

SPATIAL PREDICTION OF RELATIVE BUILDING ENERGY EFFICIENCY IN SLOVENIA USING NON-ENERGY PREDICTORS FROM THE NATIONAL EPC REGISTRY

Luka Lah

Faculty of Organizational Sciences, University of Maribor, Kidričeva cesta 55a, 4000 Kranj, Slovenia

luka.lah5@um.si

Abstract

National registries of energy performance certificates have become a standard empirical basis for analyzing building stocks, but most predictive studies built on such registries rely on direct energy indicators that closely mirror the certification output. This article develops a spatially aware machine learning framework for Slovenia that predicts relative building energy efficiency using only non energy predictors, isolating the contribution of administrative location to predictive performance. The analysis uses a public snapshot of the Slovenian EPC registry covering 58,310 records, 16 variables and 2,360 unique cadastral community codes between 2017 and 2025. The target is constructed as a year normalized binary indicator, equal to one when a building's primary energy exceeds the national median primary energy in its issue year. After target construction all direct energy variables are dropped from the predictor set, leaving conditioned floor area, certificate type, EPBD category, issue year, building part number and the cadastral community code. Three logistic regression models are compared. Model A uses only the six non spatial predictors. Model B augments Model A with smoothed out of fold target encoding of the cadastral community code. Model C further adds a neighborhood based spatial lag computed from cadastral adjacency. On a stratified 80/20 split, Model A reaches ROC AUC 0.730 and F1 0.699, Model B reaches 0.748 and 0.708, and Model C reaches 0.748 and 0.708. A temporal holdout that trains on 2017 to 2024 and tests on 2025 produces the same ordering with lower absolute scores (ROC AUC 0.640, 0.661 and 0.661). The paper contributes a year relative target formulation that is robust to the structural break in the primary energy unit observed between 2022 and 2023 in the public registry, an empirically grounded quantification of the predictive value of cadastral location when no direct energy variable is available, and a negative finding that a neighborhood spatial lag adds no measurable value once smoothed target encoding has been applied to the same cadastral feature. The framework is implemented in Python and the full notebook, processed cadastral support files and per model output figures are released for reproducibility.

Keywords: building energy efficiency, energy performance certificates, spatial machine learning, target encoding, spatial lag, Slovenia, registry data

**PROSTORSKA NAPOVED RELATIVNE ENERGETSKE
UČINKOVITOSTI STAVB V SLOVENIJI Z UPORABO
NEENERGETSKIH NAPOVEDOVALNIKOV IZ NACIONALNEGA
REGISTRA EPC**

Izvleček

Nacionalni registri energetskega izkaznic so postali standardna empirična podlaga za analizo stavbnega fonda, vendar večina napovednih študij, ki temeljijo na takih registrih, slone na neposrednih energetskega kazalnikov, ki tesno odražajo same rezultate certificiranja. V tem članku razvijemo prostorsko ozaveščen okvir strojnega učenja za Slovenijo, ki napoveduje relativno energetske učinkovitost stavb izključno z neenergetskimi napovedovalci in tako izoliramo prispevek upravne lokacije k napovedni zmogljivosti. Analiza uporablja javen posnetek slovenskega registra EI z 58.310 zapisi, 16 spremenljivkami in 2.360 unikatnimi šiframi katastrskih občin v obdobju 2017 do 2025. Ciljna spremenljivka je definirana kot po letih normalizirani binarni indikator, ki je enak ena, kadar primarna energija stavbe presega nacionalno mediano primarne energije v letu izdaje izkaznice. Po konstrukciji cilja iz nabora napovedovalcev odstranimo vse neposredne energetske spremenljivke, ostanejo pa kondicionirana površina stavbe, tip izkaznice, kategorija EPBD, leto izdaje, številka dela stavbe in šifra katastrske občine. Primerjamo tri logistične regresijske modele. Model A uporablja le 6 neprostorskih napovedovalcev. Model B razširi model A s ciljnim kodiranjem glajenih gub (angl. out of fold) šifer katastrske občine. Model C dodatno doda prostorski zamik (angl. spatial lag), izpeljan iz katastrske sosednosti. Na stratificirani razdelitvi 80/20 model A doseže ROC AUC 0,730 in F1 0,699, model B 0,748 in 0,708 in model C 0,748 in 0,708. Časovno vzdržljivostjo (angl. hold-out) z učenjem na obdobju 2017 do 2024 in testiranjem na letu 2025 ohrani isti vrstni red ob nižjih absolutnih vrednostih (ROC AUC 0,640, 0,661 in 0,661). Članek prispeva letno relativno formulacijo cilja, ki je robustna na strukturni prelom v enoti primarne energije, opažen med letoma 2022 in 2023 v javnem registru, empirično kvantifikacijo napovedne vrednosti katastrske lokacije v odsotnosti vsake neposredne energetske spremenljivke in negativno ugotovitev, da prostorski zamik soseske ne prinese opazne dodatne vrednosti, ko je na isti spremenljivki katastrske občine že uporabljeno glajeno ciljno kodiranje. Ogrodje je implementirano v Pythonu, celoten notebook, predprocesirane katastrske datoteke in slike izhodov po modelih pa so objavljeni za ponovljivost.

Ključne besede: energetska učinkovitost stavb, energetske izkaznice, prostorsko strojno učenje, ciljno kodiranje, prostorski lag, Slovenija, podatki javnega registra

1 INTRODUCTION

Buildings account for a substantial share of energy use and greenhouse gas emissions in Europe, and energy performance certificates have become the standardized instrument for documenting building energy characteristics across the European Union [1], [2], [3]. The maturation of national EPC registries has in turn enabled a fast growing literature on data driven prediction of building energy use and energy labels, often based on statistical learning, tree ensembles and neural networks applied to administrative tabular data [4], [5], [6], [7], [8], [9]. A recurring pattern in this literature is that the most accurate models are those that include direct energy indicators among their inputs, which mechanically tracks the outcome they aim to predict because the EPC label itself is a deterministic function of energy indicators in most national methodologies [10], [11], [12]. When direct energy variables are admitted as predictors, the resulting models tend to reproduce the certification formula rather than reveal structural drivers of efficiency.

