Approach to Imputation Multivariate Missing Data of Urban Buildings by Chained Equations Based on Geospatial Information

Khrulkov A.A.[1], Mishina M.E.[1] and Mytagin S.A.[1]

[1] Saint-Petersburg National Research University of information Technologies, Mechanics and Optics (ITMO University), 197101 Saint-Petersburg, Russia

Abstract. Accurate information about real estate in the city, and about residential buildings in particular, is the most important resource for managing the development of the urban environment. Information about residential buildings, for example, the number of residents, is used in the inventory and digitization of the urban economy and subsequently becomes the basis of digital platforms for managing urban processes. Inventory of urban property can be carried out independently by different departments within the framework of official functions, which leads to the problem of conflicting information and missing values in urban data, in building data in particular. These problems are especially pronounced when information from different sources is combined to create centralized repositories and digital twins of the city. This leads to the need to develop approaches to filling missing values and correcting distorted information about residential buildings. As part of this work, the authors propose an approach to data imputation of residential buildings, including additional information about the environment. The analysis of the effectiveness of the approach is based on data collected for St. Petersburg (Russia).

Keywords: urban data, residential buildings, multiple imputation, machine learning, geospatial information.

Introduction

One of the main information resources of the city is information about residential buildings. The importance of this resource is determined by the fact that residential real estate accounts for a significant proportion of all real estate in the city and largely forms the urban environment. Buildings, as atomic city objects, have spatial and attribute characteristics. Spatial characteristics determine the location of a building in space in a given coordinate system, and attributive characteristics describe such features as number of storeys, population, area, etc.

Information about buildings posted in open Internet resources (for example, OpenStreetMap) forms the basis of many parametric methods for assessing the territory used in urban studies [1–3]. In addition, information about residential buildings and the citizens living in them is used simultaneously in many business processes of the city: from managing public transport services to the population to providing citizens with social

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

infrastructure facilities. This leads to the formation of parallel processes of inventory and monitoring of the parameters of residential buildings, and therefore with which errors inevitably arise when combining these data. The most common manifestation is the incompleteness and inaccuracy of the information used [4]. Urban building data often contains omissions and erroneous values for one or more of the considered features, this makes it difficult to process and analyze these data. This leads to problems at the level of intersectoral interaction and management of complex projects for the development of territories, as well as creating or integrating city management systems based on digital twins and models of the urban environment.

Such a situation requires the development of approaches to imputation and correction of information about residential buildings and other objects of the urban environment. Within the framework of this article, we consider a general approach to filling missing data in the residential building features.

1 Techniques for handling the missing data

1.1 Missing data types

The existing approaches to imputation missing data are based on the fundamental theory of Donald Rubin [5], which classifies such data into three categories depending on the reason for their absence: missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR).

The reason for the absence of data of the MCAR category is a random event that does not depend on any factors. For example, some information may be lost due to a failure in the storage system or as a result of incorrect actions when working with data. In practice, missing data of the MCAR category lead to a decrease in the analyzed sample without introducing a systematic error.

On the contrary, data of the MAR category are characterized by the dependence of the missing value on the known value of another feature of the object. This case is the most common, in particular, it includes the processing of missing data about urban buildings - missing values are often observed in relation to certain types of buildings, for example, new residential complexes or individual housing construction projects. As a rule, more advanced techniques are used to obtain unbiased statistical estimates when imputing MAR data than simply excluding them from the sample.

In the data feature of the MNAR category there is the systematic association of missing values with the reason for their absence. Thus, in surveys conducted in the course of urban studies, the presence or absence of a respondent's answer often depends on his position on the issue under consideration [6]. The processing of data of the MNAR category is the most problematic case, leading to an analysis of the reasons for the occurrence of missing values and to a change in the data collection strategy.

Determining the category of missing data (MCAR, MAR, or MNAR) allows us to choose the most appropriate imputation method that provides correct aggregated results and reliable statistical inferences.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

1.2 Overview of techniques for handling the missing data

Standard way for missing data processing is to exclude them from the sample. In case of this approach the following methods are used: listwise deletion and pairwise deletion, which could be used for MCAR data. In addition, applying these methods for strongly damaged samples leads to significant data losses.

