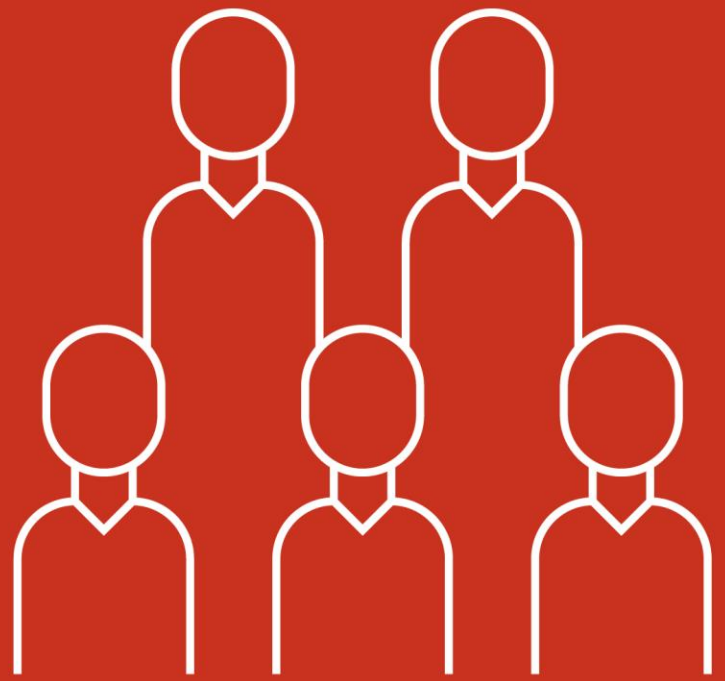


11th European Conference on Quality in Official Statistics (Q2024)

"Data mining techniques on the administrative data system to enhance the accuracy of the population census counts"

Antonella Bernardini, Nicoletta Cibella, Giampaolo De Matteis, Gerardo Gallo, Antonio Laureti Palma, Fabrizio Solari - Permanent Population Census Division

CENSIMENTI PERMANENTI L'ITALIA, GIORNO DOPO GIORNO. POPOLAZIONE E ABITAZIONI

Istat Istituto Nazionale di Statistica

Data mining in the signs-of-life database used for the 2021 Italian population census

Summary:

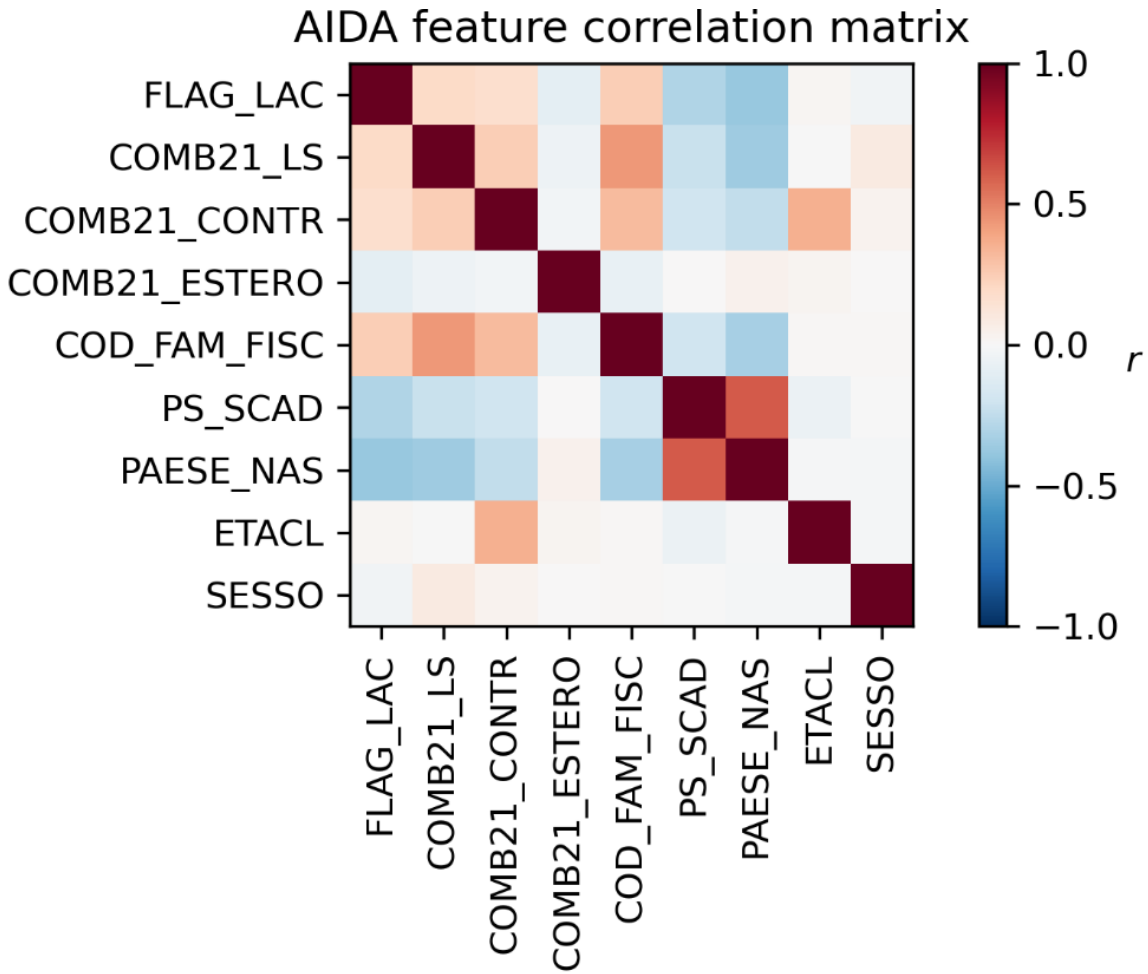
1. signs of life used for determining the usual resident or non-usual resident population
2. data preparation: derivation of a supervised training set
3. the neural network model used
4. data analysis of the neural network output
5. evaluating the most probable place of usual residence based on utilities
6. identifying household consumption patterns
7. data preparation: derivation of unsupervised training sets
8. definition of a forecasting model
9. data analysis of the forecasting outputs

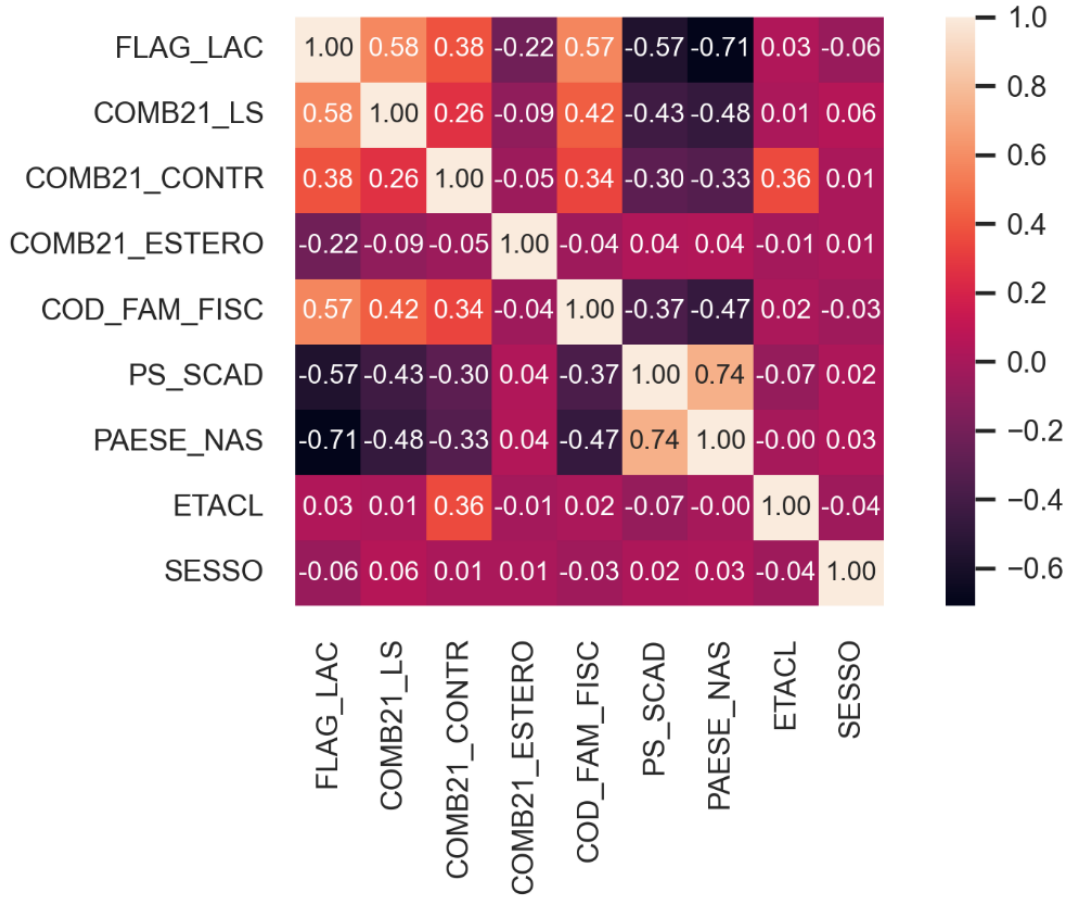
1 - Signs of life used for determining the usual resident or non-usual resident population: used features

- COMBI21_LS: work and study signals
- FLAG_LAC: presence or absence of information in the population register
- COMB21_CONTR: active contracts (about car, house rental, real estate)
- COMB21_ESTERO: signal of presence abroad (Income Database, Consular population register of Italians abroad)
- PS_SCAD: permits to stay from 2012 to 2021
- COD_FAM_FISC: household ID in the Tax register
- PAESE_NAS: country of birth in the Tax register
- ETACL: age in the Tax register
- SESSO: sex in the Tax register

1 - Signs of life used for determining the usual resident or non-usual resident population: analysis of the used features

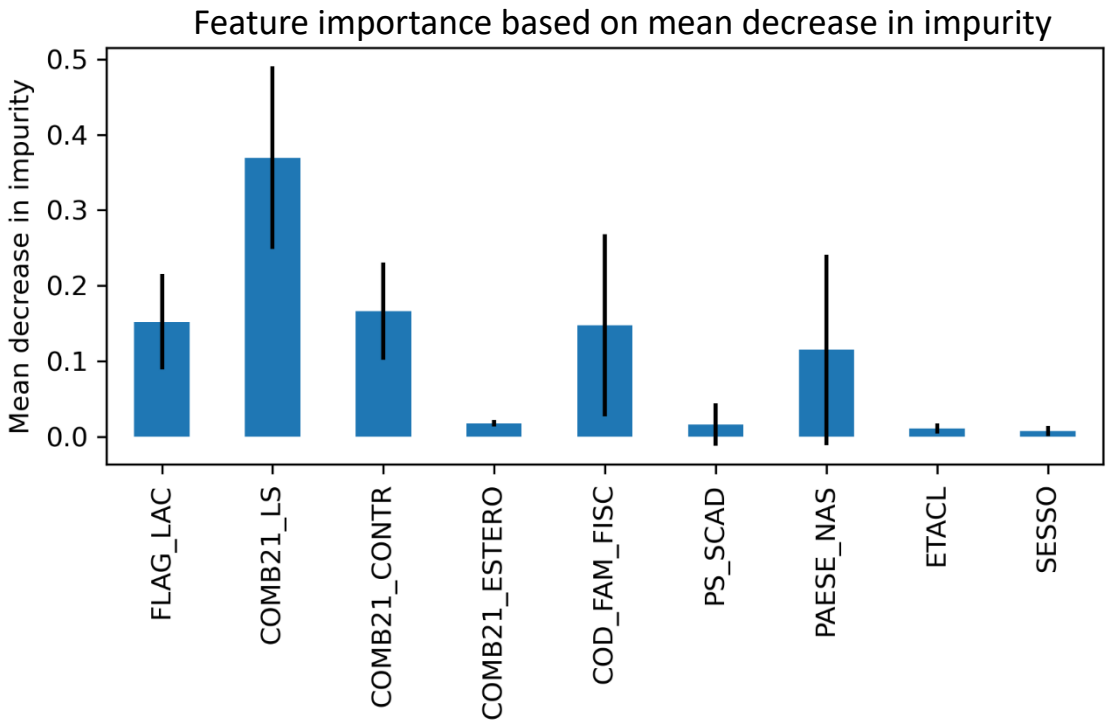
Correlation matrix of the features considered