A separate strand of research treats building energy outcomes as spatially embedded phenomena that reflect climate zone, settlement morphology, construction vintage, renovation intensity and local socioeconomic conditions [13], [14], [15], [16], [17], [18]. The relevance of location for energy outcomes is supported both at the urban scale, where GIS based EPC analyses have repeatedly documented spatial heterogeneity [14], [16], and at the regional scale through spatial econometric models that exploit dependence among neighboring

units [17], [19]. In a registry that links every certificate to a cadastral community code, location is operationalized through a high cardinality categorical variable, which is challenging for standard one hot encoding but well suited to smoothed target encoding [20], [21], [22], [23].

This article connects the two strands by asking whether a national EPC registry can predict relative building energy efficiency when every direct energy variable has been removed from the predictor set and the only spatial signal comes from the cadastral community code. The motivation for this restriction is twofold. It separates predictive value that can be attributed to building geometry, certificate type and location from predictive value that comes from mirroring the certification output, and it produces a model that remains usable in screening situations where building level energy measurements are unavailable but cadastral location is always known. The Slovenian public EPC registry maintained by the Ministry of the Environment, Climate and Energy [24] is well suited to this question because it covers nine years of issued certificates, 2,360 cadastral communities and a balanced mixture of whole building and building part certificates, and because Slovenia provides public geodetic infrastructure that supplies the cadastral adjacencies and centroids needed to build neighborhood features [25].

A practical complication that emerges directly from the data is a structural break in the primary energy variable between 2022 and 2023. National yearly medians of primary energy hover around 13,000 kWh in the first six years of the registry and then drop by two orders of magnitude in 2023 and onward, which is consistent with a change in the reported unit or normalization rather than a real change in building stock performance. Absolute thresholds therefore do not transfer across years and a year relative target is the only robust formulation. This observation simultaneously motivates one of the contributions of this paper and constrains the design space of any predictive study on this registry.

The contributions of this paper are three. First, we propose a year relative efficiency target that is invariant to changes in the absolute scale of primary energy, which we show is necessary on the Slovenian registry because of the 2022 to 2023 break and is more generally applicable to any registry where reporting conventions evolve. Second, we quantify how much predictive power can be extracted from the registry when every direct energy variable is removed, using a transparent logistic regression baseline and out of fold smoothed target encoding of the cadastral community code, and we provide the first such estimate for Slovenia. Third, we report a negative methodological result, namely that adding a neighborhood spatial lag on top of smoothed target encoding does not improve predictive performance under either a random split or a temporal holdout, which we interpret as evidence that, when a high cardinality location variable has already been smoothed into the model, an additional neighborhood aggregate becomes effectively collinear with that smoothing. This finding qualifies the common expectation that spatial lags routinely add value to tabular machine learning models with administrative location features.

The remainder of the article is organized as follows. Section states the problem and research questions. Section describes the registry, the cadastral support files and the inclusion criteria. Section presents the target construction, feature restrictions, spatial feature engineering and the three model configurations. Section reports the empirical metrics on both the random and temporal evaluation regimes without interpretive language. Section interprets those numbers, connects them to the wider literature and limits each claim to what the data support. Section concludes.

2 PROBLEM DEFINITION AND RESEARCH QUESTIONS

The empirical question is whether the relative energy efficiency of a building, defined as the position of its primary energy above or below the national median for its issue year, can be

predicted from non energy attributes alone. The intended use case is a screening tool that ranks cadastral areas by expected efficiency for retrofit planning, where the energy variables that define the EPC label are not available as inputs, only administrative records and cadastral geography. The question is operational rather than mechanistic, in the sense that a positive answer would mean that location and basic descriptors carry information predictive of relative efficiency, not that any specific physical mechanism has been identified.

Three concrete sub questions follow from this framing. The first asks whether a non spatial baseline built from six non energy attributes already exceeds chance level on the chosen metrics. The second asks whether smoothed out of fold target encoding of the cadastral community code measurably improves over that baseline. The third asks whether adding a neighborhood based spatial lag on top of that encoding yields a further measurable improvement. These three questions translate into a direct numerical comparison among Model A, Model B and Model C, each evaluated on the same held out test data under matched preprocessing.

3 DATA AND INCLUSION CRITERIA

The empirical basis is the Slovenian public register of issued energy performance certificates maintained by the Ministry of the Environment, Climate and Energy [24], accessed as a snapshot of 58,310 records and 16 columns covering certificates issued between January 2017 and the cut-off in 2025. The variables include the certificate identifier, the date of issue and validity, the cadastral community code, the building number, the building part number, the certificate type, the required heating energy, delivered energy for operation, total energy, delivered electricity, primary energy, CO₂ emissions, conditioned floor area, official energy class and the EPBD category. Cadastral support is supplied by two files derived from public geodetic services [25], namely a centroid table with 2,698 cadastral communities in WGS84 and D96/TM coordinates and an adjacency table with 15,628 ordered neighbor pairs that defines the cadastral graph used for the spatial lag.

Inclusion criteria are stated explicitly. A record is retained for modeling if it carries a parseable issue date, a non missing primary energy value used only for target construction, a non missing conditioned floor area, a non missing cadastral community code, a non missing certificate type and a non missing EPBD category. Slovenian European number formatting in the raw file is normalized to floating point before any filtering. Records that fail any of these criteria are dropped before target construction. Applying these rules leaves 57,841 records and 2,357 unique cadastral community codes, which removes 469 records and three cadastral communities from the raw snapshot. The remaining sample is approximately balanced by construction, with 50.05 percent of observations in the lower class and 49.95 percent in the higher class.

