Modeliranje potrošnje toplinske energije u sustavima daljinskog grijanja zasnovano na strojnom učenju

Maljković, Danica

Doctoral thesis / Disertacija

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture / Sveučilište u Zagrebu, Fakultet strojarstva i brodogradnje**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:235:388647

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

Repository / Repozitorij:

Repository of Faculty of Mechanical Engineering and Naval Architecture University of Zagreb







Sveučilište u Zagrebu

FAKULTET STROJARSTVA I BRODOGRADNJE

Danica Maljković

MODELIRANJE POTROŠNJE TOPLINSKE ENERGIJE U SUSTAVIMA DALJINSKOG GRIJANJA ZASNOVANO NA STROJNOM UČENJU

DOKTORSKI RAD

Zagreb, 2020.



Sveučilište u Zagrebu

FAKULTET STROJARSTVA I BRODOGRADNJE

Danica Maljković

MODELIRANJE POTROŠNJE TOPLINSKE ENERGIJE U SUSTAVIMA DALJINSKOG GRIJANJA ZASNOVANO NA STROJNOM UČENJU

DOKTORSKI RAD

Mentori:
prof. dr. sc. Igor Balen
prof. dr. sc. Bojana Dalbelo Bašić

Zagreb, 2020.



University of Zagreb

FACULTY OF MECHANICAL ENGINEERING AND NAVAL ARCHITECTURE

Danica Maljković

MODELLING OF ENERGY CONSUMPTION IN DISTRICT HEATING SYSTEMS BASED ON MACHINE LEARNING

DOCTORAL DISSERTATION

Supervisors:
Professor Igor Balen, PhD
Professor Bojana Dalbelo Bašić, PhD

Zagreb, 2020

BIBLIOGRAFSKA KARTICA

UDK:	**697 : 620.9**
Ključne riječi:	daljinsko grijanje, strojno učenje, predviđanje potrošnje energije, energetska učinkovitost, razdjelnici troškova toplinske energije
Key words:	district heating, machine learning, energy forecasting, energy efficiency, heat cost allocators
Znanstveno područje:	TEHNIČKE ZNANOSTI
Znanstveno polje:	STROJARSTVO
Institucija u kojoj je izrađen:	Sveučilište u Zagrebu Fakultet strojarstva i brodogradnje
Mentori rada:	prof. dr. sc. Igor Balen prof. dr. sc. Bojana Dalbelo Bašić
Broj stranica:	150
Broj slika:	91
Broj tablica:	33
Broj korištenih bibliografskih jedinica:	105
Datum obrane:	8. srpnja 2020.
Povjerenstvo:	1. prof. dr. sc. Dražen Lončar, predsjednik 2. prof. dr. sc. Jan Šnajder, član 3. doc. dr. sc. Mladen Zeljko, član
Institucija u kojoj je rad pohranjen:	Sveučilište u Zagrebu Fakultet strojarstva i brodogradnje

O MENTORIMA

Igor Balen rođen je 1967. godine u Zagrebu. Studij strojarstva na Fakultetu strojarstva i brodogradnje Sveučilišta u Zagrebu završio je 1991. Magistrirao je 1995. godine, a doktorirao 1999. u polju strojarstva, na Sveučilištu u Zagrebu, Fakultetu strojarstva i brodogradnje.

Od 1991. zaposlen je na Zavodu za termodinamiku, toplinsku i procesnu tehniku Fakulteta strojarstva i brodogradnje. 1995. godine boravio je 30 dana na znanstvenom usavršavanju na Technische Universität Wien, kao stipendist Ministarstva znanosti i tehnologije Republike Austrije. 1999. godine boravio je 2 mjeseca na znanstvenom usavršavanju u razvojnom institutu tvornice GEA-Happel Klimatechnik u Herneu, Savezna Republika Njemačka. 2003. godine boravio je 60 dana kao gost istraživač na znanstvenom projektu Low Energy Heating and Cooling System for Deep Plan Buildings na Dublin Institute of Technology, Republika Irska. Sudjeluje u izvođenju nastave niza kolegija iz područja grijanja i klimatizacije te energetske učinkovitosti i optimizacije sustava u zgradama. Ovlašteni je inženjer za projektiranje i stručni nadzor sustava grijanja, ventilacije, klimatizacije i plinskih instalacija te je član Hrvatske komore inženjera strojarstva. Član je ASHRAE-a (American Society of Heating, Refrigerating and Air Conditioning Engineers). Član je Hrvatskog zavoda za norme, Tehničkog odbora 541 (Sustavi grijanja, ventilacije i klimatizacije). Kao član povjerenstva Ministarstva graditeljstva i prostornog uređenja Republike Hrvatske sudjelovao je u izradi niza propisa i pravilnika iz područja energetske učinkovitosti u zgradarstvu, vezano uz racionalnu uporabu energije, energetske preglede i certificiranje zgrada. Bio je mentor studentima na 80 diplomskih i završnih radova i mentor jednom doktorandu koji je uspješno završio doktorski studij. Recenzent je u 12 inozemnih i domaćih međunarodnih znanstvenih časopisa. Recenzent je jednog sveučilišnog udžbenika i jedne znanstvene knjige te je urednik jedne stručne knjige. Objavio je više od 40 znanstvenih i stručnih radova iz područja strojarstva te dva priručnika i dva poglavlja u Hrvatskoj tehničkoj enciklopediji. Surađivao je u izradi niza stručnih projekata, analiza i studija iz područja toplinske i procesne tehnike.

Bojana Dalbelo Bašić rođena je u Zagrebu 1958. godine. Diplomirala je matematiku na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu 1982., a magistrirala i doktorirala u polju računarstva na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu 1993., odnosno 1997. godine.

Od 1982. do 1986. radila je u Građevinskom institutu u Zagrebu, od 1986. do 1990. u tvrtki Iskra Delta, a od 1990. do 2000. na Šumarskom fakultetu Sveučilišta u Zagrebu. Od 2000. godine radi na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave Fakulteta elektrotehnike i računarstva (FER). U prosincu 2013. godine izabrana je u trajno zvanje redovite profesorice. 2000. godine znanstveno se usavršavala na Sveučilištu Erlangen-Nürnberg (DAAD stipendija) te je bila je gostujuća znanstvenica na institutu INRIA u Rennesu, Francuska, 2005. godine. Bila je voditeljica četiriju međunarodnih i pet domaćih znanstvenoistraživackih projekata te suradnica na jedanaest drugih projekata. Osnivačica je i voditeljica Laboratorija za analizu teksta i inženjerstvo znanja na FER-u. Bila je mentorica petorici doktoranada i mentorica studentima na 90 diplomskih i završnih radova. Objavila je više od 95 radova u časopisima i na znanstvenim konferencijama iz područja obrade prirodnog jezika, pretraživanja informacija, strojnog učenja, inteligentnih sustava i multivarijatne statistike. Prof. Dalbelo Bašić članica je suradnica Hrvatske akademije tehničkih znanosti (HATZ), senior članica udruge IEEE, članica i suosnivačica Hrvatskog društva za jezične tehnologije, članica Hrvatskog biometrijskog društva i Hrvatskog matematičkog društva. Sudjelovala je u više međunarodnih programskih odbora međunarodnih znanstvenih konferencija i bila recenzentica u većem broju međunarodnih časopisa i konferencija. Bila je voditeljica tima kojem je dodijeljena nagrada "Zlatno Teslino jaje" 2007. i 2009. godine te „Premijerkina WWW nagrada“ 2009. za projekt CADIAL. Za svoj rad 2015. dobila je priznanje "Zlatna plaketa Josip Lončar" koje dodjeljuje FER.

ZAHVALA

Hvala mojim mentorima. Profesorici Bojani Dalbelo Bašić hvala što mi je otvorila jedan novi svijet u tijeku pisanja ovog rada i što je uvijek imala strpljenja i vremena za rasprave oko razdjelnika. Profesoru Igoru Balenu hvala što uvijek želi istraživati grijanje, što me naučio svemu što znam o grijanju još od studentskih dana i što je bio otvoren za to da u sektoru daljinskog grijanja pokušamo istraživati neke nove metode.

Zahvaljujem članovima Povjerenstva što šire znanja o ovom području, kao i na svemu što su i mene naučili u tom procesu.

Hvala mom ravnatelju, Goranu Graniću, što me uvijek podržavao u svim mojim istraživanjima i što je uživao u našim raspravama o toplinarstvu. Hvala Karmen što me naučila da ne gledam energetiku kao čistu tehničku disciplinu, nego da je smjestim u tamo neke zakonske okvire i da uvijek mislim na krajnje potrošače.

Hvala mojoj obitelji što su čak i sada bili tu. Vlatki, što je nepogrešivi navigator i uvijek zna kada smo u najboljem kursu. Mojoj mami, zato što mi je otkrila što je lijepo u čitanju i što mi je uvijek pružala pogled koji je drugačiji. Mom tati, zato što me uvijek puštao da naučim sama i što nikad nije htio opterećivati mlade generacije. Mojim tetkama, zato što su uvijek podržavale sve moje čudnovate ideje. Život s tetkama je bogatstvo. Tetki Jelki posebno hvala na poticanju na istraživanje svih oblika. Mom bratu i mojim sestrama, zato što su od mene uvijek očekivali da napišem doktorat - i Petri i Irini što su zbog njih smanjili intenzitet pitanja kada će biti gotov.

Na kraju, posebno zahvaljujem gospođi Mariji Ževrnji i gospodinu Saši Topiću što su nebrojeno sati proveli sa mnom i što su mi dostavili neprocjenjive podatke o potrošnji toplinske energije u njihovim zgradama. Nadam se da će rezultati ovog rada uljepšati i njihovu svakodnevicu, barem u sezoni grijanja. Hvala Ivi što misli da stručne radove ne treba lektorirati, ali svejedno to čini.

SAŽETAK

Na putu prema postizanju energetskih ciljeva Europska unija je kroz zakonodavstvo propisala zemljama članicama provođenje mjera energetske učinkovitosti od kojih je jedna uvođenje individualnog mjerenja kod svih krajnjih potrošača u sustavima daljinskog grijanja. Cilj ove disertacije je ocijeniti stvarni učinak ugradnje razdjelnika troškova toplinske energije (RTTE) kao individualnih mjerila u smislu ocjene postignutih ušteda i predviđanja potrošnje nakon ugradnje. Modeliranje je provedeno metodama strojnog učenja – višestrukom linearnom regresijom, regresijskim stablima, slučajnim šumama i strojem potpornih vektora. Razvijeni su modeli s visokom točnošću predviđanja na razini stana, zgrade, distribucijskog sustava i na nacionalnoj razini. Točnost predviđanja na razini stana kreće se od +/- 4 kWh/m^2 do +/- 16 kWh/m^2 na godišnjoj razini, a na razini zgrade od +/- 13 kWh/m^2 do +/- 16 kWh/m^2. Razvijeni modeli interpretirani su i sa stajališta parametara koji utječu na potrošnju na potrošnju. Tako je na razini stana dominantan utjecajni parametar postojanje individualnog mjerenja i stanovi koji su ugradili RTTE mogu očekivati uštedu na razini od 40 %. Drugi utjecajan parametar na razini stana, a prvi utjecajan parametar na razini zgrade jest stopa ugradnje RTTE koja označava u kojem se postotku grijane površine zgrade potrošnja topline mjeri s RTTE. Važnost ovog parametra daje indikaciju koliko je bitno da se ugradnja provede u svim stanovima. Zgrade koje ugrade RTTE mogu očekivati apsolutnu uštedu do 46 kWh/m^2 godišnje, ovisno o stopi ugradnje. Očekivane uštede na razini distribucijskog područja i na nacionalnoj razini jednake su kao u slučaju zgrade. Razvijeni modeli mogu služiti u predviđanju potrošnje topline za stanove i zgrade koje još nisu ugradile RTTE te se njima mogu predviđati i uštede na razini distribucijskog područja i na nacionalnoj razini. Rezultati ovog rada pokazuju da se metode strojnog učenja mogu uspješno primjenjivati u predviđanju potrošnje energije s visokom razinom točnosti u usporedbi s klasičnim simulacijskim metodama te se u budućnosti očekuje veća primjena strojnog učenja u predviđanju u sektoru energetike.

Ključne riječi:

daljinsko grijanje, strojno učenje, predviđanje potrošnje energije, energetska učinkovitost, razdjelnici troškova toplinske energije

EXTENDED ABSTRACT

In recent years energy planners consider district heating to be the most sustainable option for the future of urban heating. The most important characteristic of DH is the use of an energy source that provides a significant cost differential in generating heat compared with conventional heating using boilers or direct electric heating. Mainly due to the flexibility of the primary energy sources that could be fully reliable on renewables, district heating has a significant potential to be one of the most promising heat supply options in the future.

At the same time, the growing trend of energy consumption in the building sector has been observed for the last few decades. The building sector in the EU consumes approximately 40% overall final energy, while heating and hot water preparation alone account for almost 80% of total energy use. The EU has set a common legal framework in order to ensure that the Union's 2020 headline targets on energy efficiency of 20% and its 2030 headline targets on energy efficiency of at least 32.5 % are met and paves the way for further energy efficiency improvements beyond those dates.

According to the Energy Efficiency Directive, Member States shall ensure that, for district heating, district cooling, and domestic hot water, final customers are provided with competitively priced meters that accurately reflect their actual energy consumption. Measuring heat delivered to each apartment can be easily performed with heat cost allocators (HCA). Introduction of HCA is a technique which can be used to ensure that the consumer pays only for the heat supplied to an individual apartment. Although numerous studies and analyses have confirmed that by installing HCAs at the building level annual heat savings can be expected from 15% to 30%, in the case of Croatia these goals were not reached as expected. This indicated that there might be a significant impact of behavioural and other non-technical factors on the energy consumption in district heating. Recently, a growing research focus has been placed on the non-technical factors affecting energy consumption.

The research goal of this thesis is to evaluate the impact of HCA instalment in district heating systems on energy savings, develop energy prediction models for the purpose of prediction of energy consumption and evaluation of impact of energy efficiency measures on apartments', buildings', district heating systems' and nationwide level with machine learning algorithms. Models developed by machine learning methods thus have a dual function, they are predictive

and serve to predict outcomes based on past data, but they are also descriptive because they serve to acquire new knowledge about the process of household heat consumption.

Although it is custom to use numerous simulation tools developed for calculating building energy loads these simulation tools are often costly, require a great deal of prior knowledge for efficient use and lack in terms of prediction accuracy of actual energy consumed. Several previous researches indicated that machine learning methods can mitigate these obstacles and develop models with high accuracy of prediction when considering energy consumption. The research results showed that machine algorithms are superior in building energy calculations in comparison to conventional statistical and engineering methods (simulation tools), while there are ready to use software and programming languages for implementation like Matlab, Python or R. Analysis made in this research are developed using R programming language.

Within this research chosen machine learning algorithms were implemented on a data set consisting of 3,854,000 observations of 50 variables of actual billing data in the period of 7 years for two district heating systems in Croatia. Additionally, a on-field research of behavioural aspects of final consumers in district heating was undertaken on a set of three multiapartment buildings in Zagreb via questionnaires and interviews.

The models were developed by data pre-processing, learning on the training data subset, and accessing model accuracy on the verification set, following by the interpretation of the influence of individual parameters on the heat consumption for each model. Given the type of data available in this analysis models were developed using these machine learning methods: regression analysis, regression trees, random forest, and support vector machine (support vector machines, SVM). The results obtained were interpreted in such a way that the prediction accuracy of each method was evaluated, and a recommendation was made as to which model is the most suitable for the prediction of consumption in district heating systems.

The development of high accuracy models for the prediction of heat consumption will also stimulate an increase in the integration of renewable energy sources (RES) for the production of thermal energy in district heating systems, thus enabling the integration of such a developed model with existing models for the prediction of energy consumption. This primarily refers to the conversion of electricity from renewable sources to thermal energy during periods of high availability of intermittent RES such as solar or wind, which would contribute to greater integration of RES into the energy system

Objective and hypothesis of the research

The objective of the research is to develop machine learning models of heat consumption in district heating systems, for the purpose of energy planning and evaluating the impact of energy efficiency measures. From this goal, the following research hypotheses emerge:

1. Available data are sufficiently representative, and by appropriate pre-processing, using descriptive multivariate statistical analysis and exploratory analysis, it is possible to obtain a more detailed insight into the character of the data, their representativeness and their interrelation and the appropriate assessment of suitability for incorporation into the predictive model.

2. Based on previous analyses, it is possible to build a machine learning based models that will predict energy consumption more accurately than current models.

Scientific contribution

The scientific contribution of the research is the development of models for predicting energy consumption and assessing the impact of energy efficiency measures in district heating systems based on machine learning methods. The model has been developed at several levels, sequentially (i) at the level of the individual apartment in the district heating system, (ii) at the level of the individual building connected to the district heating system, (iii) at the level of each distribution area, and (iv) at country level.

Also, evaluation of each machine learning algorithms considered for this application is given, namely multiple linear regression, regression trees, random forests, and support vector machine as supervised methods, and grouping as an unsupervised method.

Additionally, as a result of the undertaken machine learning analyses, the level of influence of each factor of the model will is determined and its influence on the heat consumption in district heating systems is interpreted.

Conclusion

The developed regression models at the apartment level show that apartments with built-in HCAs can expect 40% less energy consumption than apartments that did not install them, in

case they do not account more than 4% of all impulses in the building on an annual basis. An additional interpretability is obtained by the random forest model, which shows that the lowest average annual consumption is in those apartments that have installed HCAs and which account for less than 3.6% of all impulses in the building annually. There is a total of 10% of such apartments, and their specific consumption ranges from 35 to 88 kWh/m^2. The highest consumption in the group of apartments that have installed HCA are those that count more than 8.1% of all HCA impulses on an annual basis and their average consumption is around 125 kWh/m^2. Such apartments make 3% of all analysed apartments. At the same time, the specific consumption of those apartments which did not install HCAs, and which are in the buildings where the installation was partly carried out, ranges from 97 to 176 kWh/m^2. Therefore, the existing allocation method allows some of the apartments in the same building that did not install HCA to be allocated less heat than the apartments that have installed them.

The reason for the large number of HCAs' impulses in some apartments may be due to behavioural characteristics of the consumers or in the technical defects of the heating system in a building or in an apartment, which should be separately identified for each such case. As part of this research, surveys with questionnaires and interviews were undertaken to determine the impact of behavioural parameters. The analysis of data obtained from the questionnaire was performed by the clustering method, an unsupervised machine learning method. The results indicated that there is an influence of behavioural parameters on specific heat consumption, especially the parameters such as the desired level of thermal comfort, daily occupation time and number of unheated rooms.

At the building level, all analysed buildings achieved savings after HCAs were installed. The resulting regression model gives an indication that 100% installation of HCAs in the building will result with an absolute specific savings of 46 kWh/m^2 annually on the building level. The model of regression trees gives additional interpretability on the building level, indicating that the lowest consumption (93 kWh/m^2) is in those buildings that have a rate of installation of HCAs higher than 62 % of heated area. The prediction accuracy of each of the 4 models at the apartment and building levels is high compared to conventional simulation methods. The highest accuracy is obtained by the random forest method (apartments +/- 4.27 kWh/m^2, buildings 13.24 kWh/m^2) and the lowest by multiple linear regression (apartments +/- 16.44 kWh/m^2, buildings 16.59 kWh/m^2).

When selecting a model, a trade-off between level of interpretation and accuracy should be found. If the goal of modelling and forecasting is to obtain as accurate model as possible with a certain level of interpretability, then it is suggested to use a regression tree model. On the other hand, if our sole purpose of forecasting is to obtain highest accuracy, as it would be in the case of evaluating the impact of energy efficiency measure of HCA instalment, it is suggested to use a random forest as a machine learning algorithm.

Considering the degree of influence of individual variables, it is concluded that in the apartments the variables related to the metering of energy consumption (ratio of counted impulses in the apartment and counted impulses in the building, and metered heat on the building level) are dominant along with variables that describe whether the apartment has installed HCA and how high is the rate of installation in a building. If the target is to reduce the average energy in all apartments, these variables should be affected. In order to maximise the energy savings it is recommended to install HCA in all apartments (measures related to the implementation of the secondary energy efficiency measure), to implement all cost-effective measures on the building envelope (measures related to the implementation of primary energy efficiency measures), and to implement end-user training measures.

The positive effects of HCA installation are inevitable at the building level, DH systems and broader national contexts, but at the apartment level, large scattering and appereance of outliers occur mostly due to the behavioural parameters identified in this research. The models developed in this research can serve as a basis for forecasting consumption in district heating systems with high-accuracy and can be used as a tool to evaluate the performance of some of the energy efficiency measures.

Key words:

district heating, machine learning, energy forecasting, energy efficiency, heat cost allocators

SADRŽAJ

Bibliografska kartica I

O mentorima II

Zahvala IV

Sažetak V

Extended abstract VI

Sadržaj XI

Popis slika XIV

Popis tablica XX

Popis oznaka XXII

1. Uvod 1

1.1. Područje istraživanja 2

1.1.1. Sustavi daljinskog grijanja 5

1.1.2. Razdjelnici troškova toplinske energije 14

1.1.3. Predviđanje potrošnje energije metodama strojnog učenja 17

1.2. Metodologija istraživanja 19

1.3. Znanstveni doprinos 21

1.4. Hipoteza 21

1.5. Struktura rada 22

2. Metode strojnog učenja 23

2.1. Grupiranje 25

2.1.1. Grupiranje k-srednjim vrijednostima 25

2.1.2. Određivanje optimalnog broja grupa za podatkovni skup 26

2.1.3. Mjere sličnosti 27

2.2. Procjena točnosti modela 28

2.2.1. Određivanje kvalitete odabira 28

2.2.2. Kompromis između pristranosti i varijance 30

2.3. Višestruka linearna regresija 31

2.3.2. Vizualizacija linearnih modela ... 35
2.3.3. Potencijalni problemi ... 36
2.4. Metode ponovnog uzrokovanja ... 39
2.4.1. Metoda skupa za provjeru ... 39
2.4.2. Unakrsna provjera - zadržavanje jedne vrijednosti sa strane ... 40
2.4.3. *K*-struka unakrsna provjera ... 41
2.4.4. Bootstrap ... 42
2.4.1. Kompromis između odstupanja i varijance k-struka unakrsne provjere ... 42
2.5. Odabir linearnog modela i regularizacija ... 43
2.5.1. Odabir najboljeg podskupa ... 44
2.5.2. Regularizacija ... 47
2.6. Regresijska stabla odluke ... 47
2.6.1. Osnove odlučivanja stablima ... 48
2.6.2. Skraćivanje stabla ili rezidba ... 49
2.6.3. Bootstrap ... 51
2.6.4. Usporedba metoda stabala i linearnih modela ... 52
2.7. Slučajne šume ... 53
2.7.1. Boosting ... 54
2.8. Stroj potpornih vektora ... 55
3. Provedeno istraživanje ... 60
3.1. Analizirani skupovi podataka ... 60
3.1.1. Podaci iz sustava za naplatu ... 62
3.1.2. Podatkovni skup za razinu stana ... 64
3.1.3. Podatkovni skup za razinu zgrade i više razine ... 66
3.2. Analiza podataka dobivenih intervjuima i upitnicima ... 67
3.3. Analiza grupiranjem ... 72
3.3.1. Utjecaj modela obračuna energije ... 73
3.3.2. Utjecaj položaja vlasničke jedinice unutar zgrade ... 74
3.3.3. Utjecaj broja negrijanih prostora ... 76

3.3.4. Utjecaj broja stanara ... 77
3.3.5. Utjecaj tipa prozora ... 79
3.3.6. Utjecaj temperature u prostorijama ... 81
3.3.7. Utjecaj više faktorskih varijabli ... 82
3.3.8. Grupiranje na razini zgrada ... 84
3.4. Strojno učenje za razinu stana ... 90
3.4.1. Višestruka linearna regresija ... 90
3.4.2. Regresijska stabla ... 112
3.4.3. Slučajne šume ... 119
3.4.4. Stroj potpornih vektora ... 122
3.5. Strojno učenje na razini zgrade ... 124
3.5.1. Višestruka linearna regresija ... 126
3.5.2. Nelinearni modeli strojnog učenja ... 127
3.6. Strojno učenje na višim razinama ... 129
4. Prikaz rezultata ... 131
4.1. Predviđajući model za razinu stana ... 133
4.2. Predviđajući model na razini zgrada ... 140
4.3. Predviđajući model na višim razinama ... 143
4.4. Vrednovanje algoritama strojnog učenja ... 144
4.5. Određivanje stupnja utjecaja pojedinih faktora na potrošnju toplinske energije ... 145
4.5.1. Utjecaj na razini stanova ... 145
4.5.2. Utjecaj na razini zgrada ... 148
4.5.3. Utjecaj na razini distribucijskog sustava i nacionalna razina ... 148
5. Zaključak ... 149
Popis literature ... 151
Životopis ... 162
Prilozi ... 164

POPIS SLIKA

Slika 1-1 Sustavi daljinskog grijanja u Republici Hrvatskoj (Izvor: [37]) 11

Slika 1-2 Razdjelnik troškova toplinske energije na radijatoru 14

Slika 1-3 Shema ugradnje razdjelnika troškova toplinske energije unutar zgrade (Izvor: Honeywell) 15

Slika 3-1 Histogramski prikaz raspodjele izabranih varijabli za *psStan* 65

Slika 3-2 Zgrade u kojima su provedeni intervjui i upitnici 67

Slika 3-3 Primjer ispunjenog upitnika 69

Slika 3-4 Postotak članova kućanstva prema dobnim skupinama zgrada A,B i C 71

Slika 3-5 Vrijeme prosječnog boravka u stanovima zgrada A,B i C 71

Slika 3-6 Tip prozora zgrada A, B i C 71

Slika 3-7 Razina temperature u prostorijama zgrada A,B i C 71

Slika 3-8 Broj negrijanih prostorija zgrada A, B i C 71

Slika 3-9 Prosječno vrijeme prozračivanja stana zgrada A, B i C 71

Slika 3-10 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *ModelEnergija* 73

Slika 3-11 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *ModelEnergija* 73

Slika 3-12 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *Kat* 75

Slika 3-13 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *Kat* 75

Slika 3-14 Grafički prikaz grupiranje po varijablama *SpecTopSDG* i *NegrijaneProstorije* 76

Slika 3-15 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *NegrijaneProstorije* 77

Slika 3-16 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *UpitnikOsobe* 77

Slika 3-17 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *UpitnikOsobe* 78

Slika 3-18 Pravokutni dijagram neskaliranih podataka, ovisnosti potrošnje stana o broju osoba i vremenu boravka 78

Slika 3-19 Pravokutni dijagram neskaliranih podataka ovisnosti potrošnje o vremenu boravka 79

Slika 3-20 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *TipProzora*..... 80

Slika 3-21 Pravokutni dijagram neskaliranih podataka, *SpecTopSDG* i *TipProzora* 80

Slika 3-22 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *ToplinskaUgodnost*..... 81

Slika 3-23 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *ToplinskaUgodnost* 81

Slika 3-24 Višedimenzionalno grupiranje utjecaja bihevioralnih parametara na potrošnju 82

Slika 3-25 Specifična potrošnja po pojedinim grupama 82

Slika 3-26 Provjera utjecaja stršećih vrijednosti *psStan* 85

Slika 3-27 Dijagram rasipanja neskaliranih podataka potrošnje u analiziram zgradama 86

Slika 3-28 Grafički prikaz grupiranja u zavisnosti od broj grupa 86

Slika 3-29 Prikaz rasipanja SpecTopSDG za 2017. godinu za oba načina obračuna energije (2EG i 3EG) 91

Slika 3-30 Utjecaj stršećih vrijednosti na analizirani podatkovni skup za regresijski model na razini stana..... 92

Slika 3-31 Korelacija između numeričkih varijabli u početnom regresijskom modelu na razini stana..... 94

Slika 3-32 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za početni regresijski model za skup svih stanova 95

Slika 3-33 Ovisnost log($\boldsymbol{\lambda}$) i MSE za unakrsnu provjeru u 10 koraka za skup svih stanova ... 96

Slika 3-34 Promjena vrijednosti regresijskih koeficijenata u ovisnosti o log $\boldsymbol{\lambda}$ za skup svih stanova 96

Slika 3-35 Rezultati *LASSO* regularizacije za skup svih stanova 96

Slika 3-36 Dijagram reziduala i prilagođenih vrijednosti za konačni regresijski model za skup svih stanova 99

Slika 3-37 Utjecaj stršećih vrijednosti na analizirani skup za regresijski model na stanove s ugrađenim RTTE 100

Slika 3-38 Korelacija između numeričkih varijabli u početnom regresijskom modelu za stanove koji su ugradili RTTE 102

Slika 3-39 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za početni regresijski model za stanove koji su ugradili RTTE 102

Slika 3-40 Ovisnost log($\boldsymbol{\lambda}$) i MSE za unakrsnu provjeru u 10 koraka za stanove sa ugrađenim RTTE 103

Slika 3-41 Promjena vrijednosti regresijskih koeficijenata u ovisnosti o log $\boldsymbol{\lambda}$ za stanove s RTTE 104

Slika 3-42 Rezultati *LASSO* regularizacije za stanove sa ugrađenim RTTE 104

Slika 3-43 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za konačni regresijski model za psStan, model 3EG 105

Slika 3-44 Utjecaj stršećih vrijednosti na analizirani podskup skup za regresijski model za stanove bez RTTE 106

Slika 3-45 Korelacija između numeričkih varijabli početnog regresijskog modela za stanove bez RTTE 108

Slika 3-46 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za početni regresijski model za stanove bez RTTE 108

Slika 3-47 Ovisnost log($\boldsymbol{\lambda}$) i MSE za unakrsnu provjeru u 10 koraka za stanove bez RTTE 109

Slika 3-48 Promjena vrijednosti regresijskih koeficijenata u ovisnosti o log $\boldsymbol{\lambda}$ za stanove bez RTTE 110

Slika 3-49 Rezultati LASSO regularizacije za stanove bez RTTE 110

Slika 3-50 Dijagram reziduala konačni regresijski model za stanove bez RTTE 111

Slika 3-51 Postupak provođenja *bagging* metode za regresijska stabla 113

Slika 3-52 Rezultati unakrsne provjere za učenje regresijskim stablima za sve stanove 114

Slika 3-53 Prikaz grananja regresijskog stabla prema varijablama i definirana pravokutna područja prema modelu obračuna energije i potrošnji na razini zgrade 115

Slika 3-54 Prikaz ovisnosti izbrojanih impulsa u stanu sa specifičnom potrošnjom (y-os) ... 115

Slika 3-55 Rezultati unakrsne provjere za učenje regresijskim stablima za stanove s RTTE 116

Slika 3-56 Prikaz grananja regresijskog stabla prema varijablama i definirana pravokutna područja prema modelu obračuna energije 117

Slika 3-57 Rezultati unakrsne provjere za učenje regresijskim stablima za stanove bez RTTE 118

Slika 3-58 Prikaz grananja regresijskog stabla prema varijablama specifične potrošnje na razini zgrade i korekcijskom faktoru 118

Slika 3-59 Ovisnost zavisne varijable o modelu za obračun energije i potrošnji zgrade 119

Slika 3-60 Ovisnost zavisne varijable o broju impulsa i potrošnji zgrade 119

Slika 3-61 Ovisnost zavisne varijable površini i potrošnji zgrade 120

Slika 3-62 Ovisnost zavisne varijable o temperaturi i potrošnji zgrade 120

Slika 3-63 Prikaz ovisnosti zavisne varijable o potrošnji zgrade i korekcijskom faktoru 121

Slika 3-64 Prikaz ovisnosti zavisne varijable o potrošnji i površini zgrade 121

Slika 3-65 Predviđanje na razini svih stanova. X= SpecTopOMM_SDG, Y=StopaUgr, Z=SpecTopSDG 123

Slika 3-66 Klasifikacija strojem potpornih vektora ovisnosti potrošnje o modelu obračuna energije i korekcijskom faktoru..... 123

Slika 3-67 Predviđanje na razini stanova s ugrađenim RTTE. X= SpecTopOMM_SDG, Y=StopaUgr, Z=SpecTopSDG..... 123

Slika 3-68 Predviđanje na razini stanova bez ugrađenih RTTE. X= SpecTopOMM_SDG, Y=StopaUgr, Z=SpecTopSDG..... 123

Slika 3-69 Dijagram rasipanja specifične godišnje potrošnje topline u zgradama prema ugrađenosti RTTE 124

Slika 3-70 Utjecaj ugradnje RTTE na potrošnju na razini zgrade za odabrani skup zgrada.. 125

Slika 3-71 Dijagram reziduala i normalni Q-Q dijagram za razinu zgrade..... 126

Slika 3-72 Predviđanje na razini zgrada za X=StopaUgr, Y=Temp, Z=SpecTopOMM_SDG 127

Slika 3-73 Predviđanje na razini zgrada za X=PovZgrada, Y=Temp, Z=SpecTopOMM_SDG 127

Slika 3-74 Predviđanje na razini zgrada za X=PovZgrada, Y=StopaUgr, Z=SpecTopOMM_SDG 127

Slika 3-75 Klasifikacijski prikaz stroja potpornih vektora koji pokazuje vezu između StopeUgr i SpecTopOMM_SDG..... 128

Slika 3-76 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za model na razini distribucijskog područja 130

Slika 3-77 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za model na nacionalnoj razini 130

Slika 4-1 Regresijsko stablo za razinu stanova za čitav podatkovni skup 137

Slika 4-2 Regresijsko stablo za razinu stanova s ugrađenim individualnim mjerenjem 138

Slika 4-3 Regresijsko stablo za razinu stanova bez ugrađenog individualnog mjerenja 139