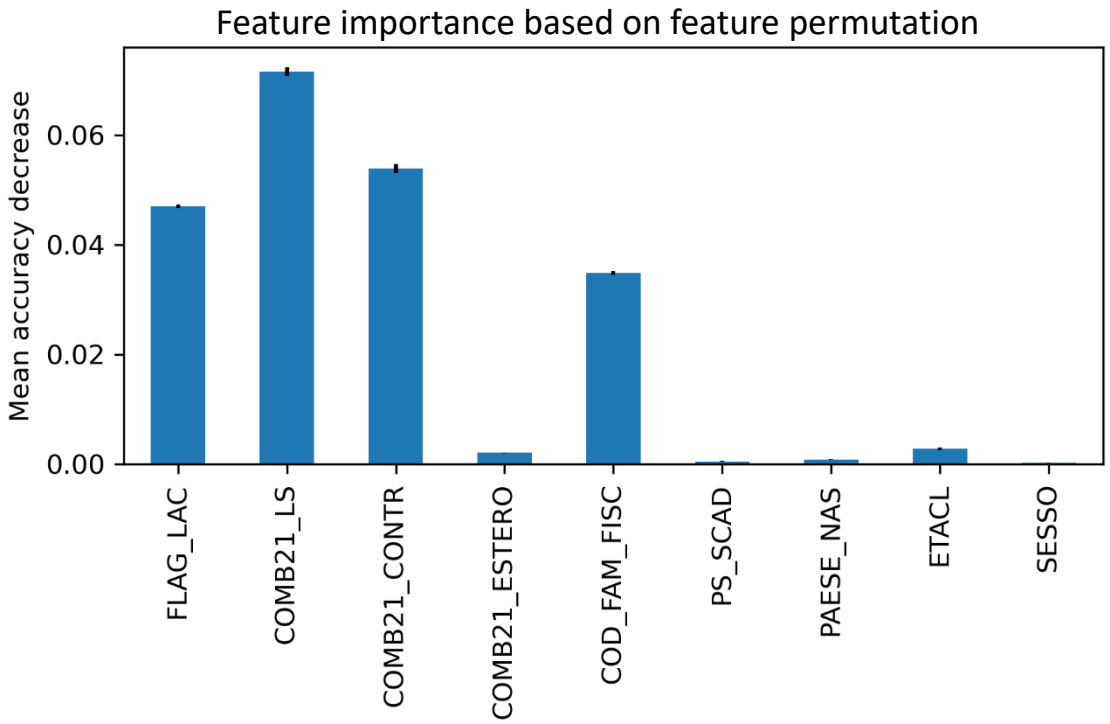




1 - Signs of life used for determining the usual resident or non-usual resident population: analysis of the used features

- ✓ Random forest algorithm to evaluate feature importance: blue bars represent the importance of forest characteristics, along with the variability between trees represented by the error bars.
- ✓ Random forest with permutation feature to overcomes limitations of the impurity-based feature importance.





usual residents set): identified through the sample survey of the annual census

1,300,000 raw records

non-usual residents set): two hypotheses were considered

i. feedback information from respondents obtained through the 2021 field survey surveyors

46,000 raw records → 1,800 cleaned records + 10% of non cleaned records

ii. information from the AIDA (SoL database) production process

1,800,000 *raw records → 137,000 cleaned records*

3- Machine Learning model used: model analysis

- ✓ The class-weighted Support-Vector Machines (SVM) classifier model was used.

- ✓ The performance of the SVM was evaluated using different kernels:
  - ✓ linear SVM
  - ✓ radial basis function kernel SVM (SVM-RBF)
  - ✓ polynomial SVM

- ✓ From the performance analysis with respect to success percentages on a test set and taking into account the problems of over and underfitting, the SVM – RBF proved the best model

- ✓ The choice of the best model and the best fitting parameters was identified through:
  - ✓ model selection using nested grid search
  - ✓ measurement checks of under/over fitting via the learning, validation and ROC (Receiver Operator Characteristics) curves

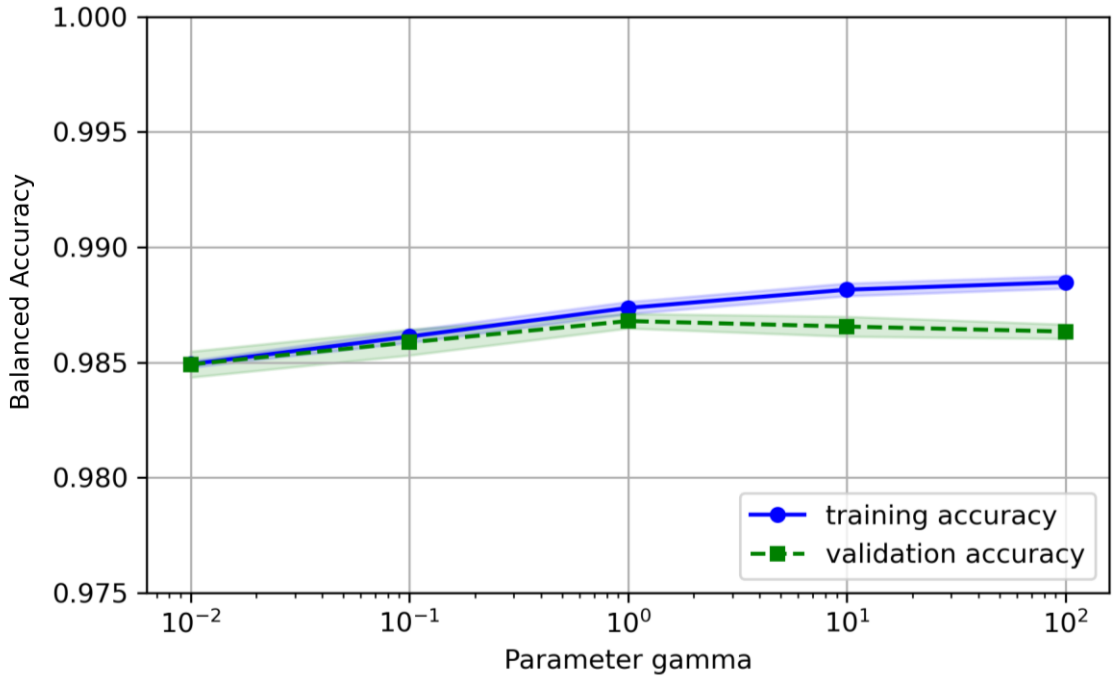
3 - Machine Learning model used: hyperparameter evaluations

case A, hypothesis i): non-usual residents from sample surveys - sample size: 85,365 records (non-residents 2%)

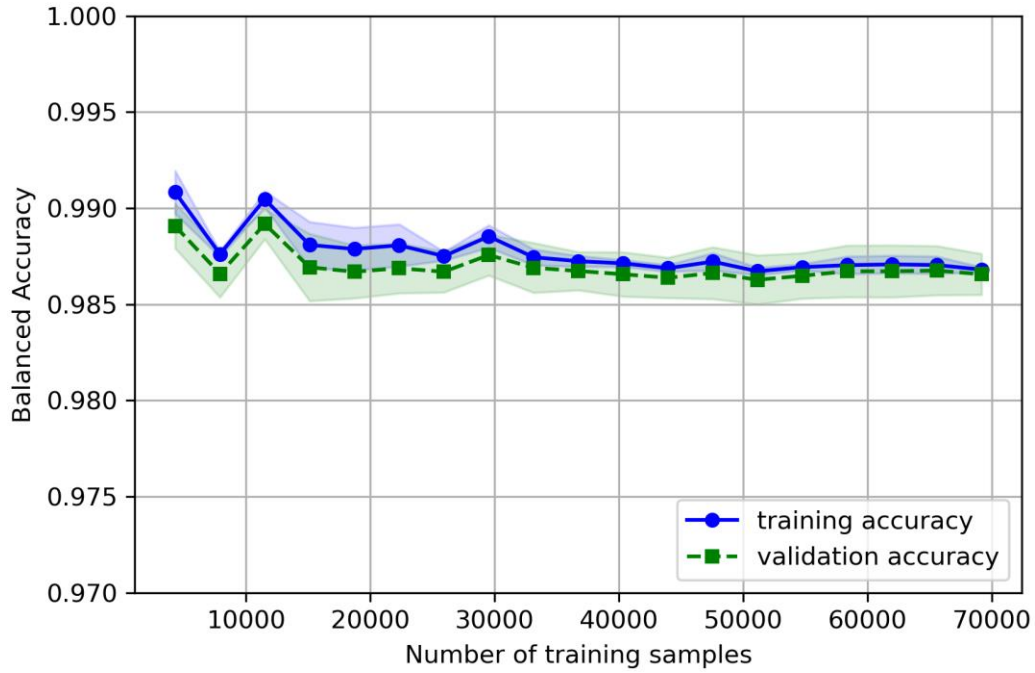
SVM-RBF kernel: checking the over/under fitting and sample size dimension

C is the regularization parameter, γ is the inverse of the radius of influence of samples selected