The yearly distribution of records reveals the structural break referenced in the introduction. National medians of primary energy in kWh, rounded, take the values 12,820 in 2017, 12,736 in 2018, 14,209 in 2019, 15,867 in 2020, 14,149 in 2021 and 13,483 in 2022, and then move sharply to 245, 200 and 190 in 2023, 2024 and 2025. The most plausible reading of this change is a unit or normalization revision in the registry rather than a physical collapse in primary energy demand. Any predictive target based on a fixed national threshold of primary energy would therefore be incoherent across years, while a target that compares each building to the median of its own issue year remains well defined regardless of the underlying unit, which is the design adopted here.

The sampling frame is the population of issued certificates, not the population of buildings in Slovenia. Issued certificates are a non random subset of the building stock that reflects sale

events, regulatory triggers and voluntary requests. The conclusions of this article therefore apply to records in the EPC registry and not directly to the broader Slovenian building stock, a caveat that is consistent with the documented heterogeneity of EPC schemes across Europe [1], [2], [3].

4 METHODOLOGY

The pipeline follows the CRISP-DM cycle of business understanding, data understanding, data preparation, modeling, evaluation and deployment [26], implemented in Python 3.13 with pandas, numpy, scikit-learn 1.4, matplotlib and seaborn. All randomness is controlled by a single seed set to 42 and a five fold stratified cross validation scheme used for out of fold computations. The full notebook is released alongside this paper.

The dependent variable is a binary indicator of year relative efficiency. For each issue year the national median of primary energy is computed over all valid certificates issued in that year, and each record is labeled one if its primary energy strictly exceeds that median and zero otherwise. Primary energy enters the pipeline exclusively for target construction and is dropped from the predictor matrix immediately afterwards together with the required heating energy, delivered energy, total energy, delivered electricity and CO₂ emissions. The retained predictors are conditioned floor area as a continuous physical descriptor, issue year as a temporal control, building part number with a derived binary feature equal to one for whole buildings, certificate type and EPBD category as administrative categorical descriptors, and the cadastral community code as the only spatial variable. This restriction is the central design decision of the paper and produces an inference time pipeline that consumes no direct energy variable.

The cadastral community code has 2,357 unique values in the modeling sample, which makes one hot encoding both sparse and unstable for rare categories. We instead use smoothed target encoding in the style of Micci-Barreca [23], implemented as a sigmoid mixture between the global training mean of the target and the per cadastral community mean weighted by sample size with parameters `min_samples` 20 and `smoothing` 20, and computed inside an out of fold loop on the training set to prevent leakage [20], [21]. The neighborhood spatial lag is computed analogously. For each cadastral community we average the per cadastral community target rate of its neighbors under the public adjacency graph, falling back to the average of its ten nearest centroids in the D96/TM projection when adjacency information is missing. The neighborhood average is also produced out of fold on the training set and is filled with the global training mean for isolated communities, which makes the construction well defined for the few cadastral communities with no neighbors in the training set.

The three model configurations share the same scikit-learn pipeline, namely median imputation and standard scaling for the four continuous features, most frequent imputation and one hot encoding for the two categorical features, and a logistic regression classifier with the L2 penalty, the lbfgs solver, balanced class weights, a tolerance of 1e-4 and a maximum of 5,000 iterations. Model A consumes the six non spatial predictors. Model B adds the target encoded cadastral community code. Model C adds the neighborhood spatial lag on top of Model B. Holding the algorithm and hyper-parameters fixed across the three configurations isolates the marginal contribution of each spatial feature.

Evaluation uses a single stratified 80/20 split with seed 42, producing 46,272 training records and 11,569 test records, and reports four metrics, namely the area under the ROC curve, the F1 score for the positive class, balanced accuracy and standard accuracy. A stratified dummy classifier provides the chance reference. Variability across resampling is not reported in the main split, which is a deliberate limitation of the current evaluation regime, and is instead approximated by a temporal holdout in which the model is trained on certificates issued

between 2017 and 2024 and tested on certificates issued in 2025 only, producing 55,034 training records and 2,807 test records. The released notebook also contains an optional bootstrap routine that recomputes 95 percent confidence intervals on the test metrics by resampling test predictions 200 times.

5 RESULTS

This section reports numerical outcomes only. Interpretive content is deferred to Section .

Figure 5.1 shows the share of records above the year median by issue year, which is close to 0.5 in every year as expected from the year relative construction, and the kernel density of conditioned floor area split by the target, which shows that the upper efficiency class concentrates at smaller floor areas while the lower class has a heavier tail at larger floor areas.

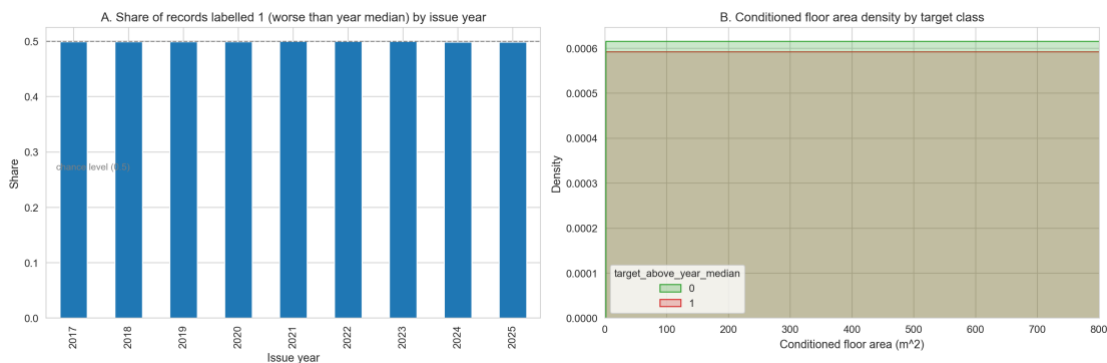


Figure 1: Exploratory analysis on the modeling sample ($n = 57,841$ records). Left panel A shows the share of records labeled one (primary energy above the national median of the issue year) by issue year. Right panel B shows the density of conditioned floor area in square meters for the two target classes on a common density scale. ROC AUC denotes the area under the receiver operating characteristic curve and F1 denotes the F1 score for the positive class throughout.