In most cases filling the omissions are more rational. Common ways to fill the missing data consist in imputation with the: median, mean, random value from the sample or constant value such as zero. More complicated approach is to use an EM-algorithm based on iterative computation of maximum likelihood estimates. For a wide class of problems, the maximum likelihood method is consistent and asymptotically efficient, however, to use it, it is necessary to have an idea of the laws of distribution of values in the observed features of an object. This shortcoming is devoid of widely used methods of regression and classification, which involve fitting a model based on known data.

Multiple imputation is the most universal approach to processing missing data, providing unbiased estimates and reliable confidence intervals, including for data of the MAR and MNAR category [7]. The concept of multiple imputation of missing data is based on the idea of combining several results obtained by a single imputation, reflecting the uncertainty of the imputed values. An increase in the number of iterations at which a single data imputation is carried out contributes to the achievement of convergence and stability of the final values. However, given the limited computing resources, in practice, in most cases, 3 to 5 iterations are sufficient.

1.3 Machine learning methods used in data imputation

Since the data contains both quantitative and categorical features of objects, impute missing values involves solving both regression and classification problems. At the moment, there are many software-implemented mathematical methods, the success of which is determined by the available data and the dependencies that exist in them. The most common methods are: cluster analysis, K-nearest neighbors, Decision Trees and Random Forests [8].

An important step in the use of the described mathematical methods of data prediction is the hyperparameter tuning that are set before starting training, which affect the accuracy of the results and the cost of computing resources. Hyperparameter optimization methods are conditionally divided into two groups.

The first group includes methods based on enumeration of various combinations of hyperparameters. A reliable way to identify the optimal combination is to enumerate all possible hyperparameter values. A significant disadvantage of this method is its exponential complexity, which limits its use on large grids of parameters. Random and Successive Halving [9, 10] are considered effective modifications of the classical Grid Search algorithm, based on the reduction of the considered combinations due to various heuristics.

Another group includes sequential optimization methods based on a surrogate model that take into the results obtained after running in previous iterations (also called Bayesian optimization methods). The most common surrogate models are Gaussian processes

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

and Parzen trees. The efficiency of Bayesian optimization based on Gaussian processes depends on the space of hyperparameters and decreases greatly when the dimension of the space increases, equating to the efficiency of Random Grid Search [11]. Bayesian optimization based on the Parzen window is considered to be less expensive in terms of the required computational resources, but this approach does not take into account the possible interaction of the hyperparameters under consideration [12].

The choice of an appropriate hyperparameter optimization method largely depends on the problem being solved and the available computing resources. For simple models built on a small amount of data, Grid Search guarantees accurate and unambiguous results in a reasonable time. On the contrary, in the case of scalable models, it is advisable to use Bayesian optimization or Successive Halving.

2 Approach to imputation multivariate missing urban buildings data based on geospatial information

2.1 Description of the developed approach

Considering that city buildings data has multivariate missing values, when more than one attribute has missing data, applying single imputation methods cannot provide a predictable result. In this regard to impute missing data of residential buildings multiple imputation by chained equations is used [13]. The imputation of missing values at each iteration is carried out by regression and classification methods.

Cities consist of territorial units characterized by the prevailing evolutionary-morphological type of development, and having characteristic values for a number of features (number of storeys of tasks, density characteristics, etc.) [14]. The developed approach assumes that the use of additional information about the features of neighboring objects, as well as, if available, other aggregated information about the spatial environment (for example, about services located nearby or calculated spatial indices [15, 16]) provides more accurate results for data imputation process.

According to the given block diagram (Figure 1):

1. At the first stage, the input data is preprocessed: categorical features are binarized and the known values are reduced to a numerical type. Positions of missing values are written to a new variable, and initial values are initialized in their place, which are subsequently imputed by predicted values. The initial values are the distance-weighted average values of other objects (1).

$$x_{init}^{t} = \frac{\sum_{i=1}^{k-1} w_i \cdot x_i^t}{\sum_{i=1}^{k-1} w_i}, \quad w_i = \frac{1}{d} \tag{1}$$

where x_{init}^{t} — the calculated initial value of the object on the feature t; x_i^t — the initial value of another object on the feature t; w_i — coefficient of remoteness of another

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

object; k — the number of objects in the dataset; d — the Euclidean distance between objects (2).