[*C(non res)=30, C(res)=1*]
validation curve



C=1 [*C(non res)=30, C(res)=1*], *γ=0.1*
learning curve

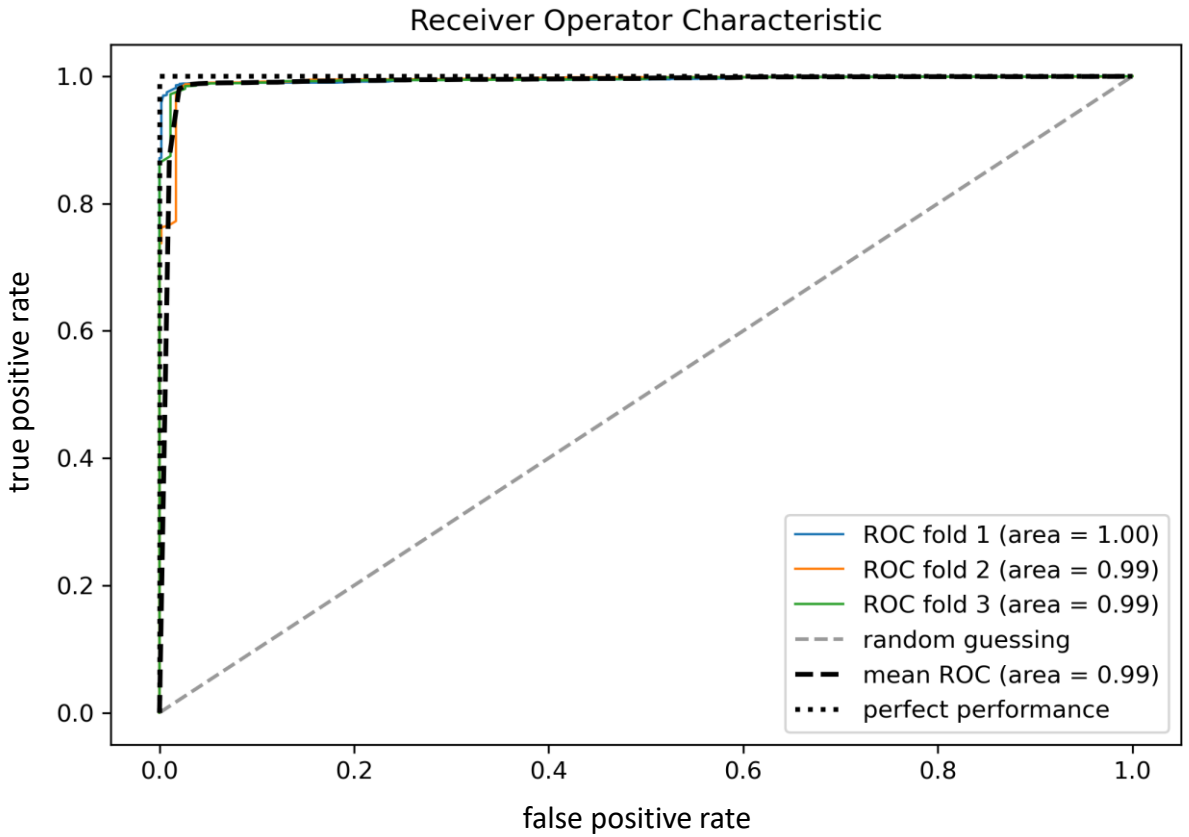


3 - Machine Learning model used: hyperparameter evaluations

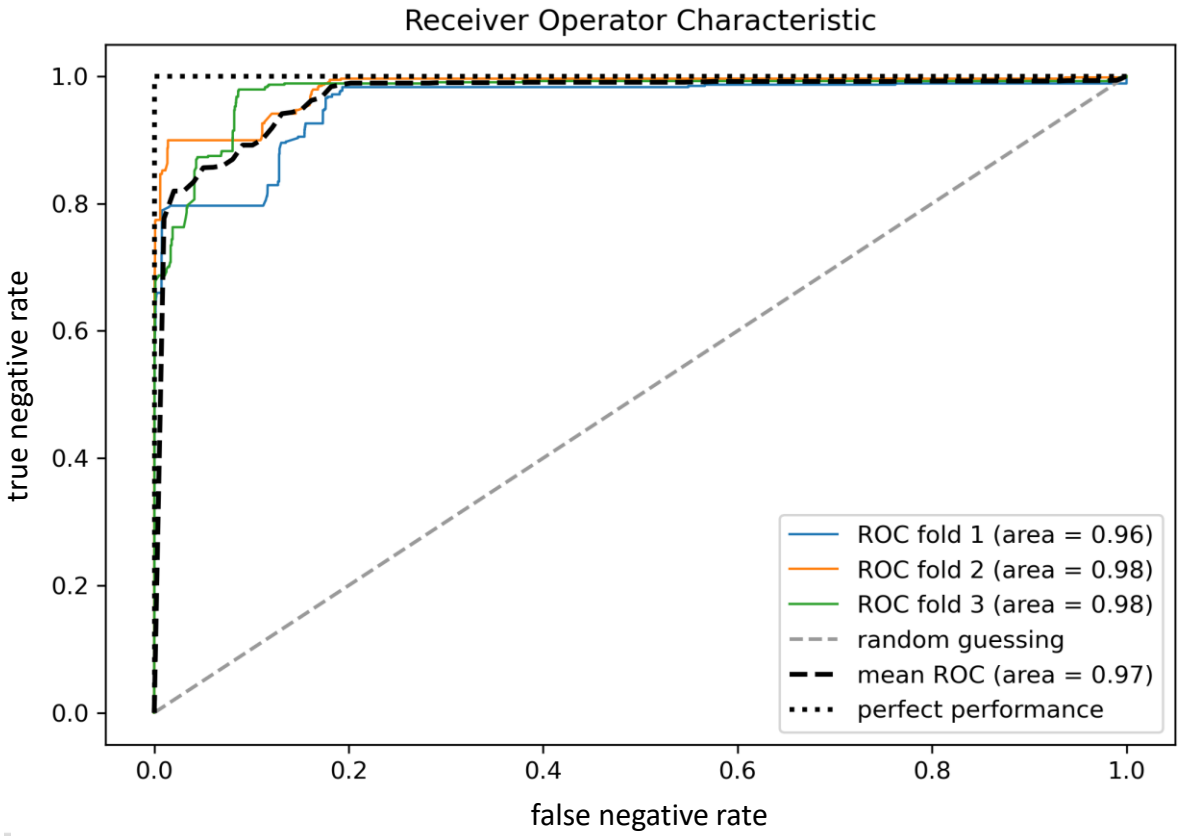
case A, hypothesis i): non-usual residents from sample surveys - sample size: 85,365 records (non-residents 2%)

SVM-RBF kernel: *C=1* [*C(non res)=30, C(res)=1*], *γ=0.1*

validation curve *C=1*



learning curve *C=1, γ=1*



4 - data analysis of the neural network output: case A/i: non-usual residents from sample surveys

- ✓ SVM - RBF kernel: *C=1.0 Gamma= 0.1* sample size: *85,365* record (non-residents *2%*)

- ✓ We identified four sub groups:

  A-B) The resident and non-resident are the two groups of individuals where the SVM output and the official population register are coherent,

  C) the overcovered population indicates non-residents for the SVM output who are classified as residents in the official population registers

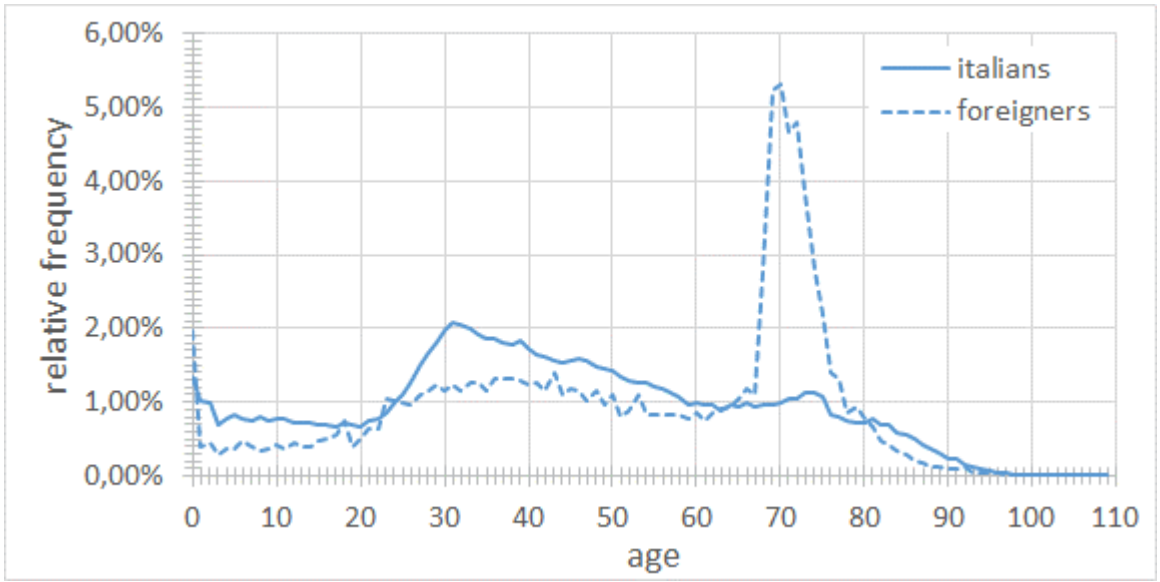
  D) the undercovered population indicates residents for the SVM output but non-resident in the municipal population registers.

ML vs Population Register: relative frequency of the population

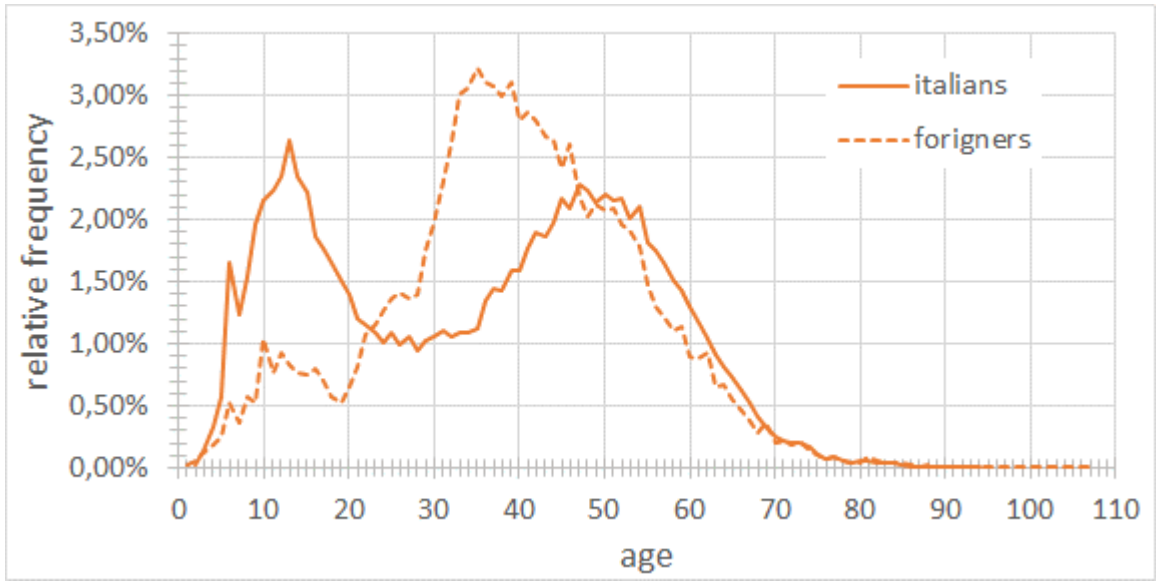
	Population Labels	ML	PR	Percentage
A	Not Resident (Matched)	non resident	non resident	1.76%
B	Resident (Matched)	resident	resident	97.81%
C	Overcovered	non resident	resident	0.33%
D	Undercovered	resident	non resident	0.10%

✓ Relative frequency of the citizenship in the population register compared to the age of the individuals:

over-coverage



under-coverage



5 -Evaluating the most probable place of usual residence based on utilities

- ✓ Place of residence discrepancies can cause misplacement errors that produce over and under-counting, simultaneously within two different municipalities.
- ✓ Through the association between the information on households included in statistical registers and the utility consumption patterns it might be possible to assess the usual place of residence
- ✓ We use utility consumption, electricity and gas, as data sources to identify the monthly consumption patterns associated with each point of delivery.
- ✓ The data on electricity and gas consumption are provided by the Regulatory Authority for Energy, Networks and the Environment (ARERA).
- ✓ All identification data have been anonymized, both for the contract holder and the place of supply. Therefore, It was only possible to link an energy-contract with an anonymous person, belonging to an anonymous family, simply at the municipal level.

