n
Two distinct values in the TS	u_j, u_l	A pair of users
Two distinct values of the shifted TS	i_i, i_j	A pair of items
Number of times a distinct value in the TS and its shifted version co-occurs	r_{jk}	Rating of user u_j on item i_k
TS value to which perform a forecasting	u_a	Active user to which recommend an item

Application 1: a price recommendation system

Table 2. TS characteristics used in the methodology testing (P: percentile; min: minimum value, max: maximum value; in €).

Abbreviation	Length	min	P50	max	P75-P25	Distinct values (m)
TS-1	2169	13	84	142	14	82
TS-2	1224	7	17	36	8	22

Evaluation

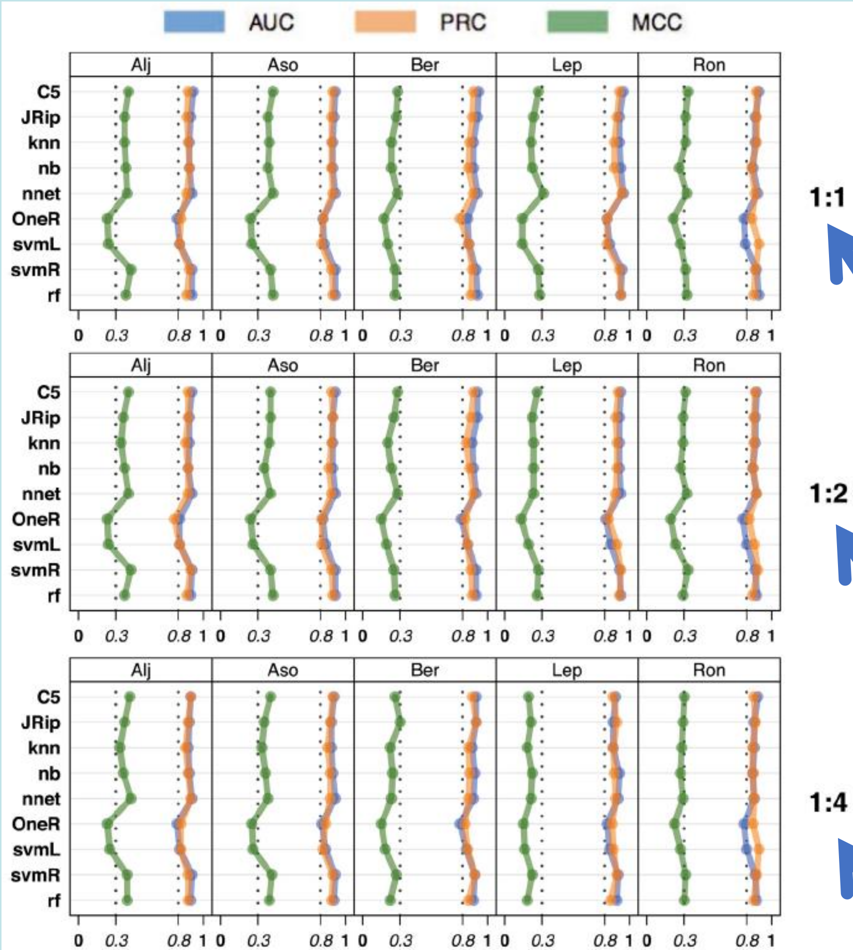
Table 3. MAE performance on both TS and the different similarity measures (s_1,s_2 and s_3: Pearson correlation, cosine and Otsuka-Ochiai similarities, respectively), in €.

	TS-1			TS-2		
	s_1	s_2	s_3	s_1	s_2	s_3
MAE	6.2	6.5	5.5	6.0	3.2	3.0

Application 2: a forecasting system for ozone in cities

- Source of data: air quality monitoring network.
- Objective: forecasting high levels of ozone to prevent population
- Size of data: **~ 100 MB**
- Methodology: supervised classification + stacking of 10 classifiers
- Technology: *R language*
- Characteristics: time series with marked patterns
- Dissemination: *Environmental Modelling and Software* 2018, 110: 52-61