Table 5.1 reports the four evaluation metrics on the random 80/20 test set for the dummy classifier and the three substantive models. The dummy classifier reaches 0.503 on every metric. Model A reaches ROC AUC 0.730, F1 0.699, balanced accuracy 0.691 and accuracy 0.691. Model B reaches 0.748, 0.708, 0.700 and 0.700. Model C reaches 0.748, 0.708, 0.700 and 0.700. The difference between Model B and Model C on every reported metric is smaller than 0.001 in absolute value.

Table 1: Comparison of the dummy classifier and Models A to C on the stratified 80/20 test split ($n = 11,569$ records, seed 42). ROC AUC is the area under the ROC curve. F1 is the F1 score for the positive class. Balanced accuracy is the arithmetic mean of class wise recall. Accuracy is the standard accuracy. Higher is better for every metric. Differences are reported to three decimal places to match the resolution at which Model B and Model C separate.

Model	ROC AUC	F1	Balanced acc.	Accuracy
Dummy (stratified)	0.503	0.499	0.503	0.503
Model A: non spatial baseline	0.730	0.699	0.691	0.691
Model B: A + KO target encoding	0.748	0.708	0.700	0.700
Model C: B + neighborhood spatial lag	0.748	0.708	0.700	0.700

Figure 5.2 compares the four metrics across the four models in two panels, the first showing ROC AUC and F1 and the second showing balanced accuracy and standard accuracy. Figure 5.3 overlays the ROC curves of Models A, B and C with the chance diagonal.

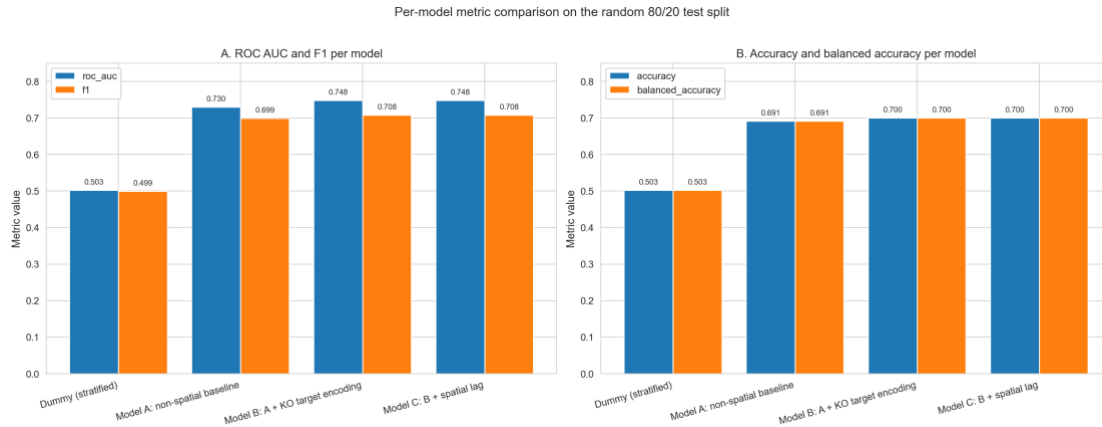


Figure 2: Per model metric comparison on the random 80/20 test split. Bars are labeled with the model identifiers Model A, Model B and Model C. Panel A reports ROC AUC and F1 score. Panel B reports balanced accuracy and standard accuracy. The dummy classifier is included as the chance reference. All bars share the same y axis range from 0 to 0.8 across both panels for direct visual comparison.

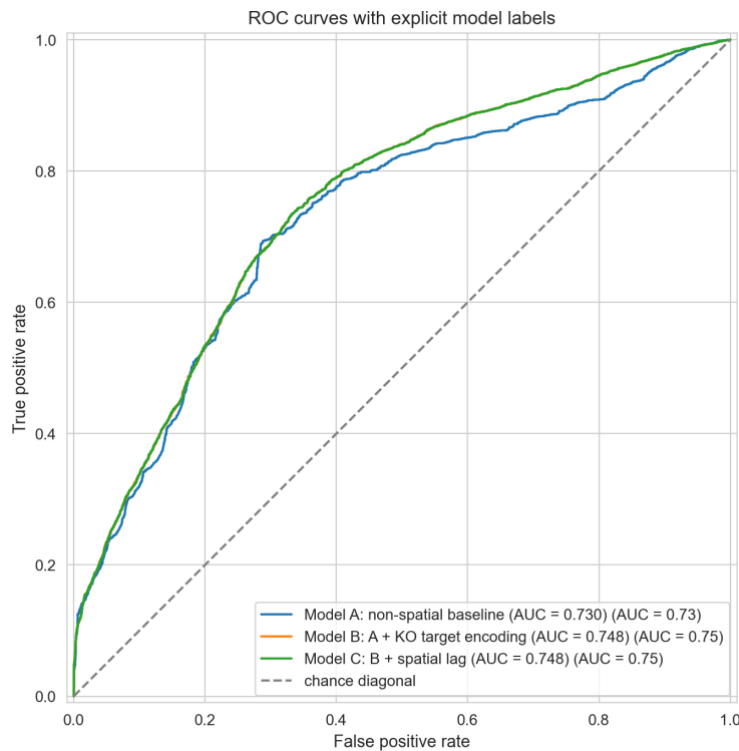


Figure 3: ROC curves for Model A, Model B and Model C on the stratified 80/20 test set ($n = 11,569$ records). The dashed line is the chance diagonal. Curves for Model B and Model C overlap to within the plotting resolution and the legend reports the AUC of each model to three decimal places.

The classification report for Model C on the test set shows precision 0.712 and recall 0.672 for the lower class, precision 0.689 and recall 0.728 for the upper class, and weighted F1 0.701. The corresponding confusion matrix is shown in [Figure 5.4](#).