  Therefore we can distinguish two groups:

  A. Households with a single contract;
  B. Households with multiple contracts.

5 - Evaluating the most probable place of usual residence based on utilities

- ✓ Through each home's energy consumption model, it was possible to estimate effective household location.

- ✓ We carried out an analysis of the possible consumption models that emerged from the data using simple k-means cluster analysis, hypothesizing eight different possible centroids

- ✓ The k-means metric was based on the Euclidean distance between different normalized monthly consumptions and up to the second-order derivative of the monthly trend

- ✓ Two types of consumption profiles emerged from the analysis:
  - i) consumption apparently for home, i.e. patterns with limited consumption variations during the observed year (max variations of the order of 100%)
  - ii) consumption apparently for not usual or residential homes; i.e. patterns with high consumption variations during the observed year (variations of the order of 1000% or more)

5 - Evaluating the most probable place of usual residence based on utilities

- ✓ We worked on:
  22,001,036 electricity contracts (*13,776,499* home units; *8,224,537* business units)
  14,516,785 gas contracts

- ✓ We built two groups:
  A. Households with a single contract: 13,136,120 electricity contracts

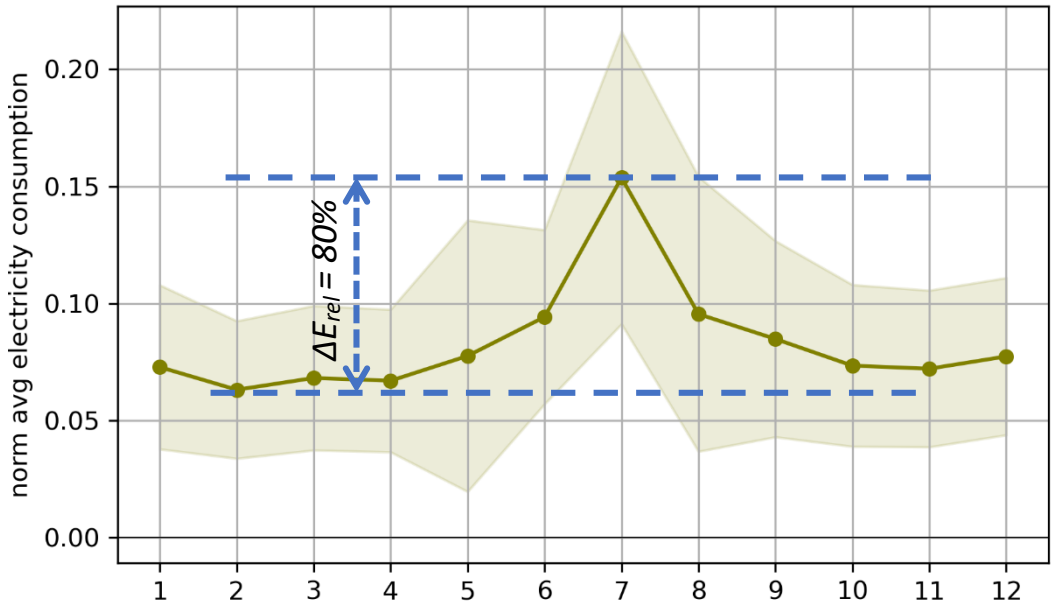
  B. Households with multiple contracts: 8,864,916 electricity contracts, for 3,383,369 families

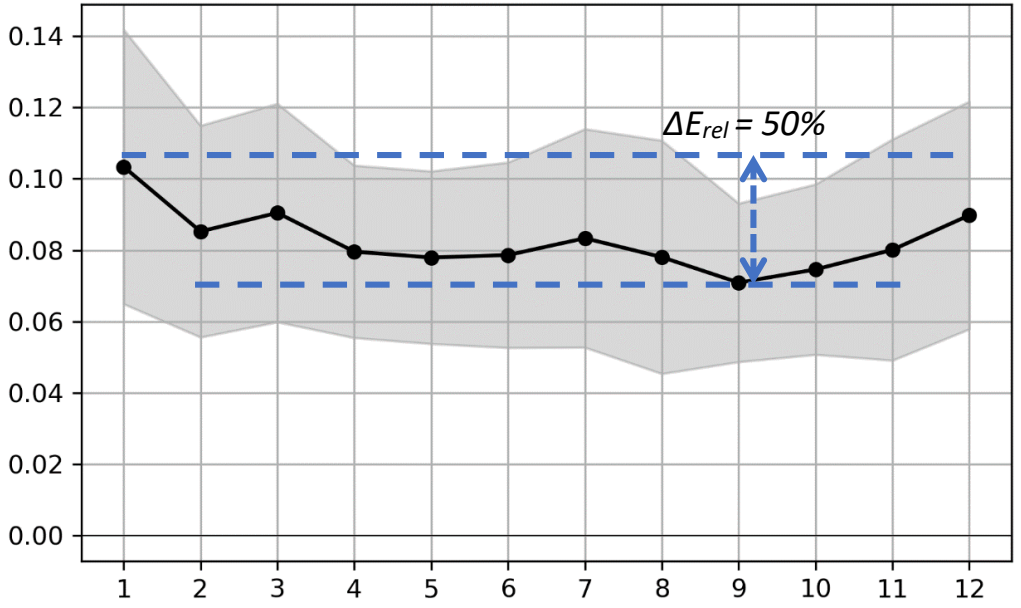
- ✓ We carried out a k-means clustering using the above mentionated Euclidean distance

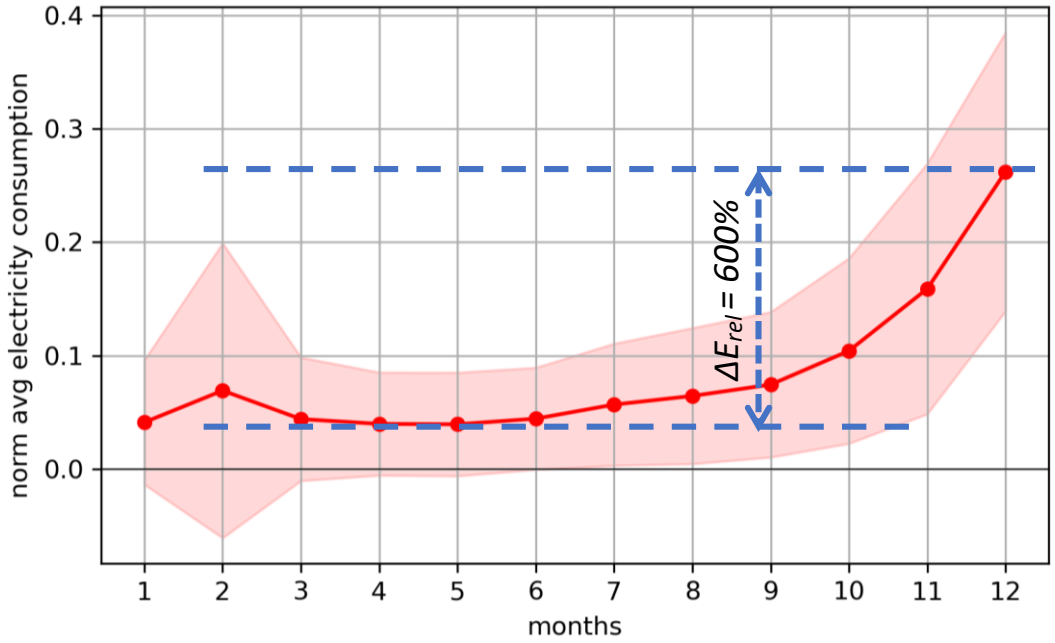
- ✓ In the context of an unsupervised classifier, we used group A as the training set and then applied the prediction to group B.

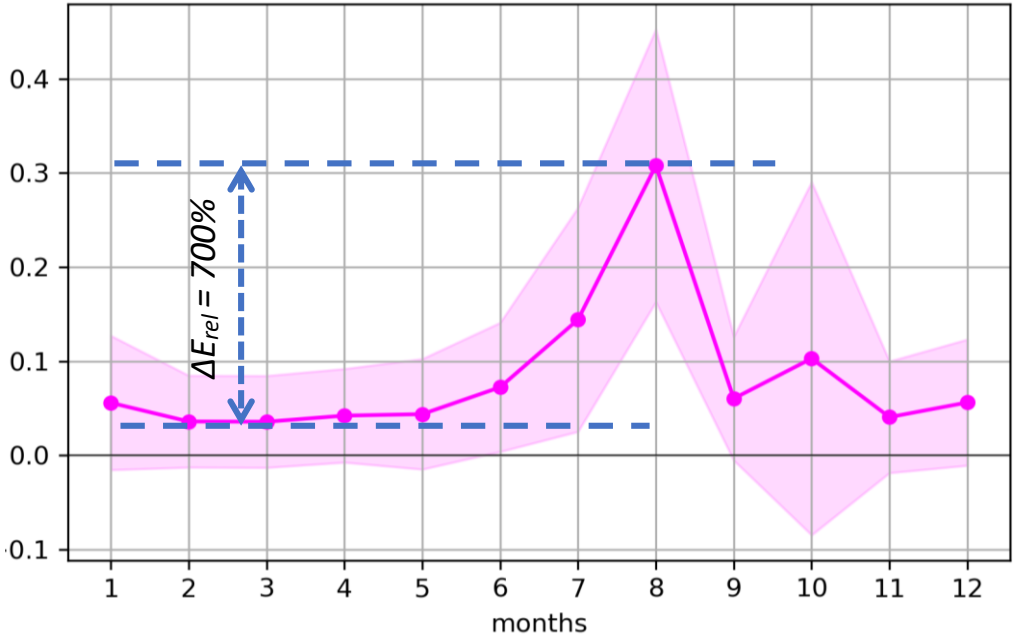
6 - identifying household consumption patterns: clusters of energy consumption patterns