Slika 4-4 Komparativni prikaz potrošnje pojedinih zgrada s obzirom na način alokacije topline 141

Slika 4-5 Komparativni prikaz potrošnje stanova s obzirom na način alokacije topline 141

Slika 4-6 Regresijsko stablo za razinu zgrade 142

Slika 4-7 Dijagram rasipanja i regresijski pravac za (a) najveće analizirano distribucijsko područje i (b) nacionalnu razinu 143

Slika 4-8 Utjecajne varijable za svaki podskup na razini stanova za metodu regresijski stabala 146

Slika 4-9 Utjecajne varijable za svaki podskup na razini stanova za metodu slučajnih šuma 146

Slika 4-10 Utjecajne varijable na razini zgrada za metodu regresijskih stabala i slučajnih šuma 148

Slika 0-1 Pravokutni dijagram i distribucija varijable (Izvor: By Jhguch at en.wikipedia, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=14524285) 166

POPIS TABLICA

Tablica 2-1 Kriteriji za odabih optimalnog modela 44

Tablica 3-1 Opis analiziranih kvantitativnih varijabli 62

Tablica 3-2 Opis analiziranih kvalitativnih varijabli 63

Tablica 3-3 Sumarna statistika za *psStan* 65

Tablica 3-4 Sumarna statistika za *psZgrada* 66

Tablica 3-5 Osnovni podaci o analiziranim zgradama 68

Tablica 3-6 Sumarna statistika izabranih varijabli za *psGrad* 84

Tablica 3-7 Sumarna statistika varijable *SpecTopSDG*, prema modelu obračuna 84

Tablica 3-8 Grupa 1 prema utjecaju stope ugradnje na promjenu potrošnje u zgradama 87

Tablica 3-9 Grupa 2 prema utjecaju stope ugradnje na promjenu potrošnje u zgradama 87

Tablica 3-10 Grupa 3 prema utjecaju stope ugradnje na promjenu potrošnje u zgradama 87

Tablica 3-11 Regresijski parametri i statistički pokazatelji početnog regresijskog modela za čitav podatkovni skup stanova 93

Tablica 3-12 Pokazatelji točnosti početnog regresijskog modela za skup svih stanova 93

Tablica 3-13 Regresijski parametri za regularizirani regresijski model svih stanova 97

Tablica 3-14 Regresijski parametri i statistički pokazatelji konačnog regresijskog modela za skup svih stanova 98

Tablica 3-15 Pokazatelji točnosti početnog regresijskog modela 98

Tablica 3-16 Regresijski parametri i statistički pokazatelji početnog regresijskog modela za stanove koji su ugradili RTTE 101

Tablica 3-17 Pokazatelji točnosti početnog regresijskog modela stanove s RTTE 101

Tablica 3-18 Regresijski parametri za regularizirani regresijski model stanova s RTTE 103

Tablica 3-19 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za stanove s RTTE nakon *LASSO* regularizacije 104

Tablica 3-20 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za stanove s RTTE nakon *LASSO* regularizacije nakon izbacivanja varijable *KorekcijskiFaktor* 105

Tablica 3-21 Pokazatelji točnosti finalnog regresijskog modela za stanove s RTTE 105

Tablica 3-22 Regresijski parametri i statistički pokazatelji za početni regresijski model za stanove bez RTTE 107

Tablica 3-23 Pokazatelji točnosti početnog regresijskog modela za stanove bez RTTE 107

Tablica 3-24 Regresijski parametri za regularizirani regresijski model stanova bez RTTE .. 109

Tablica 3-25 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za stanove bez RTTE 110

Tablica 3-26 Pokazatelji točnosti finalnog regresijskog modela za stanove bez RTTE 111

Tablica 3-27 Regresijski parametri i statistički pokazatelji regresijskog modela za zgrade .. 126

Tablica 3-28 Pokazatelji točnosti regresijskog modela z arazinu zgrade 126

Tablica 3-29 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za razinu distribucijskog područja 129

Tablica 3-30 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za nacionalnu 129

Tablica 4-1 Vrednovanje RMSE točnosti predviđanja razmatranih algoritama strojnog učenja na razini stana 144

Tablica 4-2 Vrednovanje RMSE točnosti predviđanja razmatranih algoritama strojnog učenja na razini zgrade 144

POPIS OZNAKA

Oznaka	Opis	Mjerna jedinica
Temp	Srednja vanjska temperatura	°C
Povrsina	Površina pojedinog krajnjeg kupca	m^2
Povrsina3EG	Suma površina svih stanova u zgradi na modelu 3EG [56]	m^2
Povrsina2EG	Suma površina svih stanova u zgradi na modelu 2EG [56]	m^2
ToplinaOMM	Ukupno izmjerena ili obračunata isporučena toplina za grijanje prostora na obračunskom mjernom mjestu	MWh
PTV_OMM	Ukupno izmjerena ili obračunata isporučena toplina za pripremu potrošne tople vode na obračunskom mjernom mjestu	MWh
EnergijaOMM	Ukupno izmjerena isporučena toplina na obračunskom mjernom mjestu	MWh
PROSJEK_TV	Količina topline za PTV utvrđena računskim prosjekom za toplu vodu	MWh
OSOS2EV	Ukupan broj članova kućanstva	-
BrojOsoba	Broj članova kućanstva kod krajnjeg kupca	-
ImpusliOMM	Ukupan broj impulsa izmjeren na obračunskom mjernom mjestu	-
ImpulsiKK	Broj impulsa izmjeren kod krajnjeg kupca	-
UR	Izabrani udio impulsa s razdjelnika u obračunskoj formuli [56]	-
UPOV	Izabrani udio površine pojedinog krajnjeg kupca u obračunskoj formuli [56]	-
SnagaKK	Instalirana toplinska snaga kod krajnjeg kupca	kW
ToplinaKK	Toplina za grijanje prostora krajnjeg kupca	kWh
PTV_KK	Toplina za pripremu potrošne tople vode kod krajnjeg kupca	kWh
EnegijaKK	Ukupna toplina alocirana krajnjem kupcu	kWh
ToplinaSpec	Specifična potrošnja topline za grijanje kod krajnjeg kupca	kWh/m^2
GrupaKK	Oznaka grupe krajnjeg kupca	-
ModelEnergija	Oznaka modela obračuna topline	2EG, 3EG
KorekcijskiFaktor	Korekcijski faktori kod alokacije	%
SpecTopSDG		kWh/m^2
SpecTopOMM		kWh/m^2
SpecTopOMM_SDG		kWh/m^2
Imp		

PovrZgrada		m^2
StopaUgr	Omjer između površine stanova koji su ugradili RTTE i površine svih stanova	-
UgradilaZgr	Oznaka je li zgrada ugradila RTTE	-
RTTE	Razdjelnik troškova toplinske energije	-
RMSE	Engl. *root mean square error*	kWh/m^2
MSE	Engl. *mean square error*	
SU	Strojno učenje	-
SDG	Stupanj dan grijanja	-
$\boldsymbol{\beta}$	Regresijski parametar	-
$\boldsymbol{\lambda}$	Parametar za regularizaciju	-
IQR	Interkvartilni raspon	-
Q1	Donji kvartil	-
Q3	Gornji kvartil	-

1. UVOD

Direktiva o energetskoj učinkovitosti [1] prenesena je u hrvatsko zakonodavstvo Zakonom o tržištu toplinske energije [2] i Zakonom o energetskoj učinkovitosti [3]. Zakonom o tržištu toplinske energije te relevantnim podzakonskim aktima propisano je da se u sve stambene/poslovne prostore spojene na zajedničko mjerilo toplinske energije u određenim rokovima moraju ugraditi elektronički razdjelnici troškova toplinske energije (u nastavku teksta skraćeno RTTE) ili mjerila toplinske energije i termostatski radijatorski ventili. RTTE prema normi HRN EN 834:2014 definira se kao virtualni osjetnik za evidentiranje razmjernog dijela toplinske energije odane preko ogrjevnih tijela unutar stanova [4].

Temeljni princip odnosa krajnjeg kupca i tvrtke koja naplaćuje energiju je da svaki kupac samostalno odlučuje o načinu korištenja usluge iz umreženih sustava. To je u prošlosti bio standard za električnu energiju i plin, dok je za toplinsku energiju i vodu takav standard postojao samo u privatnim samostojećim kućama i u novijim višestambenim zgradama. U starijim višestambenim zgradama, izgrađenim prije 1995. godine, obračun se uglavnom temeljio na principu solidarnosti, te je i tehnička izvedba sustava izvedena po istom principu kako bi se minimizirali troškovi.

Za korištenje usluge grijanja iz centralnog toplinskog sustava ili iz zajedničke kotlovnice u zgradama koje su većinom izgrađene do 1995. godine, koje imaju vertikale na koje se spajaju radijatori na svim katovima te nemaju jedan ulazni/izlazni priključak za svaki stambeni/poslovni prostor, nije moguće osigurati izravno mjerenje potrošnje topline. Rješenja za mjerenje potrošnje toplinske energije putem RTTE (tehnički aspekt opremanja) i matematičke formule samo su nadomjesci pravog i nedvojbenog mjerenja, ali neusporedivo bolji i vjerodostojniji način raspodjele troškova od izračuna putem omjera površina.

Novim zakonodavstvom Hrvatska se opredijelila da svaki stambeni prostor, neovisno o tome radi li se o višekatnici ili obiteljskoj kući, ima li vlastiti pristup svim umreženim uslugama i vlastito mjerenje potrošnje. Za učinkovito korištenje usluga nužno je da vlasnik stambenog prostora može odlučivati o načinu korištenja, da se njegova potrošnja može izmjeriti te da se plaća prema potrošnji..

Zakon o gradnji [5], određuje da svaka nova zgrada, ovisno o vrsti i namjeni, mora biti projektirana, izgrađena i održavana tako da može zadovoljiti propise energetske učinkovitosti

te da se bez znatnih troškova može osigurati individualno mjerenje potrošnje toplinske energije, energenata i vode s daljinskim očitanjem za pojedine posebne dijelove zgrade. Isti zakon preporučuje da se i u zgradama koje su izgrađene prije stupanja ovog zakona na snagu provedu nužne mjere i radovi kako bi se osiguralo individualno mjerenje potrošnje svih energenata i vode.

Direktiva o energetskoj učinkovitosti u točkama (28) i (29) preambule i u članku 9. navodi da je uporaba pojedinačnih mjerila ili RTTE za mjerenje individualne potrošnje topline u zgradama s više stanova spojenim na centralni toplinski sustav korisna kada krajnji kupci mogu nadzirati vlastitu potrošnju. Prvi korak ka nadziranju vlastite potrošnje ugradnja je termostatskih ventila na svako ogrjevno tijelo (radijator). Također, u Direktivi se konstatira da se u slučajevima kada je ugradnja individualnih mjerila toplinske energije skupa ili tehnički složena ugrađuje RTTE.

Brojne studije i analize potvrdile su da se ugradnjom RTTE na razini toplinske stanice (tj. cijele zgrade) mogu očekivati godišnje uštede topline na razini od 15 do 30 %. Ta ostvarena ušteda, tj. smanjena potrošnja topline, dalje se raspodjeljuje po svakom pojedinom potrošaču određenom metodologijom. Osim same ugradnje mjerenja/raspodjele individualne potrošnje, u metodu obračuna nužno je uvesti i različite kriterije kako bi raspodjela potrošnje i ostvarenih ušteda bila tehnički i troškovno korektna.

Cilj ovog rada je odrediti stvarni učinak ugradnje uređaja za individualno mjerenje potrošnje toplinske energije u sustavima sa zajedničkim izvorom topline ili s priključkom na sustav daljinskog grijanja, kao i utvrditi koji su utjecajni parametri na potrošnju toplinske energije u sustavima daljinskog grijanja. Rezultati istraživanja dalje se mogu koristiti za predviđanja potrošnje i ocjenu učinka provedbe mjera energetske učinkovitosti te modifikaciju postojećeg sustava obračuna u svrhu postizanja sustava koji će vrednovati provedene mjere energetske učinkovitosti i pravilnije alocirati potrošenu toplinu kod krajnjih potrošača.

1.1. Područje istraživanja

Sektor zgradarstva u Europskoj uniji predstavlja potrošnju od oko 40 % ukupne konačne energije u EU-u, pri čemu u tom sektoru oko 80 % energije otpada na grijanje i pripremu potrošne tople vode (ostatak čine hlađenje s 0,6 % i električna energija s 19,40 %)[1]. Prvi veći poticaj za povećanje energetske učinkovitosti i smanjenje potrošnje energije dogodio se nakon potpisivanja Protokola u Kyotu, što je potaknulo razne pristupe i metodologije povećanja

energetske učinkovitosti u zgradama, posebice stambenim [7]. U dosadašnjim istraživanjima razvijeni su brojni modeli za ocjenu učinka primjene mjera energetske učinkovitosti i predviđanja potrošnje u sektoru zgradarstva, bazirani na tradicionalnim regresijskim metodama [8], metodama koje koriste neuronske mreže [9] te različitim simulacijskim metodama [10]. Međunarodna energetska agencija (engl. *International Energy Agency* – IEA) prepoznaje dva osnovna pristupa ocjene potrošnje i uštede energije, a to su „*top-down*" i „*bottom-up*" pristupi za čije je proračune IEA dao i detaljne smjernice [11]. Pristup „*bottom-up*" naglasak stavlja na pojedinačne tehnologije pružanja energetskih usluga u kućanstvima, sektoru usluga, industriji i poljoprivredi. Za takve tehnologije procjenjuju se troškovi i koristi povezane s ulaganjima u povećanu energetsku učinkovitost, često u kontekstu smanjenja emisije stakleničkih plinova ili drugih utjecaja na okoliš. Metoda „*top-down*" pretpostavlja opću ravnotežu ili makroekonomsku perspektivu, gdje su troškovi definirani u smislu gubitaka od ekonomske proizvodnje, prihoda ili bruto domaćeg proizvoda (BDP), obično od uvođenja poreza na energiju ili emisiju [12]. U EU-u, model „*bottom-up*" predominantno se koristi pri određivanju poticajnih zakonodavnih okvira za postizanje energetske učinkovitosti, ali njegov glavni nedostatak leži u činjenici da su takvi modeli neadekvatni za opis netehničkih utjecajnih parametara te uvode veći broj modelskih pretpostavki vezanih za bihevioralne aspekte potrošnje energije kao što su demografski faktori, dob potrošača, dnevni raspored korištenja prostora, spremnost potrošača da plati i sl. [13].

U posljednje vrijeme istraživači sve veći fokus stavljaju na spomenute brojne netehničke faktore koji utječu na potrošnju energije. Tako Yang i suradnici razmatraju ponašanje korisnika prostora i razinu toplinske ugodnosti [14]. Nadalje, Nguyen i suradnici analiziraju inteligentne sustave praćenja načina korištenja i kontrole potrošnje energije u zgradama [15]. S povećanjem raspoloživih podataka o netehničkim vidovima potrošnje, kao i razvojem tehnologija za prikupljanje tih podataka, relativno novo područje analize velikih količina podataka (engl. *Big Data Analyses*) ima potencijal omogućiti bolje razumijevanje i modeliranje potrošnje energije na temelju niza netehničkih faktora. U analizi velikih količina podataka primjenjuju se metode statističkog strojnog učenja, kao Olofsson i suradnici, uz uporabu osnovne koordinatne analize i regresijske metode parcijalnih najmanjih kvadrata za određivanje glavnih utjecajnih faktora na potrošnju energije u sustavima daljinskog grijanja, potrošnje električne energije, potrošnje pitke vode i gubitaka topline [16]. Algoritmi strojnog učenja, predominantno stroj potpornih vektora, pokazuju se prikladnim za ocjenu potrošnje energije u zgradarstvu [17].

Promjenom energetske politike u EU-u sustavi daljinskog grijanja identificirani su kao jedan od glavnih instrumenata za postizanje ciljeva povećanja energetske učinkovitosti i povećanja udjela obnovljivih izvora energije u energetskom miksu [18][19]– čime se nameće potreba za preciznim metodama procjene potrošnje i ušteda energije u tim sustavima. U veljači 2016. Europska komisija izdaje prvu strategiju za optimiziranje grijanja i hlađenja u zemljama članicama, čime je prepoznata važnost ovih sustava u budućnosti [20]. Krajem 2012. Europski parlament i Vijeće Europske unije donose Direktivu 2012/27/EU o energetskoj učinkovitosti kojoj je primarni cilj ostvarivanje energetskih ušteda u neposrednoj potrošnji od 20 % do 2020. godine i otvaranje novih mogućnosti za poboljšanje energetske učinkovitosti nakon tog razdoblja te je u članku 9. Direktive uvođenje individualne raspodjele potrošnje toplinske energije u sustavima daljinskog grijanja identificirano kao osnovni preduvjet za postizanje energetskih ušteda promjenom ponašanja korisnika [21]. U usporedbi s potrošnjom energije u svim vrstama zgrada, dodatni utjecajni faktori na zgrade u sustavima daljinskog grijanja su postojanje ili nepostojanje individualnog mjerenja potrošnje, postojanje pripreme potrošne tople vode, mogućnost zamjene izvora primarne energije, iznos i upravljanje toplinskim gubicima u distribucijskom sustavu te valorizacija pasivnih toplinskih dobitaka u višestambenim zgradama [22]. RTTE je mjerni uređaj koji mjeri površinsku temperaturu radijatora i temperaturu zraka u prostoru, a njegova prednost je u tome što nisu postavljeni preveliki zahtjevi prilikom njegove ugradnje [23].

Ugradnjom RTTE očekuje se smanjenje potrošnje topline zbog same činjenice da krajnji korisnici nakon ugradnje imaju mogućnost utjecati na svoju potrošnju [24]. Korisnici koji ne mogu utjecati na svoje račune za toplinu individualnim mjerenjem potrošnju temelje na željenoj razini toplinske ugodnosti te provjetravaju prostore bez ograničenja, dok oni potrošači koji mogu utjecati na svoje račune uglavnom to i čine nauštrb toplinske ugodnosti i razine provjetravanja [25].

U Poljskoj [26] je u sedamnaestogodišnjem razdoblju provedeno eksperimentalno istraživanje procjene učinka ugradnje individualnog mjerenja potrošnje toplinske energije primjenom RTTE. Potrošnja se mjerila u dvije identične skupine stanova (jednaki ulazi u istoj zgradi) tako da su u jednoj skupini ugrađeni razdjelnici troškova topline, dok u drugoj nisu. U skupini s ugrađenim razdjelnicima izmjerene su godišnje uštede na razini od približno 27 %. Ovako procijenjena ušteda ne može se jednoznačno pripisati samo utjecaju ugradnje razdjelnika troškova toplinske energije budući da na samu potrošnju energije utječu brojni faktori. Imajući

u vidu sve širu primjenu i dobre rezultate analize predviđanja potrošnje energije primjenom određenih algoritama strojnog učenja, kao što je npr. stroj potpornih vektora [27], predlaže se analiza učinka mjera energetske učinkovitosti na smanjenje potrošnje energije primjenom metoda strojnog učenja, s usporednom analizom i identifikacijom najprikladnije metode za procjenu učinka mjera energetske učinkovitosti i planiranja potrošnje toplinske energije u zgradarstvu u sustavima daljinskog grijanja. Kod predviđanja potrošnje električne energije metode strojnog učenja do sada su pokazale bolje karakteristike nego klasični regresijski modeli [28]. Također, hibridne metode strojnog učenja pokazale su robusnost pri predviđanju ukupnih potreba za energijom na nacionalnoj razini s odstupanjem od 2 do 4 % [29].

1.1.1. Sustavi daljinskog grijanja

Sustavi daljinskog grijanja ili centralni toplinski sustavi, kako se definiraju u Zakonu o tržištu toplinske energije [2], predstavljaju sustave koji se sastoje od proizvodnog postrojenja toplinske energije i distribucijske mreže te putem kojeg se odvija opskrba toplinskom energijom i prijenos toplinske energije od toplinskog izvora prema krajnjim kupcima.

Termodinamički gledano, sustavi daljinskog grijanja mogu se opisati na sljedeći način: sekundarnu uporabu viška topline iz nekog energetskog sustava u sustavu daljinskog grijanja omogućuje drugi glavni stavak termodinamike. Veliki dio eksergije goriva iskoristi se u primarnoj industrijskoj proizvodnji ili prilikom proizvodnje električne energije u kogeneracijskim postrojenjima. Spajanjem na centralizirani toplinski sustav preostala eksergija i sva ostatna energija iskorištava se za potrebe grijanja potrošača. Na taj način, sadržaj eksergije za grijanje prilagođen je temperaturnom režimu koji zadovoljava potrebe potrošača. Kada se energija oslobođena izgaranjem goriva odmah koristi za grijanje na niskim temperaturnim režimima, velik dio eksergije nepovratno se gubi jer se toplina nije koristila i na višim temperaturnim režimima. Energetski održiv sustav mora maksimalno iskoristiti eksergiju sadržanu u gorivu kako bi se povećala iskoristivost i što je više moguće smanjio utjecaj na okoliš. Većina sustava daljinskog grijanja funkcionira prema navedenom principu.

Prvi sustavi daljinskog grijanja spominju se još u 14. stoljeću i prvim sustavom distribucije tople vode za grijanje kojom se opskrbljuje više potrošača smatra se sustav geotermalnog daljinskog grijanja 30 privatnih kuća u mjestu Chaudes-Aigues u Francuskoj [30]. Ako se pogleda još dalje u prošlost, također je poznato da su se toplom geotermalnom vodom grijali i

stari Rimljani, bilo da su daljinsko grijanje koristili za balneološke svrhe, bilo da su toplinsku energiju koristili za uzgoj poljoprivrednih kultura u staklenicima.

Počeci modernog korištenja daljinskog grijanja bilježe se 1877. godine u Sjedinjenim Američkim Državama, kada Birdsill Holly, inženjer strojarstva i inovator, projektira, gradi i uspješno stavlja u pogon prvi komercijalni sustav daljinskog grijanja u Lockportu u saveznoj državi New York. Nekoliko godina poslije susrećemo prve sustave daljinskog grijanja u Europi, iako europski sustavi daljinskog grijanja s početka stoljeća nisu bili striktno komercijalne namjene. Prvi veći sustav izgrađen je 1900. godine u Dresdenu, dok su manji sustavi koji su se u pravilu sastojali od kotlovnice iz koje je grijano više zgrada, bili građeni i prije. Tako se bilježe sljedeći sustavi daljinskog grijanja kao prvi na području Europe [31]:

- 1878. na području bolnice Sabbatsberg u Stockholmu,
- 1884. na Tehničkom sveučilištu u Berlinu,
- 1893. Gradska vijećnica u Hamburgu spojena na energetsko postrojenje Poststrasse te se koristila otpadnom toplinom iz proizvodnje električne energije,
- 1896. Lord Kelvin gradi prvi londonski sustav kojim je spojio spalionicu otpada Shoreditch s krajnjim potrošačima,
- 1898. bolnica Sahlgrenska u Göteborgu gradi sustav daljinskog grijanja kao druga takva bolnica u Švedskoj,
- 1903. prvi sustav daljinskog grijanja u Danskoj izgrađen je za potrebe bolnice u Fredriksbergu, dijelu Kopenhagena.

U današnje vrijeme cilj centraliziranih toplinskih sustava je upotreba lokalno nastalog goriva ili topline (koji bi inače ostali neiskorišteni) za zadovoljavanje lokalnih toplinskih potreba. Toplina se transportira distribucijskom mrežom koja spaja proizvođače i potrošače, pritom tvoreći tržište toplinske energije.

Gledajući od početka razvoj tehnologija primjenjivanih u sustavima daljinskog grijanja, do današnjeg dana prepoznaju se četiri tehnološke generacije [32], s naznakama razvoja pete u današnjim najmodernijim sustavima.

Prema [32], tehnološki se pojedine generacije klasificiraju kako slijedi:

1. Prva generacija („1DH“) predstavlja sustave koji su prvi put razvijeni 1880.-ih godina u SAD-u, a koji kao sekundarni medij koriste paru koja se proizvodi u termoenergetskim postrojenjima koja su bila, ili jesu, isključivo na ugljen. Sustavi s ovom tehnologijom razvijani su do 1930.-ih godina, bili su vrlo neučinkoviti a distribucijske mreže gradile su se polaganjem betonskih kanala i cijevi, bez toplinske izolacije i s vrlo velikim energetskim gubicima.
2. Druga generacija („2DH“) sustava daljinskih grijanja predstavlja sustave koji su se gradili od 1930-ih do 1970-ih. Toplinska energija i dalje se dominantno proizvodi u termoenergetskim postrojenjima, s time da osim ugljena tada susrećemo i termoelektrane na naftu i prirodni plin. Sustavi su se razvijali uz termoenergetska postrojenja u većim i gušće naseljenim gradovima te su se koristili otpadnom energijom iz kogeneracijske proizvodnje te su nerijetko bili rashladni sustav za ona energetska postrojenja koja nisu imala neku drugu mogućnost hlađenja (kao što su rijeke, jezera i sl.). Sekundarni medij više nije para već vrela voda. Temperature polaza u distribucijskom vrelovodu više su od 100°C, distribucija se provodila vodovodnim, neizoliranim cijevima, položenima u betonske kanale. Ovakvi sustavi intenzivno su se gradili u drugoj polovici 20. stoljeća u zemljama bivšeg Sovjetskog Saveza, te srednje i istočne Europe, među kojima je i Hrvatska.
3. Treća generacija („3DH“), koja se često naziva i „skandinavska tehnologija daljinskog grijanja jer je većina proizvođača opreme za 3DH sustave nalazi u skandinavskim zemljama, predstavlja sustave koji se grade od 1970-ih godina do danas. Sustavi koriste predizolirane i predfabricirane cijevi koje su obično direktno položene u tlo te su distribucijske mreže na pogonskim temperaturama nižim od 100 °C. Glavni motiv razvoja ovakvih sustava nastaje nakon naftne krize 1970.-tih godina, kada se u želji za povećanom sigurnošću energetske opskrbe stvaraju začeci globalnih nastojanja za povećanje energetske učinkovitosti u svim vidovima potrošnje. Također, u ovim sustavima dolazi do diverzifikacije primarnih izvora na strani proizvodnje te do tranzicije s fosilnih goriva na obnovljive izvore energije kao što su biomasa, geotermalna i solarna energija.
4. Četvrta generacija („4DH“) predstavlja moderne, niskotemperaturne, integrirane i pametne sustave, koji moraju imati ove karakteristike [33] kako bi se kao takvi klasificirali:

a. niskotemperaturne pogonske temperature (manje od 70 °C), dobavu energije kroz distribucijske sustave niskotemperaturnog grijanja za nove niskoenergetske zgrade i energetski obnovljene postojeće zgrade,
b. distribuciju toplinske energije uz male gubitke,
c. prihvat otpadne niskotemperaturne topline iz različitih, većinom industrijskih ali i drugih, procesa i integraciju obnovljivih izvora topline kao što su geotermalna i solarna energija, energija iz biomase te spalionica otpada,
d. sposobnost integracije u pametne energetske sustave (engl. *smart energy systems*), što uključuje i sustave daljinskog hlađenja,
e. predstavljati integralni dio šireg sustava održivih energetskih, prostornih i gospodarskih razvojnih planova.

U posljednje vrijeme sve se češće spominju i sustavi pete generacije („5DH“) koji imaju pogonske parametre u distribucijskom sustavu na razini temperature tla, čime se minimiziraju gubici u distribuciji. Krajnji potrošači priključeni na ovakve distribucijske sustave, umjesto toplinske stanice s izmjenjivačem topline imaju energetsko postrojenje s dizalicom topline koju kojom se i za grijanje i za hlađenje. Sustav na taj način može istovremeno raditi u oba pogonska načina, i grijanja i hlađenja.

1.1.1.1. Prednosti i nedostaci sustava daljinskog grijanja

Današnji znanstveni pogled na sustave daljinskog grijanja je takav da su to lokalni energetski sustavi koji se koriste lokalno dostupnim gorivom ili otpadnom toplinom koja bi inače bila neiskorištena, s ciljem zadovoljavanja toplinskih potreba krajnjih potrošača na tom istom prostoru [33]. Dodatno, kako se navodi u [32], sustavi daljinskog grijanja „4DH“ generacije omogućuju veće udjele obnovljivih izvora energije u elektroenergetskom sustavu s obzirom na to da kao takvi mogu služiti (i) za uravnoteženje sustava te predstavljaju i (ii) skladište proizvedene energije iz obnovljivih izvora u periodima kada je niža potražnja za tom energijom preko električnih kotlova i spremnika (akumulatora) toplinske energije.

Na trenutnoj tehnološkoj razini prepoznaje se pet pogodnih i ekonomski isplativih izvora topline za sustave daljinskog grijanja, a to su:

- otpadna toplina koja nastaje prilikom proizvodnje električne energije, odnosno u kogeneracijskim procesima, bez obzira na primarno gorivo,

- otpadna toplina dobivena u spalionicama otpada,
- otpadna toplina iz industrijskih procesa,
- toplina iz goriva, poput biomase, posebno u višestambenim zgradama i u naseljenim područjima,
- toplina iz obnovljivih izvora energije.

S povećanjem udjela obnovljivih izvora energije u ukupnom energetskom miksu na strani proizvodnje električne energije u budućnosti se očekuje povećanje udjela dizalica topline velikih kapaciteta. Povećanje udjela obnovljivih izvora energije i udjela dizalica topline očituje se u dva bitna elementa. Prvi element je taj da je cilj da korištena energija bude iz obnovljivih izvora kako se ne bi koristila električna energija proizvedena u postrojenjima sa znatnim štetnim emisijama u okoliš. Drugi element je taj da su s povećanjem udjela intermitentnih obnovljivih izvora energije dizalice topline dobar sustav koji može služiti kao kontinuirani potrošač električne energije sa sposobnošću uravnoteženja elektroenergetske mreže.

Kao i svi tehnički sustavi, i sustavi daljinskog grijanja imaju svoje prednosti i nedostatke [34]. Glavne prednosti korištenja centraliziranog toplinskog sustava za potrošača su:

- jednostavna, kontinuirana i pouzdana isporuka topline te visoka razina toplinske ugodnosti na strani krajnjih potrošača,
- manja potreba za zauzimanjem korisnog stambenog prostora i manji investicijski troškovi u usporedbi s individualnim sustavom grijanja,
- manja opasnost od požara i eksplozije jer se izgaranje ne događa u stanu ili kući,
- mogućnost promjene instaliranog toplinskog kapaciteta prema stvarnim potrebama koji se mogu mijenjati za pojedinu zgradu,
- upotreba lokalnog i obnovljivog izvora topline bez potrebe za dodatnim ulaganjima potrošača,
- sustavi daljinskog grijanja imaju mali utjecaj na okoliš jer ispuštaju neznatne količine NO_2 i čestica te uzrokuju manje zvučno onečišćenje okoliša u usporedbi sa sustavima koji se koriste fosilnim gorivima,
- povećana sigurnost opskrbe toplinom zbog fleksibilnosti primarnog energenta (u sustavima četvrte generacije, manji dio energije osigurava se iz fosilnih goriva),
- sustavi daljinskog grijanja imaju visok stupanj fleksibilnosti u smislu da pružatelj usluge grijanja može trenutno promijeniti izvor iz kojeg će se dobavljati toplina, ako bi se na

tržištu promijenila cijena pojedinog izvora topline. Na cijenu energenata utječu promjene cijene nafte i električne energije, ali i bijeli i zeleni certifikati, zajamčena otkupna cijena električne energije (engl. *feed-in-tariff*), visina poreza na CO_2 i energiju, dopuštene kvote ispuštanja CO_2 itd. Visok stupanj fleksibilnosti čini ovakve sustave robusnima, ali omogućuje regulaciju cijene, što pruža realne i manje tržišno osjetljive cijene toplinske energije. Zasebni sustavi ne mogu pružati ovakav stupanj fleksibilnosti jer su različiti izvori topline najčešće dostupni samo u sustavima daljinskog grijanja.

Nedostaci centraliziranog toplinskog sustava:

- potrošači nisu u mogućnosti individualno pregovarati o cijenama i uvjetima isporuke toplinske energije u prirodnim monopolima.
- rizik od nestajanja prednosti koje pruža prirodni monopol zbog smanjenja efikasnosti sustava uzrokovanog slabim razvojem. Razlog slabog razvoja sustava najčešće je nepostojanje tržišne konkurencije.
- visoki troškovi i druge teškoće prilikom izlaska potrošača iz sustava, što može uzrokovati njegovo „zarobljavanje“ u sustavu.
- prekidi i teškoće u radu sustava utječu istovremeno na sve korisnike na nekom području.

1.1.1.2. Sustavi daljinskog grijanja u Republici Hrvatskoj

Korištenje toplinske energije u sustavima daljinskog grijanja prepoznato je kao bitan faktor europske energetske politike učinkovitog korištenja energije [35]. Sustavi daljinskog grijanja smatraju se bitnim elementom za povećanje energetske učinkovitosti na nacionalnoj razini, a promicanje izgradnje i razvoja sustava daljinskog grijanja, visokoučinkovite kogeneracije i korištenje obnovljivih izvora energije kao izvora toplinske energije proglašeno je interesom Republike Hrvatske [2].