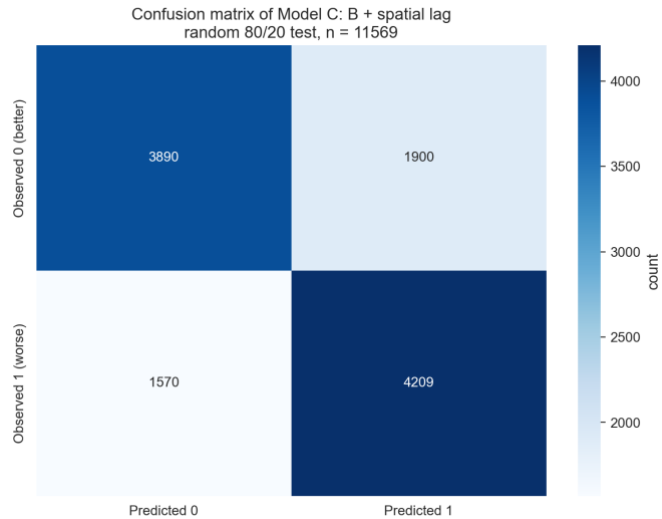


Figure 4: Confusion matrix of Model C on the random 80/20 test set, $n = 11,569$ records. Rows are the observed class and columns are the predicted class. Class 0 denotes records at or below the national median of the issue year and class 1 denotes records above that median.

Table 5.2 reports the five cadastral communities with the highest and the five with the lowest smoothed target encoded efficiency among those with at least 20 records in the training set. The highest values lie between 0.711 and 0.849 and the lowest values lie between 0.146 and 0.211. Figure 5.5 compares the fifteen most and fifteen least inefficient cadastral communities on a shared x axis. Figure 5.6 maps the encoded efficiency over the cadastral centroids.

Table 2: Top five and bottom five cadastral communities by smoothed target encoded efficiency, restricted to communities with at least 20 records in the training set. n is the number of training records. Observed share is the empirical share of records above the year median. Encoded efficiency is the smoothed target encoding used as input to Model B and Model C. KO denotes a cadastral community (slovenian: katastrska občina).

KO name (code)	n	Observed share	Encoded efficiency
DOLENJI LOGATEC (2017)	167	0.850	0.849
LJUTOMER (259)	97	0.804	0.788
ŠOŠTANJ (959)	100	0.770	0.762
VRTOJBA (2316)	46	0.870	0.757
NAKLO (2096)	33	0.939	0.751
ŠTEPANJA VAS (1732)	292	0.226	0.225
HUJE (2122)	156	0.212	0.212
NOVE JARŠE (2680)	227	0.189	0.190
SPODNJE RADVANJE (678)	907	0.189	0.188
GORENJA STRAŽA (1447)	138	0.145	0.148

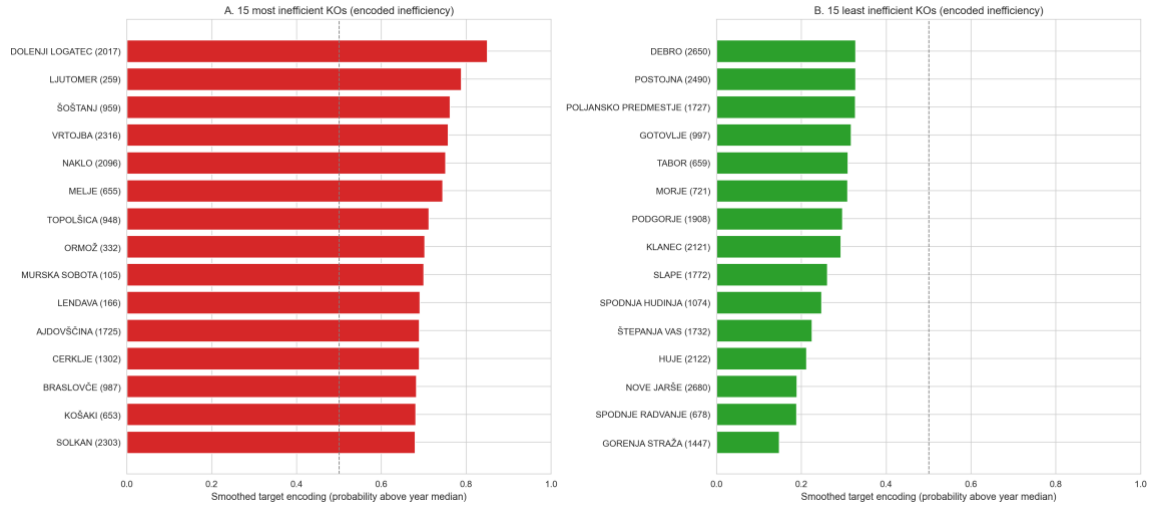


Figure 5: Top fifteen and bottom fifteen cadastral communities by smoothed target encoded efficiency among communities with at least 20 records in the training set. Bars are sorted within each panel. Panel A on the left shows the most inefficient cadastral communities and panel B on the right shows the least inefficient cadastral communities. Both panels share an x axis ranging from 0 to 1.