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2}$$

where x, y — coordinates of two points.

2. The next step is to expand the original feature space. For each object from the input dataset, a search is made in the metric space for K-nearest neighbors in a given radius r (3), based on the Euclidean distance measured between the centers of objects (2). The result of the search is two matrices containing the indices of the nearest neighbors and the distance to them for each sample object. The obtained indices are used to calculate the average values of neighboring objects, which supplement the input data array.

$$kNN(q) = \{R \subseteq X,\ |R| = k \cap \forall n \in R,\ o \in X - R : d(q,n) \le d(q,o),\ d(q,n) \le r\} \tag{3}$$

where X — the set of all objects in the dataset; k — the number of required neighbors; q — the object for which the search for neighbors is carried out; d — a function of distance; r — the specified neighbor search radius.

If additional data is used (for example, on the calculated spatial indices of territorial units), they are aggregated and attached to the initial dataset by a given attribute (for example, by a building or block identifier).

3. Imputation of missing values at each iteration begins with the selection of dependent and independent variables in the prepared data. The dependent variable is the first in order feature that has missing values in the initial dataset for any objects of the sample (the positions of missing values before the initialization of the initial values were written to the variable at step 1). Then, a training sample is selected in the data, consisting of objects that do not have missing values in the original dataset according to the attribute assigned to the dependent variable (4).

$$X^{raw}_{m \times n} = X^{obs_t}_{p \times n} \cup X^{mis_t}_{k \times n} \quad u \quad \varnothing = X^{obs_t}_{p \times n} \cap X^{mis_t}_{k \times n} \tag{4}$$

where $X^{raw}_{m \times n}$ — the prepared dataset; $X^{obs_t}_{p \times n}$ — data that does not contain missing values in the initial dataset on the feature t; $X^{mis_t}_{k \times n}$ — data containing missing values in the initial dataset on the feature t.

4. On the selected sample, the regression and classification models are trained. The selection of a combination of hyperparameters that provide the best predictive ability of the method is performed by the Grid Search with Successive Halving [10]. The assessment of the quality of a model trained on a certain combination of hyperparameters is carried out using the k-fold cross-validation. The loss function for regression models is the Mean Square Error and for classification models is the Log Loss function. Prediction of missing values on the rest of the data is made with the best combination of hyperparameters. The calculated values impute the missing values.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

5. Having imputed the missing values of the dependent variable, steps 3 and 4 are repeated cyclically for each variable that has omissions in the values.

6. Steps 3-7 constitute one complete iteration of data imputation, the total number of which is specified by the user, based on available computing resources. According to [17], upon imputation of at least 2 iterations, the coefficients of the models become stable, which makes it possible to avoid the dependence of the result obtained on the order in which the target variables are assigned.

7. The result of steps 1-6 is a complete dataset containing the imputed values. According to the concept of multiple data imputation, steps 1–6 are repeated several times (from 3 to 5 are considered sufficient [18]), after which the values in independently imputed datasets are averaged. The final dataset and the average quality score obtained by the k-fold cross-validation in step 4 are the output of the method.