Prema zadnjem godišnjem izvještaju Hrvatske energetske regulatorne agencije [36], energetskim djelatnostima proizvodnje i distribucije te opskrbe toplinske energije tarifnih kupaca u 2018. godini bavilo se 11 tvrtki u 17 gradova u Republici Hrvatskoj, koje su pružale usluge grijanja prostora i pripreme potrošne tople vode za više od 155 850 kupaca toplinske energije u većim kontinentalnim gradovima i Rijeci, pri čemu više od 95 posto ukupnog broja

kupaca pripada kategoriji kućanstva. Toplinska se energija proizvodi u kogeneracijskim postrojenjima u gradovima Zagrebu, Osijeku i Sisku ili u toplanama, blokovskim i kućnim kotlovnicama za pojedina naselja, te se vrelovodima / toplovodima / parovodima ukupne duljine oko 425 km distribuira do zgrada u kojima se u toplinskim stanicama predaje krajnjim korisnicima. U gradovima Zagrebu, Osijeku i Sisku proizvodi se i isporučuje i tehnološka para za potrebe industrije, a dijelom i za potrebe grijanja prostora. U 2018. godini u Republici Hrvatskoj isporučeno je oko 2 TWh toplinske energije.

Sustavi daljinskog grijanja po gradovima i njihova relativna veličina, prema publikaciji Energija u Hrvatskoj – godišnji energetski pregled [37], prikazani su na slici 1-1.



Slika 1-1 Sustavi daljinskog grijanja u Republici Hrvatskoj (Izvor: [37])

1.1.1.3. Toplinsko opterećenje

Zbog višeg toplinskog opterećenja smatra se ekonomski isplativijim spajati višestambene zgrade na sustav jer one iziskuju manje investicijske troškove izgradnje distribucijske mreže za razliku od zasebnih kućanstava. Tako se za sustave treće generacije smatra ekonomski isplativim spajati područja s toplinskim opterećenjem većim od 120 TJ/km^2, dok je za niskotemperaturne sustave četvrte generacije ta granica 20-48 TJ/km^2 [38].

Toplinsko opterećenje vremenski je promjenjiva veličina. Ono je prikazano kao prosječni toplinski tok u određenom vremenskom razdoblju. Najčešće se razmatraju prosječno satno i prosječno dnevno toplinsko opterećenje.

Osim o fizikalnim karakteristikama zgrade, tj. projektnom toplinskom opterećenju, potrošnja topline unutar zgrade ovisi o nizu parametara, a mogu se kategorizirati kako slijedi [39]:

- fizikalne karakteristike zgrade,
- čimbenici koji utječu na potrošnju, a to su:
    - način korištenja sustava grijanja,
    - način i tip provjetravanja,
    - korištenje uređaja i unutarnji toplinski dobici,
    - način korištenja prostora, tj. dnevni režim boravka te
- bihevioralni parametri, kao što su:
    - način ponašanja stanara, kao što su željena razina temperature u prostoru, učestalost prozračivanja i sl.,
    - način života, učestalost izvankućnih aktivnosti i sl. te
    - opća percepcija stanara vezano za uštedu energije.

Pri analizama i simulacijama toplinskih opterećenja u zgradama, uobičajeno su dostupni, i nužni, podaci o fizikalnim karakteristikama zgrada, dok informacije o drugim čimbenicima koji utječu na potrošnju i bihevioralnim parametrima obično nisu dostupne [40]. Monicair, nizozemski istraživački projekt kojem je cilj razvoj modela za točno predviđanje potrošnje u zgradama, pokazao je da rezultati modela koji se uobičajeno koriste za predviđanje potrošnje u zgradama u stvarnosti znatno odstupaju od izmjerenih vrijednosti, što se u jednom dijelu pripisuje pogrešnim pretpostavkama fizikalnih karakteristika zgrada, a u drugom dijelu neuključivanju utjecaja bihevioralnih parametara u modele [41].

Utjecajni bihevioralni parametri

Kao što je već spomenuto, potrošnja toplinske energije za grijanje prostora ovisi o ponašanju korisnika. Pojedinačno mjerenje omogućava svakom stanu plaćanje vlastitog korištenja grijanja i potrošne tople vode, ali je važno dati odgovarajuće informacije korisnicima o pojedinačnim mjerenjima koja mogu povećati svijest korisnika o troškovima i, u većini slučajeva, rezultirati uštedama. Naime, individualno mjerenje povećava svijest korisnika o energetskim pitanjima i potiče stanovnike da promijene svoje navike kako bi smanjili potrošnju energije [42]. Brojna su istraživanja pokazala da se u sličnim zgradama potrošnja energije može uvelike razlikovati zbog osobina stanara [43]. Promjenom načina ponašanja stanara mogu se postići znatne uštede energije – one iznose i do 22,9 % za pojedinačne mjere i do 41,0 % za integrirane mjere [44]. Prilikom prethodnih analiza potrošnje energije u zgradama utvrđeno je da se 4,2 % varijabilnosti potrošnje svih oblika energije u kućanstvima, a 20 % varijabilnosti u potrošnji energije za grijanje objašnjava bihevioralnim karakteristikama stanara [45].

S obzirom na to da ponašanje korisnika ima ključnu ulogu u postizanju ušteda, nije dovoljno samo nametnuti individualno mjerenje troškova grijanja kao jednu od mjera povećanja energetske učinkovitosti, jer to može izazvati nezadovoljstvo, kao što je slučaj u Hrvatskoj, već je potrebno osmisliti sustav koji će učinkovito funkcionirati i dodatno informirati krajnje korisnike na koji način mogu promijeniti svoje ponašanje kako bi postigli što veće uštede [46].

Vrlo važan bihevioralni parametar koji utječe na potrošnju topline u prvom je redu samo postojanje mogućnosti da krajnji kupac utječe na svoju potrošnju sustavom individualnog mjerenja [47], tj. nekim oblikom mjerenja (npr. razdjelnici toplinske energije) i postojanjem termostatskih ventila [48].

Željena razina toplinske ugodnosti također se smatra vrlo važnim bihevioralnim parametrom [49], čiji je utjecaj dokazan modeliranjem u simulacijskom programu EnergyPlus [50]. Zatim slijedi parametar koji opisuje način i učestalost otvaranja prozora, što uzrokuje ventilacijske gubitke i direktno utječe na toplinsko opterećenje [51].

U konačnici, Delzende i suradnici izradili su pregled dosadašnjih istraživanja utjecaja bihevioralnih parametara na potrošnju energije u zgradarstvu, prema dostupnim časopisima u Science Direct i Scopus bazi, sa zaključkom da postoji alarmantna razlika između projektno predviđene i stvarne potrošnje energije u zgradama (ponekad i do 300 %), što se pripisuje

nedovoljnom pridavanju važnosti bihevioralnim utjecajnim parametrima i nepostojanju adekvatne metode kojim bi se oni uključili u modeliranje i predviđanje potrošnje energije u zgradarstvu [52].

1.1.2. Razdjelnici troškova toplinske energije

Prema normi HRN EN 834:2014 razdjelnik troškova toplinske energije, tj. virtualni osjetnik topline definira se kao uređaj za evidentiranje razmjernog dijela toplinske energije odane preko ogrjevnih tijela unutar pojedinih stanova. Njihova ugradnja korak je prema individualizaciji potrošnje energije, što je dobar put prema racionalizaciji potrošnje topline te prema poimanju važnosti uštede energije kod krajnjih korisnika. Iako, razdjelnik troškova toplinske enegije ne predstavlja sredstvo direktne uštede, ali mjerenje individualne potrošnje dokazano utječe na ponašanje korisnika navodeći ih na promjene navika, što u konačnici rezultira smanjenom potrošnjom energije [53]. Razdjelnici troškova toplinske energije, prikazani slikom 1-2, ne mjere isporučenu toplinsku energiju, već odbrojavaju elektroničke impulse. Broj impulsa u obračunskom razdoblju s pripadajućih uređaja koristi se za određivanje udjela pojedinog grijanog prostora u ukupno isporučenoj toplinskoj energiji zgradi.

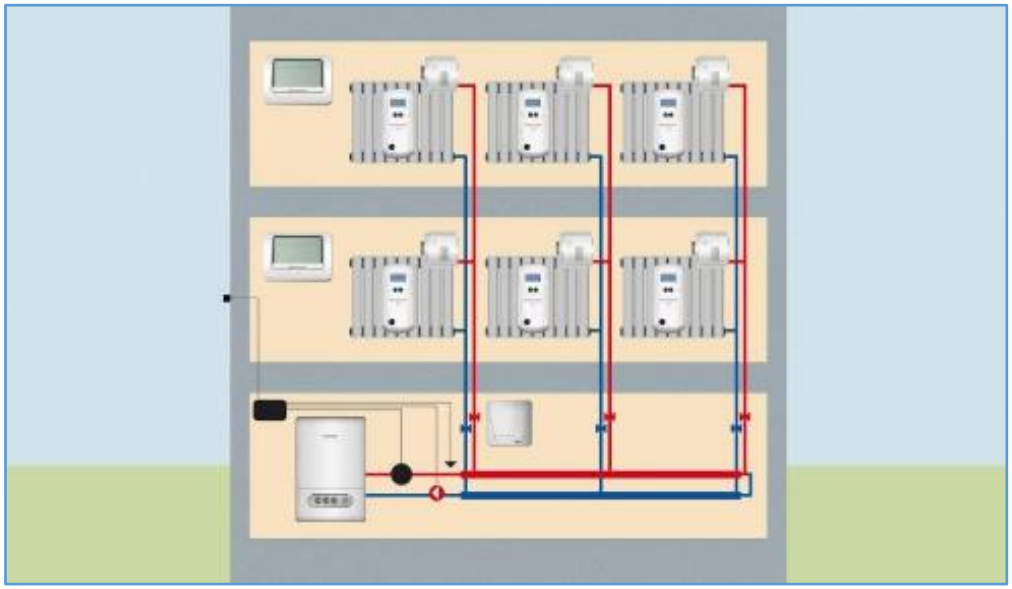


Slika 1-2 Razdjelnik troškova toplinske energije na radijatoru

Zato je bitno da sva ogrjevna tijela, tj. radijatori, u jednoj višestambenoj zgradi imaju ugrađene ove uređaje. Ipak, uređaji za lokalnu razdiobu topline mjerni su uređaji koji mjere temperaturu površine radijatora i temperaturu zraka u prostoru [54]. Brzina odbrojavanja impulsa povezana je matematičkim algoritmom s izmjerenom razlikom temperature površine i zraka. Algoritam se razlikuje od proizvođača do proizvođača uređaja, zbog čega je nužno da se na svim

ogrjevnim tijelima u jednoj zgradi sa zajedničkim obračunskim mjerilom toplinske energije ugrade uređaji istog tipa/proizvođača kako bi se osigurala konzistentna metoda očitanja impulsa u svim grijanim prostorima.

Svaki ugrađeni uređaj za lokalnu razdiobu topline na ogrjevnom tijelu (radijatoru) u zgradi mora se programirati za ispravan rad [55], a načelni način ugradnje prikazan je na slici 1-3.



Slika 1-3 Shema ugradnje razdjelnika troškova toplinske energije unutar zgrade (Izvor: Honeywell)

Za pravilno funkcioniranje RTTE potrebno je provesti određene tehničke mjere koje se odnose na toplinske podstanice, cijevni razvod, ogrjevna tijela i druge elemente sustava. Bitne mjere koje je potrebno provesti za funkcionalnost sustava grijanja su sljedeće:

- provjera ispravnosti elemenata instalacije (ventili, pumpe, ekspanzijska posuda),
- provjera začepljenja, propuštanja (curenja),
- ispiranje stare instalacije na strani potrošača – sekundarni krug (cjevovodi, ogrjevna tijela),
- provjera ispravnosti izvedbe primjerice spajanja polaznog i povratnog cjevovoda,
- odzračivanje sustava u sekundarnom krugu potrošača nakon ponovnog punjenja sustava ogrjevnim medijem.

Mjere u vezi s toplinskim izvorom, odnosno toplinskom podstanicom zgrade su:

- regulacija polazne temperature sustava grijanja u sekundarnom krugu potrošača prema vanjskoj temperaturi,
- upotreba cirkulacijskih pumpi s elektronički reguliranom brzinom vrtnje, povezivanje regulacije pumpi s regulacijom izvora topline (toplinske podstanice).

Pravilna hidraulička distribucija toplinske energije u zgradi od velike je važnosti, kako je i definirano mjerodavnom normom za ugradnju razdjelnika HRN EN 834 [4]. Spomenuta norma, u Aneksu A, jasno definira preporuke i nužne preduvjete za ugradnju razdjelnika troškova toplinske energije kako bi sustav ispravno funkcionirao. Tako norma, u dijelu „A.2 Sustavi grijanja“, točka (d), govori „sustavi grijanja trebaju biti hidraulički balansirani, tj. protoci ogrjevnog medija su prilagođeni izvedbenom stanju strojarskih instalacija“. Iz navedenog se zaključuje da bi svakako trebalo provesti sljedeće mjere u vezi cijevnog razvoda za grijanje (sekundarni krug potrošača):

- ugradnja automatskih balansirajućih ventila u cijevnom razvodu,
- provođenje automatskog hidrauličkog uravnoteženja cijevnog razvoda.

Mjere u vezi s ogrjevnim tijelima odnose se na ugradnju RTTE na radijatore u stanovima kao i ugradnju mjerila toplinske energije koji moraju biti povezani sa zamjenom konvencionalnih zapornih ventila s termostatskim regulacijskim ventilima. Time se omogućuje kontrola osobne potrošnje u pojedinim stanovima. Stoga obje sljedeće mjere treba provesti zajedno, a to su:

- ugradnja termostatskih ventila na radijatorima
- ugradnja RTTE na radijatorima (alternativno, mjerilo toplinske energije na ulazu u pojedine stanove)

Način raspodjele troškova i tehnički preduvjeti ugradnje RTTE u Hrvatskoj propisani se Pravilnikom o načinu raspodjele i obračunu troškova za isporučenu toplinsku energiju [56].

Treba napomenuti da će samo u slučaju ispunjenja svih navedenih mjera u jednoj zgradi ugradnja razdjelnika troškova toplinske energije i raspodjela troškova grijanja funkcionirati i dovesti do očekivanog učinka kontrole potrošnje.

1.1.3. Predviđanje potrošnje energije metodama strojnog učenja

Mehmood i sur. vjeruju da će umjetna inteligencija (engl. *artificial intelligence* – AI) i obrada velikih skupova podataka (engl. *big data* – BD) imati dominantnu ulogu u budućim energetskim sustavima i ilustriraju kako se AI i BD mogu primijeniti na energetsku učinkovitost u zgradama kako bi zgrade bile energetski učinkovite i pružile visoku razinu toplinske ugodnosti u prostoru [57]. Prethodna istraživanja pokazala su da su metode strojnog učenja (SU), interdisciplinarnog područja koje služi pri istraživanju pojedinih znanstvenih pojava koristeći računalne i statističke metode, prikladne za procjenu potrošnje energije u zgradama [58].

SU je podskup AI polja, koje ima za cilj izgraditi modele koji mogu učiti iz dostupnih podataka, a obično se koristi u tri slučaja: (i) kada postoje složeni problemi za koje ne postoji ljudsko znanje o procesu ili nema nedvosmislenog objašnjenja za postupak, (ii) kada je potrebna zahtjevna obrada podataka, posebno ako se iz tih podataka može dobiti novo znanje, i (iii) pri razmatranju dinamički promjenjivih sustava [59].

Dodatno, budućnost energetike vidi se kroz integraciju različitih energetskih sustava putem pametnih mreža, čime integrirani sustavi neizbježno postaju nelinearni što predstavlja ograničenja pri planiranju rada takvih sustava [60]. Problem nelinearnosti može se riješiti primjenom modela SU, pa se tako pokazalo da se upotreba regresijskog stroja potpornih vektora može uspješno primijeniti na višestruki vektorski prostor tako da nelinearne karakteristike aproksimira linearnim pretpostavkama [61].

U izračunima toplinskog opterećenja i potrošnje energije uobičajeno se koriste brojni alati razvijeni za izračunavanje energetskih opterećenja zgrade [62]. Takvi simulacijski alati imaju neka ograničenja u smislu činjenice da su u pravilu skupi i da zahtijevaju određeno predznanje o pojedinom slučaju (zgradi) kako bi se učinkovito primijenili [63]. Tako je u [64] provedeno istraživanje s pretpostavkom da se korištenjem algoritama SU ta ograničenja mogu izbjeći, te su razvijeni modeli s algoritmima regresijskih stabala i slučajnih šuma, čija je točnost uspoređena s rezultatima dobivenim korištenjem simulacijskog alata EnergyPlus. Rezultati su pokazali da su algoritmi SU superiorniji u predviđajućim modelima potrošnje energije u zgradama u usporedbi s konvencionalnim simulacijskim modelima, a danas su dostupni *open-source* programi i programski jezici za razvoj ovakvih modela, kao što su, npr. Python [65] i R [66]. Analiza napravljena u ovom istraživačkom radu bit će razvijena korištenjem programskog jezika R [66].

Nadalje, pokazuje se da se i pri korištenju SU algoritama zajedno sa simulacijskim alatima dobiva viša točnost predviđanja, što je demonstrirano na predviđanju kombiniranjem TRNSYS alata za simulaciju i algoritma umjetnih neuronskih mreža (engl. *artificial neural networks* - ANN) i rezultiralo je točnim, jednostavnim i brzim proračunima za nestambene zgrade na europskoj razini [67].

Na razini sustava daljinskog grijanja izrađene su prethodne analize SU algoritmima, specifično algoritmom stroja potpornih vektora i *firefly searching* algoritma, te se pokazalo da algoritam stroja potpornih vektora daje točnije predviđajuće modele, a time i poboljšava ukupnu učinkovitost rada sustava daljinskog grijanja omogućavajući bolje planiranje [68]. Jedan od analiziranih algoritama u ovom radu bit će upravo stroj potpornih vektora.

Napravljene su analize koji su od SU algoritama dominantni u smislu točnosti predikcija u sektoru potrošnje energije u zgradama, te je donesen zaključak da iako se najčešće koristi linearna regresija, ANN i stroj potpornih vektora, najprikladniji model ne može se odabrati *a priori* već pojedini algoritmi moraju biti testirani za različite probleme [69]. Stoga je u ovom radu napravljena analiza podataka s više algoritama, višestrukom linearnom regresijom, regresijskim stablima, slučajnom šumom i strojem potpornih vektora, među ostalim s ciljem identifikacije najprikladnijeg algoritma za ovu vrstu problema. Slučajne šume danas su jedan od najučestalijih i učinkovitijih metoda strojnog učenja [70].

S obzirom na to da točnost algoritama ovisi i o ekspertizi osobe koja provodi modeliranje, napravljena je i analiza točnosti modela temeljenih na linearnoj regresiji, ANN, stroju potpornih vektora i regresijskim stablima za predviđanje toplinskog opterećenja u sustavima daljinskog grijanja te se pokazalo da su modeli linearne regresije davali najlošija predviđanja, zatim stroj potpornih vektora, dok su ANN i regresijska stabla davala najtočnije predviđajuće modele [71].

Osim ekspertnog znanja osobe koja izrađuje model, za točnost modela bitan je i podatkovni skup na kojem se model razvija. Tako se npr. visoka preciznost predviđanja pokazala na primjeru predviđanja toplinske potrošnje stambene zgrade u Kanadi na temelju ulaznih podataka vanjske temperature, sunčeva zračenja, dnevnog vremena boravka stanara te dijela tjedna (vikend ili radni dan) [72]. Osim u predviđanju sustavima daljinskog grijanja, u sektoru energetike SU algoritmi posebice se koriste u predviđanju potrošnje električne energije, gdje se superiornim pokazuje algoritam slučajnih šuma, zbog svoje robusnosti i malog zahtjeva na pripremu podataka prije modeliranja [73].

1.2. Metodologija istraživanja

Istraživanjem je obuhvaćena izrada modela potrošnje toplinske energije u sustavima daljinskog grijanja zasnovanog na strojnom učenju kroz eksperimentalno istraživanje u postojećim sustavima daljinskog grijanja.

Istraživanje se sastoji od prikupljanja podataka o stvarnoj potrošnji iz sustava za naplatu toplinarskih tvrtki za dva distribucijska područja u sedmogodišnjem nizu (od 2011. do 2017.), iz kojeg se izdvajaju samo podaci za kućanstva. Dodatno, provedeni su upitnici i intervjui u tri zgrade u Zagrebu 2018. i 2019. Cilj je bio prikupiti bihevioralne parametre potrošnje te ih obraditi i interpretirati. U sklopu provedbe intervjua više puta je rađen terenski obilazak i kratki energetski audit te su fizički pregledane one stambene jedinice koje su pokazale neočekivane iznose potrošnje.

Na temelju prikupljenih podataka i intervjua, a prema stručnoj procjeni, definirani su utjecajni faktori, koji su zatim pripremljeni u posebne podatkovne skupove na kojima će se provoditi modeliranje metodama strojnog učenja.

Obrada podataka, deskriptivna analiza, grupiranje i modeliranje napravljeni su u programskom jeziku R [66], koristeći program RStudio [74].

Eksperimentalno istraživanje provedeno je na skupu stvarnih podataka o potrošnji energije u višegodišnjem nizu. Definirali su se svi faktori koji utječu na potrošnju, u smislu tehničkih i bihevioralnih faktora. Pod tehničkim faktorima podrazumijevaju se parametri iz sustava za naplatu te klimatske karakteristike kao što su stupanj dani grijanja. Pod netehničkim faktorima podrazumijevaju se bihevioralni faktori vezani za korisnike prostora (kao što su broj stanara u pojedinom stanu/zgradi, životna dob stanara), način obračuna za potrošenu energiju i cijena energenta, razina željene toplinske ugodnosti u prostoru, način korištenja prostora tj. dnevni raspored boravka u pojedinim stanovima. Definirani kriteriji koristit će se u modelu, s ciljem pronalaska funkcijskih vrijednosti među parametrima i određivanja stupnja utjecaja i važnosti pojedinih parametara na potrošnju topline u sustavima daljinskog grijanja.

Izrađeni su modeli postupkom prikupljanja i predobrade podataka, učenjem modela na skupu za učenje, provjerom modela na skupu za provjeru te ocjenom točnosti modela uz intepretaciju utjecaja pojedinih parametara na potrošnju topline za svaki pojedini model. Modeli razvijeni

metodama strojnog učenja tako imaju dvojaku funkciju, oni su predviđajući i služe za predviđanje ishoda na temelju prošlih podataka, ali su i deskriptivni jer služe za stjecanje novih znanja o procesu potrošnje toplinske energije u kućanstvima.

Prikupljeni parametri su analizirani te je interpretiran njihov međusobni utjecaj kao i utjecaj svakog od njih na potrošnju topline u sustavima daljinskog grijanja. S obzirom na tip i vrstu podataka koji su bili dostupni u ovoj analizi razvijeni su modeli koristeći ove metode strojnog učenja:

- regresijska analiza,
- regresijska stabla (engl. *regression trees*),
- slučajna šuma (engl. *random forest*) i
- stroj potpornih vektora (engl. *support vector machines, SVM*).

Dobiveni rezultati interpretirani su na način da je ocijenjena točnost predviđanja svake metode te je dana preporuka koji je model najprikladniji za predviđanje potrošnje u sustavima daljinskog grijanja. Ovako dobiveni modeli mogu se koristiti za predviđanje, ali i za kvantitativno vrednovanje primjene mjera energetske učinkovitosti, u ovom slučaju ugradnje individualnog mjerenja u sustave daljinskog grijanja, na način da će se utjecaj moći egzaktno izraziti.

Strojno učenje znanstvena je metoda koja se sve više primjenjuje u raznim znanstvenim područjima zahvaljujući razvoju računalnih tehnologija te mogućnosti prikupljanja i skladištenja velikih količina podataka na kojima je moguće raditi predviđajuće analize. U sektoru energetike dostupnost podataka o karakteristikama potrošnje također raste, postoje brojne baze podataka koje sadržavaju podatke o potrošnji u sektorima umreženih energenata daljinskog grijanja, električne energije, prirodnog plina i vode. Ovi vrijedni podaci danas se agregirano koriste za predviđanja u strateškim dokumentima na nacionalnoj razini i u energetskoj statistici, ali se u budućnosti prepoznaje i njihov potencijal za predviđanja na nižim razinama (zgrada), kao i snažnija integracija između sektora energetike i dijela računalne znanosti i statistike koja se bavi strojnim učenjem, kako bi se razvili točniji modeli predviđanja potrošnje, ali i ocjene utjecaja provedbe energetskih aktivnosti na postizanje strateških energetskih ciljeva.

Razvojem modela visoke točnosti za predviđanje potrošnje toplinske energije potaknut će se i povećanje integracije obnovljivih izvora energije (OIE) za proizvodnju toplinske energije u sustavima daljinskog grijanja, jer će se na taj način omogućiti povezivanje ovako razvijenog modela s postojećim modelima za predviđanje potrošnje energije. Ovdje se u prvom redu misli na pretvorbu električne energije iz obnovljivih izvora u toplinsku energiju u razdobljima velike dostupnosti intermitentnih OIE-a kao što su solarna ili energija vjetra, čime bi se pridonijelo većoj integraciji OIE-a u energetski sustav [75].

1.3. Znanstveni doprinos

Znanstveni doprinos istraživanja predstavlja razvoj modela za predviđanje potrošnje i procjene utjecaja mjera energetske učinkovitosti zasnovane na metodama strojnog učenja. Model je razvijen na više razina, redom (i) na razini pojedinog stana (tj. krajnjeg potrošača) u sustavu daljinskog grijanja, (ii) na razini pojedine zgrade spojene na sustav daljinskog grijanja, (iii) na razini pojedinog distribucijskog područja te (iv) na razini zemlje.

Također, doprinos će se dati u vrednovanju svakog razmatranog algoritma strojnog učenja za ovu primjenu, a to su višestruka linearna regresija, regresijska stabla, slučajne šume i stroj potpornih vektora od nadziranih metoda te grupiranje kao nenadzirana metoda.

Dodatno, kao rezultat provedenih analiza strojnog učenja odredit će se stupanj utjecaja svakog faktora modela te će se interpretirati njegov utjecaj na potrošnju topline u sustavima daljinskog grijanja.

1.4. Hipoteza

Cilj istraživanja je izrada modela potrošnje toplinske energije u sustavima daljinskog grijanja radi planiranja potrošnje i ocjene učinka mjera energetske učinkovitosti. Iz navedenog cilja proizlaze hipoteze istraživanja:

1. Prikupljeni podaci koji su na raspolaganju, dovoljno su reprezentativni te je odgovarajućim pretprocesiranjem, uporabom deskriptivne multivarijatne statističke analize i eksplorativne analize moguće dobiti detaljniji uvid u karakter podataka, njihovu reprezentativnost i njihov međusobni odnos te odgovarajuću procjenu pogodnosti za ugradnju u predviđajući model.

2. Temeljem prethodnih analiza moguće je izgraditi model zasnovan na strojnom učenju koji će potrošnju energije predviđati točnije nego dosadašnji modeli.

1.5. Struktura rada

Rad je strukturiran u pet međusobno zavisnih poglavlja, kako bi se odgovorilo na istraživačka pitanja kolika se ušteda može očekivati na više razina uvođenjem individualnog mjerenja u stanove spojene na sustave daljinskog grijanja i koji su utjecajni parametri na potrošnju primjenom metoda strojnog učenja. Kako bi se odgovorilo na ta pitanja razvijeni su modeli visoke točnosti temeljeni na metodama strojnog učenja, što je predstavljeno kroz sljedeću strukturu rada.

Uvod: Opisana je motivacija i kontekst istraživanja, predstavljena su recentna istraživanja u polju predviđanja potrošnje energije u sustavima daljinskog grijanja i problematike ugradnje individualnog mjerenja i primjene drugih mjera energetske učinkovitosti. Predložena je metodologija provedbe istraživanja te izložena hipoteza i očekivani znanstveni doprinosi.

Metodologija: Metodologija istraživanja temelji se na provedenom istraživanju na podacima iz sustava za naplatu toplinarskih tvrtki, prikupljenim upitnicima i provedenim intervjuima. Poglavlje sadržava i teoretske i metodološke osnove svake primijenjene metode strojnog učenja u ovom radu.

Provedeno istraživanje: Opisano je provedeno istraživanje na podacima iz sustava za naplatu toplinarskih tvrtki, predstavljeni su rezultati analize podataka prikupljenih upitnikom koji su dodatno analizirani metodom grupiranja kao nenadziranom metodom strojnog učenja te su razvijeni modeli primjenom opisanih metoda na više razina.

Prikaz rezultata: Prikazani su rezultati i predstavljen je predviđajući model na razini stana, zgrade, distribucijskog područja i na nacionalnoj razini. Primijenjeni algoritmi strojnog učenja vrednovani su za ovu namjenu te je dan pregled utjecajnih parametara na potrošnju na svim razinama.

Zaključak: Pregled rada s rezultatima istraživanja.

2. METODE STROJNOG UČENJA

Metode strojnog učenja koje su korištene u ovom radu, a za koje je metodologija opisana u ovom poglavlju, temelje se na recentnoj relevantnoj stručnoj literaturi za područje statistike, strojnog učenja i predviđanja potrošnje u sektoru energetike.

Literatura je analizirana s ciljem identifikacije najprikladnijih metoda za istraživački problem predviđanja potrošnje u sustavima daljinskog grijanja metodama strojnog učenja, relativno novog znanstvenog područja koji nalazi sve širu primjenu u energetskom sektoru.

Provedene analize i opisana metodologija u ovom radu se temelji na odabranim metodama čije su teoretske osnove usvojene i preuzete iz relevantnih publikacija [76-83].

Strojno učenje i njegove metode koriste se kako bi se razumjeli odnosi između varijabli od interesa, tj. odnosi između ulaznih i izlaznih varijabli. Ulazne varijable se označavaju s X (X_1, X_2, …, X_3) te se nazivaju i nezavisne varijable, prediktori, značajke ili samo varijable. Izlazne varijable se nazivaju odzivi ili zavisne varijable, a označavaju se s Y.

Povezanost između ulaznih i izlaznih varijabli se može modelirati kao:

$$Y = f(X) + \epsilon \tag{1}$$

Ovdje f predstavlja funkciju ovisnu o X i slučajnu pogrešku ϵ koja je neovisna o X i ima očekivanje 0, tj. $E(\epsilon) = 0$. U ovoj formuli f predstavlja sustavnu informaciju koju X daje o Y.

U većini slučajeva funkcijska ovisnost između varijabli X i Y nije poznata, već su poznati apsolutni iznosi parova ovih vrijednosti za neki broj opažanja, čime se dobije površina u kojem se nalazi analizirani skup podataka.

Metode strojnog učenja mogu se podijeliti na parametarske i neparametarske metode. Parametarske metode provode se u dva koraka:

i. Prvo, pretpostavlja se oblik funkcije f:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{2}$$

Pretpostavka o linearnosti modela uvelike ga pojednostavljuje, tako da umjesto procjene *p*-dimenzijske funkcije treba odrediti *p+1* koeficijent $\beta_0, \beta_1, \beta_2, \dots, \beta_p$.

ii. Drugo, pomoću ulaznih podataka ispituje se valjanost modela, te se određuju parametri linearne jednadžbe tako da se zadovolji:

$$Y \cong \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \qquad (3)$$

Parametarska metoda svodi problem određivanja *f* na problem određivanja parametara te funkcije. Određivanje parametara koji odgovaraju funkciji lakše je od određivanje same funkcije.

Kod korištenja neparametarskih metoda ne pretpostavlja se eksplicitni oblik funkcije već se procjenjuje *f* koja je što bliža danim podacima. Prednost ovakvih metoda, u usporedbi s parametarskim metodama, je to da one mogu točnije odrediti funkciju jer ne pretpostavljaju njezin oblik. Nedostatak ove metode je to što se mora poznavati velika količina podataka za točnu pretpostavku *f*. Ova metoda je izrazito varijabilna s obzirom da ne predviđa izgled funkcije, nego se prilagođava danim podacima. Varijabilnost je nepoželjna karakteristika jer stvara greške.

Osim podjele metoda strojnog učenja na parametarske i neparametarske, pojedine metode dijelimo i po tipu učenja na nadzirano i nenadzirano učenje. Kod nadziranog učenja za svako promatranje $i=1,\dots,n$ poznata je ulazna varijabla x_i i uz nju pridruženi odziv y_i. Cilj ovakvih metoda je točno predviđanje odziva i bolje razumijevanje odnosa između ulaznih i izlaznih varijabli.

Kod nenadziranih metoda, za svaki primjer $i=1,\dots,n$ poznat je vektor x_i, ali bez povezanog odziva y_i. Provedba ovakvih analiza teža je jer nije poznato kako se izlazne varijable ponašaju s obzirom na ulazne varijable. Kod takvih slučajeva koristi se analiza grupiranja. Cilj analize grupiranjem je zaključiti da li se zapažanja (temeljena na $x_1, x_2, \dots, x_n$) grupiraju u relativno različite skupine.

Postoje slučajevi kod kojih su poznate ulazne i izlazne varijable za jedan dio zapažanja, a za druga zapažanja poznate su samo ulazne varijable. Tada se primjenjuju polunadzirane metode gdje se u analizu uvrštavaju sve vrijednosti.

2.1. Grupiranje

Grupiranje (engl. *clustering*) predstavlja metodu strojnog učenja kojom se unutar podatkovnog skupa od interesa nalaze podgrupe koji imaju određene zajedničke karakteristike. Ovo su analize nenadziranog učenja s obzirom da je im je cilj pronalazak strukture u podacima bez prethodnog određivanja ciljanog parametra, tj. analizom se istražuje veza između *n* opservacija bez prethodnog učenja. Grupiranjem se identificiraju homogene podgrupe unutar podatkovnog skupa.

Razlikujemo dvije osnovne metode grupiranja: grupiranje *k*-srednjim vrijednostima i hijerarhijsko grupiranje.

2.1.1. Grupiranje k-srednjim vrijednostima

Metodom *k*-srednjih vrijednosti (engl. *k-means*) grupiramo opservacije u k predefiniranih podgrupa u koje algoritam pridodaje svaku od opservacija.