✓ maximum relative variation in average monthly consumption



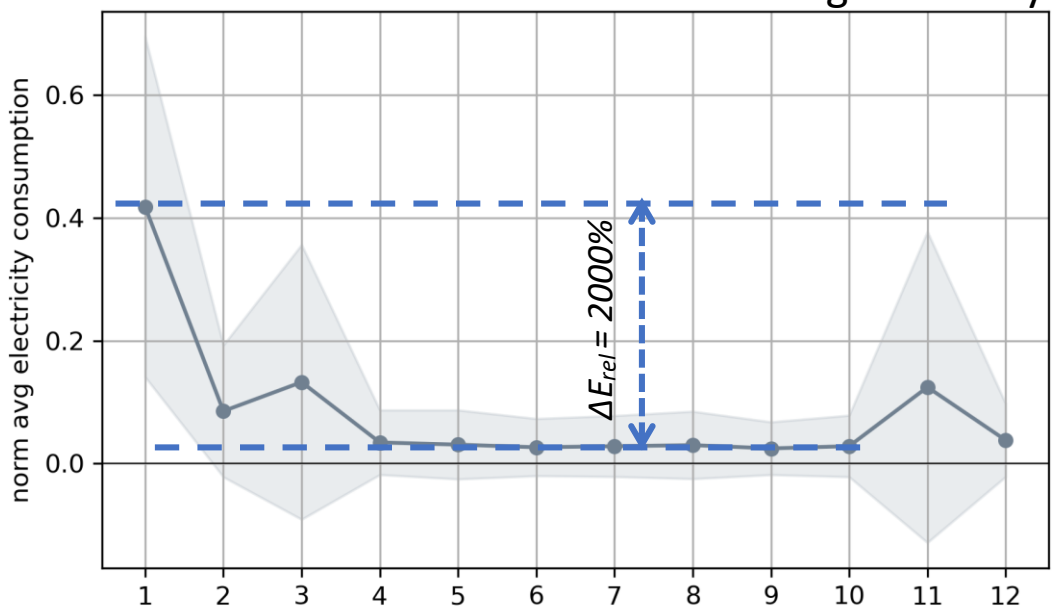


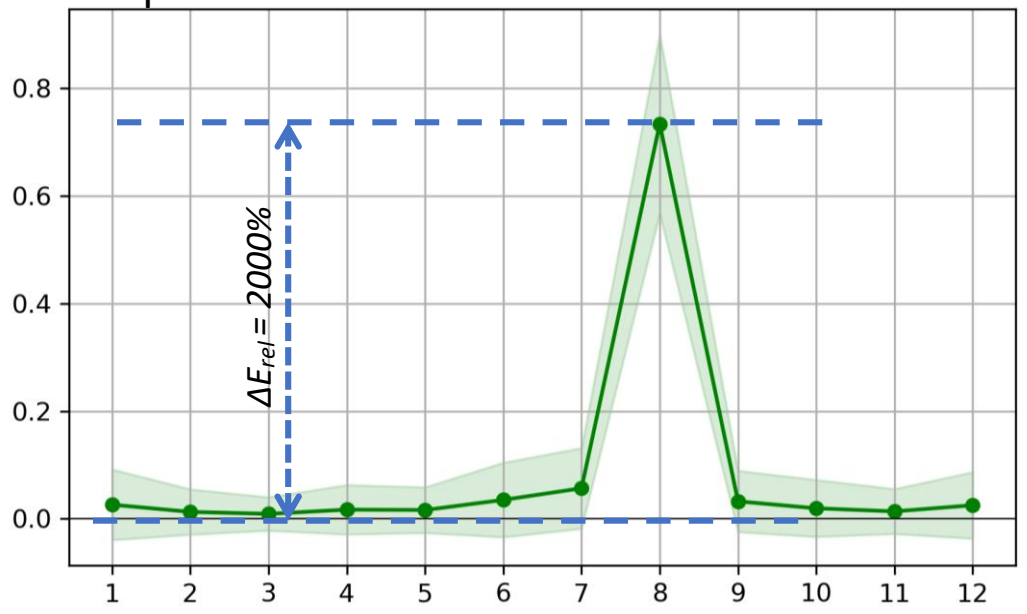


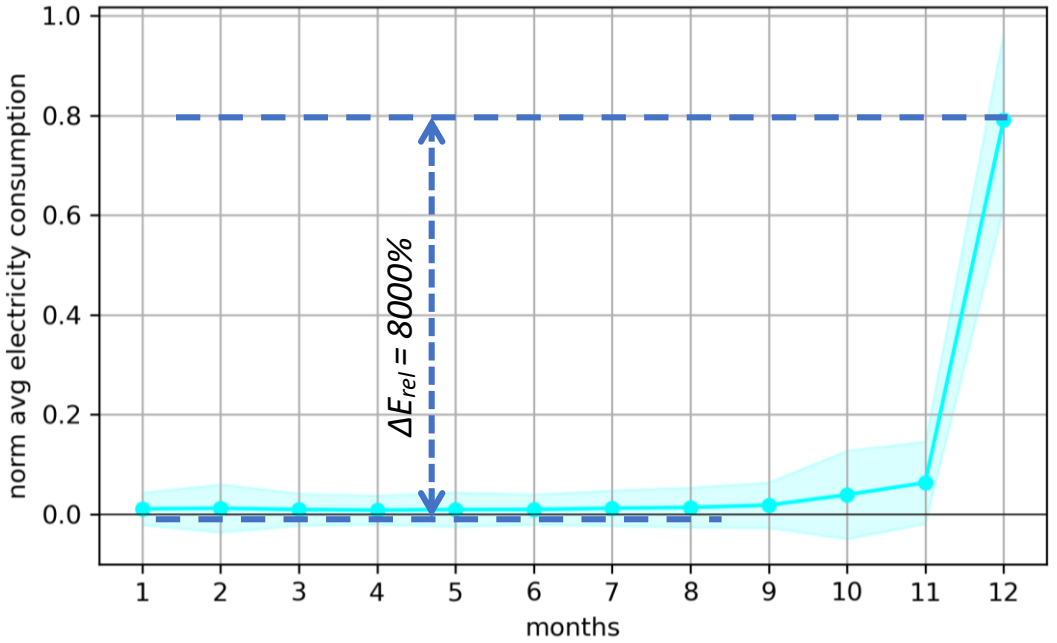


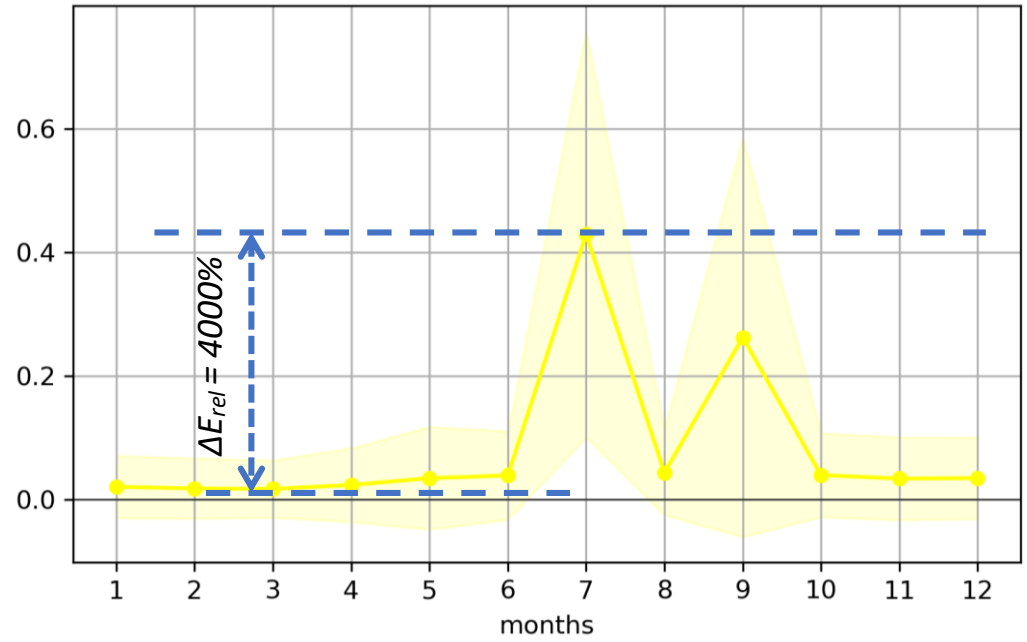
6 - identifying household consumption patterns: clusters of energy consumption patterns

✓ maximum relative variation in average monthly consumption









6 - data preparation: derivation of unsupervised training sets

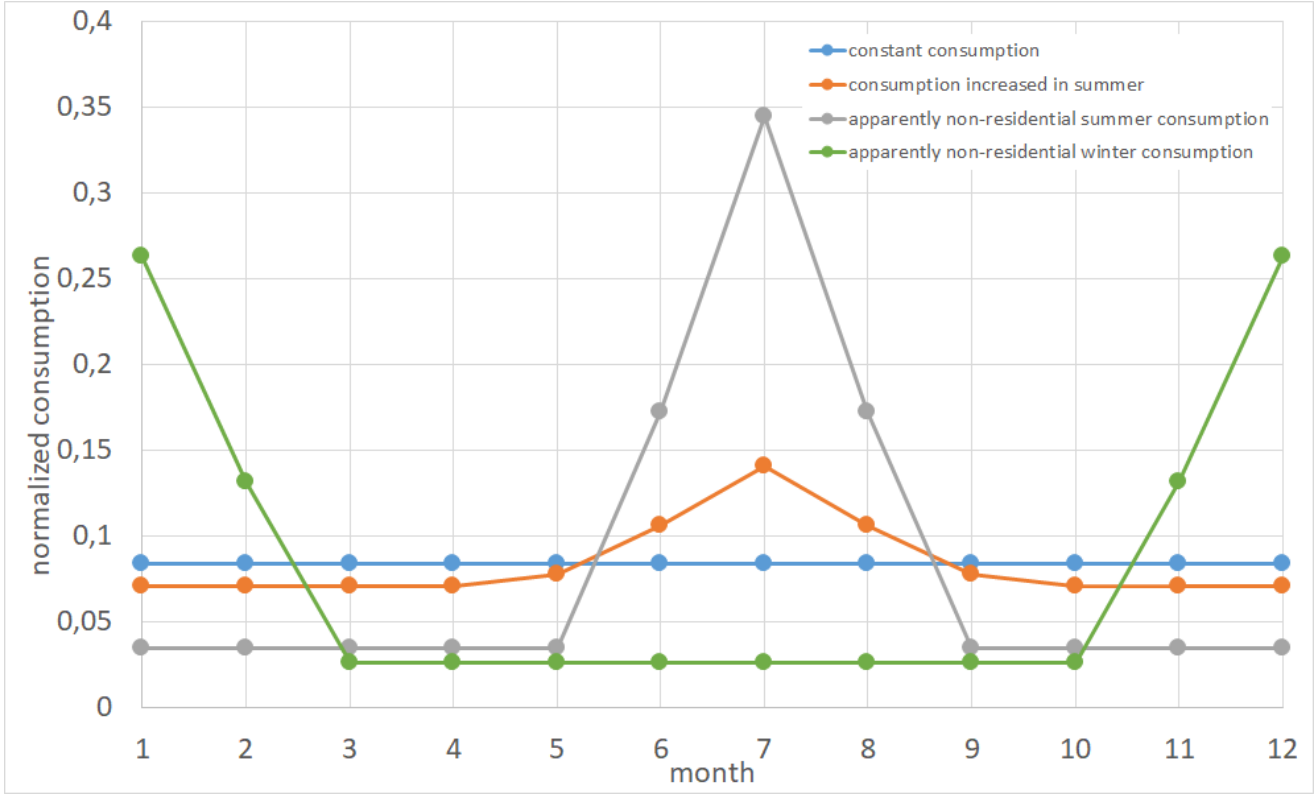
✓ We predefined four centroids to represent the two types of consumption profiles:

i) consumption for residential homes: centroids with limited variations in monthly consumption:

- constant monthly consumption
- monthly consumption that depends on the season

ii) consumption for apparently not usual or residential homes: centroids with high variations in monthly consumption:

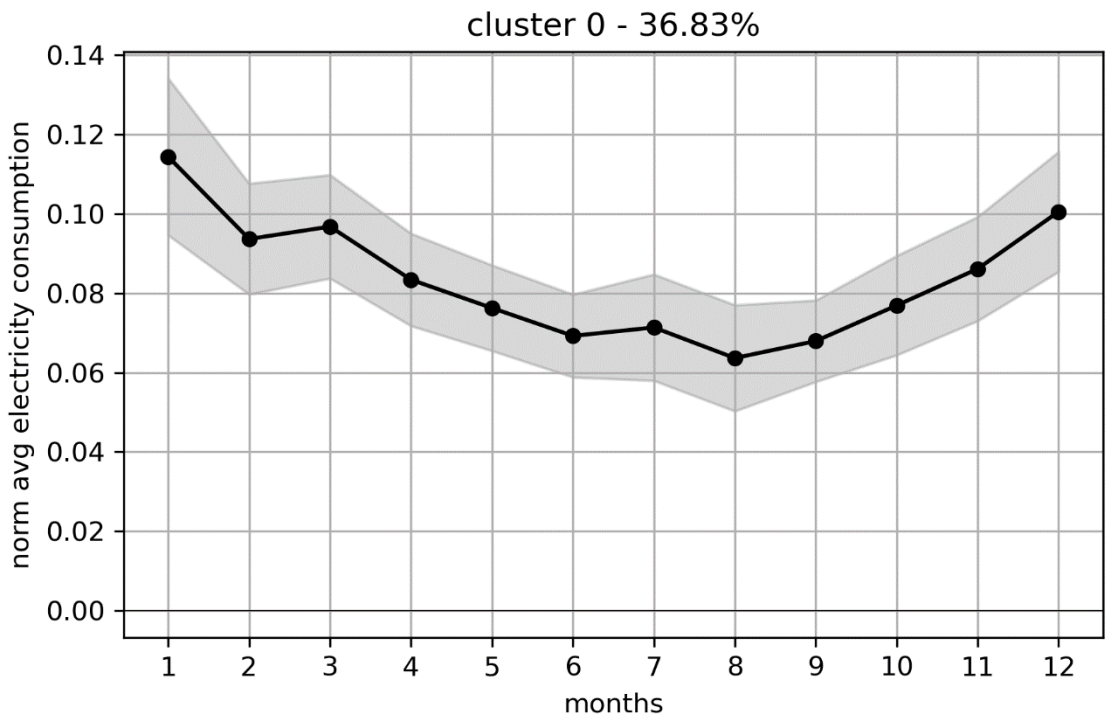
- non-residential summer consumption
- non-residential winter consumption



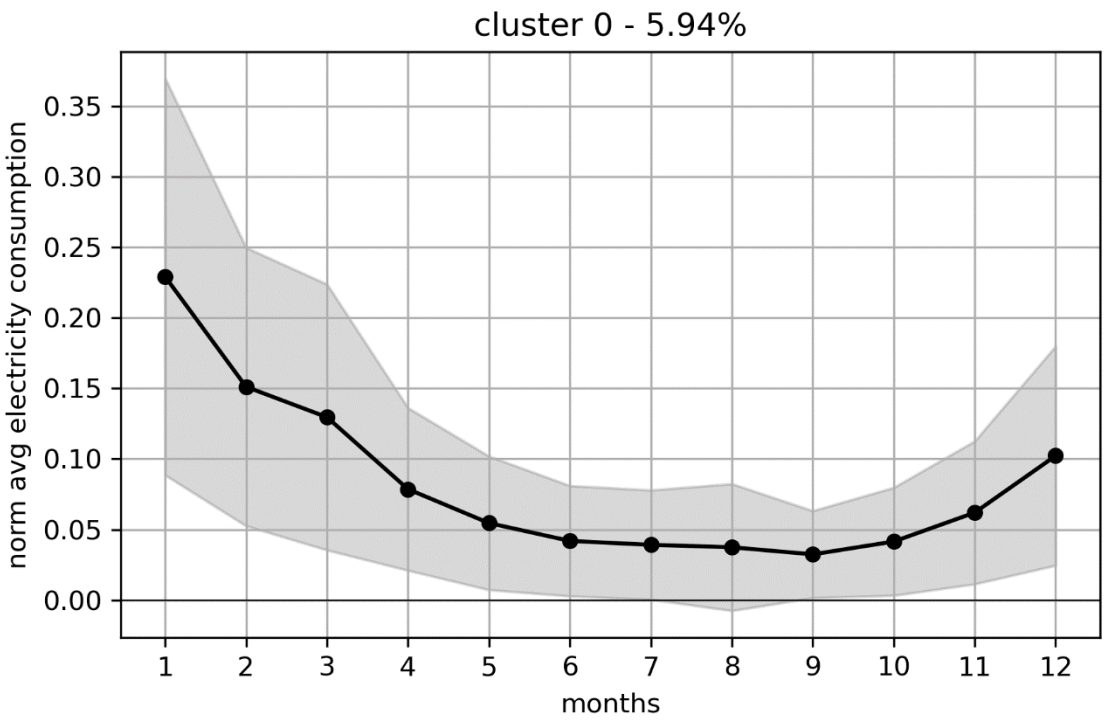
7 - data preparation: derivation of unsupervised training sets

✓ Centroid: <u>constant monthly consumption</u>

«group A»



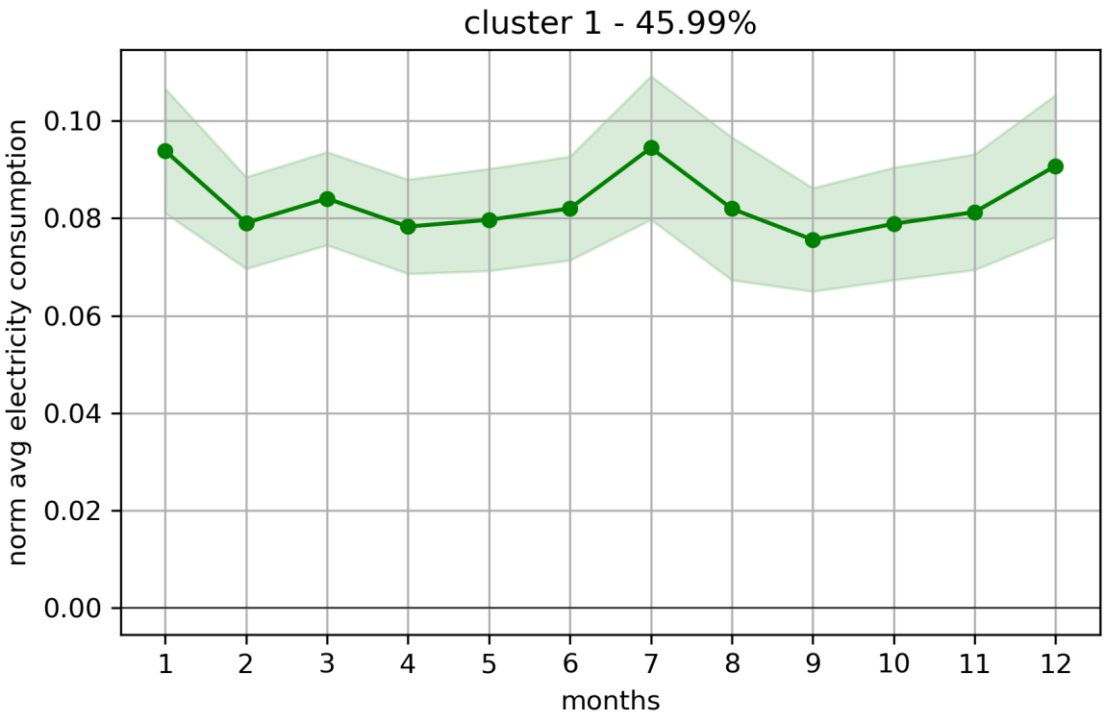
«group B »



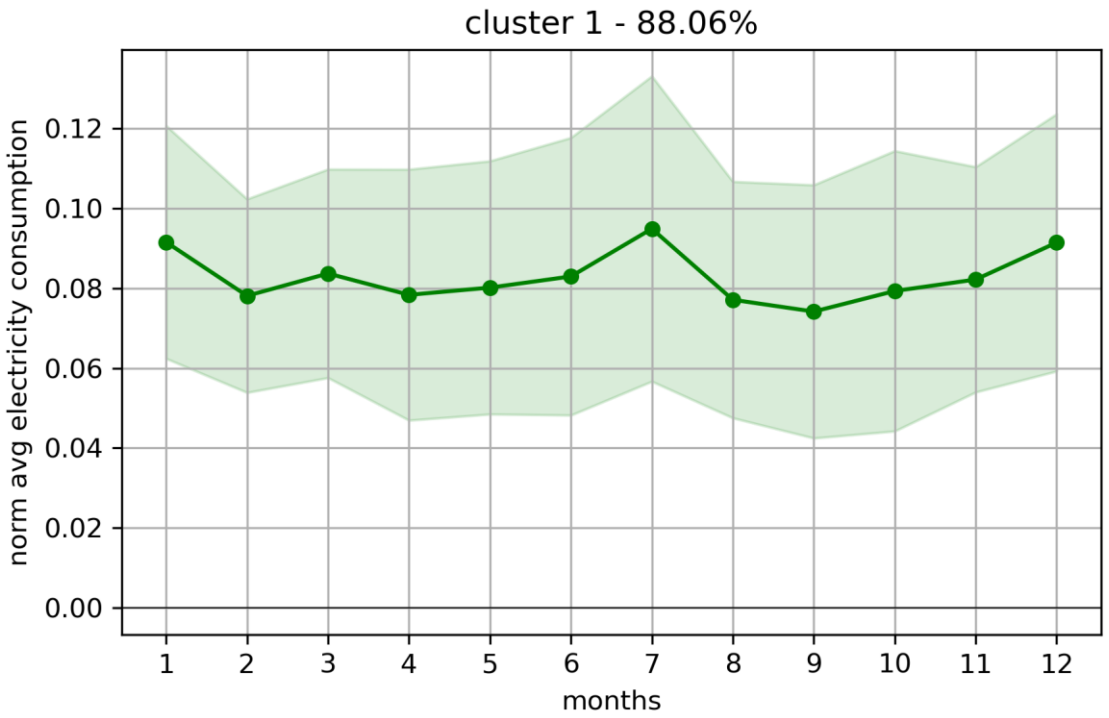
7 - data preparation: derivation of unsupervised training sets

✓ Centroid: <u>monthly consumption that depends on the season</u>

«group A»



«group B »