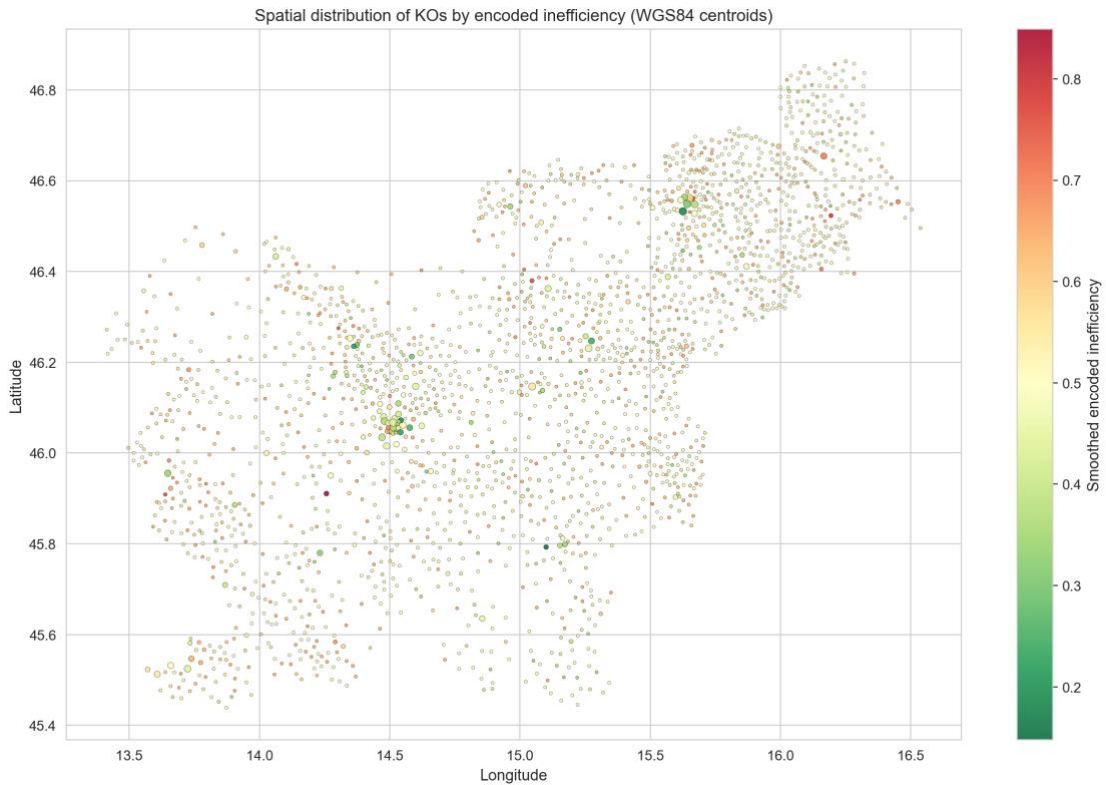


Figure 6: Spatial distribution of cadastral communities coloured by smoothed target encoded efficiency on WGS84 centroids. Marker size is proportional to the square root of the number of training records. Red denotes higher encoded probability that a record exceeds the year median and green denotes lower encoded probability. KO denotes a cadastral community.

The temporal holdout uses certificates issued in 2017 to 2024 for training and certificates issued in 2025 for testing. Under this regime, Model A reaches ROC AUC 0.640, F1 0.628, balanced accuracy 0.598 and accuracy 0.598. Model B reaches 0.661, 0.646, 0.634 and 0.633.

Model C reaches 0.661, 0.646, 0.634 and 0.633. The ordering of the three models is preserved and the difference between Model B and Model C remains below 0.001 on every metric.

6 DISCUSSION

The first observation is that the non spatial baseline already separates the year relative efficiency target from chance. Model A delivers ROC AUC 0.730 on the random split using only conditioned floor area, certificate type, EPBD category, issue year and building part information, which means that the basic geometry and administrative description of a building carry a measurable signal about whether its primary energy is above or below the national median of its issue year. This is consistent with the broader finding in the data driven building energy literature that compact physical and administrative attributes can sustain useful prediction quality even without engineering grade inputs [7], [9], [12]. The absolute level of Model A is modest, however, which limits its standalone usefulness and motivates the spatial extension.

The second observation is that smoothed out of fold target encoding of the cadastral community code adds 0.018 to ROC AUC and 0.009 to F1 over the non spatial baseline on the random split, with the same relative gain preserved on the temporal holdout. The magnitude is small in absolute terms but stable across evaluation regimes, which is consistent with the interpretation that cadastral location carries information that is not reducible to floor area, issue year or certificate type. This is in line with prior spatial EPC analyses that document neighborhood level heterogeneity in energy outcomes [13], [14], [15], [16], [17] and with the spatial econometric expectation that nearby units share latent drivers [19]. Our results are consistent with such latent drivers being present in the Slovenian registry, but the design of this study cannot identify any specific mechanism, and we accordingly do not claim that any specific factor such as climate, settlement age or renovation programme has been measured.

The third observation is that adding a neighborhood spatial lag on top of smoothed target encoding does not change any of the four metrics to three decimal places, on either the random split or the temporal holdout. The most parsimonious reading of this finding is collinearity. Both features are constructed from the same per cadastral community target rate, the target encoding is already smoothed toward the global mean for small cadastral communities, and the neighborhood lag averages those rates over adjacent communities that frequently share construction vintage, climate and settlement type. In such a setting, the neighborhood aggregate adds essentially no information that is not already contained in the smoothed cadastral community feature itself. This is a useful negative result for two reasons. It documents an empirical ceiling on what a neighborhood spatial lag can contribute once a high cardinality location variable has been smoothed into the model, and it cautions against the common practice of treating spatial lags as default add-ons in tabular machine learning pipelines that already use target encoding of administrative location.

The fourth observation concerns the temporal evaluation. The random split metrics are systematically higher than the temporal holdout metrics, with the gap reaching almost 0.1 in ROC AUC, which is consistent with the structural break in the primary energy variable between 2022 and 2023 and with broader concerns in the EPC literature about temporal drift in registry data [18]. The year relative target absorbs the change in the reported unit because the median is recomputed each year, but the joint distribution of cadastral communities and certificate types observed in 2025 differs enough from the earlier years to make 2025 a more demanding test set. The ordering of the three models is nevertheless preserved, which strengthens the qualitative finding that cadastral location is informative and the quantitative finding that the neighborhood lag does not add value on top of it.