Algoritam traži da u podatkovnom skupu nema nedostajućih vrijednosti te da se podaci standardiziraju kako bi različiti parametri bili usporedivi. Standardiziranje ili normizacija, podrazumijeva takvu transformaciju parametara na način da se za svaku varijablu izračuna srednja vrijednost i standardna devijacija, te se zatim svaka opservacija pojedinog parametra umanji za srednju vrijednost i podijeli za vrijednost standardne devijacije [87]. Na ovaj način dobivamo standardizirane vrijednosti koje predstavljaju koliko su standardne devijacije svake opservaciju udaljene od srednje vrijednosti.

$$StandardiziranaVrijednost = \frac{NestandardiziranaVrijednost - \mu}{\sigma} \qquad (4)$$

gdje je μ srednja vrijednost, a σ standardna devijacija [88].

Tendencija je da se pojedine grupe međusobno ne preklapaju. Postupak grupiranje se provodi s ciljem da se odredi grupa za koji je varijanca W unutar grupe C_k minimalna korištenjem sljedeće formule:

$$min_{C_1,\dots,C_k}(\sum_{k=1}^{k}\frac{1}{|C_k|}\sum_{k=1}^{k}\sum_{x_i\in C_k}(x_i-\mu_k)^2) \quad (5)$$

gdje je x_i opservacija u jednoj grupi C_k, a μ_k srednja vrijednost svih opservacija u grupi C_k.

Sama varijanca unutar grupe C_k se može izračuniti prema donjoj formulu:

$$W(C_k) = (\sum_{k=1}^{k}\frac{1}{|C_k|}\sum_{k=1}^{k}\sum_{x_i\in C_k}(x_i-\mu_k)^2) \quad (6)$$

Pri grupiranju potrebno je provesti ove korake (algoritam):

1. Odrediti broj grupa kao ulazni podatak.
2. Nasumično pridružiti broj grupa k svakoj opservaciji.
3. Za svaku grupu izračunati pripadajući centroid. Centroid *k*-te grupe je vektor duljine p (parametara) koji sadrži srednju vrijednost za sve opservacije u *k*-toj grupi.
4. Pridodati svaku opservaciju grupi čiji centroid je najbliži vrijednosti opservacije u smislu euklidske udaljenosti.

Koraci 3. i 4. se ponavljaju sve dok se ne postigne konvergiranje i sve opservacije su pridodane nekoj od C_k grupa. Ukoliko u podatkovnom skupu postoji više od dva parametra provest će se metoda glavnih komponenti (engl. *principal component analysis*) te će se podaci grafički prikazati prema prve dvije glavne komponente koje objašnjavaju većinu varijanci.

2.1.2. Određivanje optimalnog broja grupa za podatkovni skup

Gap statistics

Metoda *gap statistics* uspoređuje ukupnu varijancu unutar pojedine grupe za različite vrijednosti k s njihovim očekivanim vrijednostima u slučaju kada grupiranje nije provedeno, tj. kada se uspoređuje s originalnim podatkovnim skupom. Referentni podatkovni skup se određuje Monte Carlo simulacijom pri uzorkovanju, tj. za svaki parametar (x_i) u originalnom podatkovnom skupu se izračuna raspon $[min(x_i), max(x_j)]$ i odredi se vrijednost za svaku opservaciju u intervalu.

Ova metoda je razvijena od strane Tibshiranija, Walthera i Hastie 2001. [89], koji definiraju *gap statistics* za k kao:

$$Gap_n(k) = E'_n log(W(C_k)) - log(W(C_k)) \quad (7)$$

gdje je $W(C_k)$ definiran u jednadžbi (*7*), dok E'_n označava očekivanje iz originalnog podatkovnog skupa i distribucija, a određuje se *bootstrap* metodom kojom se određuje N različitih podskupova originalnog podatkovnog skupa. Metoda mjeri razliku $W(C_k)$ prema očekivanoj vrijednosti u nultoj hipotezi. Optimalni broj grupa će biti onaj koji maksimizira vrijednost $Gap_n(k)$.

Metoda lakta

Metoda lakta (engl. *elbow method*) je osnovna metoda u grupiranju, te se njom traži takav k koji će dati minimum za svaku varijancu unutar svake grupe, tj.:

$$min(\sum_{k=1}^{k} W\,(C_k)) \quad (8)$$

Ovaj algoritam računa sumu kvadrata unutar svakog od predefiniranog broja grupa (engl. *within-cluster sum of squares* - WSS), ispisuje grafički krivulju, te se načelno prva infleksija na krivulji smatra optimalnim brojem grupa.

Metoda prosječne siluete

Metoda prosječne siluete (engl. *average silhouette method*) mjeri kvalitetu grupiranja, tj. mjeri koliko je dobro pojedina opserevacija smještena unutar pojedine grupe. Visoka vrijednost daje indikaciju dobrog smještaja. Analiza se provodi za svaki pojedinu k grupu, te se najboljim definira onaj s najvišom vrijednosti [83].

2.1.3. Mjere sličnosti

Odabir prikladne mjere sličnosti je od iznimne važnosti za analizu grupiranjem, te je uvjetovana tipom varijabli koje se žele grupirati. Kod kontinuiranih varijabli mjere sličnosti koje su koriste

su: euklidska udaljenosti, Manhattan udaljenost i Minkowski udaljenost [83]. Minkowski udaljenost je generalizacija euklidske i Manhattan udaljenosti.

Ako kažemo da su M podatkovni skup koji se sastoji od p parametar, te da su $x, y \in N$, svaki čini $x = (x_1, \ldots, x_n)$ i $y = (y_1, \ldots, y_n)$, gdje je n opservacija, a svaki x_p i y_p, $1 \leq p \leq n$

Minkowski udaljenost definirana je kao:

$$d(x,y) = \sqrt[r]{\sum_{i=1}^{n} |x_i - y_i|^r}, r \geq 1 \qquad (9)$$

Što je vrijednost r viša jači je utjecaj razlika na mjeru sličnosti. Ostale dvije mjere sličnosti, euklidska udaljenost i Manhattan udaljenost, predstavljaju posebne slučajeve Minkowski udaljenosti. Euklidska udaljenost predstavlja slučaj u kojem je $r = 2$, a Manhattan udaljenost slučaj u kojem je $r = 1$. Postoji i četvrta mjera koja je poseban slučaj Minkowski udaljenosti, a to je slučaj kada je $r \to \infty$, te se naziva Čebišeljevojom udaljenosti.

U većini analiza koristi se euklidska udaljenost kao mjera sličnosti.

2.2. Procjena točnosti modela

Odabir odgovarajuće metode statističke analize jedan je od važnijih zadataka jer jedna metoda može odlično funkcionirati u određenom primjeru, dok u sličnim primjerima ona neće pokazivati dobre rezultate.

2.2.1. Određivanje kvalitete odabira

Kako bismo procijenili valjanost statističke analize danih podataka, moramo odrediti koliko se predviđanja poklapaju sa stvarnim podacima, odnosno moramo kvantificirati raspon u kojemu predviđeni odzivi odgovaraju stvarnim vrijednostima promatranog slučaja. Najčešće korištena metoda za određivanje valjanosti analize je srednja kvadratna pogreška (engl. *mean square error* - MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad (10)$$

gdje je $\hat{f}(x_i)$ predviđanje koje $\hat{f}$ daje za *i*-to promatranje. Vrijednost MSE bit će mala ako su predviđeni odzivi vrlo blizu stvarnih odziva.

Prilikom primjene funkcijskih veza nije bitna točnost kojom metoda predviđa podatke koje su nam služili za njezino određivanje, već nam je bitna točnost budućih predviđanja koja ćemo dobiti korištenjem te funkcije. Drugačije rečeno, važno nam je da na temelju novih ulaznih varijabli, funkcija koju smo odredili uz pomoć testnih varijabli, daje pouzdane izlazne vrijednosti.

U primjeru s vrijednostima $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ pomoću kojih je dobivena $\hat{f}$, može se odrediti $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_n)$. Ako su te određene vrijednosti približno jednake $y_1, y_2, \dots, y_n$ onda je MSE mala, ali u analizi nije cilj odrediti $\hat{f}(x_i) \cong y_i$, nego $\hat{f}(x_0)$ koji je približno jednak y_0 i gdje su (x_0, y_0) prije neupotrijebljene vrijednosti. U takvim primjerima cilj je dobiti što manju vrijednost MSE i ako je dostupan veliki broj novih vrijednosti može se izračunati:

$$Ave(y_0 - \hat{f}(x_0))^2 \qquad (11)$$

koji predstavlja prosječno kvadratnu pogrešku predviđanja za vrijednosti (x_0, y_0), za čiji iznos želimo da bude čim manja.

Kako bi se utvrdila pogreška koja se javlja u stvarnim primjerima, a ne primjerima koji su poslužili za određivanje funkcijske veze koriste se vrijednosti dobivene promatranjem, a koje nisu prethodno korištene u analizi. Ako takve vrijednosti nisu dostupne, pogrešku će biti teže utvrditi jer stvarna pogreška nije ista pogrešci koja se javlja pri korištenju varijabli koje su služile za aproksimaciju funkcije.

Srednja kvadratna pogreška ovisi o stupnju slobode funkcije koji pokazuje koliko je ona fleksibilna. Što je metoda fleksibilnija MSE dobiven iz vrijednosti kojima je određena funkcija je manji. Može se dogoditi da stvarni MSE pada s povećanjem fleksibilnosti, ali sve do neke granice nakon koje raste. Razlog tomu je to što će prilikom velike fleksibilnosti analize aproksimacija slijediti inicijalne vrijednosti koje vjerojatno odstupaju od stvarnih kretanja te

funkcije. Stvarni MSE će uvijek biti veći od onog dobivenog vrijednostima korištenim za dobivanje funkcije.

2.2.2. Kompromis između pristranosti i varijance

Stvarna vrijednost MSE, za zadanu vrijednost x_0, može se prikazati kao suma triju osnovnih veličina: varijance od $\hat{f}(x_0)$, kvadratnog pomaka od $\hat{f}(x_0)$ i varijance pogreške ϵ:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Pomak(\hat{f}(x_0))\right]^2 + Var(\epsilon) \qquad (12)$$

Gdje $E\left(y_0 - \hat{f}(x_0)\right)^2$ predstavlja očekivane vrijednosti stvarne MSE, odnosno prosječnu vrijednost stvarne MSE koja se računa određujući *f* pomoću velikog broja stvarnih točaka u testnom uzorku.

Iz jednadžbe se vidi da je greška manja što su varijanca i pomak manji. Također se vidi da varijanca ne može biti negativna, kao ni kvadrat pomaka, što znači da da očekivana stvarna MSE ne može biti manja od $Var(\epsilon)$.

Pojam varijanca odnosi se na iznos promjene $\hat{f}$ kad bi pri njezinom određivanju koristili različite uzorke podataka. Ako metoda ima veliku varijancu, male promjene ulaznih podataka mogu rezultirati velikim promjenama $\hat{f}$. Fleksibilnije metode imaju veću varijancu od restriktivnih.

Pristranost predstavlja pogrešku koja je uzrokovana aproksimacijom stvarnog kompleksnog modela s jednostavnim modelom. Fleksibilnije metode u pravilu pokazuju manja odstupanja od stvarnih kretanja funkcija.

Fleksibilnije metode imaju veću varijancu i manja odstupanja. Iznos tih vrijednosti utjecat će na iznos stvarne MSE. S povećanjem fleksibilnosti odstupanje pada brže od porasta varijance što za posljedicu ima manji MSE, no nakon određenih vrijednosti odstupanje pada sporije, a varijanca značajno raste što uzrokuje porast MSE.

Cilj izbora dobre metode učenja su mala odstupanja i mala varijanca, što je teško postići jer najniže vrijednosti odstupanja za sobom povlače velike vrijednosti varijance i obratno.

2.3. Višestruka linearna regresija

Linearna regresija spada u skupinu nadziranih statističkih analiza te je pogodna za predviđanje kvantitativnih odziva sustava. Ukoliko zavisna varijabla ovisi o više nezavisnih, opcija je provoditi više jednostrukih regresija ili provesti višestruku linearnu regresiju. U višestrukoj linearnoj regresiji svakoj varijabli dodjeljuje se koeficijent nagiba te je njezin opći oblik:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \qquad (13)$$

gdje X_j predstavlja *j*-tu ulaznu varijablu i gdje β_j kvantificira povezanost između ulazne i izlazne varijable. Regresijski parametar β_j interpretiramo kao učinak jediničnog povećanja X_j na prosječni *Y* (držeći ostale ulazne varijable fiksnima).

Koeficijenti $\beta_0, \beta_1, \ldots, \beta_p$, u jednadžbi (*13*) su nepoznati. Procjenom $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ računa se procjena odziva koristeći:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \qquad (14)$$

Parametri su procijenjeni pomoću metode najmanjih kvadrata. Regresijski parametri $\beta_0, \beta_1, \ldots, \beta_p$ se određuju tako da se suma kvadrata ostataka (engl. *residual sum of squares* - RSS) svede na minimum:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \qquad (15)$$

Vrijednosti $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ predstavljaju procijene koeficijenata višestruke regresije koji su dobiveni metodom najmanjih kvadrata.

R^2 statistički podatak daje informaciju koliko model dobro odgovara stvarnosti. Ova vrijednost predstavlja omjer varijanci te je stoga njegova vrijednost uvijek između 0 i 1 (neovisan je o *Y* skali). Za izračun R^2 koristi se izraz:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \qquad (16)$$

Gdje je $TSS = \sum(y_i - \bar{y})^2$ ukupna suma kvadrata (engl. *total sum of squares* - TSS). Ukupna suma kvadrata (TSS) mjeri ukupnu varijancu u odzivu *Y* i može se smatrati količinom varijabilnosti koja je karakteristična odzivu prije provedbe regresije. Nasuprot tome, suma kvadratnih ostataka (RSS) mjeri količinu varijabilnosti u odzivu koja je opisana (ili uklonjena) provođenjem regresije, a R^2 mjeri omjer varijabilnosti *Y* koji se može opisati s *X*. Iznos R^2 blizu 1 pokazuje da je velik dio varijabilnosti odziva opisan regresijom, a 0 označava da regresija ne opisuje dobro varijabilnost odziva čemu je najvjerojatniji uzrok loš odabir linearnog modela ili visoka inherentna pogreška σ^2.

U višestrukoj linearnoj regresiji s *p* prediktora provjerava se jesu li svi koeficijenti jednaki 0 ($\beta_1 = \beta_2 = \cdots = \beta_p = 0$), tj. provjerava se nulta hipoteza:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad (17)$$

i njoj alternativna:

$$H_a: barem\ jedan\ \beta_j\ nije\ jednak\ 0 \qquad (18)$$

Provjera pretpostavke provodi se pomoću *F*-statistike:

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)} \qquad (19)$$

gdje je $TSS = \sum(y_i - \bar{y})^2$, a $RSS = \sum(y_i - \hat{y}_i)^2$. Ako su pretpostavke linearnog modela točne, može se pokazati da:

$$\{RSS/(n - p - 1)\} = \sigma^2 \qquad (20)$$

i da je, točan dobiven H_0:

$$\{(TSS - RSS)/(p)\} = \sigma^2 \quad (21)$$

Stoga, ako ne postoji veze između prediktora i odziva, vrijednost *F*-statistike se približava 1. Ako vrijedi H_a, onda je$\{(TSS - RSS)/(p)\} > \sigma^2$ pa je vrijednost *F* veća od 1.

Koliko vrijednost *F* mora biti veća od 1, da bi se odbacila H_0 i zaključilo da postoji veza između prediktora i odziva, ovisi o vrijednostima *p* i *n*. Kad je *n* veliki, *F*-statistika je nešto veća od 1, što još uvijek može ukazivati na postojanje veze između ulaznih i izlaznih varijabli. Kada se utvrdi da postoji H_0 i kada pogreške ϵ_i imaju normalnu distribuciji, *F*-statistika slijedi *F*-distribuciju.

U jednadžbi (*17*) provjerava se H_0 tako da se ispituje da li se svi koeficijenti jednaki 0. U slučajevima u kojima je cilj provjeriti samo je li podskup *q* koeficijenata jednak 0, nulta hipoteza se određuje prema pretpostavci:

$$\boldsymbol{H_0: \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0} \quad (22)$$

U ovom slučaju koristi se drugi model koji koristi sve varijable osim zadnjih varijabli *q*. Pri pretpostavci da je suma kvadratnih ostataka tog modela RSS_0, odgovarajući *F*-statistički podatak iznosi:

$$\boldsymbol{F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)}} \quad (23)$$

Prvi korak analize višestrukom regresijom je određivanje *F*-statističkog podatka i ispitivanje s njim povezane *p*-vrijednosti. Ako se na temelju *p*-vrijednosti zaključi da je najmanje jedan prediktor povezan s odzivom, onda je potrebno odrediti koji je taj prediktor.

2.3.1.1. Uklapanje modela i predviđanja

Preostala standardna pogreška (RSE) i R^2 su numerički postupci koje nam pokazuje koliko model dobro odgovara problemu. U jednostavnoj linearnoj regresiji R^2 predstavlja kvadrat korelacije između odziva i ulazne varijable. U višestrukoj linearnoj regresiji jednak je $Cor(Y, \hat{Y})^2$, odnosno kvadratu korelacije između odgovarajućeg modela i izlazne varijable. Vrijednost R^2 blizu jedan znači da model dobro opisuje varijancu izlazne varijable. Drugi

numerički postupak, preostala standardna pogreška (RSE), pokazuje da model dobro opisuje pojavu ako je njegov iznos malen:

$$RSE = \sqrt{\frac{1}{n-p-1}RSS} \tag{24}$$

Nakon odabira odgovarajućeg model linearne regresije, upotrebom jednadžbe (*14*), može se odrediti odziv *Y* na temelju vrijednosti prediktora $X_1,X_2,...,X_p$. No, pri predviđanju odziva javljaju se tri problema:

1. Koeficijenti $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ su procjena $\beta_0, \beta_1, \ldots, \beta_p$. Odnosno ravnina najmanjih kvadrata:

   $$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{25}$$

   je samo procjena stvarne regresijske ravnine skupa:

   $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{26}$$

   Ova netočnost procjene koeficijenata javlja se zbog nereducirajuće pogreške. Ono što se može izračunati je interval pouzdanosti koji nam govor koliko je $\hat{Y}$ blizu $f(X)$.

2. Pretpostavka da $f(X)$ ima linearni oblik je najčešće pogrešna zbog koje se javlja reducirajuća pogreška naziva pomak modela (otklon). Najčešće se ta pogreška zanemaruje i uzima da je linearan model točan opis stvarne pojave.
3. Čak i da je poznat pravi oblik $f(X)$, tj. prave vrijednosti $\beta_0, \beta_1, \ldots, \beta_p$ predviđanje odziva neće u potpunosti biti točno zbog nereducirajuće greške. Koliko će $\hat{Y}$ odstupati od *Y* možemo vidjeti iz intervala predviđanja. On je uvijek veći od intervala pouzdanosti jer ga čine pogreška procjene $f(X)$ i odstupanje pojedine točke od stvarne regresijske ravnine populacije.

2.3.1.2. Kvalitativni prediktori

Predviđanje odziva moguće je i za kvalitativne varijable. Ako kvalitativni prediktor (faktor) ima samo dvije moguće razine, odnosno dvije moguće vrijednosti, onda se na njega lako primjeni regresijski model. Postupak ide tako da najprije odredi indikator ili fiktivna varijabla koja poprima dvije moguće numeričke vrijednost. Definira se nova varijabla koja ima oblik:

$$x_i = \begin{cases} 1 & ako\ je\ i-ti\ input\ vrijednost\ A \\ 0 & ako\ je\ i-ti\ input\ vrijednost\ B \end{cases} \qquad (27)$$

Ta varijabla služi kao prediktor u regresijskoj jednadžbi, što rezultira modelom:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & ako\ je\ i-ti\ input\ vrijednost\ A \\ \beta_0 + \epsilon_i & ako\ je\ i-ti\ input\ vrijednost\ B \end{cases} \qquad (28)$$

Umjesto fiktivne varijable 0/1 koja pokazuje da li neka vrijednost jest ili nije, može se kvantificirati i svaka vrijednost koja se razmatra, npr.:

$$x_i = \begin{cases} 1 & ako\ je\ i-ti\ input\ vrijednost\ A \\ -1 & ako\ je\ i-ti\ input\ vrijednost\ B \end{cases}$$

Što uvrštenjem u regresijsku jednadžbu rezultira s:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & ako\ je\ i-ti\ input\ vrijednost\ A \\ \beta_0 - \beta_1 + \epsilon_i & ako\ je\ i-ti\ input\ vrijednost\ B \end{cases}$$

Sličan je postupak kada kvalitativni prediktor ima više od dvije rezane. Tada jedna fiktivna varijabla ne može prikazati sve moguće vrijednosti, te se uvodi još jedna fiktivna varijabla.

2.3.2. Vizualizacija linearnih modela

Pored poznavanja kvantitativnih vrijednosti koje se odnose na koeficijente i točnost modela, potrebno je razumjeti i metode vizualizacije pri određivanju modela. Gdje god je to moguće, uvijek je prvo potrebno vizualizirati model unutar danih podataka. To omogućuje usporedbu linearnog modela s modelom neparametarske lokalne polinomijalne regresije (engl. *locally*

estimated scatterplot smoothing - LOESS). Ako se LOESS nalazi unutar intervala pouzdanosti može se pretpostaviti da linearni model dobro opisuje stvarno kretanje.

Prikaz ostataka i odgovarajućih vrijednosti upozorava na dvije stvari:

1. Nelinearnost: ako je na prikazu vidljiv prepoznatljiv uzorak to upućuje na nelinearnost ili da karakteristike uzorka nisu dobro opisane.
2. Heteroskedastičnost: bitna pretpostavka linearne regresije je ta da veličine koje označavaju grešku imaju konstantnu varijancu $Var(\epsilon_i) = \sigma^2$. Ako prikaz ostataka ima oblik lijevka to upućuje na to da određeni model odstupa od te pretpostavke. To se ponekad može riješiti logaritmiranjem ili korjenovanjem Y u modelu.

Još jedan način za procjenu normalnosti ostataka je Q-Q prikaz. Taj prikaz pokazuje distribuciju dobivenih ostataka u odnosu na teoretsku normalnu distribuciju. Što točke više padaju na pravac i u njegovoj neposrednoj blizini tada se za te ostatke može reći da su normalno distribuirani. Ako točke dosta vijugaju i odstupaju od pravca tada ostaci nisu normalno distribuirani.

Linearni modeli se u pravilu vizualiziraju u ova dva grafička prikaza:

1. Reziduali naspram uklopljenih vrijednosti
   Ovaj graf prikazuje imaju li ostaci nelinearno ponašanje (uzorak). Ako postoji nelinearna ovisnost između prediktorskih varijabli i izlazne varijable, ponašanje ostataka može pokazati opisuje li model dobro tu nelinearnu ovisnost. Ako su ostaci ravnomjerno raspoređeni oko horizontalne linije bez nekog uzorka to je dobar znak da ne postoji nelinearna ovisnost.
2. Normalni Q-Q dijagram
   Ovaj graf prikazuje jesu li reziduali normalno distribuirani. Ako se želi saznati prate li ostaci ravnu liniju ili značajno odstupaju od nje, kao dobar pokazatelj koristi poravnatost ostataka u odnosu liniju normalnosti.

2.3.3. Potencijalni problemi

Kada pridružujemo model dobiven linearnom regresijom nekom skupu podataka mogu se javiti sljedeći problemi:

1. Nelinearnost veze prediktora i odziva.

Dijagrami ostataka koristan su grafički alat za određivanje nelinearnost, jer pokazuju koliko u pojedinom segmentu model odstupa od stvarnosti. Idealno, dijagram ostataka neće imati nikakve vrijednosti koje ukazuju na problematičnost odabranog linearnog modela za promatranu pojavu. Ako se iz dijagrama vidi da se radi o velikoj nelinearnosti, u modelu se moraju transformirati prediktori (logaritmirati, korjenovati ili kvadrirati).

2. Korelacija veličina koje označavaju pogrešku.

   Važna pretpostavka linearne regresije je pretpostavka da su pogreške $\epsilon_1, \epsilon_2, \dots, \epsilon_n$, nepovezane. Činjenica da je ϵ_i pozitivna, ne daje informaciju o predznaku ϵ_{i+1}. Standardne pogreške, koje se izračunavaju za procijenjene koeficijente linearne regresije, temelje se na pretpostavci da pogreške nisu povezane. Ako postoji povezanost između njih, onda će procijenjena standardna pogreška imati manju vrijednost od one koju bi ona zapravo trebali imati.

   Ako postoji korelacija između veličina koje označavaju pogreške, u dijagramu se može opaziti da pojedine susjedne vrijednosti pogrešaka imaju slične vrijednosti, tj. da se ponašaju po određenom obrascu.

3. Promjenjiva varijanca (odstupanja) veličina koje označavaju pogrešku.

   Još jedna od pretpostavka modela linearne regresije je to da pogreške imaju konstantnu varijancu ili odstupanje $Var(\epsilon_i) = \sigma^2$. Postojanje promjenjive varijance pogreške, ili heteroskedastičnosti, može se utvrditi ako dijagram ostataka ima oblik lijevka. Heteroskedastičnost se smanjuje ako Y transformiramo upotrebom konkavnih funkcija poput $logY$ ili √Y.

4. Stršeće vrijednosti.

   Stršeća vrijednost je točka za koju je y_i daleko od vrijednosti koju je predvidio model. Stršeće vrijednosti javljaju se iz različitih razloga, npr. netočnog mjerenja tijekom promatranja, te najčešće ne utječu na odgovarajuće najmanje kvadrate, no one mogu uzrokovati druge probleme. Ako se ne uključe u regresiju, oni uzrokuju smanjenje MSE s kojim se izračunavaju intervali pouzdanosti i p-vrijednosti.

Kako bi se ustvrdilo postojanje stršećih vrijednosti koriste se dijagrami ostataka. Problem kod tog utvrđivanja je nepoznavanje vrijednosti za koju možemo reći da je stršeća vrijednost. Kako bismo eliminirali taj problem, koristimo dijagram studentiziranog ostatka koji se dobije dijeljenjem svakog ostatka e_i s procijenjenom standardnom pogreškom. Za one vrijednosti, gdje je studentizirani ostatak veći od 3, možemo reći da su netipične vrijednosti.

5. Točke velikog utjecaja.

   Točke velikog utjecaja imaju neobične vrijednosti x_i. Ako te točke zanemarimo, značajno ćemo utjecati na liniju najmanjih kvadrata.

   Pronalazak točaka utjecaja, u jednostavnoj linearnoj regresiji, jednostavan je postupak kojem je cilj naći vrijednosti prediktora koje su izvan normalnog raspona u kojem se oni nalaze. U višestrukoj linearnoj regresiji moguće je da je pojedina vrijednost iz promatranja unutar raspona vrijednosti pojedinog prediktora ali da odstupa u odnosu na cjelokupan skup prediktora. Kako bismo odredili utjecaj pojedinog promatranja, izračunavamo statistički podatak utjecaja. Velik iznos ovog podatka upućuje na veliki utjecaj pojedinog promatranja.

6. Kolinearnost.

   Pojam kolinearnost odnosi se na slučaj kada su dva prediktora usko vezana jedan uz drugog. To stvara problem pri provedbi linearne regresije jer se ne može točno odrediti kako pojedina varijabla koja kolinearira utječe na odziv. Kolinearnost uzrokuje porast standardne pogreške od $\hat{\beta}_j$, koja uzrokuje pad vrijednosti t-statističkog podatka (t-statistički podatak se može dobiti za bilo koji prediktor tako da se podijeli $\hat{\beta}_j$ sa standardnom greškom prediktora). Zbog toga što je t- statistički podatak manji od stvarnog nismo u mogućnosti točno odrediti koeficijente koji iznose nula.

   Postojanje kolinearnosti može se otkriti u korelacijskoj matrici prediktora, po tome što visoka apsolutna vrijednost u matrici označava par varijabli s visokom korelacijom. Ipak, kolinearnost se može javiti i kod parova varijabli koji nemaju veliki iznos. Takva pojava naziva se multikolinearnost te se iz tog razloga za provjeru kolinearnosti koristi faktor inflacije varijance (engl. variable inflation factor - VIF). VIF je omjer između

varijance od $\hat{\beta}_j$ kada je ona uklopljena u model i varijance $\hat{\beta}_j$ kada je ona izvan modela. Najmanja vrijednost VIF koja se može postići je +/- 1 što znači da kolinearnost ne postoji. U većini slučajeva postoji mala kolinearnost između prediktora ali ako VIF postigne vrijednosti više od 5 ili 10 to označava da je ona vrlo visoka. Za svaku varijablu VIF se može izračunati prema izrazu:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_j|x_{-j}}} \qquad (29)$$

Gdje je $R^2_{x_j|x_{-j}}$ R^2 dobiven regresijom X_j na sve ostale prediktore. Ako je $R^2_{x_j|x_{-j}}$ blizu 1, prisutna je kolinearnosti.

Za rješavanje kolinearnosti služe nam dvije metode: izbacivanje jedne od kolinearnih varijabli (to ne šteti vjernosti modela jer sama kolinearnost upućuje na to da je jedna varijabla, u kolinearnom paru, nepotrebna za određivanje odziva) i spajanje dviju varijabli u jedan prediktor.

2.4. Metode ponovnog uzrokovanja

Stopa pogreške uzorka razlikuje se od stvarna stope pogreške ako razmatramo pogreška koja se javlja pri korištenju vrijednosti koje nisu služile za određivanje modela. Jedan dio metoda matematički korigira stopu pogreške uzorka kako bi se dobila stvarna stopa pogreške. Drugi dio metoda, određeni broj vrijednosti dobivenih promatranjem ne koristi za određivanje funkcije s ciljem utvrđivanja stvarne stope pogreške modela s tim vrijednostima.

2.4.1. Metoda skupa za provjeru

Metoda skupa za provjeru jedna je od metoda za određivanje stvarne stope pogreške. Ona podrazumijeva odvajanje dobivenih vrijednosti promatranjem u dvije skupine. U jednoj skupini se nalaze vrijednosti za određivanje modela, a u drugoj se nalaze vrijednosti za određivanje stvarne stope pogreške. Ova druga skupina zove se skup za provjeru. Model koji je određen pomoću podataka za njegovo određivanje služi za dobivanje odziva vrijednosti skupa za provjeru. Rezultirajuća stopa pogreške skupa za provjeru (koja se u slučaju kvantitativnog odziva izražava kao srednja kvadratna pogreška) predstavlja procjenu stvarne stope pogreške.

Proizvoljno se određuje koje će se vrijednosti svrstati u koju skupinu. Model skupa za provjeru je jednostavna metoda, ali ima nekoliko nedostataka:

- Provjera stvarne stope pogreške vrlo je varijabilna, odnosno ovisi o tome koje se vrijednosti nalaze u skupini za određivanje modela, a koje u skupini za određivanje stvarne stope pogreške.
- Zbog određivanja modela samo s dijelom vrijednosti dobivenih promatranjem, dobivamo manje točan model. Stoga vrijednosti koje se nalaze u skupu za provjeru mogu precijeniti stvarnu stopu pogreške.

Kako bi se eliminirali navedeni problemi koristi se metoda unakrsne provjere.

2.4.2. Unakrsna provjera - zadržavanje jedne vrijednosti sa strane

Kod metode unakrsne provjere (engl. *cross-validation* – CV) sa zadržavanje jedne vrijednosti sa strane, vrijednosti se također dijele u dvije skupine. Razlika ove metode, u usporedbi s prethodnom, je to što skupine nisu sačinjene od jednakog broja članova. To znači da se u skupu za provjeru nalazi samo jedna vrijednost dobivena promatranjem (x_1, y_1), a u skupini za određivanje modela nalaze se vrijednosti promatranja $\{(x_2, y_2), \dots, (x_n, y_n)\}$. Na tako dobivenom modelu s $n-1$ promatranja dobiva se procjena $\hat{y}_1$ pomoću izostavljene vrijednosti promatranja x_1. Zato što (x_1, y_1) nisu upotrjebljeni za određivanje modela, srednja kvadratna pogreška $MSE_1 = (y_1 - \hat{y}_1)^2$ predstavlja procjenu stvarne pogreške. Ta pogreška vrlo je varijabilna jer se temelji na samo jednom promatranju.

Postupak se provodi i dalje, tako da se i ostale vrijednosti, jedna po jedna, izostave i da se izračunavaju njihove srednje kvadratne pogreške. To znači da se postupak ponavlja n puta, dajući n srednjih kvadratnih pogrešaka. Procjena stvarnog srednjeg kvadrata pogreške putem ove metode je prosjek procjena n testnih pogrešaka. To je prikazano jednadžbom:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \qquad (30)$$

Prednost ove metode, u usporedbi s prethodnom, je to da ova metoda pokazuje manja odstupanja (nepouzdanosti) modela zato što se prilikom određivanja modela koriste gotovo sve vrijednosti dobivene promatranjem. Iz tog razloga ova metoda ne precjenjuje stvarnu stopu

pogreške. Druga prednost je to što se primjenom ove metode više puta na isti skup dobivaju isti rezultati, za razliku od metode skupa za provjeru.

Primjena unakrsne provjere-zadržavanje jedne vrijednosti sa strane zahtjeva provođenje velikog broja računskih operacija. Kako bi se smanjila kompleksnost ovakve metode koristi se jednadžba:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2 \qquad (31)$$

Gdje je $\hat{y}_i$ *i*-ta određena vrijednost iz izvorno određenog najmanjeg kvadrata, a h_i točka velikog utjecaja. Jednadžba *(31)* slična je srednjoj kvadratnoj pogreški, samo što je *i*-ti ostatak podijeljen s $1 - h_i$. Točka velikog utjecaja se nalazi između $1/n$ i 1 i pokazuje vrijednost kojom promatranje utječe na metodu koja ga opisuje. Zato što se ta točka nalazi u tom rasponu vrijednosti, ova jednadžba daje dobre rezultate.