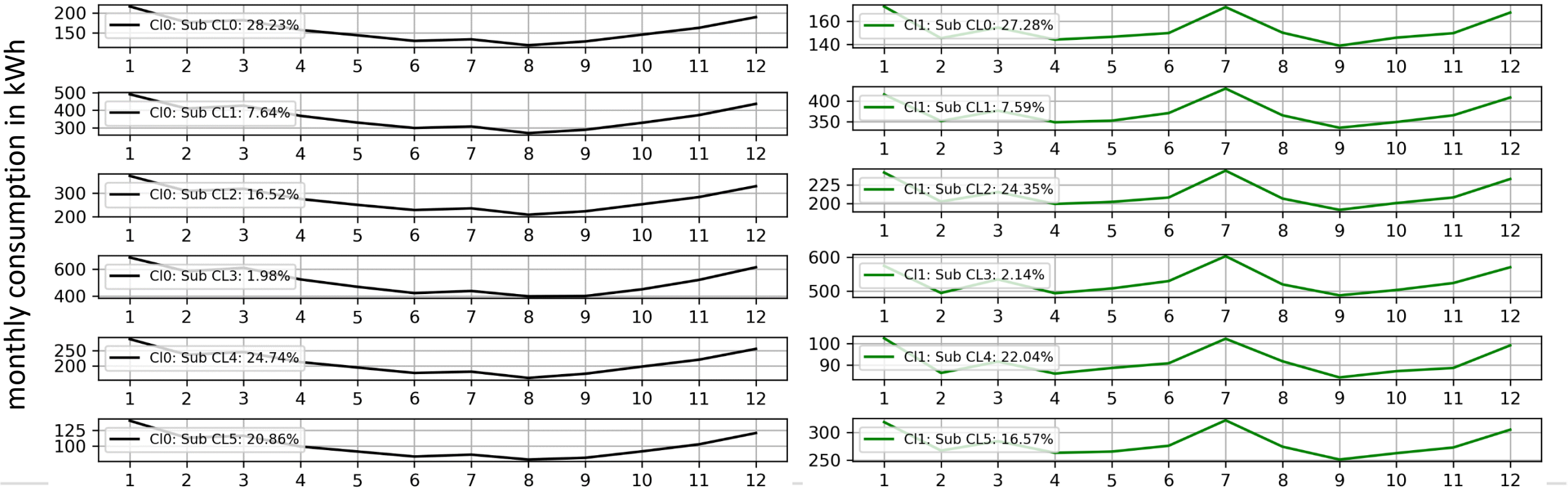
7 - data preparation: derivation of unsupervised training sets

- ✓ Centroid: consumption for residential homes
- ✓ We explored in depth the consumption associated with the two clusters relating to the residential housing profile

average monthly consumption

constant monthly consumption

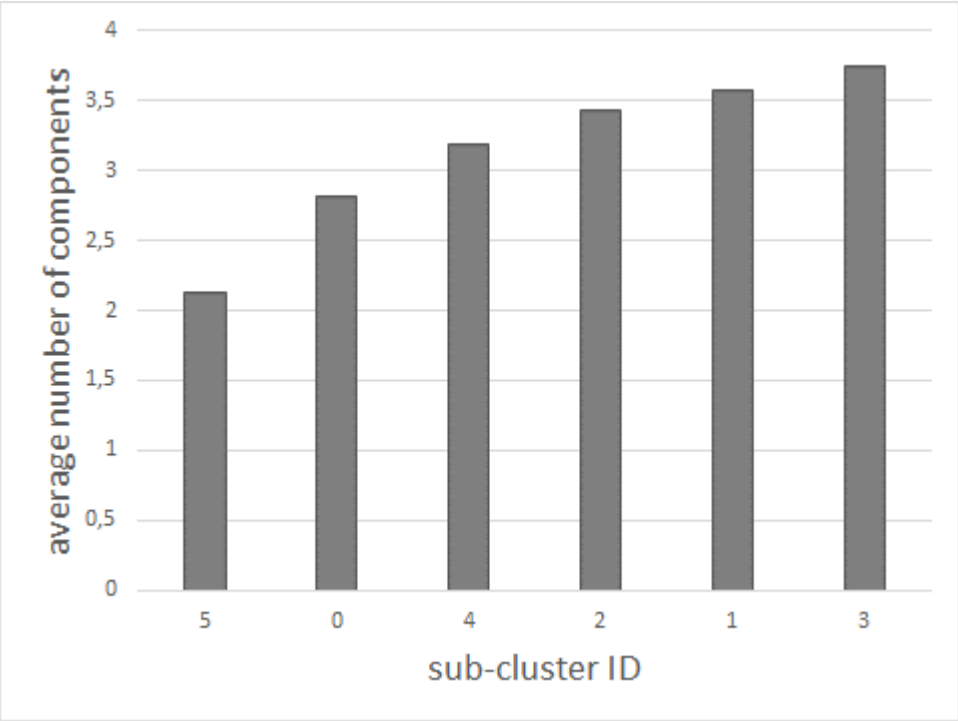
monthly consumption that depends on the season



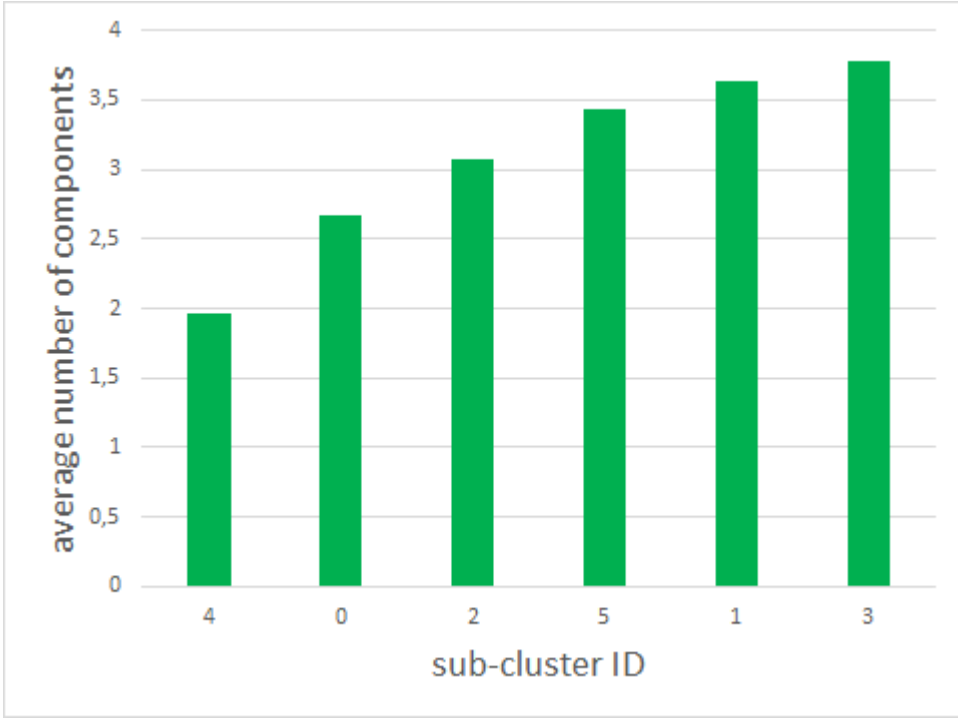
8 - definition of a forecasting model: analysis of the number of users for each cluster

- ✓ We can estimate the average number of family members to each sub-cluster of the training set
- ✓ The sub-cluster associated with an electricity meter makes it possible to calculate the probable number of the meter's users.

cluster: "constant monthly consumption"
average number of family components vs sub-cluster ID



cluster: "monthly consumption that depends on the season"
average number of family components vs sub-cluster ID



- ✓ We can define a metric for the distance between the declared number of family members and the level of associated energy consumption

- ✓ The metric can be based on three key distances:

  d_1) distance between the number of family members declared, as in the population register, and the number of users estimated from the meter's sub-cluster

  d_2) distance between the average electricity consumption associated to the electricity meter and the centroid value of the electricity meter's sub-cluster

  d_3) distance between the yearly gas consumption per gas meter and the average gas consumption of all gas meters associated with the electricity meter's sub-cluster

- ✓ The three distances contribute with different weights to the overall distance.

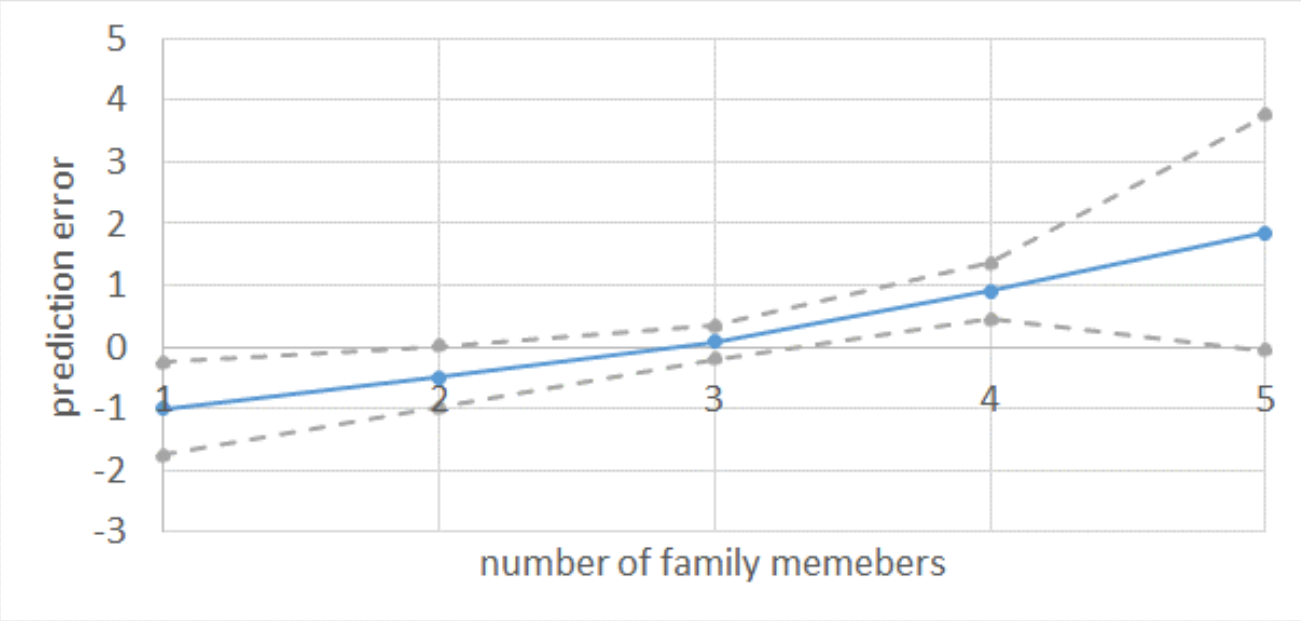
- ✓ By weighting the different contributions using the minimization of the total standard deviation, we obtain:

$$D(id) = 70\%\ d_1(id) + 15\%\ d_2(id) + 15\%\ d_3(id)$$

9 - data analysis of the forecasting outputs: evaluation of the quality of the estimate

- ✓ We calculated the distance for all the available consumption meters of group B (the 3,383,369 households with multiple contracts, in total 8,864,916 electricity meters)
- ✓ For each household with multiple meters, we chose the one with the minimum distance D as the most likely meter for daily use