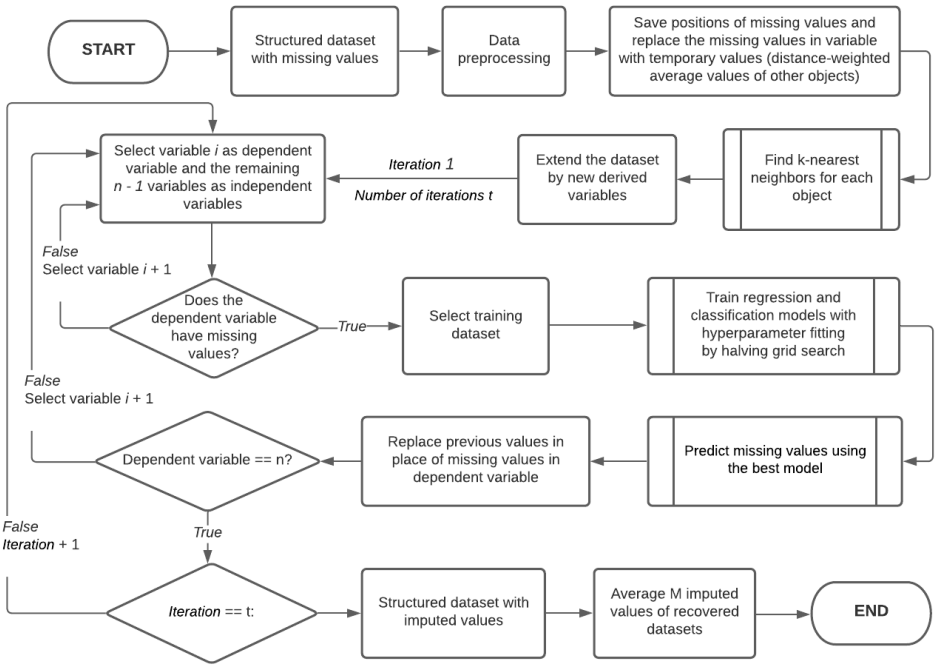

Fig. 1. Approach to imputation city building data using geospatial information

2.2 Findings of the developed approach

The issue of incomplete data imputation has a fundamental theoretical basis on which the developed approach to imputation missing data of urban buildings is based. The developed approach implements the concept of multiple data imputation, which takes into account the uncertainty of the imputed values by predicting several values for each missing value using the selected model. Since the missing values are found in data that have different internal dependencies, the presented approach preliminarily selects the best predictive model sequentially for each considered feature.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

The advantage of the developed approach in the framework of solving the applied problem of building data imputation is the presence of the stage of expanding the initial dataset with derived features containing information extracted from spatial dependencies that are initially inaccessible to the model, but known to the researcher. At the current stage, additional derived features are information about neighboring objects. In the future, it is planned to identify new spatial dependencies that exist in the urban environment and use them to obtain the most accurate imputed values.

A potential obstacle to the application of the presented approach is its resource consumption. Due to the iterative process, it takes a significant amount of time to impute a large number of missing values.

3 Imputation missing values to Saint-Petersburg buildings dataset

3.1 Collection and analysis of initial dataset

In the study, devoted to solving the practical problem of imputation the missing data of urban buildings, information about residential buildings in St. Petersburg, published in open sources, was used. A description of the information used (type of data, source of information and additional explanations) is given in table 1.

Table 1. Description of building data used.

Name	Data type	Source	Description
Floors	Integer	2GIS	Number of floors
LivingSpace	Float	reformagkh.ru,	Living space
Area	Float	reformagkh.ru, openstreetmap.org	Building foundation area
Population	Integer	data.gov.spb.ru	Number of people
Lift	Integer	reformagkh.ru	Number of lifts
Gascentral	Boolean	reformagkh.ru	Gas supply system
Hotwater	Boolean	reformagkh.ru	Hot water
Electricity	Boolean	reformagkh.ru	Electricity
Coordinates (x, y)	Float	openstreetmap.org	X, Y in 3857 ESPG projection

When collecting initial data, emphasis was placed on the residential development of the city, since at the time of the study there was the most complete database of this type of buildings in open sources of information, of which the reformagkh.ru resource was considered the most reliable. The collected data contained information on 12,867 residential buildings. The collected raw data contained no missing values. To conduct the study, the modeling of missing values was carried out by random removal of feature values. The degree of damage to the initial dataset was determined by the percentage of "sown" omissions from the total number of feature values and ranged from 10 to 90 percent in 10 percent increments. In the figure (Figure 9), the buildings of the Dekabristov Island municipality are marked in red, the data on which are deliberately distorted in this way. An example of such distorted data is given in the table 2.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

Fig. 2. Buildings with distorted data (red polygons)

Table 2. Distorted data samples of three buildings.