2.4.3. *K*-struka unakrsna provjera

K-struka unakrsna provjera predstavlja alternativu metodi unakrsne provjere sa zadržavanjem jedne vrijednosti sa strane. Ona podrazumijeva nasumičnu podjelu promatranja u k grupa približno jednakih veličina. Prva grupa služi kao skup za provjeru, a ostalih $k - 1$ grupa služi za određivanje modela. Srednja kvadratna pogreška se izračunava pomoću skupa za provjeru. Postupak se ponavlja k puta dok sve grupe jednom ne posluže kao skup za provjeru. K-struka unakrsna provjera dobije se uprosječivanjem ovih vrijednosti:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} SKP_i \qquad (32)$$

Prednost ove metode, u usporedbi s prethodnom, je to što nije kompleksna i nije potrebno provoditi puno računskih operacija. Na pouzdanost ove metode ne utječe raspodjela podataka po grupama.

Cilj unakrsne provjere je odrediti pouzdanost određenog statističkog modela kada se na njega primjenjuju stvarni podaci. Odnosno, želi se odrediti stvarna srednja kvadratna pogreška. U

nekim slučajevima se pak želi odrediti samo najmanja vrijednost točke na krivulji stvarne srednje kvadratne pogreške. U tim slučajevima provodi se unakrsna provjera više statističkih metoda različitih fleksibilnosti s ciljem utvrđivanja koja metoda daje najmanju srednju kvadratnu pogrešku.

2.4.4. Bootstrap

Ova metoda koristi se kvantificiranje nepouzdanosti koju uzrokuju statistički modeli procjene.

U stvarnosti se ne mogu generirati novi uzorci iz izvorne populacije kako bi se odredila točnost procjene. *Bootstrap* omogućuje upravo to. On oponaša stvarne procese kako bi se dobili novi skupovi uzoraka za određivanje točnosti modela tj. omogućuje ponovno uzorkovanje zapažanja izvornog skupa podataka.

Proces ponovnog uzorkovanja kreće uzimanjem n promatranja iz dostupnog skupa podataka kako bi se stvorio *bootstrap* skup podataka Z^{*1}. Uzorkovanje se izvodi sa zamjenom, što znači da se isto promatranje može pojaviti više puta u bootstrap skupu podataka. Ako je promatranje sadržano u Z^{*1}, onda su i X i Y sadržane u njemu. Zatim se koristi Z^{*1} kako bi se odredila bootstrap procjena vrijednosti α, a ta procjena se označava s $\hat{\alpha}^{*1}$. Postupak se ponavlja B puta (za velik broj B) kako bi se dobili B različiti bootstrap skupovi podataka Z^{*1}, Z^{*2},…, Z^{*B} i B procjene α vrijednosti, $\hat{\alpha}^{*1}$, $\hat{\alpha}^{*2}$,…, $\hat{\alpha}^{*B}$. Uz pomoć bootstrap podataka može se izračunati standardna pogreška SE:

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2} \qquad (33)$$

Ova jednadžba služi za procjenu standardne pogreške $\hat{\alpha}$ dobivene iz izvornog skupa podataka.

2.4.1. Kompromis između odstupanja i varijance k-struka unakrsne provjere

K-struka unakrsna provjera, zbog kompromisa između odstupanja i varijance, daje točnije procjene stope pogreške od metode skupa za provjeru i metode unakrsne provjere.

Što se tiče odstupanja, metoda skupa za provjeru može precijeniti stvarnu stopu pogreške zato što skupina podataka za određivanje modela sadrži samo polovicu dostupnih podataka. Metoda zadržavanja jedne vrijednosti sa strane ima najmanja odstupanja jer se svaki put pri određivanju modela koristi $n-1$ promatranja. Kod k-grupne unakrsne provjere javlja se umjereno odstupanje jer u svakom postupku, za određivanje modela, koristi se $(k-1)n/k$ promatranja.

Metoda zadržavanja jedne vrijednosti sa strane ima veću varijancu od metode k-grupne unakrsne provjere iako je $k < n$. Razlog tomu je to što se kod metode zadržavanja jedne vrijednosti sa strane, uprosječuju odzivi n određenih modela, od kojih je svaki određen s gotovo identičnim skupom promatranja. To uzrokuje visoku (pozitivnu) korelaciju između njih. S druge strane, kod k-grupne unakrsne provjere uprosječuju se odzivi k određenih modela koji imaju manju međusobnu korelaciju jer je preklapanje skupina podataka za određivanje modela manje. Zbog toga što visoka korelacija upućuje na veću varijancu, prvi model ima veću varijancu od drugog.

Za k-struku unakrsnu provjeru najčešće se uzimaju vrijednosti $k = 5$ i $k = 10$ jer ne pokazuju velika odstupanja ni varijancu što ih čini najpouzdanijim metodama.

2.5. Odabir linearnog modela i regularizacija

Moguće je da svi prediktori utječu na odziv, ali najrealniji slučaj je onaj, u kojemu samo dio prediktora utječe na njega. Pri analizi koji model sadrži samo te značajne varijable postoji 2^p mogućih rješenja između kojih moramo odabrati model. Stoga, ako p nije mali broj i ako ne možemo ispitati svaki slučaj posebno, moramo naći efikasan način odabira modela koje ćemo ispitati. U tomu nam pomažu:

- Odabir unaprijed. Započinje s nultim modelom-model koji sadrži odsječak ali ne i prediktore. Onda na tom modelu provodimo p jednostavnih linearnih regresija i pridružujemo mu varijable s najmanjom sumom kvadratnih ostataka (RSS). Postupak se nastavlja dok ne dobijemo zadovoljavajući model.
- Odabir unatrag. Započinjemo s modelom koji sadrži sve varijable i zatim izbacujemo onu varijablu s najvećim p iznosom (odnosno ona koja je najmanje značajna za model). Postupak provodimo onoliko puta dok p-vrijednost preostalih varijabli ne padne ispod neke granične vrijednosti.

- Mješovit odabir. On je kombinacija odabira unaprijed i unatrag. Postupak određivanja započinje s modelom bez varijabli, zatim dodajemo jednu po jednu varijablu. Prilikom dodavanja nove varijable provjeravamo *p*-vrijednost svake varijable koju smo već uvrstili u model, ako *p*-vrijednost neke varijable naraste preko zadovoljavajuće granice, tu varijablu izbacujemo iz modela. Postupak se nastavlja dok u modelu nemamo sve varijable s dovoljno malim *p*-vrijednostima, a izvan modela varijable s visokim *p*-vrijednostima.

Odabir unatrag se ne može koristiti ako je $p > n$. Odabir unaprijed može u postupku obuhvatiti varijable koje su nevažne za model. Iz tog razloga najbolje je koristiti mješovit odabir.

2.5.1. Odabir najboljeg podskupa

Za usporedbu modela na temelju testne pogreške, testna pogreška procjenjuje se na temelju:

1. Testna pogreška može se neizravno odrediti tako da se promijeni pogreška učenja kako bi se u obzir uzelo odstupanje koje se javlja zbog preodređivanja modela.
2. Testna pogreška može se izravno odrediti putem skupa za provjeru ili metodom unakrsne provjere.

Za odabir najboljeg podskupa koristi se regresija najmanjih kvadrata za svaku moguću kombinaciju p prediktora, odnosno uklapaju se svi p modeli koji sadrže točno jedan prediktor, svi $\binom{p}{2} = p(p-1)/2$ modeli koji sadrže dva prediktora itd. Nakon toga, provodi se analiza dobivenih modela i zaključi se koji je najbolji. Razvijeni su statistički kriteriji koji se koriste pri odabiru najboljeg podskupa a prikazani su u tablici 2-1 zajedno s jednadžbom i opisom cilja.

Tablica 2-1 Kriteriji za odabih optimalnog modela

Kriterij	Cilj	Jednadžba
Cp–Mallowsov pokazatelj ($\boldsymbol{C_p}$)	Smanjenje	$\boldsymbol{C_p = \frac{1}{n}(RSS + 2d\hat{\sigma})}$
Akaike informacijski kriterij (AIC)	Smanjenje	$\boldsymbol{AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)}$
Bayesov informacijski kriterij (BIC)	Smanjenje	$\boldsymbol{BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)}$
Prilagođeni $\boldsymbol{R^2}$	Povećanje	$\boldsymbol{pr.R^2 = 1 - \frac{RSS/n - d - 1}{TSS/(n-1)}}$

U tablici 2-1 d je broj prediktora, a σ^2 procjena varijance pogreške (ϵ) koja je povezana s mjerom odziva regresijskog modela. Svaki od ovih kriterija dodaje kaznu RSS-u učenja kako bi se prilagodila činjenici da pogreška učenja ima težnju podcjenjivanja testne pogreške. Što je broj prediktora veći, to je kazna veća.

Postupak određivanja najboljeg podskupa provodi se u 3 sljedeća koraka:

1. s M_0 označi se nulti model koji ne sadržava prediktore, već on samo predviđa prosječnu vrijednost uzorka za svako promatranje.
2. za $k = 1, 2, \ldots, p$: treba odrediti sve $\binom{p}{k}$ modele koji sadrže točno k prediktora. Zatim, treba odabrati najbolji od $\binom{p}{k}$ modela i nazvati ga M_k. Najbolji znači da ima najmanji ostatak sume kvadrata (RSS) ili najveći R^2.
3. odabrati najbolji model između $M_0, \ldots, M_p$ pomoću pogreške predviđanja unakrsne provjere, C_p, AIC, BIC, ili prilagođenog R^2.

Što je prostor pretraživanja veći, veće su šanse pronalaska modela koji dobro odgovara podacima za učenje iako oni možda nemaju utjecaj na predviđanje budućih rezultata. Stoga veliki prostor pretraživanja može uzrokovati preodređivanje modela i visoku varijancu pretpostavki koeficijenata. Iz tih razloga koriste se metode postupnog odabiranja kojima se provjeravaju točnosti modela s manjim brojem prediktora.

Iz računskih razloga, postupak odabira najboljeg podskupa naprijed ili unatrag (engl. *Stepwise Forward Selection* i *Stepwise Backward Selection)* ne može se primijeniti kada je broj p prediktorskih varijabli velik.

Postupak postupnog odabiranja unaprijed započinje s modelom koji ne sadrži prediktore. Nakon toga se postupno dodaje jedan po jedan prediktor modelu sve dok model ne sadrži sve prediktore. U svakom koraku, modelu se dodaje varijabla koja najviše poboljšava modela.

Taj postupak provodi se u tri koraka:

1. s M_0 označi se nulti model koji ne sadržava prediktore, nego on samo predviđa prosječnu vrijednost uzorka za svako promatranje.
2. za $k = 1,2,..,p-1$: u obzir treba uzeti sve $p-k$ modele koji povećavaju broj prediktora u M_k za jedan prediktor. Treba odabrati najbolji od $p-k$ modela i

nazvati ga M_{k+1}. Najbolji znači da ima najmanji ostatak sume kvadrata (RSS) ili najveći R^2.

3. odabrati jedan najbolji model između $M_0, \ldots, M_p$ pomoću pogreške predviđanja unakrsne provjere, C_p, AIC-a, BIC-a, ili prilagođenog R^2.

Postupak postupnog odabiranja unatrag predstavlja alternativu za postupak odabira najboljeg podskupa. No, ovaj postupak, za razliku od postupnog odabiranja unaprijed, započinje s punim modelom najmanjih kvadrata i sadrži sve p prediktore. Zatim se iterativnim postupkom, jedan po jedan, uklanja najmanje koristan prediktor u tom koraku.

Taj postupak provodi se u tri koraka:

1. s M_0 označi se nulti model koji ne sadržava prediktore, nego on samo predviđa prosječnu vrijednost uzorka za svako promatranje.
2. za $k = p, p-1, .., 1$: u obzir treba uzeti sve k modele koji sadrže sve osim jednog prediktora u M_k za ukupno $k-1$ prediktora. Treba odabrati najbolji od k modela i nazvati ga M_{k-1}. Najbolji znači da ima najmanji ostatak sume kvadrata (RSS) ili najveći R^2.
3. odabrati jedan najbolji model između $M_0, \ldots, M_p$ pomoću pogreške predviđanja unakrsne provjere, C_p, AIC-a, BIC-a, ili prilagođenog R^2.

Pri usporedbi dva linearna regresijska modela zaključuje se sljedeće:

1. R^2: što model ima veći R^2 to je on pouzdaniji. Također je bitno u obzir uzeti prilagođeni R^2 koji je modificirana verzija R^2 i koji je prilagođen za broj prediktora u modelu. On se povećava samo ako nova vrijednost poboljšava model više, nego što bi se to dogodilo zbog slučajnosti.
2. MSE: ako model ima manju vrijednost MSE to znači da on smanjuje varijancu parametra ϵ i da je taj model pouzdan.
3. F-statistika: što je ona veća, model je pouzdaniji.
4. Ostalo: za procjenu pouzdanosti modela mogu se koristiti i različite druge statistike. Neke od njih su Akaikin informacijski kriterij (AIC) i Bayesianov informacijski kriterij (BIC). Modeli s manjim vrijednostima AIC-a i BIC-a su pouzdaniji.

2.5.2. Regularizacija

U slučajevima analiza velikih podatkovnih skupova s većim brojem varijabli u interesu je naći najmanji mogući podskup koji daje jednaku ili sličnu razinu točnosti. Drugim riječima cilj je pronaći najmanji podskup nezavisnih varijabli koji ima utjecaja na zavisnu. Na ovaj način se pojednostavljuje sam model te se smanjuje parametarski prostor u kojem se provodi analiza, čime se posljedično smanjuje i varijanca regresijskih parametara koje dobivamo običnom metodom najmanjih kvadrata (engl. *ordinary least squares* – OLS). Pri modeliranju se uvijek preferiraju jednostavniji modeli. Dodatna korist pri određivanja najmanjeg mogućeg podskupa je porast interpretativnosti. Porast interpretativnosti je često negativno spregnut s razinom točnosti, tako da pri analizi treba odrediti što je od to dvoje važnije.

Regularizacija je postupak odabira podskupa na način da se smanjuje varijabilnost parametara prema nuli. Korištenjem regularizacije postiže se određena konzistentnost i reproduktivnost u redukciji parametara.

Koriste se dvije vrste regularizacije, hrbatna (engl. *ridge*) i *LASSO* (engl. *Least Absolute Shrinkage and Selection Operator* – LASSO), a razlikuju se u formi regularizacijskog člana, što rezultira time da *LASSO* regularizacija iz regresijskog modela izbacuje one parametre koje smatra da nemaju značaja za predviđanje.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=RSS+\lambda\sum_{j=1}^{p}|\beta_j| \qquad (34)$$

gdje je λ parametar podešenja, a RSS je suma kvadratnih ostataka.

2.6. Regresijska stabla odluke

Stabla odluke su statistički model koji se koristi u procesima odlučivanja. Model je visoko interpretativan, posebno zbog jednostavnih grafičkih prikaza koji su nalik dijagramu tijeka u kojoj svaki unutarnji čvor predstavlja ishod testa na nekom prediktoru, svaka grana dijagrama predstavlja ishod testa, a svaki list predstavlja konačni rezultat tj. odluku. Modeli su u strukturi hijerarhijski i neparametarski, u smislu da ne pretpostavljaju parametarski oblik prediktora i da

oblik stabala nije unaprijed određen, već stablo raste u procesu učenja, zavisno o specifičnom problemu na kojem se provodi učenje.

Algoritmi strojnog učenja stablima odluke smatraju se jednom od najboljih i najčešće korištenih metoda nadziranog učenja. Metode temeljene na stablima odluke omogućuju predviđajuće modele s visokom preciznošću, stabilnošću i lakoćom tumačenja. Za razliku od linearnih modela, imaju visoku razinu točnosti kod nelinearnih pojava te su primjenjivi i kod klasifikacijskih i kod regresijski problema (označavaju se s CART, engl. *Classification and Regression Trees*).

2.6.1. Osnove odlučivanja stablima

Stabla odluke mogu se koristiti za regresijske i klasifikacijske probleme. Proces određivanja stabla odluke provodi se u dva koraka:

1. Podjela prediktorskog prostora, odnosno mogućih vrijednosti $X_1, X_2, \dots, X_p$, u J različitih područja koja se ne preklapaju. Ta područja označavaju se s $R_1, R_2, \dots, R_J$.
2. Svakom promatranju u istom području R_j pridruži se isto predviđanje. To predviđanje predstavlja srednju vrijednost odziva koja su dobivena iz vrijednosti koje služe za određivanje modela.

Područja $R_1, R_2, \dots, R_J$ općenito mogu imati bilo kakav oblik, ali se najčešće uzima da imaju oblik pravokutnika velikih dimenzija. Takav podjela koristi se zbog jednostavne interpretacije. Cilj je pronaći pravokutnike, $R_1, R_2, \dots, R_J$, tako da je suma kvadrata ostataka (RSS) što manja. Postupak određivanja pravokutnika dan je s:

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 \qquad (35)$$

Gdje $\hat{y}_{R_j}$ predstavlja srednji odziv dobiven od vrijednosti koje služe za određivanja modela u j-tom pravokutniku. Računski je nemoguće u obzir uzeti svaku moguću podjelu prostora značajki na J kutija. Iz tog razloga koristi se pristup odozgo prema dolje (pohlepni pristup) koji je poznat kao rekurzivna binarna podjela. U ovom pristupu započinje se od vrha stabla (u tom trenutku sva promatranja pripadaju jednom području) i zatim se prediktorski prostor sukcesivno dijeli a

mjesto podjele karakterizira grananje u dvije grane. Ovaj pristup naziva se pohlepnim jer se prilikom svakog grananja pozornost obraća samo na trenutni korak, a ne na buduće korake i njihove potencijalne oblike koji bi možda rezultirali boljom podjelom.

Kod rekurzivne binarne podjele potrebno je odrediti prediktor X_j i točku podjele s koji će podjelom prediktorskog prostora na dva područja $\{X|X_j < s\}$ i $\{X|X_j \geq s\}$ rezultirati najvećim smanjenjem sume kvadrata ostataka. Odnosno, razmatraju se svi prediktori $X_1, X_2, \dots, X_p$ i njima pripadajuće točke podjele s, koje će rezultirati stablom s najmanjom sumom kvadrata ostataka. Preciznije rečeno, za svaki j i s određuje se par poluravnina:

$$R_1(j,s) = \{X|X_j < s\} \quad i \quad R_2(j,s) = \{X|X_j \geq s\} \tag{36}$$

i traže se vrijednosti j i s koje će minimizirati jednadžbu:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{37}$$

Ovdje $\hat{y}_{R_1}$ predstavlja srednji odziv ulaznih promatranja koja se nalaze u $R_1(j,s)$, a $\hat{y}_{R_2}$ predstavlja srednji odziv ulaznih promatranja koja se nalaze u $R_2(j,s)$.

Postupak se ponavlja određivanjem najboljeg prediktora i najbolje točke grananja za sljedeću podjelu podataka, s ciljem smanjenja sume kvadrata ostataka svakog novonastalog područja. Za razliku od prve podjele, ovdje se umjesto podjele cjelokupnog prediktorskog prostora, dijeli jedno od prije određenih područja. Taj proces ponavlja se dok se ne zadovolji unaprijed određeni kriterij zaustavljanja.

Kada se odrede područja $R_1, R_2, \dots, R_J$, predviđaju se odzivi danih početnih promatranja (uz pomoć njihove srednje vrijednosti) u područjima, u koja ta promatranja pripadaju.

2.6.2. Skraćivanje stabla ili rezidba

Regresijsko stablo može dati pogrešne rezultate kada se na modelu ispituju neke nove vrijednosti, jer vrijednosti koje su služile za njegovo određivanje mogu stvoriti model koji je prenaučen. Uzrok tome je složenost samog stabla. Stoga će stabla s manje podjela (s manje

područja $R_1, R_2, \dots, R_J$) imati manju varijancu (pogrešku), a podaci će se lakše interpretirati. Mala stabla imaju veća odstupanja modela, ali su oni i dalje bolji od složenih stabala s velikom varijancom. Alternativa skraćivanju je grananje stabla onoliko dugo dok smanjenje sume kvadrata ostatka ne padne ispod određene razine. Ova metoda nije pouzdana, jer smanjenje ostatka sume kvadrata kod jedne podjele na početku može biti malog iznosa što zaustavlja proces grananja. U tom slučaju, kasnija grananja koja bi imala značajno smanjenje tog ostatka onda se ne bi uzimala u obzir.

Iz navedenog se da zaključiti da je bolje razviti veliko stablo T_0 i zatim ga skratiti čime se dobiva podstablo. Podstablo se odabire tako da stvarna stopa pogreške bude što manja. Ona se može ispitati putem unakrsne provjere ili skupa za provjeru, ali te metode su prekompleksne jer se mora ispitati pogreška svakog podstabla, a njihov broj je velik. Kako bi se smanjio broj provjera u obzir se mora uzeti manji broj podstabala. Odabir koja podstabla će se ispitati naziva se skraćivanje najslabije karike. Kod tog skraćivanja umjesto svakog podstabla, razmatra se niz stabala indeksiranih pozitivnim parametrom ugađanja α.

Svakoj vrijednosti α pripada podstablo $T \subset T_0$ za koje je jednadžba u nastavku, što je moguće manja:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha|T| \qquad (38)$$

Ovdje $|T|$ označava broj krajnjih čvorova stabla, R_m je pravokutnik (tj. podskup prediktorskog prostora) koji odgovara m-tom krajnjem čvoru, a $\hat{y}_{R_m}$ je predviđeni odziv povezan s R_m-om, odnosno srednjom vrijednosti početnih promatranja u R_m-u. Parametar ugađanja α balansira između kompleksnosti podstabla i njegovog podudaranja s ulaznim podacima. Kada je $\alpha = 0$, podstablo T bit će jednako T_0, a jednadžba *(38)* služit će samo za određivanje pogreške koja se javlja kada se u model uvrštavaju podaci koji su služili za njegovo određivanje. Kako α raste, zbog velikog broja krajnjih čvorova stabla, vrijednost dana jednadžbom *(38)* bit će manja zbog manjeg stabla.

2.6.3. Bootstrap

Metoda *bootstrap* koristi se u slučajevima kada je teško izravno izračunati standardnu devijaciju veličine koje se razmatra. Kod metode stabala javlja se visoka varijanca. *Bootstrap* agregacija, ili *bagging*, smanjuju je, odnosno poboljšavaju stablo odluke.

U skupu n nezavisnih varijabli $Z_1, \dots, Z_n$, gdje svaka varijabla ima varijancu σ^2, srednja vrijednost promatranja $\bar{Z}$ ima varijancu σ^2/n. Iz navedenog se da zaključiti da uprosječivanje skupa promatranja smanjuje varijancu. Stoga, kako bi se poboljšala točnost modela uzimaju se različiti skupovi ulaznih podataka iz skupa pomoću kojih se za svaki stvaraju zasebni modeli predviđanja koji se na kraju uprosječuju. Postupak se matematički zapisuje kao:

$$\hat{f}_{sr}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \qquad (39)$$

Ovaj postupak obično nije jednostavno provesti jer u većini slučajeva različiti skupovi ulaznih podataka nisu dostupni za analizu. Umjesto toga, koristi se *bootstrap* metoda kojom se više puta ponavlja uzimanje uzoraka iz jednog skupa ulaznih podataka. Tom metodom stvara se B različitih *bootstrap* skupova ulaznih podataka. Model se zatim ispituje na B-tom skupu podataka koji je dobiven *bootstrap* metodom kao bi se odredio $\hat{f}^{*b}(x)$. Na kraju se predviđanja uprosječuju kako bi se dobio:

$$\hat{f}_{pak}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \qquad (40)$$

Opisan postupak naziva se *bagging*, poboljšava predviđanja i koristi se kod raznih regresijskih metoda, ali najviše kod stabala odlučivanja. Kako bi se *bagging* metoda primijenila na regresijsko stablo, prvo je potrebno stvoriti B regresijskih stabala upotrebom B skupova ulaznih podataka koji su dobiveni *bootstrap* metodom, a zatim je potrebno uprosječiti dobivena predviđanja. Dobivena stabla nisu skraćena i zato imaju visoku varijancu i malo odstupanje. Uprosječivanjem tih B stabala, smanjuje se varijanca i tako se poboljšava točnost modela zbog spajanja velikog broja stabala.

2.6.3.1. Procjena pogreške pomoću promatranja neiskorištenih za *bagging*

Procjena stvarne pogreške modela dobivenih *bagging* metodom vrlo je jednostavan postupak jer nije potrebno provoditi unakrsnu provjeru ili metodu skupa za provjeru. Kao što je već rečeno, *bagging* metoda zasniva se na određivanju stabala pomoću skupova promatranja dobivenih *bootstrap*-om. Može se pokazati da se za svako stablo koristi dvije trećine dostupnih promatranja. Ostatak promatranja, koja nisu iskorištena za *bagging* stabala, nazivaju se promatranja neiskorištena za *bagging*. Odziv *i*-tog promatranja može se predvidjeti pomoću stabala kod kojih to promatranje nije iskorišteno za *bagging*. To će rezultirati s $B/3$ predviđanja *i*-tog promatranja. Kako bi se dobio jedan odziv, može se uzeti srednja vrijednost svih predviđanja. Postupkom se dobiva jedno predviđanje skupa promatranja koja nisu iskorištena za *bagging i*-tog promatranja. Kada se ta vrijednost dobije za *n*-promatranja, izračunava se ukupna srednja kvadratna pogreška promatranja koja nisu iskorištena za *bagging* (za regresijske probleme) ili greška klasifikacije (za klasifikacijske probleme). Ta rezultirajuća pogreška valjana je, za procjenu stvarne pogreške *bagging* modela, zato što je odziv svakog promatranja predviđen koristeći samo stabla za čije određivanje nije korišteno to promatranje.

2.6.4. Usporedba metoda stabala i linearnih modela

Linearnom regresijom pretpostavlja se da se neka ovisnost između prediktora i odziva ponaša prema modelu:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \qquad (41)$$

Regresijska stabla pretpostavljaju model oblika:

$$f(X) = \sum_{m=1}^{M} c_m \cdot 1_{(X \in R_m)} \qquad (42)$$

Gdje $R_1, R_2, \ldots, R_M$ predstavljaju podjelu prostora značajki.

Ako postoji linearna povezanost između prediktora i odziva, za određivanje modela koristi se linearna regresija. Ako pak postoji nelinearna veza preporučuje se korištenje regresijskih stabala, ali i kod slučajeva u kojima se problem želi vizualizirati.

Prednosti metoda stabla odluke su:

- Lako su shvatljiva.
- Odluke se donose na način koji je blizak ljudskom procesu donošenja odluka.
- Prikazuju se grafički, što omogućuje dobru interpretabilnost.
- Lako se koriste za kvalitativne prediktore, bez potrebe stvaranje fiktivne varijable.

Nedostaci metoda stabla odluke su:

- Mala preciznost u usporedbi se drugim regresijskim i klasifikacijskim metodama.
- Ostvaruje se nerobustan model što može uzrokovati da mala promjena podataka uzrokuje veliku promjenu izgleda finalnog stabla.

Iako *bagging* povećava točnost modela, takav model teško je interpretirati, jer je model nemoguće prikazati pomoću jednog stabla i nije jasno koje su varijable bitne za postupak, ali mogu se odrediti najvažniji prediktori upotrebom sume kvadrata ostataka za *bagging* regresijska stabala. Najvažniji prediktori imat će velike iznose sume kvadrata.

2.7. Slučajne šume

Metoda slučajnih šuma (engl. *random forest*) je metoda bazirana na metodi stabla odluke, u principu se sastoji od nekog broj stabala odluke koji čine šumu, s razlikom da se izgradnja stabala temelji na *bagging* metodi tako da se svako stablo definirano nasumičnim izborom varijabli koje se ne moraju nužno ponavljati niti u korijenu niti u grananju.

Slučajne šume predstavljaju poboljšanje u odnosu na *bagging* metodu tako da stvara malu dekorelaciju između stabala. Kao i kod metode *bagging*, u ovoj metodi stabla grade se pomoću ulaznih uzoraka dobivenih *bootstrap*-om. Razlika je u tome da se, prilikom grananja stabla, kao uvjet dijeljenja razmatra prediktor iz slučajnog uzoraka m prediktora koji je dio skupa svih prediktora p. Jedan prediktor se koristi za jedno grananje, a kod svakog novog grananja koristi se novi uzorak prediktora. Uzorak prediktora m koji se uzimaju u obzir pri grananju računa se prema formuli $m \approx \sqrt{p}$.

Drugim riječima, prilikom stvaranja slučajnog skupa stabala pri svakom grananju algoritmu nije dopušteno razmatrati većinu dostupnih prediktora. Razlog tome je taj što bi model uvijek koristio dominantne parametre ukoliko mu se ne bi nametnulo da mora razmatati sve parametre, zbog čega bi većina stabala biti slična a odzivi bi bili visoko korelirani. Prosječna vrijednost visoko korelirani veličina dat će veliku varijancu što uzrokuje smanjenje pouzdanosti modela.

Metoda slučajnih šuma rješava taj problem tako što, za svaku podjelu, na raspolaganje stavlja samo određeni dio prediktora. Stoga, u prosjeku $(p-m)/p$ podjela u obzir neće uzeti dominantan prediktor, nego i slabije prediktore. Taj proces može se smatrati dekorelacijom stabala što čini prosjek dobivenih stabala manje varijabilnim i pouzdanijim.

2.7.1. Boosting

Boosting je tehnika koja se primjenjuje za povećanje točnosti raznih metoda. Postupak je sličan *bagging* metodi. Razlikuju se po tome što se stabla kod *boosting* metode tvore uzastopno jer se svako stablo tvori pomoću informacija od prethodno dobivenih stabala. Kod *boosting* metode podaci se ne dobivaju *bootstrap* metodom, nego se svako stablo stvara na temelju modificirane verzije originalnog skupa podataka. Postupak provedbe ove metode dan je u nastavku:

1. Postaviti da je $\hat{f}(x) = 0$ i $r_i = y_i$ za bilo koji i skupa ulaznih podataka.
2. Za $b = 1,2,\dots,B$, treba ponavljati sljedeći postupak:
    a. Odrediti stablo $\hat{f}^b$, s d dijeljenjima $(d+1\ krajnji\ čvor)$, pomoću ulaznih podataka (X, r).
    b. Ažurirati $\hat{f}$ dodavanjem smanjenje verzije novog stabla:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{43}$$

    c. Ažurirati ostatke:

$$r_i \leftarrow r_i + \lambda \hat{f}^b(x_i) \tag{44}$$

3. Izvesti *boosting* model:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \qquad (45)$$

Cilj ove metode nije određivanje velikog stabla za velik broj ulaznih podataka. Umjesto toga, s obzirom na trenutni model, određuje se stablo uz pomoć ostataka modela, tj., za određivanje stabla koriste se ostaci, a ne ishod Y. Zatim se dodaje to novo stablo odluke u funkciju kako bi se ažurirali ostaci. Svako od tih stabala određeno je s parametrom d. Takva stabla nisu velika i imaju samo nekoliko krajnjih čvora. Stvaranjem malih stabala uz pomoć ostataka, postupno se poboljšava $\hat{f}$ u područjima gdje daje neprecizne rezultate. Parametrom λ dobivaju se različiti oblici stabala kako bi oni što bolje odgovarali dobivenim ostacima. Iz ovog je vidljivo da je kod *boosting* metode tvorba stabala ovisna o postojećim stablima.

Boosting ima tri parametra za podešavanje (ugađanje):

1. Broj stabala B. Ako postoji veliki broj stabala *boosting* neće rezultirati zadovoljavajućim modelom. Za odabir broja stabala koristi se metoda unakrsne provjere.
2. Parametar smanjenja λ je mali pozitivan broj. Njime se kontrolira brzina provedbe metode *boosting*. Uobičajeno, za njegov iznos, uzimaju se vrijednosti 0,01 ili 0,001, a to ovisi o vrsti problema. Mali λ zahtjeva postojanje velikog broja B kako bi model bio precizan.

Broj dijeljenja d svakog stabla kontrolira složenost *boosting* metode. Kada je $d = 1$ stablo ima samo jedno dijeljenje te se naziva panj. Općenito, d je dubina interakcije i upravlja redoslijedom interakcija pojačanog modela, zato što d razdioba može uključivati najviše d varijabli.

2.8. Stroj potpornih vektora

Metoda stroja potpornih vektora (engl. *support vector machines* – SVM) je prvotno razvijena za klasifikacijske probleme, ali je pogodna i za regresijske probleme [84]. Prednosti kod primjene na regresijske probleme ima zbog svoje robustnosti i niske osjetljivosti na stršeće vrijednosti [85].

Metoda potpornih vektora proširenje je postupka klasifikatora potpornih vektora kod kojeg se povećava prostor značajki upotrebom jezgri.

Klasifikator potpornih vektora klasificira stvarna promatranja na temelju njihova položaja u odnosu na hiperravninu. Hiperravnina se određuje tako da se većini promatranja točno odredi razred u koji pripadaju, ali se dopušta da se neka promatranja krivo klasificiraju. Taj klasifikator je rješenje optimizacijskog problema:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M}{\text{povećava}} \quad M \tag{46}$$

$$podložnih \sum_{j=1}^{p} \beta_j^2 = 1 \tag{47}$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M(1 - \epsilon_i) \tag{48}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \tag{49}$$

Rješenje problema klasifikatora potpornih vektora uključuje samo unutarnje produkte promatranja. Unutarnji produkt dva *r*-vektora a i b definiran je s $\langle a, b\rangle = \sum_{i=1}^{r} a_i b_i$. Stoga, unutarnji produkt dvaju promatranja x_i i x_i', se može zapisati kao:

$$\langle x_i, x_{i'}\rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{50}$$

Može se pokazati da:

- Linearni klasifikator potpornih vektora može se prikazati kao:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i\rangle \tag{51}$$

gdje ima n parametara α_i, $i = 1, \ldots, n$ jedan za svako ulazno promatranje.