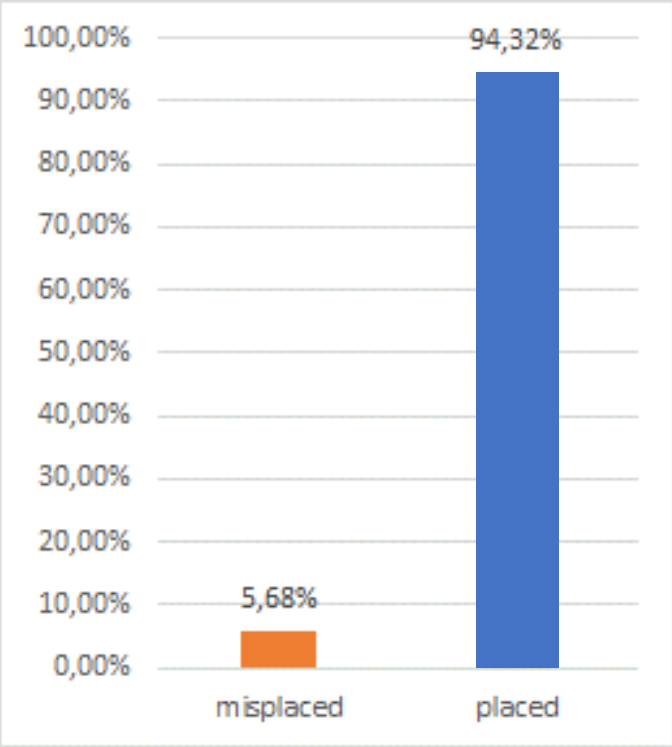
Average value of the prediction error (the number of predicted household members minus the number of registered household members) compared to the number of components in the population register



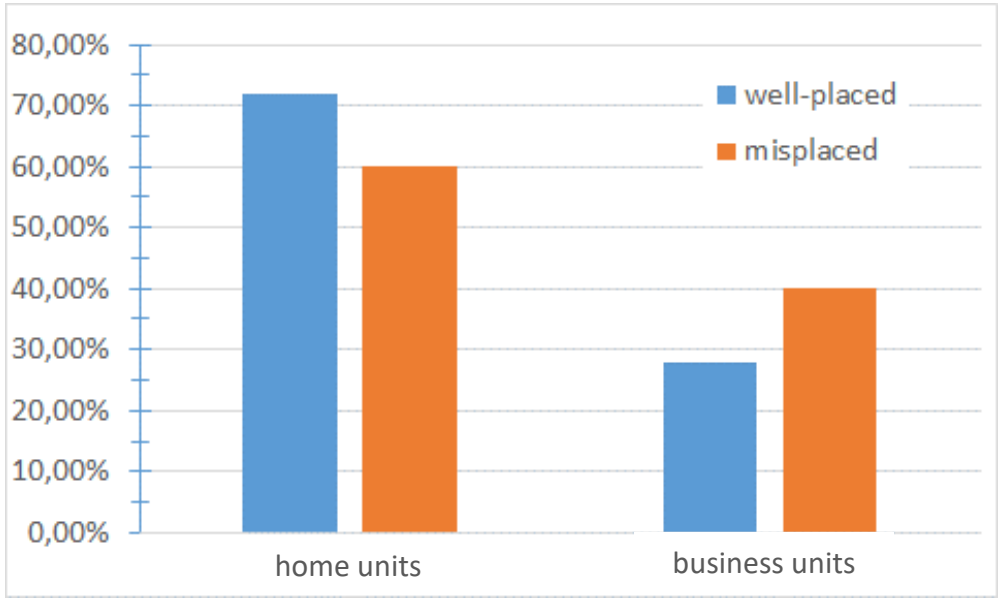
9 - data analysis of the forecasting outputs: evaluating the most probable place of usual residence

✓ Compared to the information from the population register, it is possible to estimate an error in the positioning of the usual residence of approximately 5.68%.

placement percentages

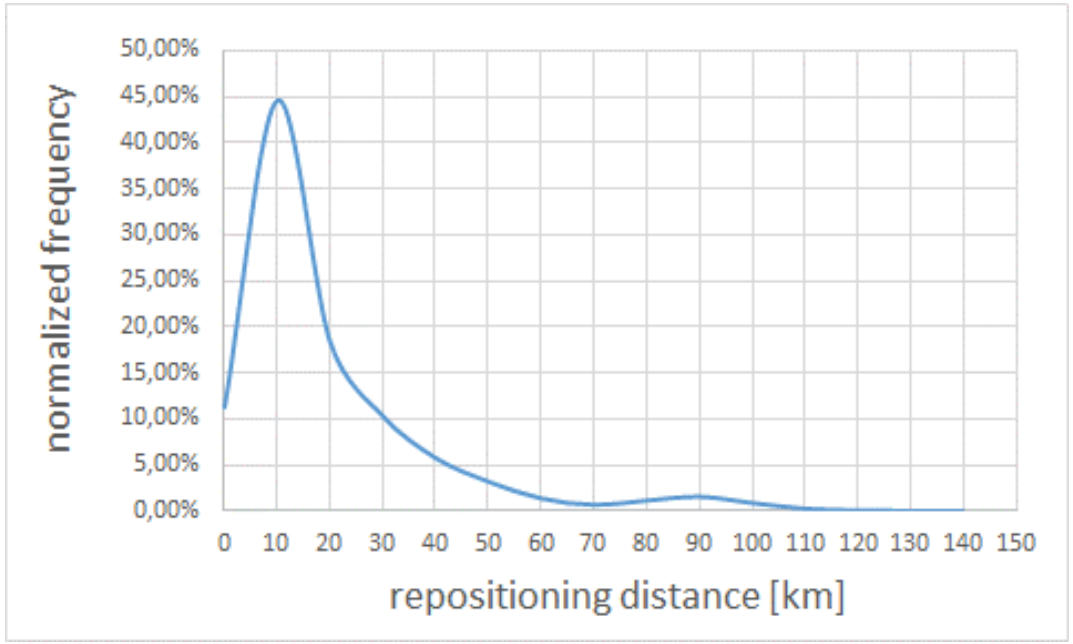


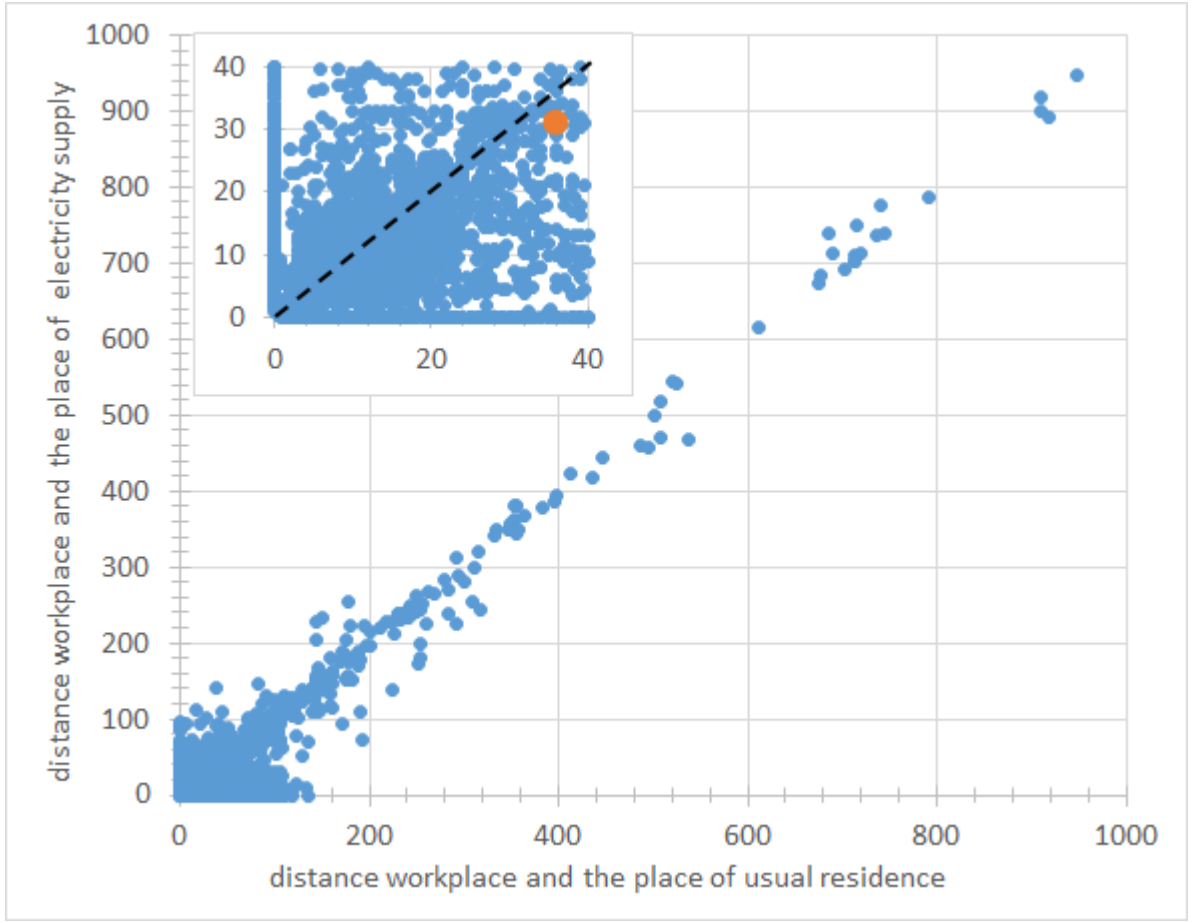
placement percentages with respect to the type of contract holder



9 - data analysis of the forecasting outputs: evaluating the most probable place of usual residence

- ✓ Misplaced normalized frequency versus distance in km of the estimated location for actual consumption of the family





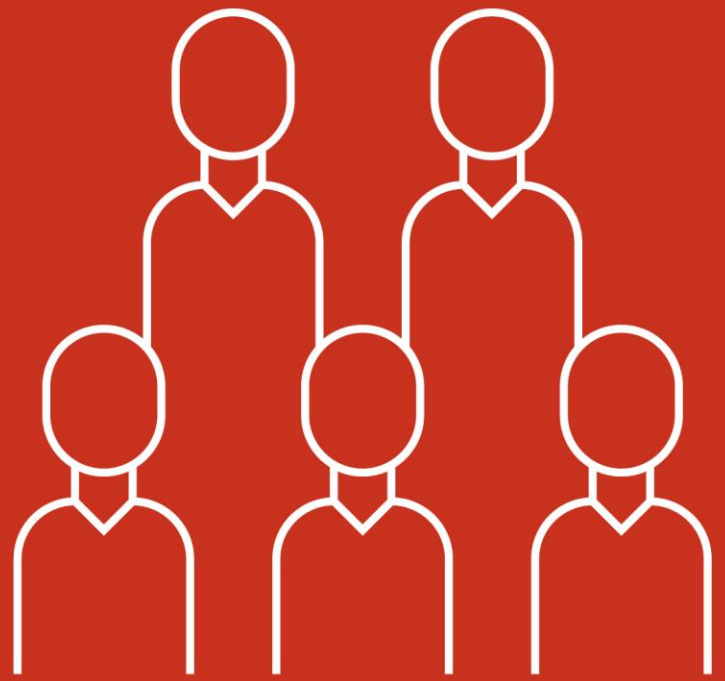
10 - Conclusion

- ✓ The use of machine learning to verify over- and under-coverage of the population register appears very promising, particularly when used on an integrated system of multiple administrative registers, such as the permanent census of the Italian population.

- ✓ The annual sample survey for the Italian census appears to be a key factor in determining the training set necessary for the development of the ML algorithm.

- ✓ Energy consumption data allows us to improve information on the distribution of individuals across the national territory and provides a quality indicator to evaluate the number of members in census data households.

- ✓ Finally, extending energy consumption data to homes could be used to further improve the quality of the information produced by the housing census



Thanks for your attention

Antonio.LauretiPalma@istat.it