The cadastral ranking in Table [5.2](#) and Figure [5.5](#) is consistent with a familiar geographic pattern in Slovenia, namely that smaller settlements with older detached housing in the eastern and south eastern part of the country tend to occupy the upper tail of relative efficiency, while compact urban neighborhoods in Ljubljana and Maribor with predominantly newer multi unit buildings occupy the lower tail. We deliberately frame this pattern as descriptive. The smoothed encoded efficiency of a cadastral community is by construction an estimate of the conditional probability that a record from that community is above the year median, given only the registry, and not a measurement of any particular physical driver. Mapping this estimate over the country, as in Figure [5.6](#), is therefore a useful triage tool for retrofit prioritization but not a substitute for engineering audits or socioeconomic surveys at the cadastral level.

Several limitations of this study are tied directly to future work. The sampling frame is the population of issued certificates, not the building stock, which limits external validity to building stock conclusions and motivates the integration of administrative building registers as a future enrichment layer. The structural break in the primary energy unit between 2022 and 2023 indicates that registry curation choices have a non trivial impact on absolute metric levels, which suggests that future evaluations should include a formal change point analysis on the registry before training. The classifier is a single logistic regression with default L2 regularization, which trades off some nonlinear interaction capacity for transparency. A natural next step is to repeat the three model comparison with gradient boosted trees under the same feature restriction, which would clarify whether nonlinear interactions among non energy attributes can compress the gap to direct energy methods. Variability across random splits is approximated only by the temporal holdout in the main paper, which justifies the bootstrap confidence interval routine released with the notebook as a near term complement and a full repeated split evaluation as a future extension. The spatial lag is constructed from cadastral adjacency or k nearest centroids, both of which ignore road network connectivity and demographic similarity. Future work can test richer neighborhood definitions, such as travel time or socioeconomic similarity, to confirm whether our negative finding on the spatial lag survives more sophisticated graph constructions.

The substantive contribution of the paper is therefore best read as three connected statements. A non energy registry view is sufficient to extract a meaningful signal about relative building energy efficiency in Slovenia. The cadastral community code, encoded as a smoothed out of fold target, adds a small but stable improvement that is consistent with location being a proxy for unobserved drivers. A neighborhood spatial lag adds no measurable value on top of that encoding under either random or temporal evaluation, which qualifies the prior literature on spatial features in tabular machine learning by showing that target encoding can absorb most of the neighborhood signal.

7 CONCLUSION

This paper has presented a spatial machine learning framework that predicts relative energy efficiency of buildings in Slovenia using only non energy attributes from the national EPC registry. By defining the target as the position of a building's primary energy relative to the national median of its issue year, the framework remains coherent across a structural break in the reported primary energy unit that occurs in the Slovenian registry between 2022 and 2023, which absolute thresholds cannot accommodate. By restricting the predictor set to conditioned floor area, certificate type, EPBD category, issue year, building part information and the cadastral community code, the framework isolates the predictive contribution of administrative location from the mechanical contribution of energy variables that drive the EPC label itself.

The empirical findings are that a non spatial baseline reaches ROC AUC 0.730 on a random test split, that smoothed out of fold target encoding of the cadastral community code raises

this to 0.748, and that adding a neighborhood spatial lag does not change any reported metric to three decimal places. The same ordering is preserved under a temporal holdout that trains on 2017 to 2024 and tests on 2025, where the three values become 0.640, 0.661 and 0.661 respectively.

The contribution to the broader literature is therefore methodological as well as empirical. The year relative target formulation generalizes beyond Slovenia to any registry that revises reporting conventions over time. The empirical bound on cadastral location as a non energy predictor is the first such estimate on the Slovenian registry. The negative result on the neighborhood spatial lag refines the common belief that spatial lags routinely improve tabular machine learning models. Future work that combines richer cadastral context layers with nonlinear learners under the same non energy restriction can extend this framework to other national registries and to higher resolution policy applications such as targeted retrofit prioritization.

8 DATA AND CODE AVAILABILITY

The Slovenian EPC registry is publicly accessible through the Ministry of the Environment, Climate and Energy portal [24]. The cadastral support data are accessible through the Slovenian national geospatial portal [25]. The full reproducibility bundle, comprising the notebook `predict_median_en.ipynb`, the processed cadastral support files `ko_centroids.csv` and `ko_neighbors.csv`, the per cadastral community output table `ko_efficiency_summary.csv` and all named figures referenced in Section , is available at the project repository at <https://github.com/Lahlukap669/Spatial-Prediction-of-Relative-Building-Energy-efficiency-in-Slovenia>.

9 APPENDICES

The appendices include all visualizations, processed data files and intermediate notebooks at <https://github.com/Lahlukap669/Spatial-Prediction-of-Relative-Building-Energy-efficiency-in-Slovenia>.

References

- [1] Y. Li, S. Kubicki, A. Guerriero, and Y. Rezgui, “Review of building energy performance certification schemes towards future improvement,” *Renewable and Sustainable Energy Reviews*, vol. 113, p. 109244, 2019, doi: [10.1016/j.rser.2019.109244](https://doi.org/10.1016/j.rser.2019.109244).
- [2] M. Sayfekar and D. P. Jenkins, “Energy performance certificate calculation methodologies across europe and accommodating new performance indicators,” *Building Services Engineering Research and Technology*, vol. 46, no. 1, pp. 71–93, 2025, doi: [10.1177/01436244241282076](https://doi.org/10.1177/01436244241282076).
- [3] Buildings Performance Institute Europe, “Energy performance certificates across the EU: A mapping of national approaches,” BPIE, 2014. Available: <https://www.bpie.eu/publication/energy-performance-certificates-across-the-eu/>
- [4] K. Amasyali and N. M. El-Gohary, “A review of data-driven building energy consumption prediction studies,” *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018, doi: [10.1016/j.rser.2017.04.095](https://doi.org/10.1016/j.rser.2017.04.095).
- [5] H. Zhao and F. Magoulès, “A review on the prediction of building energy consumption,” *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, 2012, doi: [10.1016/j.rser.2012.02.049](https://doi.org/10.1016/j.rser.2012.02.049).