- Kako bi se odredili parametri $\alpha_1, \ldots, \alpha_n$ i β_0, potrebni je imati $\binom{n}{2}$ produkata $\langle x_i, x_{i'} \rangle$ između svih parova ulaznih promatranja.

Kako bi se odredila funkcija $f(x)$ potrebno je izračunati unutarnji produkt između nove točke x i svake ulazne točke x_i. Vidi se da α_i nije nula samo za potporne vektore, a za sva ostala ulazna promatranja je nula. Ako je $\mathcal{S}$ skup svih indeksa potpornih točaka, funkcijsko rješenje dobiveno u jednadžbi *(51)*, može se zapisati kao:

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle \qquad (52)$$

Ta jednadžba sadrži manji broj članova od jednadžbe *(51)*. Zaključno rečeno, za linearni klasifikator $f(x)$ i za izračun njegovih koeficijenata, potrebno je odrediti unutarnje produkte.

Kada se unutarnji produkt iz jednadžbe *(50)* pojavi u *(51)* ili u proračunu rješenja klasifikatora potpornih vektora, zamjenjuje se s generaliziranim oblikom unutarnjeg produkta:

$$K(x_i x_{i'}) \qquad (53)$$

Gdje K predstavlja funkciju koja se naziva jezgra. Jezgra je funkcija koja kvantificira sličnost dvaju promatranja. Može se uzeti oblik jednadžbe:

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \qquad (54)$$

koji daje klasifikator potpornih vektora. Ta jednadžba poznata je kao linearna jezgra zato što je dobiveni klasifikator potpornih vektora linearan u svim svojim značajkama. Linearna jezgra kvantificira sličnost para promatranja upotrebom Pearsonove (standardne) korelacije. Može se uzeti i drugačiji oblik jednadžbe *(53)* pa se onda može svaki dio $\sum_{j=1}^{p} x_{ij} x_{i'j}$ zamijeniti s veličinom:

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d \quad (55)$$

Ovaj oblik je poznat pod nazivom polinomna jezgra stupnja d, gdje je d pozitivan broj. Upotrebom takve jezgre, gdje je $d > 1$, dobiva se fleksibilnija granica odluke. Kada se klasifikator potpornih vektora koristi u kombinaciji s nelinearnom jezgrom kao što je to *(55)*, onda se dobiva metoda potpornih vektora. U tom slučaju nelinearna funkcija ima oblik:

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i) \quad (56)$$

Polinomna jezgra dana jednadžbom *(55)* jedan je od mogućih oblika nelinearnih jezgri. Drugi popularni oblik je radijalna jezgra oblika:

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \quad (57)$$

gdje je γ pozitivna konstanta.

Ako su stvarna promatranja $x^* = (x_1^*, \dots, x_p^*)^T$ daleko od ulaznog promatranja x_i (Euklidska udaljenost), onda će $\sum_{j=1}^{p} (x_j^* - x_{ij})^2$ biti velik, a $K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_j^* - x_{ij})^2)$ malen. To znači da u jednadžbi *(56)*, x_i neće imati značajnu ulogu u funkciji $f(x^*)$. Predviđen razred promatranja x^* određuje se pomoću predznaka $f(x^*)$. Stoga, ulazna promatranja koja su daleko od x^*, neće imati značajnu ulogu u određivanju razreda od x^*. To ukazuje da radijalna jezgra ima lokalni karakter jer samo okolna ulazna promatranja imaju utjecaj na njeno klasificiranje.

Prednost upotrebe jezgri, u usporedbi sa samim povećavanjem prostora značajki, je proračunska jednostavnost jer je samo potrebno izračunati $K(x_i, x_{i'})$ za sve $\binom{n}{2}$ različite parove i, i'.

Jedan u odnosu na jedan ili klasifikacija svih parova stvara $\binom{K}{2}$ metoda potpornih vektora koje uspoređuju par razreda. Potom se stvarna promatranja klasificiraju pomoću $\binom{K}{2}$ klasifikatora i

bilježi se koliko puta je stvarno promatranje dodijeljeno svakom od K razreda. Na kraju se provodi završna klasifikacija koja dodjeljuje stvarno promatranje razredu kojem je ono najviše puta dodijeljeno u $\binom{K}{2}$ parovima.

Kod klasifikacije jedan u odnosu na sve, određuje se K metoda potpornih vektora usporedbom jednog od K razreda s ostalim $K-1$ razredima. Ako se s $\beta_{0k}, \beta_{1k}, \dots, \beta_{pk}$ označe parametri dobiveni metodom potpornih vektora, odnosno usporedbom *k*-tog razreda s ostalima, i ako se s x^* označi promatranje. Onda se promatranje dodjeljuju razredu za koji je $\beta_{0k} + \beta_{1k}x_1^* + \beta_{2k}x_2^* + \cdots + \beta_{pk}x_p^*$ najveći jer to označava visok stupanj pouzdanosti da to promatranje pripada *k*-tom razredu.

3. PROVEDENO ISTRAŽIVANJE

3.1. Analizirani skupovi podataka

Sukladno Pravilniku o načinu raspodjele i obračunu troškova za isporučenu toplinsku energiju [56] definirani su različiti modeli za raspodjelu isporučene toplinske energije za grijanje prostora na zajedničkom mjerilu toplinske energije:

- model 1EG koji se temelji na udjelu snage stana u ukupnoj priključnoj snazi u zgradi,
- model 2EG koji se temelji na udjelu površine stana u ukupnoj površini zgrade i
- model 3EG koji se temelji na udjelu broja impulsa u stanu u ukupnom broju impulsa u zgradi.

Tako će se u daljnjim analizama te u notaciji pojedinih varijabli ili interpretacijama i prikazima koristiti termini ili nastavci „2EG“ ili „3EG“ pri označavanju načina obračuna pojedinih stanova. Model 1EG nije uključen u razmatrane podatkovne skupove i u stvarnosti se rijetko pojavljuje kao model obračuna u stanovima u Hrvatskoj.

Kako bi se mogle uspoređivati potrošnje u različitim vremenskim periodima, potrebno ih je normalizirati. Stvarna godišnja potreba za toplinom te omjer meteorološkog i klimatološkog stupanj-dana grijanja daju normaliziranu godišnju potrebu za toplinskom energijom.

Postupak određivanja normalizirane godišnje potrebe za toplinskom energijom, uz korištenje godišnjih toplinskih potreba i stupanj-dana grijanja dan je izrazom [34]:

$$Q_n = \frac{Q_a}{\alpha + \beta(1-\alpha)} \tag{58}$$

$$\alpha = Q_0/Q_n \tag{59}$$

$$\beta = G_a/G_n \tag{60}$$

gdje je:

Q_n – potrošnja topline tijekom prosječne godine [kWh]

Q_a – potrošnja topline tijekom aktualne godine [kWh]

Q_0 – vremenski neovisna komponenta potrošnje topline kao što je priprema tople vode i toplinski gubici prilikom distribucije topline [kWh]

G_n – meteorološki stupanj-dan grijanja za prosječnu godinu

G_a – meteorološki stupanj-dan grijanja za aktualnu godinu

Konvencija o stupanj-danu grijanja razlikuje se od zemlje do zemlje. U Hrvatskoj se uobičajeno stupanj-danom grijanja predstavlja umnožak broja dana grijanja s temperaturnom razlikom između dogovorene srednje unutarnje temperature zraka u grijanom prostoru (najčešće 20 °C) i temperature vanjskog zraka, pri čemu se u proračun uzimaju samo oni dani u godini kod kojih je vanjska temperatura zraka niža od 12 °C [86].

3.1.1. Podaci iz sustava za naplatu

Analizirane kvantitativne varijable iz sustava za naplatu toplinarskih tvrtki na dva distribucijska područja sastoje se od 20 varijabli u 3.845.310 opservacija, prema opisu pojedinih varijabli je kako slijedi u tablici 3-1.

Tablica 3-1 Opis analiziranih kvantitativnih varijabli

R.br.	Naziv varijable	Opis varijable	Mjerna jedinica
1.	*Temp*	Srednja vanjska temperatura	°C
2.	*Povrsina*	Površina pojedinog krajnjeg kupca	m2
3.	*Povrsina3EG*	Suma površina svih stanova u zgradi na modelu 3EG [56]	m2
4.	*Povrsina2EG*	Suma površina svih stanova u zgradi na modelu 2EG [56]	m2
5.	*ToplinaOMM*	Ukupno izmjerena ili obračunata isporučena toplina za grijanje prostora na obračunskom mjernom mjestu	MWh
6.	*PTV_OMM*	Ukupno izmjerena ili obračunata isporučena toplina za pripremu potrošne tople vode na obračunskom mjernom mjestu	MWh
7.	*EnergijaOMM*	Ukupno izmjerena isporučena toplina na obračunskom mjernom mjestu	MWh
8.	*PROSJEK_TV*	Količina topline za PTV utvrđena računski prosjekom za toplu vodu	MWh
9.	*OSOS2EV*	Ukupan broj članova kućanstva	-
10.	*BrojOsoba*	Broj članova kućanstva kod krajnjeg kupca	-
11.	*ImpusliOMM*	Ukupan broj impulsa izmjeren u zgradi	-
12.	*ImpulsiKK*	Broj impulsa izmjeren kod krajnjeg kupca	-
13.	*UR*	Izabrani udio impulsa sa razdjelnika u obračunskoj formuli [56]	-
14.	*UPOV*	Izabrani udio površine pojedinog krajnjeg kupca u obračunskoj formuli [56]	-
15.	*SnagaKK*	Instalirana toplinska snaga kod krajnjeg kupca	kW
16.	*ToplinaKK*	Toplina za grijanje prostora krajnjeg kupca	kWh
17.	*PTV_KK*	Toplina za pripremu potrošne tople vode kod krajnjeg kupca	kWh
18.	*EnegijaKK*	Ukupna toplina alocirana krajnjem kupcu	kWh
19.	*ToplinaSpec*	Specifična potrošnja topline za grijanje kod krajnjeg kupca	kWh/m2
20.	*KorekcijskiFaktor*	Korekcijski faktori prema važećem zakonu	%

Analizirane kvalitativne varijable iz sustava za naplatu sastoje se od 10 varijabli u 3.845.310 opažanja, kako je prikazano u tablici 3-2.

Tablica 3-2 Opis analiziranih kvalitativnih varijabli

R.br.	Naziv varijable	Opis varijable
1.	*DistribucijskoPodrucje*	Zemljopisno distribucijsko područje
2.	*Adresa*	Adresa građevine
3.	*KrajnjiKupac*	Oznaka krajnjeg kupca
4.	*OMM*	Oznaka obračunskog mjernog mjesta, zgrade
5.	*MjesecGodina*	Razmatrano obračunsko razdoblje, mjesec/godina
6.	*GrupaKK*	Oznaka grupe krajnjeg kupca [1]
7.	*ModelSnaga*	Oznaka modela obračuna toplinske snage
8.	*ModelEnergija*	Oznaka modela obračuna topline
9.	*ModelToplaVoda*	Oznaka modela obračuna potrošne tople vode
10.	*CB_PROSJEK_TV*	Prosjek potrošne tople vode

Kvalitativne varijable nemaju kontinuiranu i mjerljivu vrijednost već opisuju neku karakteristiku, npr. varijabla *ModelEnergija* može poprimiti vrijednosti „2EG“ ili „3EG“, oznaka *KrajnjegKupca* je unikatna za svakog krajnjeg kupca, varijabla *MjesecGodina* označava obračunsko razdoblje, npr. siječanj 2019. i sl. Ovakve varijable mogu se koristiti u regresijskim i drugim provedenim metodama u ovom istraživanju.

Podaci iz sustava za naplatu prikupljeni su za dva distribucijska područja.

.

[1] Prema internoj klasifikaciji tvrtki u sektoru, npr. http://www.hep.hr/toplinarstvo/krajnji-kupci/oznake-krajnjih-kupaca/31

3.1.2. Podatkovni skup za razinu stana

Za izradu predviđajućih modela i provedbu analize potrošnje topline u stanovima na godišnjoj razini, iz čitavog podatkovnog skupa određuje se podatkovni podskup „*psStan*“, koji sadržava sljedeće varijable iz tablica 3-1 i 3-2:

1. *Godina*
2. *KrajnjiKupac*
3. *OMM*
4. *SnagaKK*
5. *Povrsina*
6. *GrupaKK*
7. *KorekcijskiFaktor*
8. *Temp*
9. *ModelEnergija*

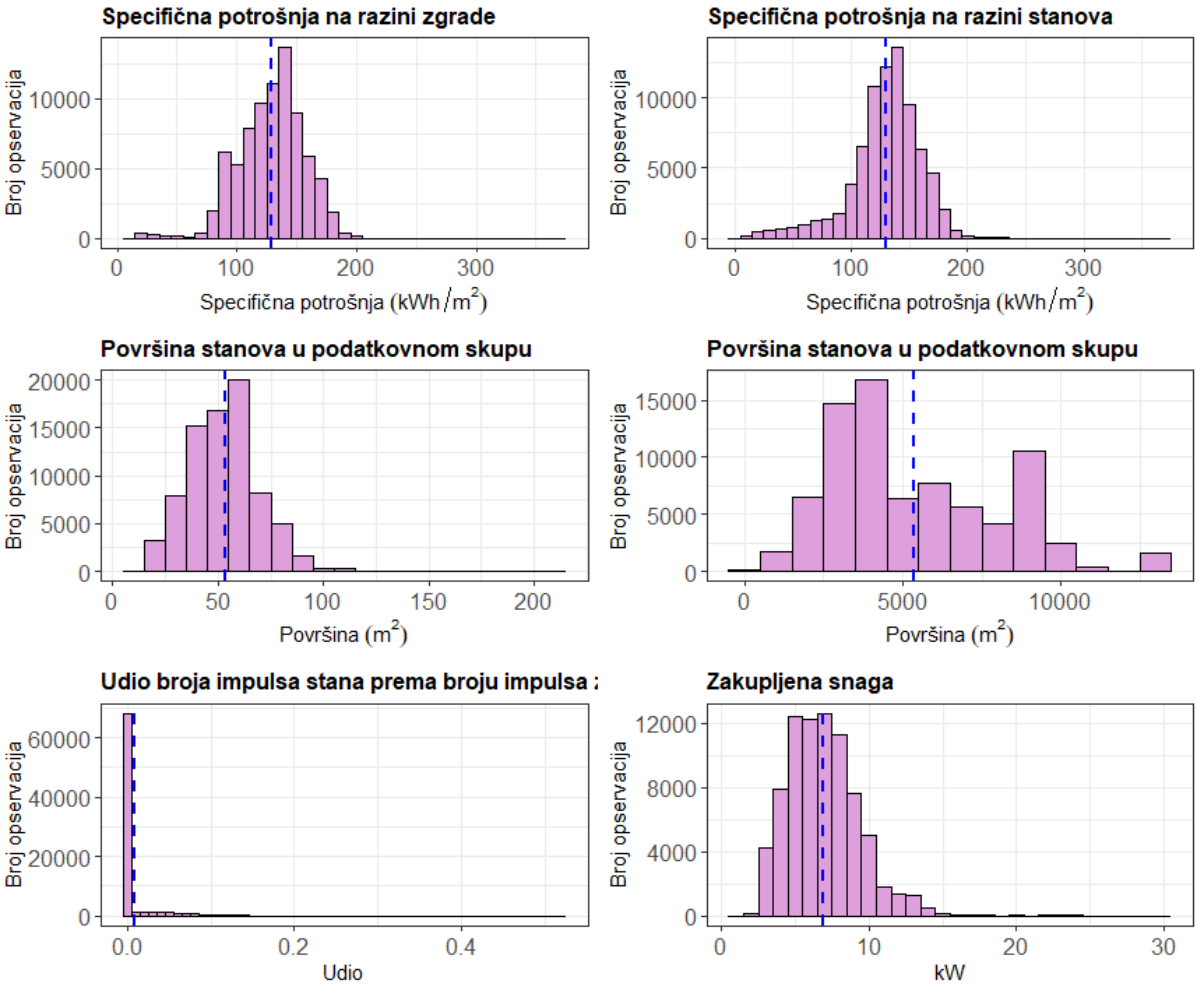
Dodatno su u podatkovni skup uključene ove izvedene varijable više osnovnih varijabli iz sustava za naplatu:

1. *SpecTopSDG* – kontinuirana varijabla, predstavlja omjer između varijable *ToplinaKK* i *Povrsina*, svedeno na stupanj-dan grijanja u razmatranom razdoblju. Mjerna jedinica je kWh/m^2.
2. *PovZgrada* – kontinuirana varijabla, predstavlja ukupnu površinu zgrada te je zbroj varijabli *Povrsina2EG* i *Povrsina3EG*. Mjerna jedinica je m^2.
3. *SpecTopOMM_SDG* – kontinuirana varijabla, predstavlja omjer varijabli *ToplinaOMM* i *PovZgrada*. Mjerna jedinica je kWh/m^2.
4. *StopaUgr* – kontinuirana varijabla, predstavlja stopu ugradnje individualnog mjerenja u pojedinoj zgradi i iznosi omjer između varijable *Povrsina3EG* i *PovZgrada*. Mjerna jedinica je m^2.
5. *Imp* – bezdimenzijska kontinuirana varijabla koja označava omjer varijabli *ImpulsiKK* i *ImpulsiOMM*.

Sumarna deskriptivna statistika za ovaj podatkovni skup dana je u tablici 3-3 i na slici 3-1.

Tablica 3-3 Sumarna statistika za *psStan*

	Min	Donji kvartil	Median	Srednja vrijednost	Gornji kvartil	Maks
SpecTopOMM_SDG	7,961	111,344	130,39	129,024	147,861	366,662
SnagaKK	0,879	5,213	6,744	6,947	8,387	30,021
StopaUgr	0	0	0	0,1597	0	1
PovZgrada	50,62	3317,03	4475,42	5365,06	7467,96	13345,1
Povrsina	6,38	39,46	54,25	53,15	62,09	212,8
Imp	0	0	0	0,008143	0	0,51921
SpecTopSDG	0	116,3	133,1	129,7	149,5	366,7
KorekcijskiFaktor	0	0	0	16,1	0	100
Temp	4,612	5,292	5,587	5,812	5,983	7,952
ToplinaKK	0	717	1130	1296	1655	295000
Povrsina2EG	0	395,1	3916,8	4169,1	7293,8	15475,5
Povrsina3EG	0	0	0	1599,34	21,26	14972,7
ToplinaOMM	0	75,33	125	140,54	192,49	584
ImpulsiKK	0	0	0	132	0	29488
ImpulsiOMM	0	0	0	17350	0	647301
ModelEnergija	2EG:67333	3EG:11564				
GrupaKK	A:43568	L: 217	N:35112			



Slika 3-1 Histogramski prikaz raspodjele izabranih varijabli za *psStan*

3.1.3. Podatkovni skup za razinu zgrade i više razine

Za izradu predviđajućih modela i provedbu analize potrošnje topline za razinu potrošnje u zgradama na godišnjoj razini, iz čitavog podatkovnog skupa se određuje podatkovni podskup „*psZgrada*“, koji sadrži sljedeće varijable iz tablica 3-1 i 3-2:

1. *Godina*
2. *OMM*
3. *Temp*
4. *DistribucijskoPodrucje*

Dodatno su u podatkovni skup uključene ove izvedene varijable više osnovnih varijabli iz sustava za naplatu:

1. *PovZgrada* – kontinuirana varijabla, predstavlja ukupnu površinu zgrada, te je zbroj varijabli *Povrsina2EG* i *Povrsina3EG*. Mjerna jedinica je m^2.
2. *SpecTopOMM_SDG* – kontinuirana varijabla, predstavlja omjer varijabli *ToplinaOMM* i *PovZgrada*. Mjerna jedinica je kWh/m^2.
3. *StopaUgr* – kontinuirana varijabla, predstavlja stopu ugradnje individualnog mjerenja u pojedinoj zgradi i iznosi omjer između varijable *Povrsina3EG* i *PovZgrada*. Mjerna jedinica je m^2.
4. *UgradilaZgr* - oznaka da li je *StopaUgr* veća od nekog broja tj. da li je zgrada krenula s ugradnjom ili ne. Kvantitativna varijabla s vrijednostima „ugradili“ i „nisu ugradili“.
5. *SpecTopDP* - kontinuirana varijabla, predstavlja omjer varijabli $\sum ToplinaOMM$ i $\sum PovZgrada$ na distribucijskom području. Mjerna jedinica je kWh/m^2.

Sumarna deskriptivna statistika za ovaj podatkovni skup dana je u tablici 3-4.

Tablica 3-4 Sumarna statistika za *psZgrada*

	Min	**Donji kvartil**	**Median**	**Srednja vrijednost**	**Gornji kvartil**	**Maks**
SpecTopDP	22,8	105,8	127,9	126,7	145,2	334,3
Temp	4,287	5,292	5,587	5,797	5,983	7,952
StopaUgr	0,000	0,000	0,000	0,203	0,000	1,000
PovZgrada	264,6	2675,5	4100,5	4611,0	6351,1	13345,1
UgradilaZgr	nisuUgradili:1412		ugradili : 857			
DistPodrucje	DP1:1820		DP2: 490			

3.2. Analiza podataka dobivenih intervjuima i upitnicima

Tijekom analize učinaka provedbe individualnog mjerenja u sustavima daljinskog grijanja u Hrvatskoj, s ciljem rješavanja problema opisanih u prethodnom tekstu, pristup anonimiziranim podacima o naplati odobren je za period od 2011. do 2017. godine. U 2018. dostavljeni su upitnici 154 stana s jednostavnim kratkim pitanjima s ciljem dobivanja informacija o pretpostavljenim utjecajnim čimbenicima na potrošnju topline za koje se ne mogu naći podaci u podacima o naplati. U bliskoj suradnji sa službenim predstavnikom stanara, upitnici su podijeljeni stanarima, a ukupno 51 stan sudjelovao je u intervjuima, što čini udio od 33,55 %. Nadalje, predstavnik stanara dostavio je anonimne podatke mjerenja za sva 154 stana. Upitnici nisu bili anonimizirani i svi su sudionici morali dati svoj pristanak za sudjelovanje u ovom znanstvenom istraživanju. Iz podataka o naplati za stanove sudionika dobiveni su svi relevantni tehnički i mjerni podaci za zgradu. Uparivanje podataka o naplati s anonimnim podacima mjerenja dobivenih od predstavnika stanara daje skup podataka za dio analize grupiranjem prikazane u ovom radu.

Ankete su provedene na 3 zgrade (označene s A, B i C), odnosno za tri različita obračunska mjerna mjesta, koje su prikazane na slici 3-2 i u tablici 3-5.



Slika 3-2 Zgrade u kojima su provedeni intervjui i upitnici

Tablica 3-5 Osnovni podaci o analiziranim zgradama

Oznaka zgrade	A	B	C
Ukupna stambena površina	3383,18	3500,04	1880,53
Ukupan broj stanova	61	61	32
Ukupan broj stanova s RTTE	58	61	31
Ukupna površina stanova s RTTE [m^2]	3246,79	3500,04	1793,56
Ukupna površina stanova bez RTTE [m^2]	136,39	0	86,97
Instalirana snaga stanova sa RTTE [kW]	410,731	394,260	227,389
Instalirana snaga stanova bez RTTE [kW]	17,254	0	11,026
Ukupan broj stanara	111	128	70
Broj katova	6	6	5

Svaka od podgrupa priključena je na zasebnu toplinsku podstanicu i ima po jedno mjerilo toplinske energije. U ovom radu napravljena je analiza za navedene tri zgrade za 2016. i 2017. godinu.

Pitanja u anketi formirana su prema primjeru sa slike 3-3, a odnose se na:

1. Informacije o starosnoj strukturi i broju stanara: ovo poglavlje razmatralo je pitanje o starosnoj strukturi stanara u godinama prije i nakon ugradnje RTTE. Smatra se da ta informacija utječe na pretpostavku da određene dobne skupine (pogotovo djeca mlađa od 7 godina i umirovljenici) provode više vremena u stanu nego dobne skupine školske djece ili studenata i zaposlenih. Osim toga, pretpostavka je da je željena razina toplinske ugodnosti za dobnu skupinu djece mlađe od 7 godina veća nego za ostale skupine. Ovo pitanje postavljeno je za godine 2010.-2017., dok je 2015. bila godina ugradnje RTTE u ove zgrade. Kategorije dobnih skupina definirane su kao:
   - Dob 0-6 godina
   - Dob 7-18 godina
   - Dob 19-65 godina
   - Dob 65 i više godina
2. Informacije o katnosti traže se radi procjene utjecaja na potrošnju lokacije određenog stana u zgradi.
3. Informacije o dnevnom vremenu boravka: ovaj odjeljak grupiran je u tri skupine - u stanu cijeli dan, u stanu veći dio dana i vrlo rijetko.
4. Informacije o postojanju renoviranih prozora.
5. Informacije o tipu prozora: drveno okno dvostruko staklo, PVC okno dvostruko staklo, aluminijsko okno dvostruko staklo.

6. Razina poželjne toplinske ugodnosti.
7. Broj negrijanih prostorija: ako postoji neka prostorija koja se ne grije prema želji stanara..
8. Prosječna dnevna stopa ventilacije: posljednje pitanje odnosi se na utjecaj ventilacije stana na potrošnju i pretpostavlja se da je parametar ponašanja znatnog utjecaja.



UPITNIK vezan za potrošnju topline u Vašoj zgradi

***1.)* Popunite tablicu na način da u kvadratić napišete broj članova kućanstva određene dobi u odgovarajućoj godini** *(Ako su 2010. godine u stanu bila 3 člana u dobi od 19-65 godina, u kvadratić se piše broj 3)*:

	0-6 godina	7-18 godina	19-65 godina	65 godina i više
2010. godine	1		2	
2011. godine	1		2	
2012. godine	1		2	
2013. godine	2		2	
2014. godine	2		2	
2015. godine	1	1	2	
2016. godine	1	1	2	
2017. godine	2	1	2	

2.) Kat: 4.

3.) Vrijeme prosječnog boravka u stanu *(zaokružiti odgovor koji najviše odgovara Vašem boravku u stanu)*:

☐ u stanu cijeli dan ☒ u stanu veći dio dana ☐ u stanu jako rijetko

4.) Da li je provedena adaptacija originalnih prozora u Vašem stanu:

☒ DA ☐ NE *(ukoliko je odgovor NE, prijeđite na 6.pitanje)*

5.) O kojem tipu prozora se radi, ako je odgovor na prošlo pitanje, pitanje 4. bilo DA?:

☐ Dvostruko ostakljenje (starija drvena stolarija)
☒ Dvostruko ostakljenje (PVC stolarija)
☐ Trostruko ostakljenje (PVC stolarija)
☐ Dvostruko ostakljenje (aluminijska stolarija)
☐ Trostruko ostakljenje (aluminijska stolarija)

6.) Koja razina toplinske ugodnosti Vam odgovara:

☐ hladno ☐ umjereno ☒ toplo

7.) Upišite broj prostorija u stanu u kojima većinom ne „palite" radijatore : 1

8.) Prosječno vrijeme prozračivanja stana (dnevno, tijekom sezone grijanja):

☐ jako rijetko ☐ povremeno ☒ često

Molimo okrenite!

Za sva pitanja nazovite 01/6326166 ili na dmaljkovic@eihp.hr

Slika 3-3 Primjer ispunjenog upitnika

Od ukupno 135 predanih upitnika, ispunjena je 51 anketa (37,8 %) koja je obrađena prema odgovorima, a rezultati obrade prikazani su u dijagramima u nastavku (zgrade A, B i C).

U analiziranim zgradama najviše je stanovnika u dobi od 19 do 65 godina, odnosno među stanarima u stanovima koji su ispunili ankete njih je 50 %.. Potom slijede stanari u dobnoj skupini 65 godina i više te djeca u dobi od 0 do 6 godina. Najmanje stanara u obuhvaćenim stanovima u dobnoj je skupini između 7 i 18 godina (slika 3-4).

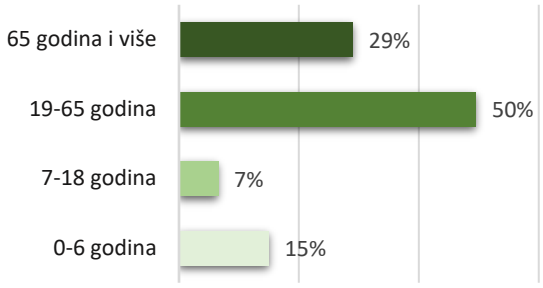
Najveći broj ispitanika u stanu provodi većinu dana (odgovori: *u stanu cijeli dan* i *u stanu veći dio dana*), dok ih je samo 14 % vrlo rijetko u stanu (slika 3-5).

Većina stanara u svojim stanovima nije zamijenila drvene prozore (njih više od 50 %), dok oni koju su ih zamijenili većinom imaju PVC prozore s dvostrukim ostakljenjem. Pretpostavka je da stariji prozori uzrokuju i veće gubitke toplinske energije, čime se javlja i veća potreba za grijanjem te se određene uštede mogu postići zamjenom prozora (slika 3-6).

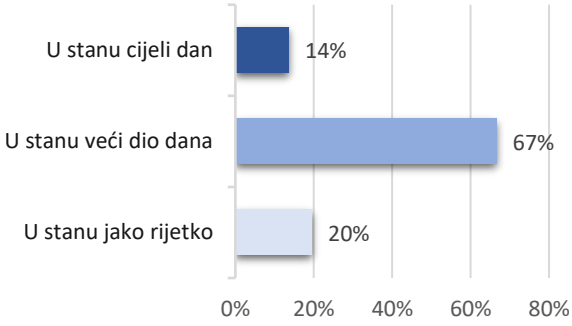
Ispitanici u zgradama A, B i C preferiraju umjerenu temperaturu u prostoru (61 %), a 39 % preferira *toplo*, dok nitko od ispitanika kao odgovor nije odabrao *hladno* (slika 3-7).

Kada je riječ o broju negrijanih prostorija, većina stanara ima jednu negrijanu prostoriju ili ni jednu. Njih 25 % ima dvije negrijane prostorije, a 2 % stanara ima tri negrijane prostorije u svojim stanovima (slika 3-8).

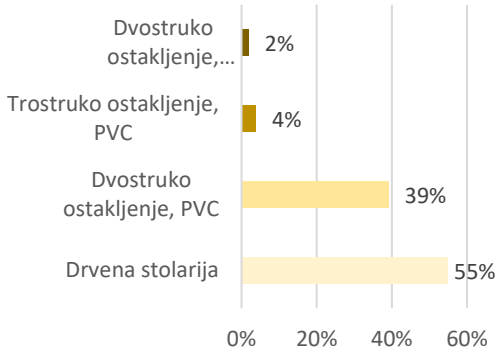
Većina ispitanika ponekad prozračuje svoje stanove, često ih prozračuje njih 22 %, dok samo 2 % stanara rijetko prozračuje svoje stanove (slika 3-9).



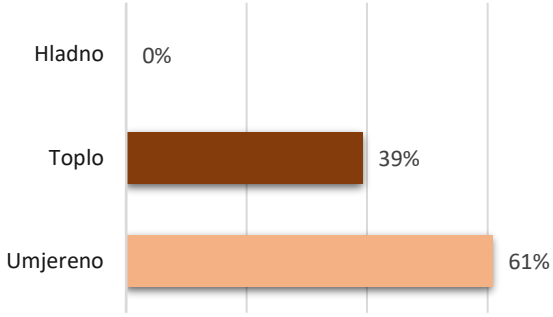
Slika 3-4 Postotak članova kućanstva prema dobnim skupinama zgrada A,B i C



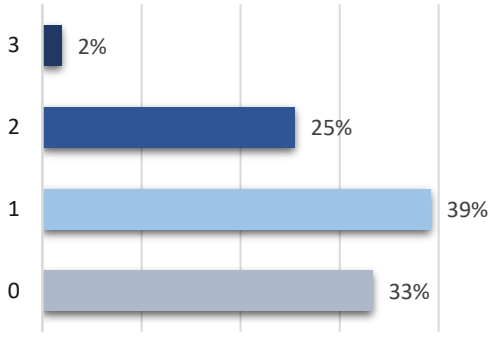
Slika 3-5 Vrijeme prosječnog boravka u stanovima zgrada A,B i C



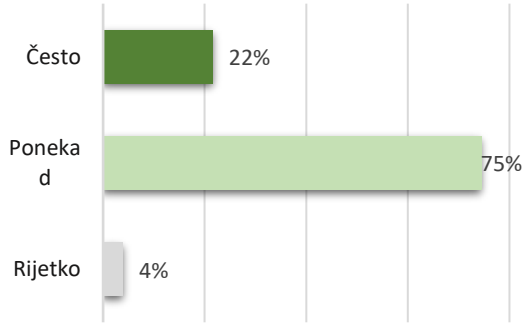
Slika 3-6 Tip prozora zgrada A, B i C



Slika 3-7 Razina temperature u prostorijama zgrada A,B i C



Slika 3-8 Broj negrijanih prostorija zgrada A, B i C



Slika 3-9 Prosječno vrijeme prozračivanja stana zgrada A, B i C

3.3. Analiza grupiranjem

Pri analizi grupiranjem podataka iz upitnika, od interesa je ocijeniti utjecaj na ciljanu varijablu specifične potrošnje topline i zaključiti može li se razaznati jasan utjecaj svake varijable iz provedenih upitnika na potrošnju, postoji li sličnost u potrošnji toplinske energije u stanovima u pojedinim grupama.