ID	Floors	Living Space	Area	Population	Lift	Gas central	Hot water	Electricity
1511	16	NULL	9378	509	8	NULL	True	True
1356	12	28873	20341	NULL	15	NULL	True	True
2314	NULL	1197	1988	0	0	True	False	True

3.2 Investigating the Effectiveness of an Approach to City Building Data imputation

The approach was implemented in the python programming language, using open libraries - numpy, geopandas and scikit-learn. At the first stage of imputation, the input data was preprocessed: all values of the features were reduced to a numeric type. For features Gascentral, Hotwater, Electricity, the binarization procedure was carried out (True value - 1, False value - 0). Initial values were initialized in place of missing values. To expand the input dataset, information about neighbors was used - the average values of the features of the three nearest neighboring buildings located within a radius of 500 m. An example of such derived data is presented in table 3.

Table 3. Average characteristics of neighbors.

ID	Floors neigh	Living Space neigh	Area neigh	Population neigh	Lift neigh	Gas central neigh	Hot water neigh	Electricity neigh
1511	16	12646	4591	506	7	0	1	1
1356	20	28302	14740	1101	13	1	1	1
2314	5	1920	2121	16	0	1	0	1

The selection of the best model for predicting missing values was carried out using the Grid Search with Successive Halving. The use of the algorithm of Successive

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

Halving for reduction of the hyperparameter combinations made it possible to get away from the exponential complexity of the basic grid search and obtain the optimal combination in less time. At the stage of determining the best predictive model, the three most common methods of regression and classification in the context of data imputation (according to Section 2.2) were investigated: K-nearest neighbors, Random Forests, and Gradient Boosting Decision Trees. The optimized hyperparameters are presented in table 4. Z-normalization of the data was preliminarily carried out.

Table 4. Optimizing hyperparameters.

Method	Hyperparameter grid
K-nearest neighbors	The number of neighbors is 3, 5, 7, 9.
Random Forests	The number of trees is 50, 100, 150, 200.
	The maximum depth of trees is 2, 3, 4, 5, 6, 7, 8.
	The minimum number of objects to split a node is 2, 5, 10.
	The subset of features for node splitting is all, sqrt.
Gradient Boosting Decision Trees	The number of trees is 50, 100, 150, 200.
	The maximum depth of trees is 2, 3, 4, 5, 6, 7, 8.
	The minimum number of objects to split a node is 2, 5, 10.
	The subset of features for node splitting is all, sqrt.
	The learning rate is 0.01, 0.1, 1.

The evaluation of the generalizing abilities of models with various combinations of hyperparameters is carried out by the k-fold cross-validation method on 5 groups of samples. The following were used as loss functions: the Mean Square Error (for regression problems) and the Log Loss function (for classification problems).

The quality metrics of the imputed data were: the coefficient of determination (for regression problems) and the F-measure (for classification problems). The quality metrics calculated for a dataset with a degree of damage of 10 percent are shown in table 5. According to the results, the best generalizing ability is possessed by the Gradient Boosting Decision Trees. With an increase in the percentage of deleted values, similar results were observed (Figure 3).

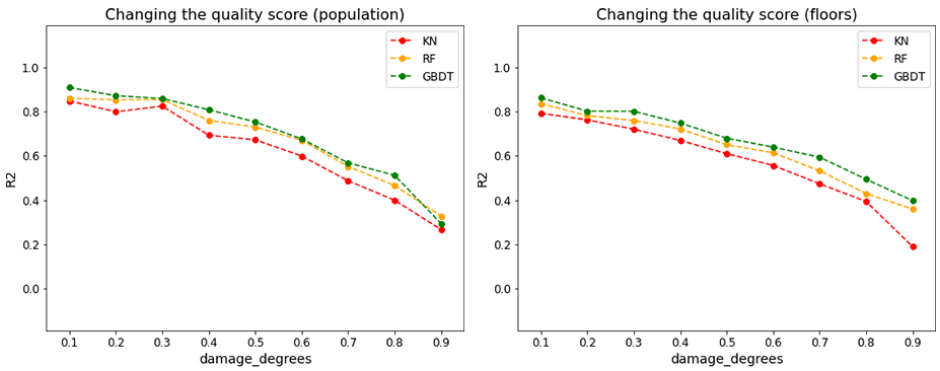

Fig. 3. Comparison of method quality metrics for the features Floors and Population