Application 2: a forecasting system for ozone in cities

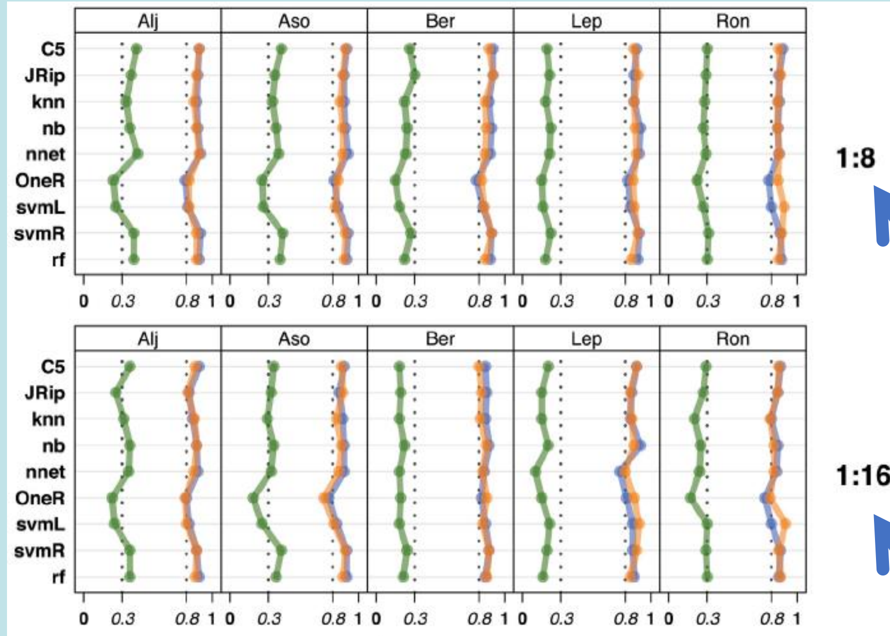


Accuracy

1:1

1:2

1:4



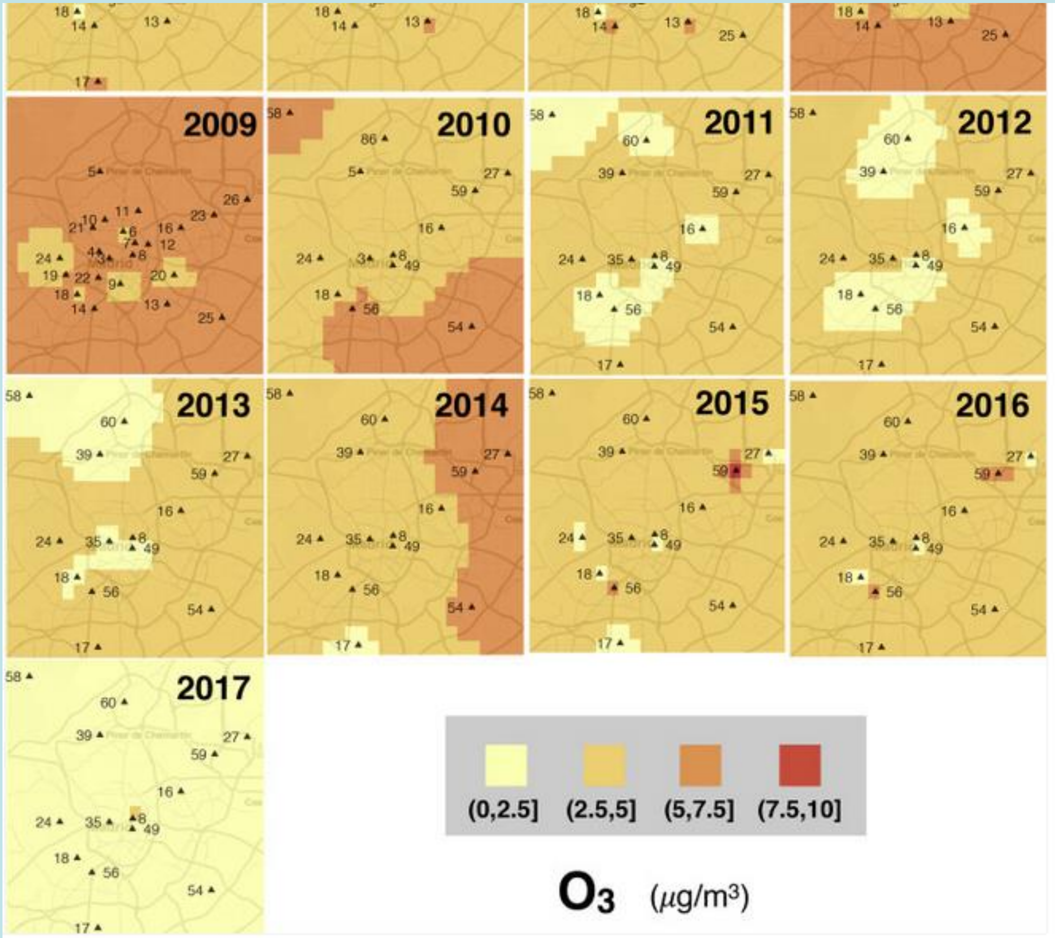
1:8

1:16

Fraction of the training data

European Commission

Application 2: a forecasting system for ozone in cities



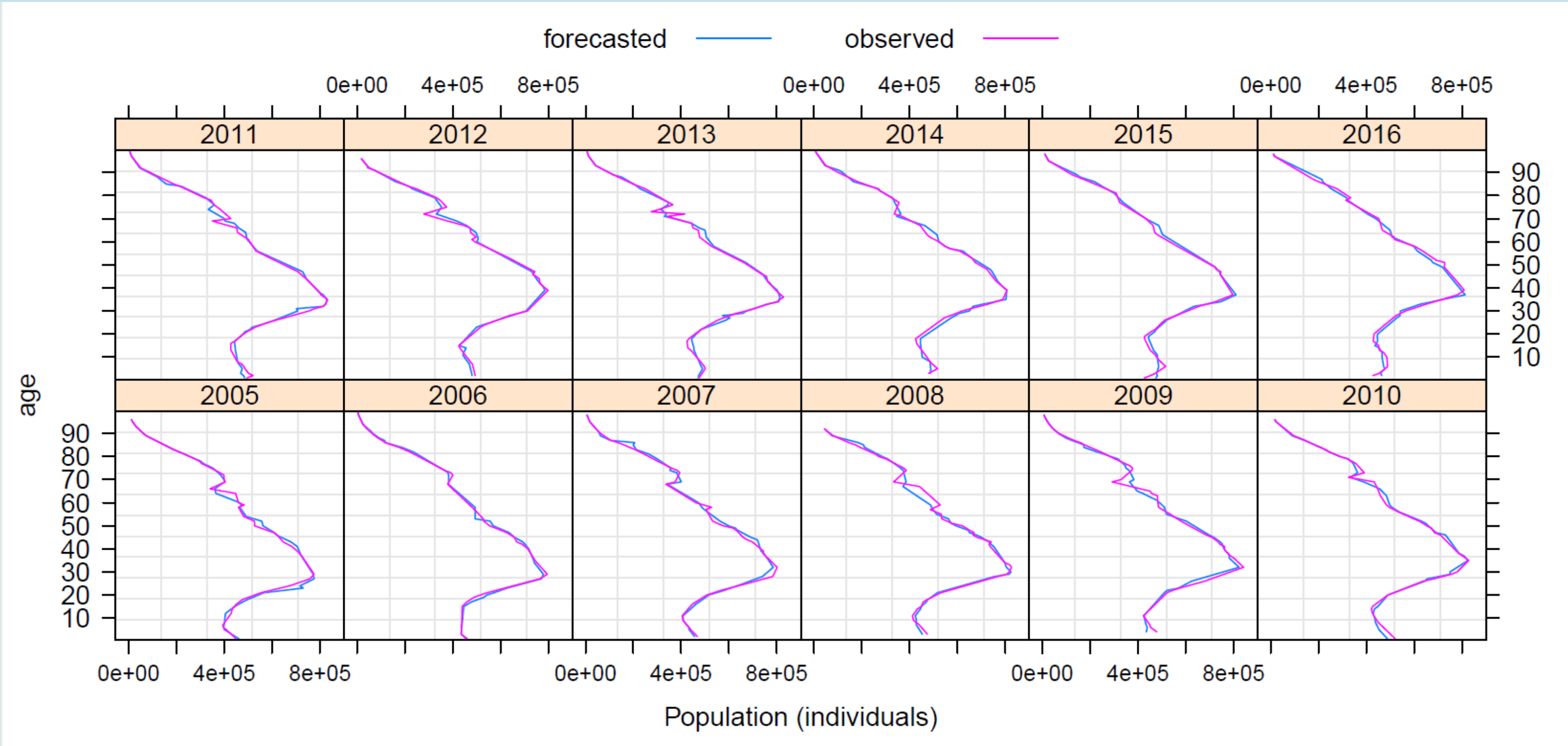
Estimation of Ozone (O3) using the original data or different fractions produce reliable estimations.

Madrid (Spain)

Application 3: a demographic forecasting system

- Source of data: Eurostat.
- Objective: forecasting Spanish population pyramides from 2005 to 2016
- Size of data: **~ 10 MB**
- Methodology: supervised learning + random forest algorithm
- Technology: *R language*
- Characteristics: marked patterns, little quantitative variations
- Dissemination: European Conference on Quality in Official Statistics (Krakow, 2018)

Application 3: a demographic forecasting system



European Commission

Application 3: a demographic forecasting system

Comparison with classical techniques:

Machine learning (ML) vs Arima & exponential smoothing (ES)

Units: population (individuals)

	ML	Arima	ES
RMSE	20318	81010	81247
MAE	14353	60596	62269

RMSE: root mean square deviation

MAE: mean absolute error

What we have learnt

In this time of Artificial Intelligence and pervasive data:

(1) These modest examples could suggest to **look back** to those small data size present in organizations: **value in small (high quality) data sets**

(2) Research is needed to adapt machine learning algorithms to small/medium data set sizes. **What amount of data do we really need?**

(3) We should ascertain the **knowledge** data sets can provide to us

(4) **Democratization of technology**

(5) Probably many information still remains **hidden** which is not analyzed due to the size of the data is not considered suitable.

(6) Lack of data analysis / machine learning skills in public administrations?

Thanks

Questions?

You can find us at:

alvaro.gomez-losada@ec.europa.eu

nestor.duch-brown@ec.europa.eu

European Commission