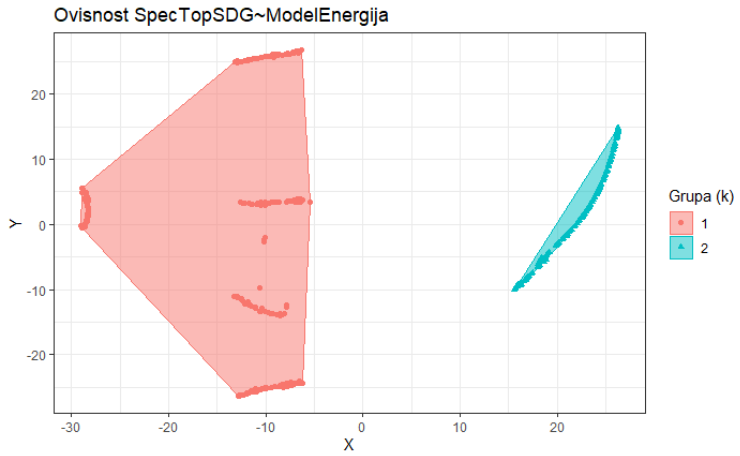
Ako postoji jasno grupiranje među podacima, grupe će imati jasnu međusobnu granicu te neće dolaziti do znatnijih preklapanja između pojedinih grupa. Osnovni cilj je pronaći prirodno grupiranje na temelju kojeg se može doći do korisnih zaključaka koja bi bila primjenjiva na buduća opažanja. Grupiranje se provodi prema bihevioralnim varijablama dobivenim iz upitnika. Podaci time pokazuju svoju unutarnju strukturu i broj podgrupa u koje analizirane podatke možemo grupirati. Mjera sličnosti koja će se koristiti pri ovoj analizi bit će euklidska udaljenost, dok se određivanje optimalnog broja grupa provodi metodom siluete.

Grupiranje se provodi za jednu zgradu i za jedan mjesec u sezoni grijanja jer je normalizacija s obzirom na stupanj-dan grijanja provedena prema razmatranim godinama (kako je opisano u poglavlju 3.1), a ne i prema mjesecima unutar godine. Tako su usporedivi pojedini mjeseci u različitima godinama (npr. siječanj sa siječnjem, veljača s veljačom) i agregirane potrošnje među godinama, ali nisu usporedivi različiti mjeseci u jednoj ili različitim godinama. Za analizu provedenu u ovom poglavlju korištene su potrošnje za mjesec siječanj u svim razmatranim godinama (2011. - 2017.). Grupiranje je provedeno određivanjem udaljenosti između pojedinih opservacija, što se čini izračunom Gowerove udaljenosti [90] i metodom vizualizacije t-SNE [91] koja se pokazuje izrazito prikladnom kod prikaza višedimenzionalnih podataka [92].

Grupiranje se provodi metodom *k*-srednjih vrijednosti. Za grupiranje koristi se R paket *factoextra* [93].

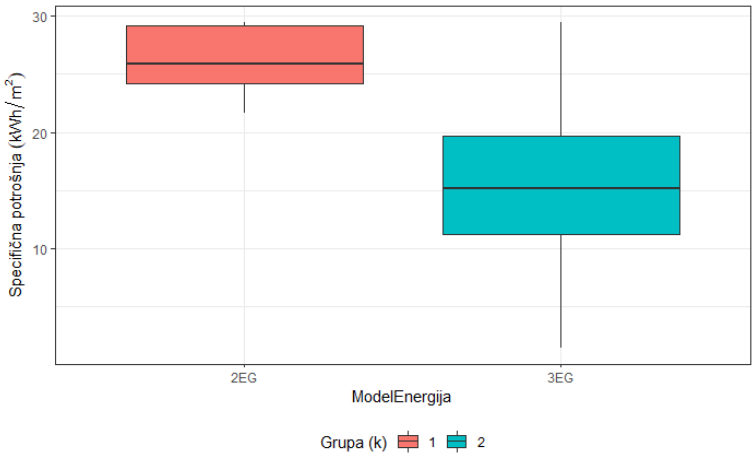
3.3.1. Utjecaj modela obračuna energije

Grupiranje specifične potrošnje topline pojedine vlasničke jedinice u ovisnosti o odabranom modelu za raspodjelu i obračun troškova za isporučenu toplinsku energiju metodom siluete dijeli podatkovni skup u dvije grupe (slike 3-10). Također, osnovna raspodjela istih podataka prije grupiranja dana je na slici 3-11.



Slika 3-10 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *ModelEnergija*

Algoritam prepoznaje dvije različite grupe te se „grupa 1“ odnosi na vlasničke jedinice koji u mjesecu siječnju u svim razmatranim godinama imaju *ModelEnergija* = 2EG, dok se „grupa 2“ odnosi na vlasničke jedinice koji u istom razdoblju imaju *ModelEnergija* = 3EG. Jedna vlasnička jedinica u jednom mjesecu može imati pridružen samo jedan od ta dva.



Slika 3-11 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *ModelEnergija*

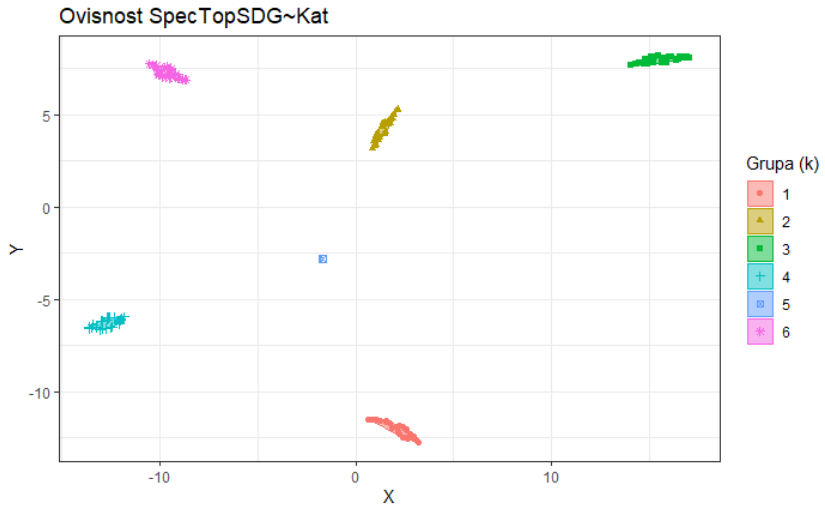
Vlasničke jedinice s modelom 3EG imaju prosječnu potrošnju za 42 % nižu od stanova s modelom 2EG. Također, za stanove s modelom 3EG moguće je ostvariti donje stršeće vrijednosti, što nije moguće za stanove s modelom 2EG. Za model 2EG minimalna vrijednost koja se obračunavala u mjesecu siječnju u promatranim godinama iznosila je 20,69 kWh/m^2, što je više od minimalne prosječne vrijednosti (15,17 kWh/m^2) koja se obračunavala vlasničkim jedinicama s modelom 3EG. Vrijednost obračunate potrošnje topline blizu nule, za stanove s modelom 3EG, upućuje na to da se u nekom mjesecu radi o praznom stanu, koji je zato što koristi mjerenje za obračun potrošnje bio u poziciji da za taj mjesec plati minimalni iznos za grijanje (alokacija dolazi iz dijela fiksnog troška zajedničke potrošnje zgrade prema načinu alokacije iz [56]), što je mogućnost koju vlasničke jedinice na modelu 2EG nemaju te one i u mjesecima u kojima se ne boravi u prostoru plaćaju iznos prema površini.

Vezano za utjecaj modela obračuna na specifičnu potrošnju pojedinog stana u razmatranoj zgradi može se reći da je on znatan. Iako su maksimalne vrijednosti jednake, srednje i najmanje vrijednosti znatno su niže kod vlasničkih jedinica koje koriste individualno mjerenje, a samim time može se reći da te jedinice postižu uštede u energetskom i financijskom smislu.

3.3.2. Utjecaj položaja vlasničke jedinice unutar zgrade

U idućem koraku napravljeno je grupiranje ovisnosti specifične potrošnje topline prema pojedinom katu kako bi se ocijenio utjecaj položaja vlasničke jedinice unutar zgrade na potrošnju, kako je prikazano na slici 3-12. Naime, u nekim zemljama EU (npr. Danskoj) taj se utjecaj smatra znatnim te se obračun provodi s korekcijskim faktorima u ovisnosti o položaju stana [94]. U razmatranim zgradama A, B i C nalazi se 6 razina, zajedno sa suterenom (razina „-1“) i prizemljem (razina „0“). Zgrade je imaju 4 kata. Metodom prosječne siluete određen je optimalan broj grupa koji iznosi $k = 6$.

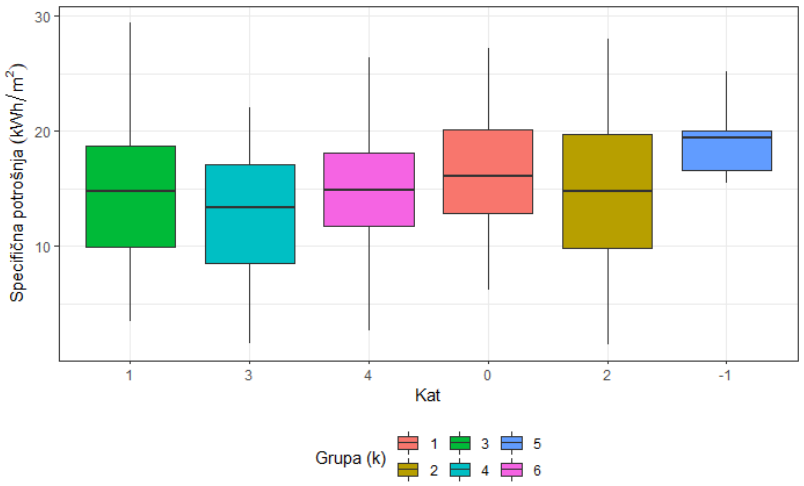
Grupiranje ovisnosti smještaja vlasničke jedinice unutar zgrade s obzirom na pojedini kat provedeno je bez preklapanja pojedinih grupa, kako se može vidjeti na slici 3-12. Za razliku od grupiranja specifične potrošnje u ovisnosti o modelu obračuna topline, ovdje je potrebno dodatno filtrirati podatke kako bi se analiza napravila samo na stanovima nakon ugradnje individualnog mjerenja (model 3EG), s obzirom na to da stanovi koji koriste model 2EG imaju jednaku specifičnu potrošnju i ne mogu se međusobno uspoređivati.



Slika 3-12 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *Kat*

Kako se može vidjeti iz slike 3-12, grupiranje je provedeno bez poklapanja. Ako se usporedi analiza grupiranja i prikaz raspodjele podataka prije grupiranja na slici 3-13, može se doći do inicijalnih zaključaka kako slijedi:

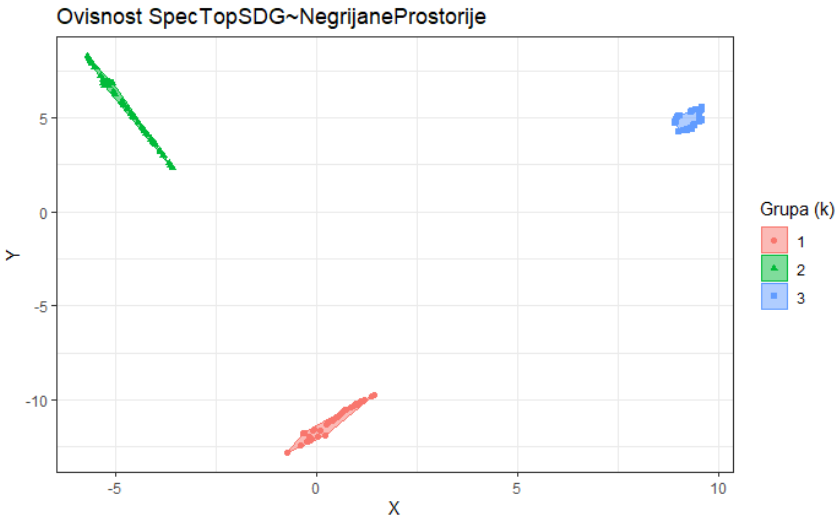
- Najnižu maksimalnu potrošnju pokazuju opservacije za 3. kat (k = 4), koji nije ruban, čime se pokazuje da centralno smješteni stanovi unutar zgrade imaju manju specifičnu potrošnju zbog svoga položaja.
- Najvišu minimalnu potrošnju imaju stanovi na razini -1 (k = 5), tj. u suterenu, čime se potvrđuje da rubni stanovi, u ovom slučaju stanovi koji potencijalno nisu dovoljno toplinski izolirani prema tlu imaju višu specifičnu potrošnju topline.



Slika 3-13 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *Kat*

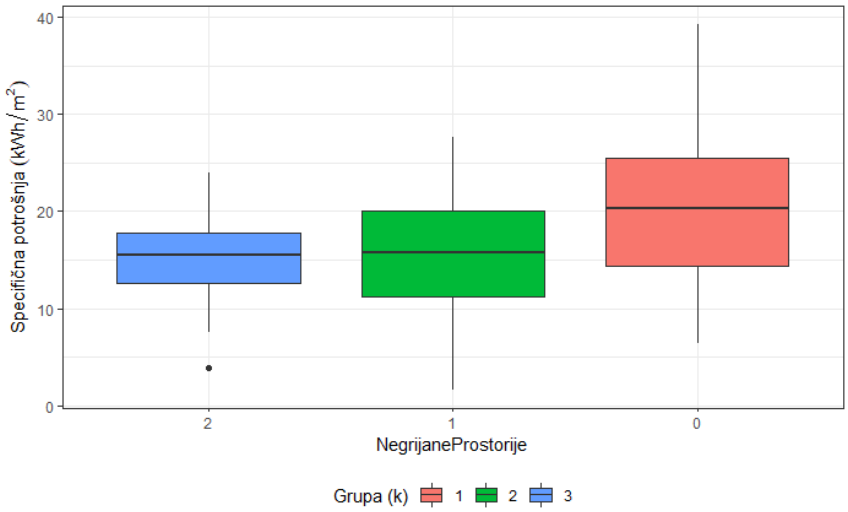
3.3.3. Utjecaj broja negrijanih prostora

Razlog da se neke prostorije u stanu ne griju može proizlaziti iz želje za uštedom ili iz činjenice da je tražena razina toplinske ugodnosti različita za različite prostorije (npr. niža temperatura u spavaćim sobama). Negrijanjem određenih prostorija direktno se utječe na stvarnu grijanu površinu, tj. vlasničke jedinice koje u stvarnosti ne griju neke prostorije prividno imaju nižu specifičnu toplinsku potrošnju. Grupiranje je provedeno u tri grupe, te je dobiveno jasno grupiranje bez preklapanja, kako je vidljivo iz slike 3-14.



Slika 3-14 Grafički prikaz grupiranje po varijablama *SpecTopSDG* i *NegrijaneProstorije*

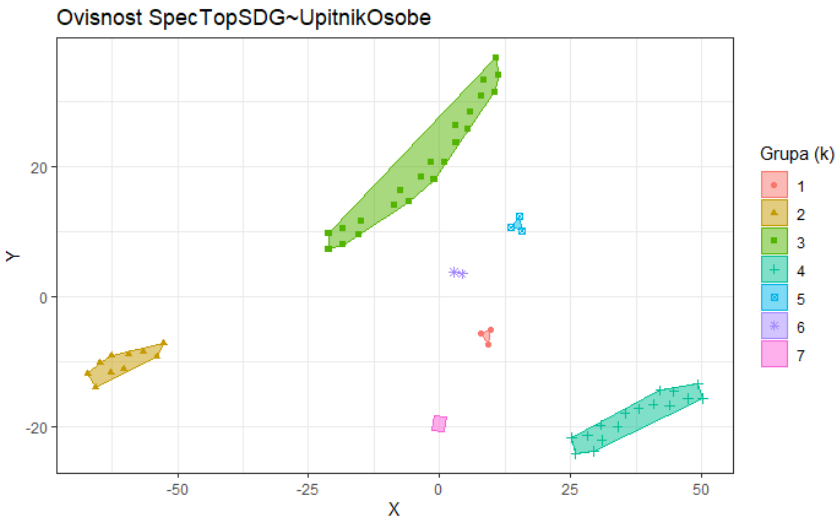
Kako je i očekivano, specifična potrošnja smanjuje se s povećanjem broja negrijanih prostorija, što se može vidjeti iz slike 3-15. Tako je srednja vrijednost specifične potrošnje za vlasničke jedinice koji koriste model 3EG i koje griju sve prostorije oko 20 kWh/m^2, dok je za one koje ne griju jednu prostoriju ta vrijednost 15,7 kWh/m^2, odnosno za one koje ne griju dvije prostorije 15,1 kWh/m^2. Načelno se može zaključiti da se negrijanjem jedne do dviju prostorija može očekivati ušteda topline 20-25 %.



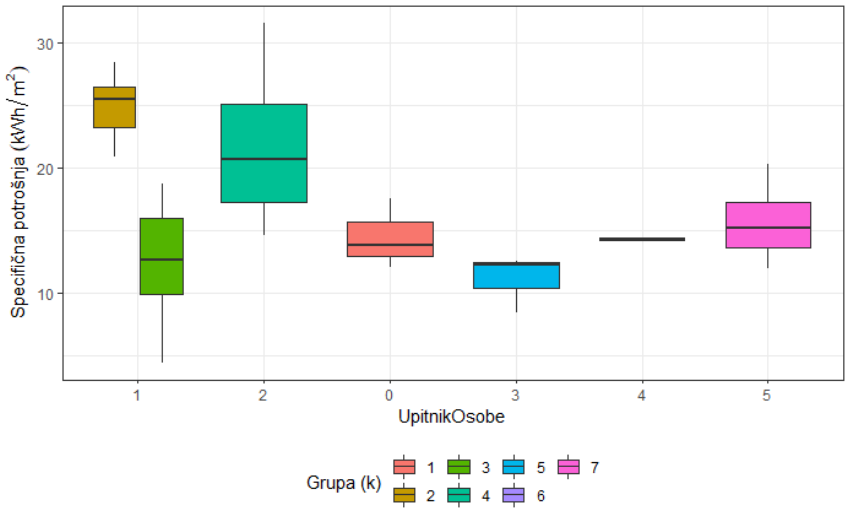
Slika 3-15 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *NegrijaneProstorije*

3.3.4. Utjecaj broja stanara

Iduća pretpostavka analize je da broj osoba koje borave u vlasničkoj jedinici ima direktan utjecaj na specifičnu potrošnju topline. Grupiranje je napravljeno prema slici 3-16 u optimalnih 7 grupa određenih metodom siluete. Iako se grupe međusobno ne preklapaju, kod ovog grupiranja imamo pojavnost toga da se vrijednosti za jednu kategoriju nalaze u dvije grupe (slika 3-17) za stanove s jednim stanarom.

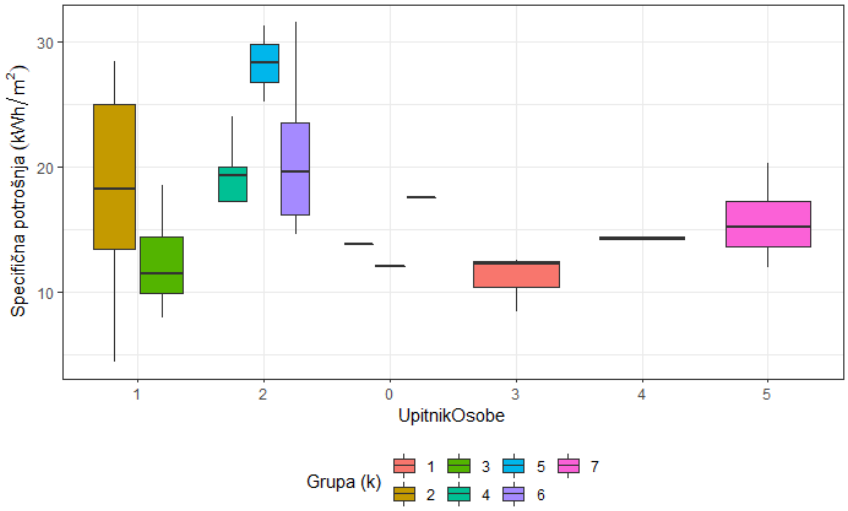


Slika 3-16 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *UpitnikOsobe*



Slika 3-17 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *UpitnikOsobe*

Razlozi pojavnosti dviju različitih grupa potrošnje kod vlasničkih jedinica mogu biti višestruki, bilo da se radi o različitim prosječnim dnevnim vremenima boravka u stanu, zahtjevima na toplinsku ugodnost, broju negrijanih prostorija, tipu prozora ili učestalosti prozračivanja. Sve od navedenog analizirano je u upitniku, ali ovdje će se analizirati samo utjecaj prosječnog vremena boravka jer se pretpostavlja da je taj parametar najutjecajniji s obzirom na to da se na njega može utjecati zatvaranjem ventila kada se ne boravi u prostoru (slika 3-18).

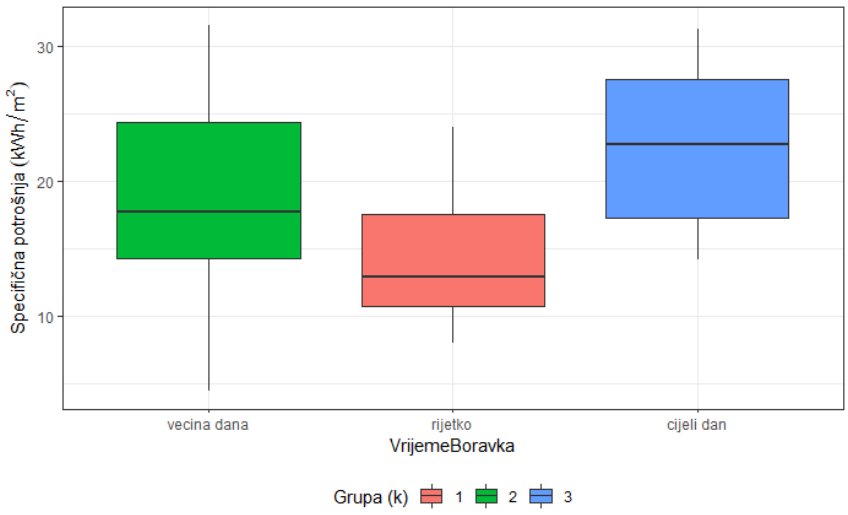


Slika 3-18 Pravokutni dijagram neskaliranih podataka, ovisnosti potrošnje stana o broju osoba i vremenu boravka

Za grupiranje zadržavamo isti broj grupa, $k = 7$, iako metoda siluete rezultira većim optimalnim brojem grupa. Može se pretpostaviti da je jedan od razloga grupiranja stanova s jednim stanarom u dvije različite grupe prema specifičnoj potrošnji u različitom prosječnom dnevnom

vremenu boravka u stanu. Tako se u ovoj analizi vlasničke jedinice grupiraju u dvije grupe, a svaka sadržava upravo jednak broj vlasničkih jedinica s jednim stanarom.

Sam utjecaj vremena boravka u svim stanovima je značajan, kako se može i očekivati. Jedinice u kojima se manje boravi imaju manju specifičnu potrošnju pod pretpostavkom da se u tom slučaju isključi grijanje, kako se može vidjeti iz slike 3-19.

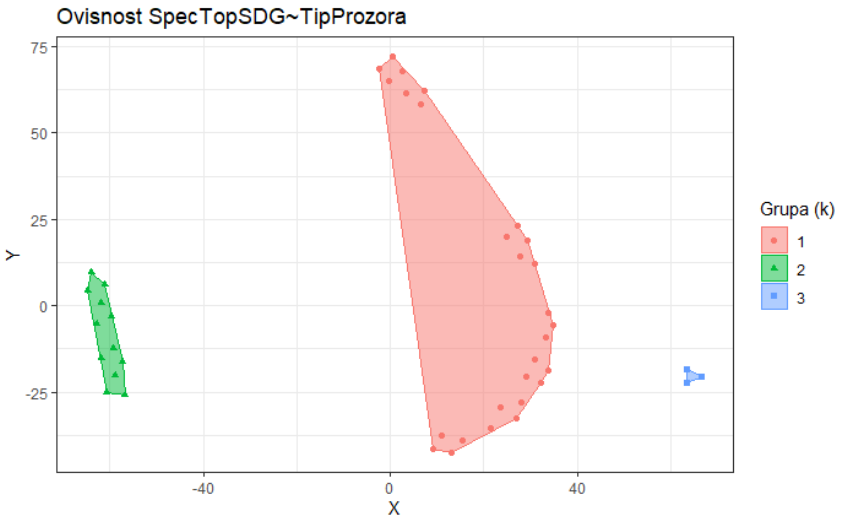


Slika 3-19 Pravokutni dijagram neskaliranih podataka ovisnosti potrošnje o vremenu boravka

Jednaku analizu bilo bi moguće napraviti za svaki od spomenutih utjecajnih parametara pojedinačno, ili sumarno za sve parametre. Pojedinačna analiza neće biti rađena za sve parametre u ovom radu, ali će se provesti sumarna analiza.

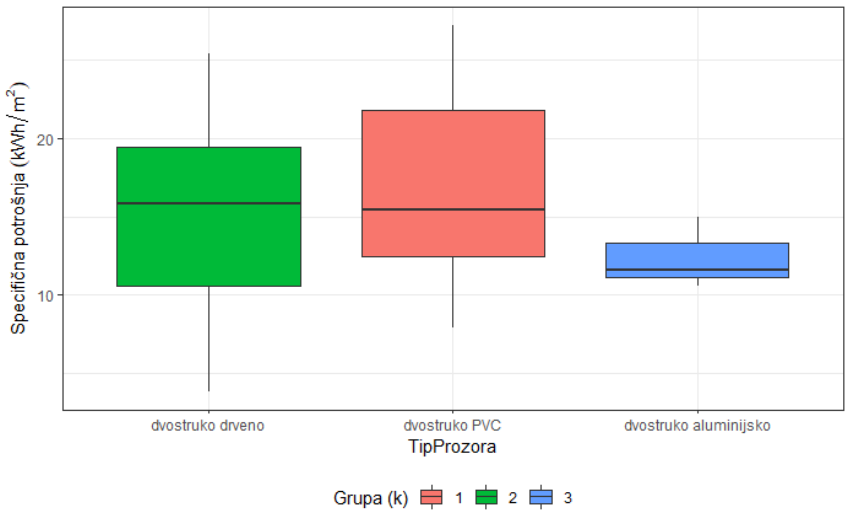
3.3.5. Utjecaj tipa prozora

Zgrade A, B i C koje su analizirane upitnicima izgrađene su 1970-ih godina te su inicijalno u zgradu ugrađeni dvostruki drveni prozori. Kroz godine su stanari u nekim stanovima zamijenili prozore energetski učinkovitijim. Svi tipovi prozora koji su prilikom terenskog obilaska prepoznati na analiziranoj zgradi su: dvostruki drveni, dvostruki PVC i dvostruki aluminijski. Grupiranje je provedeno u 3 nepreklapajuće grupe prema rezultatima analize siluete, a kako je prikazano na slici 3-20.



Slika 3-20 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *TipProzora*

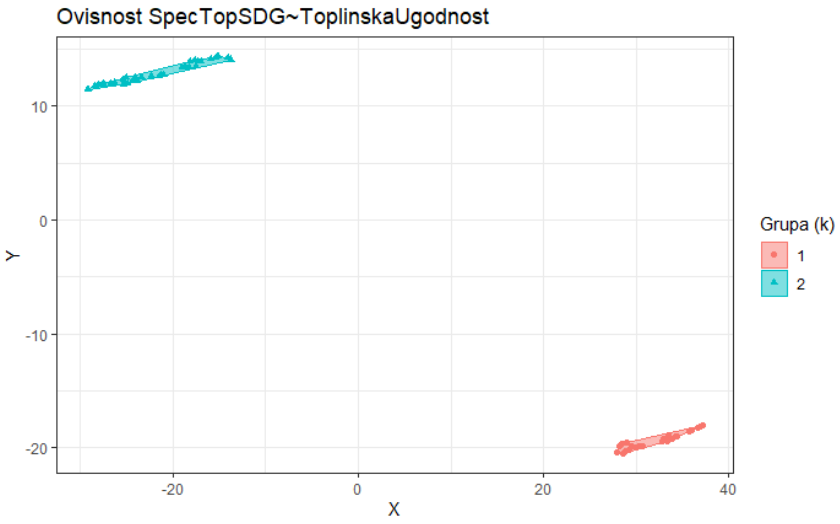
U analiziranom podatkovnom skupu ne primjećuje se znatan utjecaj zamjene inicijalno ugrađenih dvostrukih drvenih prozora s dvostrukim PVC prozorima s obzirom na to da je prosječna potrošnja slična (vidljivo iz slike 3-21). Znatnija ušteda bilježi se tek pri ugradnji aluminijskih prozora, iako bi se ušteda očekivala i kod ugradnje dvostrukih PVC prozora u usporedbi s drvenim prozorima.



Slika 3-21 Pravokutni dijagram neskaliranih podataka, *SpecTopSDG* i *TipProzora*

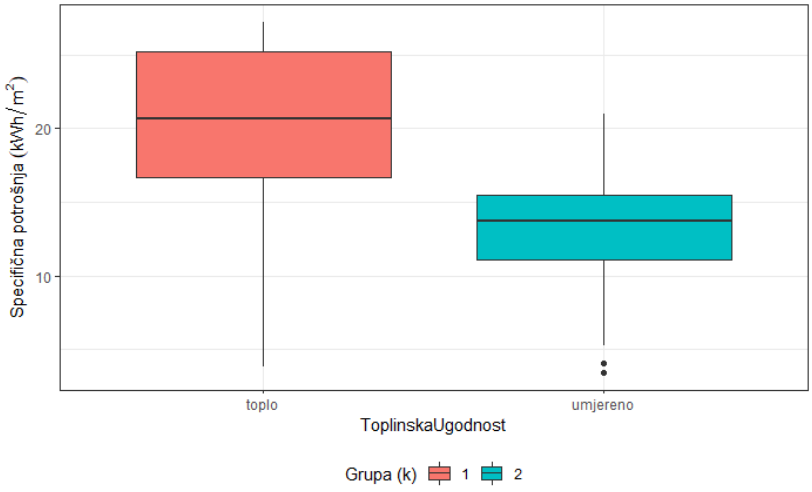
3.3.6. Utjecaj temperature u prostorijama

Posljednja analiza utjecaja bihevioralnih parametara na specifičnu potrošnju stanova razmatra utjecaj željene temperature u prostorijama. Kako je i očekivano, taj utjecaj je značajan. Srednja vrijednost specifične potrošnje za one stanove koji žele umjerenu ugodnost, u usporedbi s onim stanovima koji preferiraju toplije uvjete, niža je za približno četvrtinu (23 %). Grupiranje se provodi bez preklapanja u dvije grupe (slika 3-22).



Slika 3-22 Grafički prikaz grupiranja po varijablama *SpecTopSDG* i *ToplinskaUgodnost*

U sklopu upitnika, osim odgovora na pitanje o toplinskoj ugodnosti *toplije* i *umjereno*, ponuđen je i odgovor *hladnije*, kako se može vidjeti na slici 3-23. Ni jedan ispitanik nije izabrao taj odgovor, što pokazuje da se svi stanovi griju.

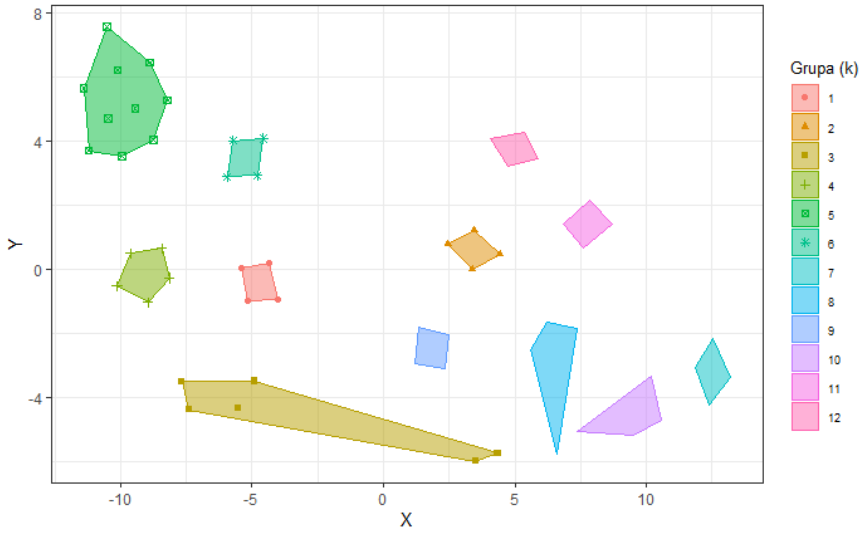


Slika 3-23 Pravokutni dijagram neskaliranih podataka za *SpecTopSDG* i *ToplinskaUgodnost*

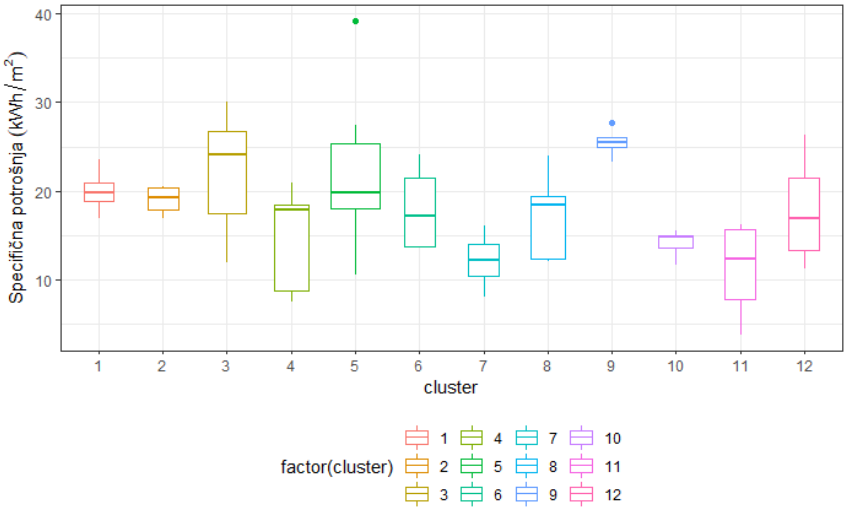
3.3.7. Utjecaj više faktorskih varijabli

U konačnici je od interesa ocijeniti utjecaj više bihevioralnih parametara u višedimenzionalnom prostoru. Odabrane bihevioralne varijable za koje će se analizirati utjecaj na specifičnu potrošnju stanova na mjesečnoj razini su *Kat*, *VrijemeBoravka* i *ToplinskaUgodnost*.