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

In order to identify the impact on the accuracy of data imputation of additional derived features, quality metrics were also calculated for data imputed using the developed approach, but without expanding the initial feature space with derived information - information about neighboring objects. The results obtained for a data set with a degree of damage of 10% are presented in table 5. According to the results of calculations, for any degree of damage to the input set, the accuracy of the restored values turned out to be higher with a preliminary expansion of the feature space. With an increase in the percentage of missing values, the influence of derived information increased (Figure 4).

Table 5. Compare accuracy results for different predictive methods.

Variable	With neighbor information			Without neighbor information
	Gradient Boosting	Random Forest	KNeighbors	Gradient Boosting
Floors	0.85	0.84	0.79	0.82
Living Space	0.95	0.94	0.89	0.93
Area	0.54	0.55	0.47	0.54
Population	0.91	0.84	0.84	0.77
Lift	0.89	0.89	0.82	0.88
Gascentral	0.96	0.96	0.95	0.96
Hotwater	0.93	0.93	0.92	0.80
Electricity	0.99	0.99	0.99	0.99

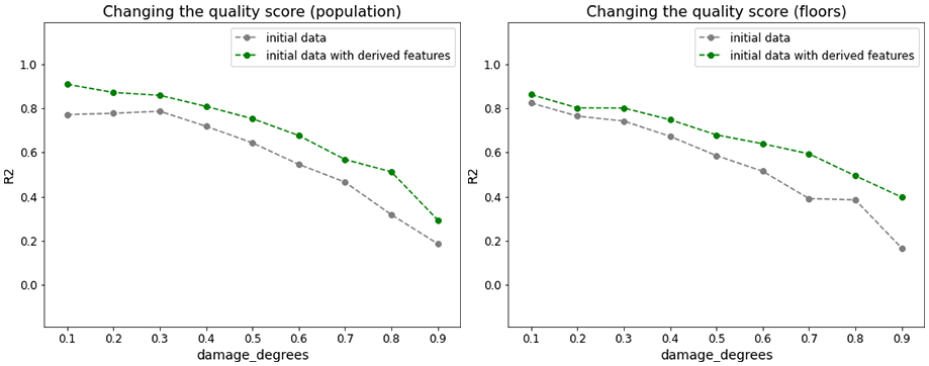

Fig. 4. Comparison of data imputation methods by characteristics Floors and population

Based on the results of experiments conducted on 9 datasets with varying degrees of damage, it was found that predicting missing values using the best model based on the dependencies available in geospatial information makes it possible to impute data on residential buildings with up to 30 percent of missing values with an accuracy higher than 0.80 for the specified metrics. In the case under consideration, a lower accuracy of predictions was observed only for the feature Area, which is justified by the absence in the existing dataset of features that are closely related to the area of the foundation of the building (for example, the area of the land plot on which the building is located).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

However, missing values for this parameter are extremely rare in practice and can be quickly restored in small quantities using the digitization and geoprocessing tools of cartographic web services (for example, OpenStreetMap or Google Maps). The highest accuracy was observed for the signs Gascentral, Hotwater, Electricity, since according to these signs, the vast majority of objects had a value of 1, which meant the presence of electricity, gas supply and hot water in most residential buildings in St. Petersburg.

For the analysis of outliers in the imputed values, absolute and relative shifts from the true values were calculated (5)-(6). As an example, Figure 5 plots the absolute and relative offsets of LivingSpace predictions for a 10 percent damaged dataset.

$$ab_i = y_{true} - y_{pred} \tag{5}$$

$$pb_i = 100 \cdot \left| \frac{(y_{true} - y_{pred})}{y_{true}} \right| \tag{6}$$

where y_{true} — the true value of the feature; y_{pred} — the predicted value of the feature.