- [6] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, vol. 147, pp. 77–89, 2017, doi: [10.1016/j.enbuild.2017.04.038](https://doi.org/10.1016/j.enbuild.2017.04.038).
- [7] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings*, vol. 49, pp. 560–567, 2012, doi: [10.1016/j.enbuild.2012.03.003](https://doi.org/10.1016/j.enbuild.2012.03.003).
- [8] A. Foucquier, S. Robert, F. Suard, L. Stéphan, and A. Jay, "State of the art in building modelling and energy performances prediction: A review," *Renewable and Sustainable Energy Reviews*, vol. 23, pp. 272–288, 2013, doi: [10.1016/j.rser.2013.03.004](https://doi.org/10.1016/j.rser.2013.03.004).
- [9] M. Bourdeau, X. Guo, E. Nefzaoui, X. Zhai, and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustainable Cities and Society*, vol. 48, p. 101533, 2019, doi: [10.1016/j.scs.2019.101533](https://doi.org/10.1016/j.scs.2019.101533).
- [10] F. Khayatian, L. Sarto, and G. Dall'O', "Application of neural networks for evaluating energy performance certificates of residential buildings," *Energy and Buildings*, vol. 125, pp. 45–54, 2016, doi: [10.1016/j.enbuild.2016.04.067](https://doi.org/10.1016/j.enbuild.2016.04.067).
- [11] T. Tsoka, X. Ye, Y. Q. Chen, D. Gong, and X. Xia, "Explainable artificial intelligence for building energy performance certificate labelling classification," *Journal of Cleaner Production*, vol. 355, p. 131626, 2022, doi: [10.1016/j.jclepro.2022.131626](https://doi.org/10.1016/j.jclepro.2022.131626).
- [12] D. Hsu, "Identifying key variables and interactions in statistical models of building energy consumption using regularization," *Energy*, vol. 83, pp. 144–155, 2015, doi: [10.1016/j.energy.2015.02.008](https://doi.org/10.1016/j.energy.2015.02.008).
- [13] C. Beltrán-Velamazán, M. Monzón-Chavarrías, and B. López-Mesa, "A new approach for national-scale building energy models based on energy performance certificates in european countries: The case of spain," *Heliyon*, vol. 10, no. 4, p. e25473, 2024, doi: [10.1016/j.heliyon.2024.e25473](https://doi.org/10.1016/j.heliyon.2024.e25473).
- [14] G. Mutani and V. Todeschi, "GIS-based urban energy modelling and energy efficiency scenarios using the energy performance certificate database," *Energy Efficiency*, vol. 14, no. 5, p. 47, 2021, doi: [10.1007/s12053-021-09962-z](https://doi.org/10.1007/s12053-021-09962-z).
- [15] S. Sisman, A. Kara, and A. C. Aydinoglu, "Leveraging spatial data infrastructure for machine learning based building energy performance prediction," *PLOS ONE*, vol. 20, no. 10, p. e0335531, 2025, doi: [10.1371/journal.pone.0335531](https://doi.org/10.1371/journal.pone.0335531).
- [16] M. McCord, D. Lo, P. T. Davis, L. Hemphill, J. McCord, and M. Haran, "A spatial analysis of EPCs in the belfast metropolitan area housing market," *Journal of Property Research*, vol. 37, no. 1, pp. 25–61, 2020, doi: [10.1080/09599916.2019.1697345](https://doi.org/10.1080/09599916.2019.1697345).
- [17] J. Nyangon and J. Byrne, "Spatial energy efficiency patterns in new york and implications for energy demand and the rebound effect," *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 16, no. 5, pp. 402–432, 2021, doi: [10.1080/15567249.2020.1868619](https://doi.org/10.1080/15567249.2020.1868619).
- [18] S. Papadopoulos, B. Bonczak, and C. E. Kontokosta, "Pattern recognition in building energy performance over time using energy benchmarking data," *Applied Energy*, vol. 221, pp. 576–586, 2018, doi: [10.1016/j.apenergy.2018.03.079](https://doi.org/10.1016/j.apenergy.2018.03.079).
- [19] L. Anselin, *Spatial econometrics: Methods and models*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1988. doi: [10.1007/978-94-015-7799-1](https://doi.org/10.1007/978-94-015-7799-1).

- [20] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *arXiv preprint arXiv:2104.00629*, 2021, doi: [10.48550/arXiv.2104.00629](https://doi.org/10.48550/arXiv.2104.00629).
- [21] F. Matteucci, V. Arzamasov, and K. Böhm, "A benchmark of categorical encoders for binary classification," in *Advances in neural information processing systems*, 2023. doi: [10.48550/arXiv.2307.09191](https://doi.org/10.48550/arXiv.2307.09191).
- [22] M. K. Dahouda and I. Joe, "A deep-learned embedding technique for categorical features encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: [10.1109/ACCESS.2021.3104357](https://doi.org/10.1109/ACCESS.2021.3104357).
- [23] D. Micci-Barreca, "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems," *SIGKDD Explorations*, vol. 3, no. 1, pp. 27–32, 2001, doi: [10.1145/507533.507538](https://doi.org/10.1145/507533.507538).
- [24] Ministry of the Environment, Climate and Energy of Slovenia, "Register izdanih energetskih izkaznic." 2025. Available: <https://www.energetika-portal.si/podrocja/energetika/energetske-izkaznice-stavb/javna-evidenca-energetskih-izkaznic/>
- [25] Surveying and Mapping Authority of the Republic of Slovenia (GURS), "Dostop do geodetskih podatkov." 2025. Available: <https://www.e-prostor.gov.si/>
- [26] C. Schröer, F. Kruse, and J. Marx Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: [10.1016/j.procs.2021.01.199](https://doi.org/10.1016/j.procs.2021.01.199).

Luka Lah is a PhD student of Organization and Management of Information Systems at the Faculty of Organizational Sciences, University of Maribor. His interests are the implementation of artificial intelligence models in sports and other disciplines.