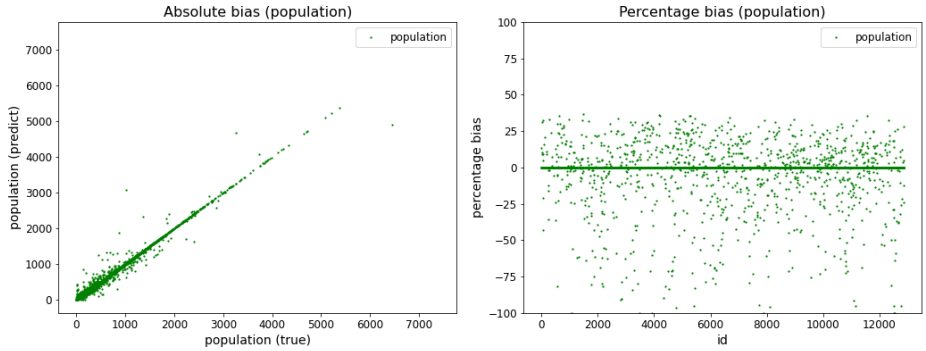

Fig. 5. Absolute and relative shift of LivingSpace trait scores

Even with a high coefficient of determination equal to 0.95, there were several significant outliers in the predicted values. The identified outliers were observed in relation to objects containing obvious errors in the known values of other features. For example, some of the buildings in the dataset were considered single-storeys, which greatly skewed the predicted values. Despite the negative impact of this effect on the accuracy of imputed values, in future studies, information on outliers can be used to identify inaccurate values in the initial data.

3.3 Comparative analysis of approaches to data imputation

An analysis of existing approaches to data imputation showed that the implemented solutions in the presence of missing values in the dataset, with the exception of their removal, can be used:

1. Filling in missing values with a point estimate (for example, an average).
2. The experimental method of the scikit-learn library is IterativeImputer.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

Singular mean filling is a common solution to the problem of missing data, which, however, tends to skew any statistical estimate of a feature other than the mean.

The IterativeImputer method implements a multiple data imputation strategy using predictive models. As input, IterativeImputer accepts one specified regression or classification method, a method for initializing initial values in positions of unknown features (mean, median, mode, or constant), the number of imputation and a number of other parameters that determine the output of data. This method is a general tool for solving the problem of missing data of various directions and does not allow considering the specifics of urban data, including identifying new dependencies by expanding the feature space. Another disadvantage is the lack of an implemented algorithm for choosing the best predictive model, which, as a result, leads to the impossibility of simultaneously imputing quantitative and qualitative values, as well as automated selection of the best combination of hyperparameters.

The comparison of the alternative solutions with the developed approach was carried out for the input dataset with different degrees of damage by calculating the coefficient of determination and the F-measure. In the IterativeImputer method, gradient boosting decision trees was used as a predictive model with default parameters in the scikit-learn library. A demonstration of the results obtained is shown in Figure 6 (similar results were observed on other grounds).

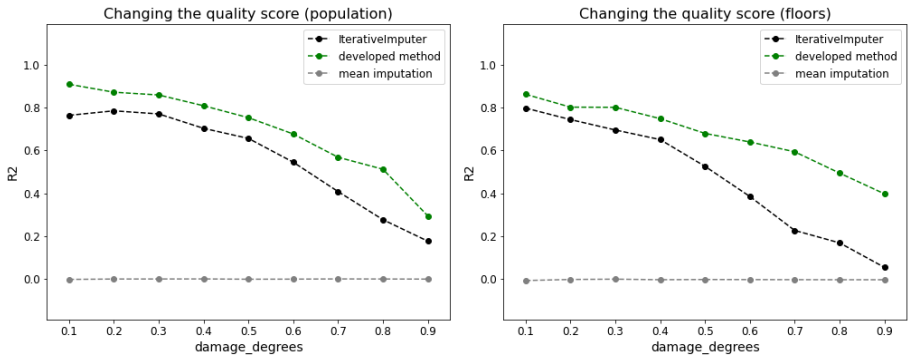

Fig. 6. Comparison of data imputation methods on the example of the features Floors and Population

The results of the calculations showed that filling the missing values of quantitative features with the average is a nominal solution that does not improve the situation - the accuracy of the imputed values remains close to zero. Using the implemented IterativeImputer method of the scikit-learn library made it possible to consider existing data dependencies and obtain an acceptable accuracy of the imputed values at the level of 0.80 for the specified metrics (with a data corruption degree of no more than 10 percent and considering that the optimal regression or classification method is known). The developed data imputation approach showed its effectiveness in comparison with alternative solutions in all cases considered (with any degree of data damage). The advantage of the developed approach was achieved by extracting additional information from the data, which is then used to impute missing values.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

Conclusion

Complete and accurate city building data is a critical resource for city management. However, such data often contain omissions and erroneous values for one or more features, which makes it difficult to process and further analyze them. This problem on a large scale leads to a decrease in the effectiveness of intersectoral interaction, management of complex projects for the development of territories and the use of integrating systems for urban management.

The developed approach for urban building data imputation is based on multiple data imputation, as well as regression and classification methods. The novelty of the developed approach lies in the presence of stages of expanding the initial feature space and determining the best model for predicting missing values. The extension of the feature space is carried out through calculations and transformations performed on the available geospatial information. The determination of the best predictive method is carried out according to the Grid Search algorithm with sequential halving of the specified combinations of hyperparameters of regression and classification methods.

It should be noted that at this stage, the presented method does not allow making confident judgments about values for individual objects, but it makes it possible to obtain correct aggregated values for territorial units and formulate reliable statistical conclusions when making decisions in the field of management and optimization of urban processes. In further studies, it is planned to conduct experiments aimed at improving the accuracy of the imputed values by identifying new relationships that exist in urban data, as well as to explore the developed approach on data containing information on all types of buildings (residential and non-residential). A separate area of work will be the analysis of the possibility of using this approach to solve the problem of inaccurate information about urban facilities by recalculating all available values.

This research is financially supported by The Russian Science Foundation, Agreement №20-11-20264.

References

1. Khrulkov A, Mityagin SA, Repkin AI (2020) Multi-factor approach to investment attractiveness assessment of urban spaces. Procedia Computer Science 178:94–102
2. van Nes A, Berghauser Pont M, Mashhoodi B (2012) Combination of Space syntax with spacematrix and the mixed-use index: The Rotterdam South test case. In: 8th International Space Syntax Symposium, Santiago de Chile, Jan. 3-6, 2012
3. Boiko D, Parygin D, Savina O, et al (2019) Approaches to Analysis of Factors Affecting the Residential Real Estate Bid Prices in Case of Open Data Use. In: International Conference on Electronic Governance and Open Society: Challenges in Eurasia. pp 360–375
4. Wang RY, Strong DM (1996) Beyond accuracy: What data quality means to data consumers. Journal of management information systems 12:5–33
5. Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21

6. Curley C, Krause RM, Feiock R, Hawkins C v (2019) Dealing with missing data: A comparative exploration of approaches using the integrated city sustainability database. Urban affairs review 55:591–615
7. van Buuren S (2018) Flexible imputation of missing data. CRC press
8. Lin W-C, Tsai C-F (2020) Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review 53:1487–1509
9. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. Journal of machine learning research 13:
10. Jamieson K, Talwalkar A (2016) Non-stochastic best arm identification and hyperparameter optimization. In: Artificial Intelligence and Statistics. pp 240–248
11. Li L, Jamieson K, DeSalvo G, et al (2017) Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research 18:6765–6816
12. Bergstra J, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. Advances in neural information processing systems 24:
13. van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in R. Journal of statistical software 45:1–67
14. Conzen MRG (1960) Alnwick, Northumberland: a study in town-plan analysis. Transactions and Papers (Institute of British Geographers) iii–122
15. Pont MYB, Haupt PA (2010) Spacematrix. space, density and urban form. NAi Publishers
16. van den HOEK JW (2009) Towards a Mixed-use Index (MXI) as a tool for urban planning and analysis. Urbanism: PhD Research 2008-2012 65
17. Templ M, Kowarik A, Filzmoser P (2011) Iterative stepwise regression imputation using standard and robust methods. Computational Statistics & Data Analysis 55:2793–2806
18. Schafer JL, Graham JW (2002) Missing data our view of the state of the art. Psychological methods 7:147

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_21