Višedimenzionalno grupiranje kategorijskih varijabli provodi se R paketom *Rtsne* [95], nakon što je metodom siluete izračunato da je optimalni broj grupa 12. Grupiranje se provodi bez preklapanja, kako je prikazano na slici 3-24. Prikaz pravokutnim dijagramima za specifičnu potrošnju po pojedinim grupama dan je na slici 3-25.



Slika 3-24 Višedimenzionalno grupiranje utjecaja bihevioralnih parametara na potrošnju



Slika 3-25 Specifična potrošnja po pojedinim grupama

Detaljni opis svih vrijednosti po svakoj grupi dan je u prilogu. Analizirajući rezultate dobivene nenadziranim učenjem metodom *k*-srednjih vrijednosti može se zaključiti sljedeće:

1. Najvišu potrošnju imaju stanovi u grupi 9. Srednja specifična potrošnja je 25,45 kWh/m^2. Prema slici 3-25, vidi se da se radi o manjem skupu stanova (uži pravokutnik u dijagramu), koji se nalaze u prizemlju zgrade, borave u prostoru većinu dana i preferiraju višu razinu toplinske ugodnosti.
2. Najmanju potrošnju imaju stanovi u grupi 11. To su stanovi koji se nalaze na 2. katu, dakle u centralnom dijelu zgrade, u prostoru borave veći dio dana i preferiraju umjerenu toplinsku ugodnost. Ova grupa daje naznake da se radi o centralno pozicioniranim stanovima koji zbog svojeg položaja imaju manje potrebe za grijanjem prostora, uz umjerenu temperaturu u prostoru.
3. Najnižu minimalnu vrijednost u grupi imaju stanovi u grupama 4 i 5, koji su u anketama izjavili da borave u stanu veći dio dana, imaju zahtjeve za toplijim prostorom, a nalaze se na 1. i 3. katu što znači da su to centralno pozicionirani stanovi. Ova minimalna potrošnja izrazito je niska, tako da je očito da su u ovom skupu i stanovi koji su u nekima od analiziranih mjeseci bili prazni.

Analize utjecaja kvalitativnih varijabli na kontinuirane varijable su korisne, ali su računski zahtjevne i visoko osjetljive na kvalitetu ulaznih parametara. U ovom radu proveden je intervju na dostupnom skupu od tri zgrade u kojem se ukupno nalazi 150 stanova. Za točnije analize bilo bi važno prikupiti veći broj ovakvih upitnika, po mogućnosti u više različitih i prostorno udaljenih zgrada.

U ovoj analizi prepoznaje se utjecaj bihevioralnih faktora na potrošnju, posebno u broju negrijanih prostorija, željene temperature i vremenu boravka u prostoru. Korekcijom nekih bihevioralnih faktora mogu se postići uštede i izbjeći visoka potrošnja te bi u budućim istraživanjima bilo korisno napraviti takve smjernice.

3.3.8. Grupiranje na razini zgrada

Grupiranje se provodi na skupu koji se sastoji od 350 zgrada u kojima se nalazi ukupno 27 102 stana za koje su dostupni podaci o potrošnji toplinske energije u razdoblju od 2011. do 2017. godine. To razdoblje pokriva godine prije i nakon uvođenja obaveze ugradnje individualnog mjerenja, na što se i odnosi sumarna statistika (tablice 3-6 i 3-7). Grupiranjem se pojedine zgrade međusobno uspoređuju.

Tablica 3-6 Sumarna statistika izabranih varijabli za *psGrad*

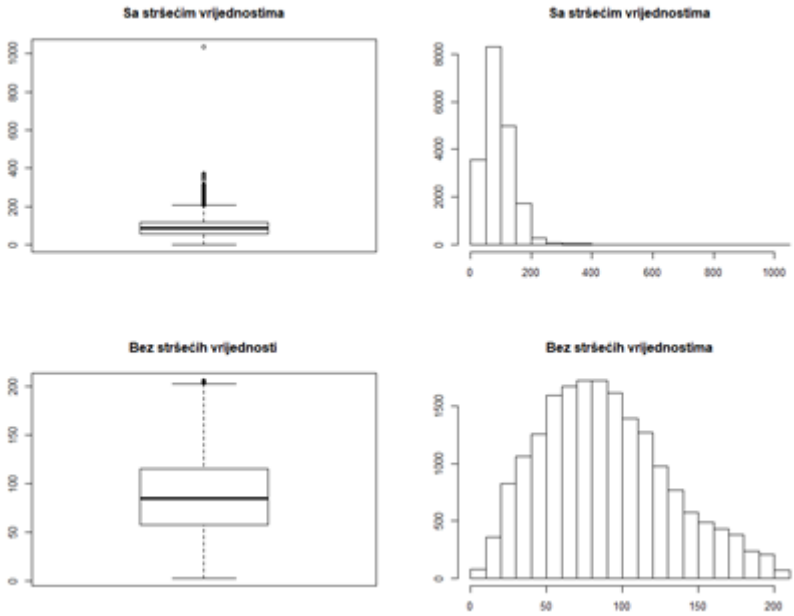
	StopaUgr	***SpecTopSDG* [kWh]**	***SpecTopOMM_SDG* [kWh]**
Min.	0,0000	0,00	6,17
Donji kvartil	0,9272	39,74	55,09
Median	0,9639	69,95	85,91
Srednja vrijednost	0,9298	76,48	75,71
Gornji kvartil	0,9835	104,87	94,92
Maks.	1,0000	1035,27	133,97

Deskriptivno se može ocijeniti jesu li zgrade koje su ugradile individualno mjerenje ostvarile uštede u odnosu na zgrade koje nisu ugradile mjerenje. Smatra se da je zgrada ugradila mjerenje ako je najmanje jedna vlasnička cjelina ugradila mjerenje.

Tablica 3-7 Sumarna statistika varijable *SpecTopSDG*, prema modelu obračuna

	Model 2EG [kWh]	**Model 3EG [kWh]**
Min.	0,0	1,40
D.kv.	118,80	37,82
Median	136,91	68,92
Sr.vr.	131,43	73,53
G. kv.	150,86	101,57
Maks.	366,72	1035,27

Stršeće vrijednosti predstavljaju grešku mjerenja, tj. u slučaju sustava daljinskog grijanja, grešku alokacijskog sustava. Sve opservacije stršećih vrijednosti predstavljaju nezadovoljne krajnje kupce i greške sustava za alokaciju te bi bilo nužno poduzeti aktivnosti kako bi se svaka od tih opservacija detaljnije analizirala i identificirao razlog zašto je greška nastala. Ako za stršeće vrijednosti možemo sa sigurnošću utvrditi da su greška mjerenja, poželjno ih je izbaciti iz podatkovnog skupa kako ne bi utjecali na rezultate analize. Utjecaj stršećih vrijednosti vidi se iz slike 3-26. Osim na raspon mjerenja, stršeće vrijednosti utječu i na distribuciju frekvencija s razredima.



Slika 3-26 Provjera utjecaja stršećih vrijednosti *psStan*

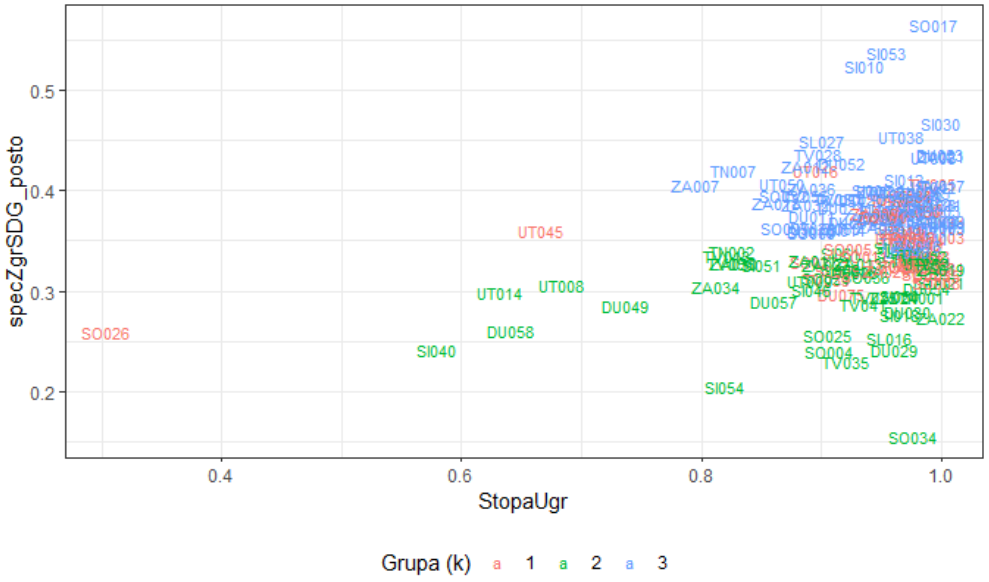
U analiziranom podatkovnom skupu identificirane su ukupno 262 stršeće vrijednosti, što čini 1,4 % opservacija. Srednja vrijednost stršećih vrijednosti iznosi 236,8 kWh/m^2 tj. srednja vrijednost svih opservacija koje uključuju i stršeće vrijednosti iznosi 90,7 kWh/m^2, dok bez njih iznosi 88,7 KWh/m^2.

Analiza je u idućem koraku provedena bez stršećih vrijednosti. Za skup zgrada od interesa je ocijeniti izdvojeni utjecaj stupnja ugradnje RTTE na potrošnju zgrade prije i poslije ugradnje individualnog mjerenja, kako bi se procijenilo koliki je utjecaj stupnja ugradnje na postignute uštede. Pretpostavka je da viši stupanj ugradnje RTTE rezultira većim uštedama.

Podatkovni skup sastoji se od 161 zgrade koje su ugradile individualno mjerenje u nekom postotku. Analiza je napravljena za 2017. godinu, sa sljedećim varijablama:

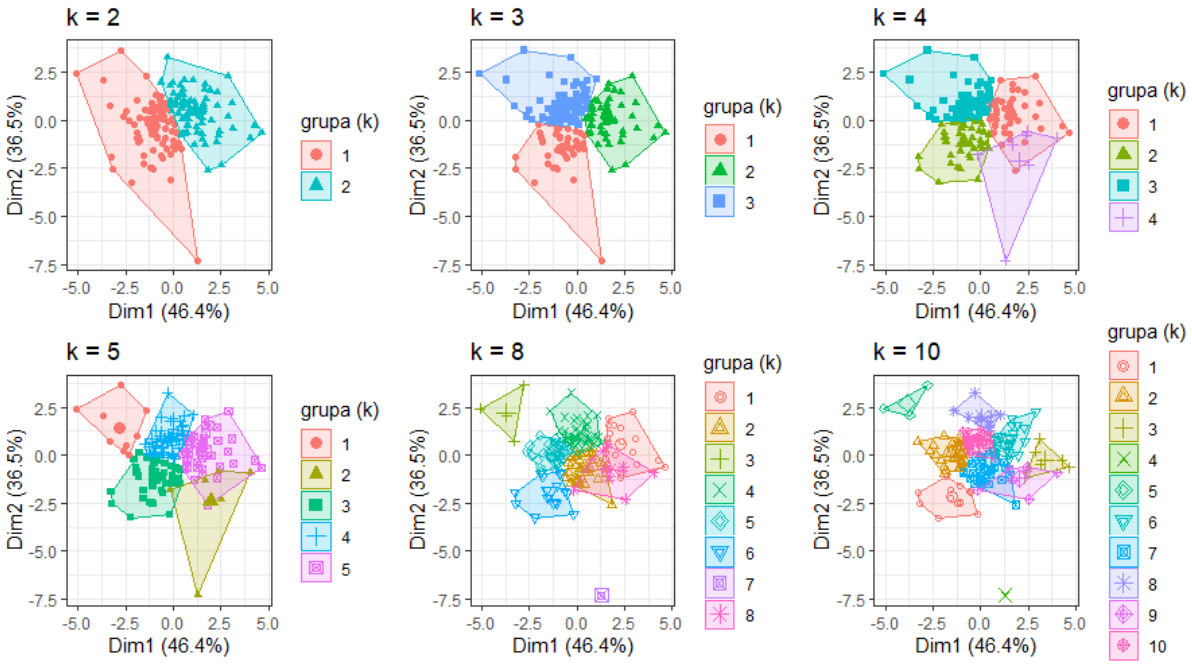
1. *OMM*
2. *StopaUgr*
3. Apsolutni iznos promjene specifične potrošnje zgrade (varijabla *specZgrSDG_delta*).
4. Postotna promjena specifične potrošnje zgrade (varijabla *specZgrSDG_posto*).

Za određivanje optimalnog broja grupa koristi se metoda siluete kao mjera sličnosti, koja daje optimalni broj $k = 3$. Grupiranje neskaliranih podataka pokazuje preklapanja ako gledamo rasipanje u ovisnosti o varijablama *specZgrSDG_posto* i *StopaUgr* (slika 3-27).



Slika 3-27 Dijagram rasipanja neskaliranih podataka potrošnje u analiziram zgradama

Provedbom višedimenzionalnog grupiranja grafički se prikazuje više opcija grupiranja koje sve daju grupe bez preklapanja, kako se vidi iz slike 3-28.



Slika 3-28 Grafički prikaz grupiranja u zavisnosti od broj grupa

Odabire se grupiranje na način da se izabere onu vrstu grupiranja koja nam daje najveći broj grupa koje su odvojene i među kojima nema preklapanja. Bira se grupa $k = 3$, i za nju se radi sumarna statistika prema tablicama 3-8, 3-9 i 3-10 .

Tablica 3-8 Grupa 1 prema utjecaju stope ugradnje na promjenu potrošnje u zgradama

	Potrošnja prije	Potrošnja poslije	StopaUgr	Potrošnja razlika	Potrošnja razlika	Br. zgrada
	kWh/m^2	kWh/m^2		kWh/m^2	%	-
Min	145,2	94,83	0,3046	47,21	0,2595	43
D. kv.	154,0	99,23	0,9516	50,39	0,3262	
Median	157,7	103,09	0,9720	56,59	0,3488	
Sr.vr.	**161,3**	**104,87**	**0,9434**	**56,44**	**0,3497**	
G.kv.	165,6	108,03	0,9838	60,44	0,3737	
Maks.	189,3	140,19	1,0000	76,24	0,4198	

Tablica 3-9 Grupa 2 prema utjecaju stope ugradnje na promjenu potrošnje u zgradama

	Potrošnja prije	Potrošnja poslije	StopaUgr	Potrošnja razlika	Potrošnja razlika	Br. zgrada
	kWh/m^2	kWh/m^2		kWh/m^2	%	-
Min	105,7	74,63	0,5801	18,40	0,1547	49
D.kv..	125,9	86,80	0,8609	36,72	0,2856	
Median	131,5	92,11	0,9299	40,88	0,3082	
Sr.vr.	**131,6**	**92,20**	**0,9007**	**39,42**	**0,2988**	
G. kv.	139,2	97,38	0,9696	43,76	0,3278	
Maks.	152,7	115,95	1,0000	48,50	0,3442	

Tablica 3-10 Grupa 3 prema utjecaju stope ugradnje na promjenu potrošnje u zgradama

	Potrošnja prije	Potrošnja poslije	StopaUgr	Potrošnja razlika	Potrošnja razlika	Br. zgrada
	kWh/m^2	kWh/m^2		kWh/m^2	%	-
Min	118,8	60,81	0,7956	43,88	0,3453	69
D. kv.	133,8	83,54	0,9167	50,77	0,3685	
Median	140,2	86,79	0,9594	54,27	0,3872	
Sr.vr.	**141,0**	**85,06**	**0,9464**	**55,93**	**0,3957**	
G.kv.	148,3	89,30	0,9835	58,87	0,4038	
Maks.	160,3	94,96	1,0000	89,43	0,5651	

Analizom pojedinih grupa, s obzirom na utjecaj stope ugradnje na smanjenje potrošnje topline nakon ugradnje RTTE, mogu se izvesti zaključci:

- Grupa 1

  Zgrade u ovoj grupi prije ugradnje imale se iznimno visoku potrošnju zoplinske enegije. Maksimalna potrošnja jedne od zgrada u grupi je 189,3 kWh/m^2. Iako je srednja vrijednost stope ugradnje u ovom skupu jednaka srednjoj vrijednosti ugradnje u grupi 3 i iznosi 94 %, u ovoj grupi nalaze se zgrade koje imaju veći

raspon između minimalne i maksimalne vrijednosti stope ugradnje i višu potrošnju topline nakon ugradnje. Srednja postotna ušteda u ovom skupu jest 35 %, a kreću se od 26 % do 42 %. Zgrade u ovoj skupini najlošijih su tehničkih karakteristika, što se vidi iz visoke potrošnje prije ugradnje. Također, pretpostavlja se da postoji i određeni problem s funkcioniranjem sustava grijanja unutar zgrade koji s jedne strane onemogućava postizanje većih ušteda, a s druge strane demotivira stanare u zgradama s nižom stopom ugradnje (minimalna vrijednost od 30 %) da ugrade RTTE, jer nemaju povjerenja u funkcioniranje sustava.

- Grupa 2

  Zgrade u ovoj grupi imale su najnižu potrošnju od svih grupa prije ugradnje RTTE, što pokazuje da se radi o tehnički boljim zgradama u skupu. Prosječna stopa ugradnje niža je nego u grupi 1 ili 3, ali je minimalna stopa ugradnje 50 %, što je više nego što je minimalna stopa ugradnje u grupi 1. Zgrade u ovoj grupi ostvarile su srednju uštedu od 30 % i prema dostupnim podacima o potrošnji iz grupiranja, pretpostavlja se da se u ovoj grupi nalaze zadovoljni krajnji korisnici.

- Grupa 3

  Zgrade u grupi 3 postižu najvišu uštedu u usporedbi s potrošnjom prije ugradnje RTTE. Postotni iznosi uštede kreću se od 35 % do 56 %. Zgrade su prije ugradnje RTTE imale malo veću potrošnju te se pretpostavlja da su nakon ugradnje poduzete sve nužne mjere da sustav grijanja s promjenjivim opterećenjem dobro funkcionira, a postoje i naznake da su napravljene neke od drugih mjera energetske učinkovitosti (obnova fasade) ili su prije individualnog mjerenja imali previsoku instaliranu snagu u toplinskoj podstanici, što se korigiralo nakon ugradnje RTTE. Zgrade u ovoj skupini smatraju se energetski najučinkovitijima.

Ovakva je analiza korisna ako se žele identificirati zgrade koje imaju dobre pokazatelje energetske učinkovitosti kako bi se njihova iskustva primijenila na onim zgradama koje ne postižu zadovoljavajuće pogonske parametre.

Pokazatelji potrošnje i energetske učinkovitosti na razini zgrade ne mogu se jednoznačno preslikati i na stanove u tim istim zgradama jer je potrošnja krajnjih korisnika pod utjecajem niza već spominjanih bihevioralnih faktora.

3.4. Strojno učenje za razinu stana

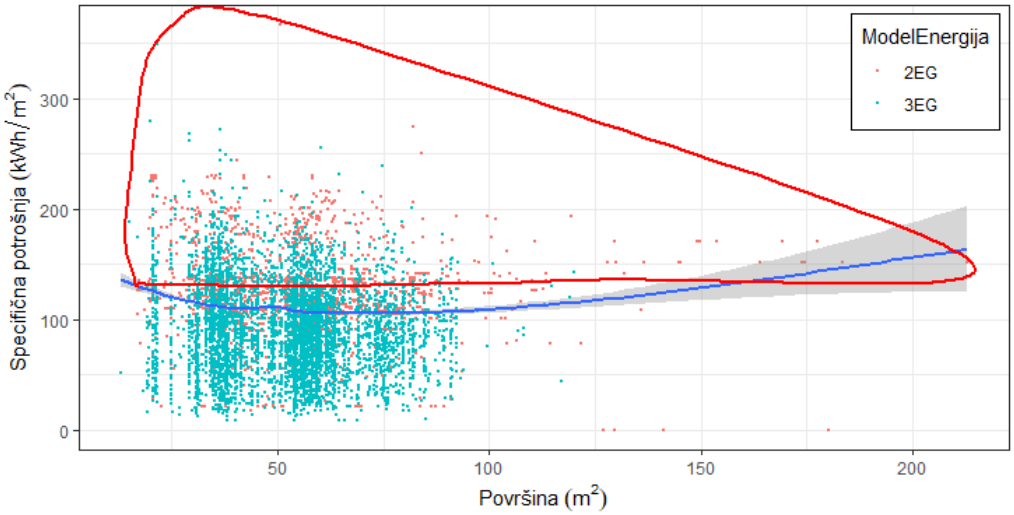
Osnovni cilj je izraditi model za predviđanje potrošnje u stanovima spojenim na sustav daljinskog grijanja. Kako je prethodno opisano u poglavlju 3.1, postoje dva načina obračuna topline, jedan koji se temelji na mjerenjima preko RTTE (3EG) i drugi prema grijanoj površini (2EG). Moguće je da u jednoj zgradi postoje oba načina obračuna. U bilo kojem od navedenih slučajeva, obračun se provodi računski, putem formula definiranih u [56].

Iz navedenog razloga model za razinu stana se izrađuje na tri razine, (i) za čitav podatkovni skup *psStan* (poglavlje 3.1.2), te (ii) za podskupove stanova koji obračun temelje na očitanjima s RTTE (3EG) i (iii) za podskupove stanova koji nisu ugradili RTTE (2EG) i kojima se obračun provodi prema grijanoj površini i korekcijskom faktoru definiranom u Pravilniku [56]. Analize se provode samo u godinama nakon stupanja na snagu obaveze ugradnje individualnog mjerenja u stanovima spojenim na zajednički sustav grijanja [2].

3.4.1. Višestruka linearna regresija

Višestruka linearna regresija (engl. *multiple linear regression* – MLR) se provodi na podatkovnom skupu koji se sastoji od približno 78 900 opservacija u 13 nezavisnih varijabli i zavisnom varijablom *SpecTopSDG*, koja označava specifičnu potrošnju topline u pojedinom stanu. Na slici 3-29 prikazan je dijagram rasipanja zavisne varijable u ovisnosti o površni stanova (samo za 2017. godinu radi preglednosti prikaza). Crvenom linijom zaokružene su stršeće vrijednosti. Plava linija predstavlja regresijski pravac, a sivo osjenčanje regresijskog pravca predstavlja interval pouzdanosti.

Stršeće vrijednosti za čitav skup podataka (model 2EG i model 3EG) predstavljaju vrijednosti više od 133,57 kWh/m^2. Za stanove s modelom 3EG su one očekivano niže i predstavljaju vrijednosti više od 116,15 kWh/m^2, dok se za stanove s modelom 2EG stršeće vrijednosti pojavljuju za specifične potrošnje stanova iznad 137,05 kWh/m^2.



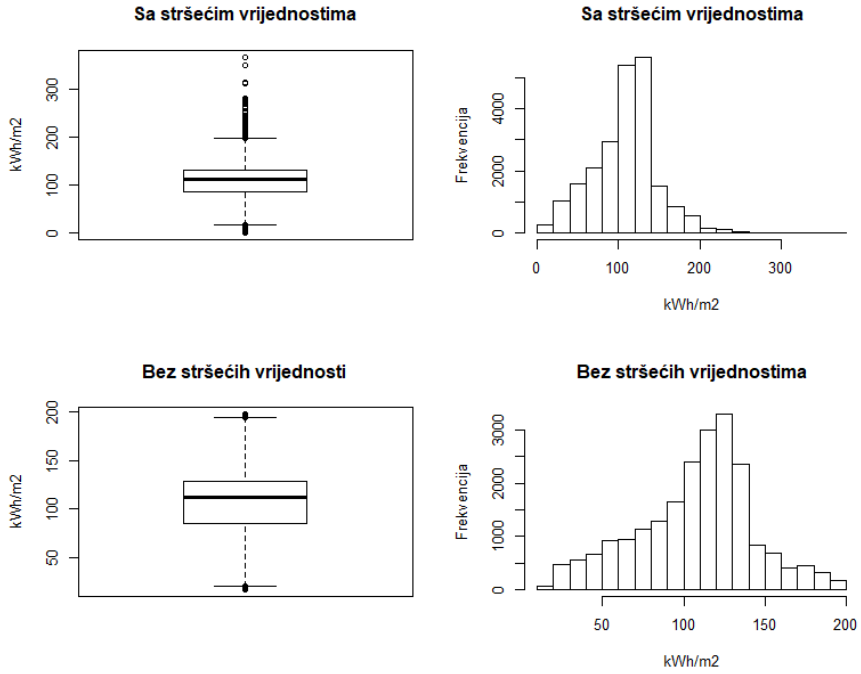
Slika 3-29 Prikaz rasipanja SpecTopSDG za 2017. godinu za oba načina obračuna energije (2EG i 3EG)

Statističko učenje metodom MLR je za sve modele opisane u ovom poglavlju provedeno na podskupu za učenje kojeg čini 60 % svih opservacija, dok je provjera točnosti predviđanja modela rađena na podskupu za provjeru kojeg čine 40 % svih opservacija.

3.4.1.1. Regresijski model za čitav podatkovni skup stanova

Kako se može vidjeti na slici 3-29 u kojoj su prikazane vrijednosti za jednu analiziranu godinu, u podatkovnom skupu postoji određeni broj stršećih vrijednosti koje želimo ukloniti iz skupa prije daljnjih analiza kako ne bi utjecale na regresijski model.

U podatkovnom skupu *psStan* 2,3 % opservacija su stršeće vrijednosti. Srednja vrijednost stršećih vrijednosti je 142,85 kWh/m^2, dok je srednja vrijednost svih opservacija zavisne varijable s uključenim stršećim vrijednostima 108,19 kWh/m^2, a nakon uklanjanja 107,39 kWh/m^2. Utjecaj stršećih vrijednosti prikazom pravokutnog dijagrama i histograma distribucije frekvencije prikazana je na slici 3-30.



Slika 3-30 Utjecaj stršećih vrijednosti na analizirani podatkovni skup za regresijski model na razini stana

Sve vrijednosti koje su više od iznosa gornjeg kvartila smatramo greškama mjerenja ili alokacije te takve opservacije više ne koristimo pri izradi modela. Vrijednosti ispod donjeg kvartila smatramo praznim stanovima te ih ostavljamo u modelu.

Početni model

Nakon što se uklone stršeće vrijednosti, određuje se početni regresijski model korištenjem funkcije *lm* iz R paketa *stats* [66]. Dobiveni rezultati su dani u tablici 3-11.

Tablica 3-11 Regresijski parametri i statistički pokazatelji početnog regresijskog modela za čitav podatkovni skup stanova

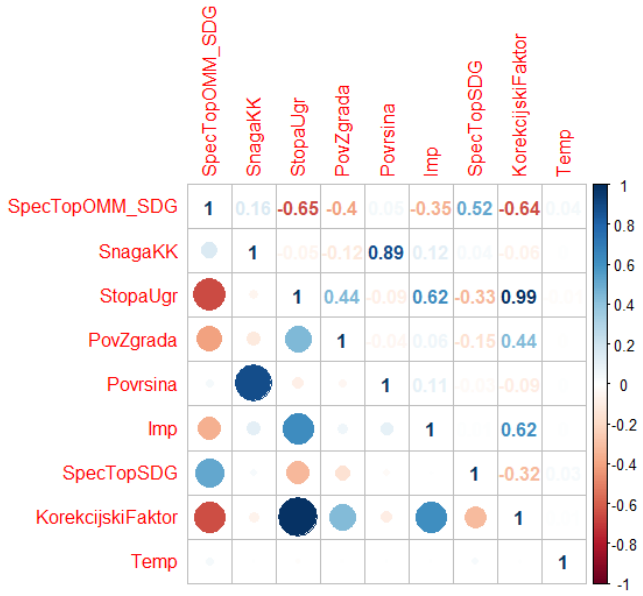
	β	St. pogreška	t- vrijednost	p - vrijednost
(odsječak)	26,167	6,756	3,873	0,000
SpecTopOMM_SDG	0,929	0,009	102,701	0,000
SnagaKK	1,119	0,135	8,306	0,000
StopaUgr	90,767	2,320	39,122	0,000
PovZgrada	0,002	0,000	38,763	0,000
Povrsina	-0,408	0,020	-20,771	0,000
Imp	643,739	5,172	124,466	0,000
KorekcijskiFaktor	-0,114	0,021	-5,517	0,000
Temp	-2,371	1,174	-2,020	0,043
ModelEnergija3EG	-123,632	0,735	-168,137	0,000

Radi usporedbe, točnost predviđanja dobivenog regresijskog modela računa se na skupu za učenje i na podskupu za provjeru. Točnost na podskupu za provjeru viša je nego točnost dobivena na skupu za učenje, i prema RMSE iznosi 16,44 kWh/m^2, tj. greška predviđanja iznosi +/- 16,44 kWh/m^2 na godišnjoj razini što je greška od približno 15 % prema srednjoj vrijednosti svih opservacija. U oba slučaja podskupa, ovaj regresijski model objašnjava približno 80 % opservacija, u skladu s vrijednošću parametra točnosti R^2 (tablica 3-12).

Tablica 3-12 Pokazatelji točnosti početnog regresijskog modela za skup svih stanova

	Podskup za učenje	Podskup za provjeru
R^2	0,799	0,793
RMSE	16,06	16,44

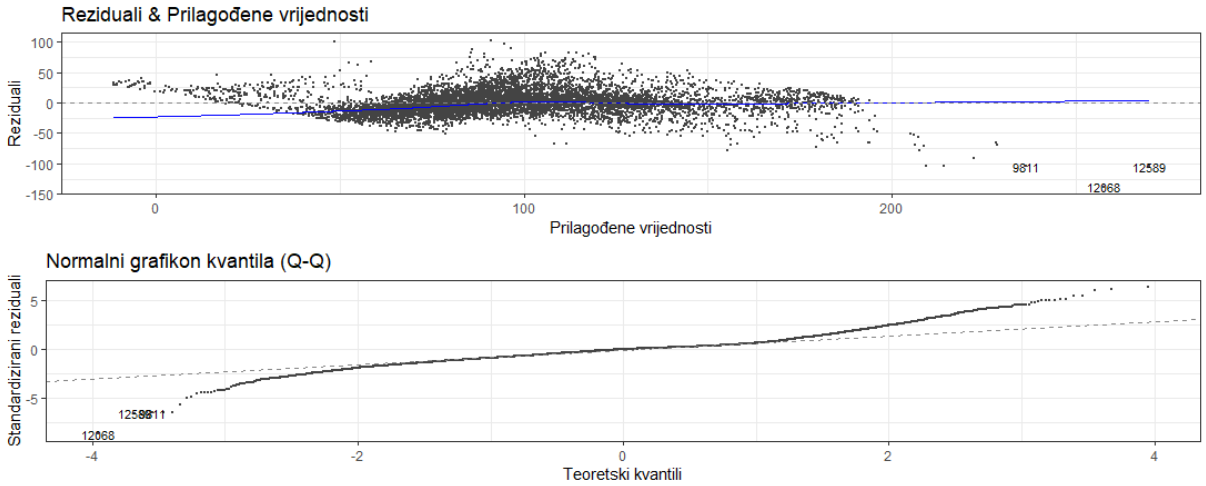
U regresijskom modelu nije poželjno imati varijable koje su međusobno visoko korelirane. Bivarijatnom analizom korelacija numeričkih varijabli u regresijskom modelu, slika 3-31, zamjećuje se visoka korelacija između varijabli *StopaUgr* i *KorekcijskiFaktor* i varijabli *SnagaKK* i *Povrsina*. Ukoliko bi se provodio odabir linearnog modela metodom odabira najboljeg podskupa (poglavlje 2.5.1), iz prije navedenih parova koreliranih varijabli bi se ekspertnom procjenom uklonila jedna varijabla iz parova koreliranih varijabli. U ovom modelu će se korelirane varijable ukloniti postupkom regularizacije, zbog čega nema potrebe za uklanjanjem varijabli metodom najboljeg podskupa prije regularizacije.



Slika 3-31 Korelacija između numeričkih varijabli u početnom regresijskom modelu na razini stana

Prema Gauss-Markovom teoremu [82], obična metoda najmanjih kvadrata (engl. *ordinary least squares* - OLS) daje najbolje rezultate predviđanja od svih drugih linearnih modela, ali samo kada su pretpostavke OLS-a točne za analizirani podatkovni skup. Što su dimenzije analiziranog podatkovnog skupa veće, to je tendencija tog podatkovnog skupa da ne zadovolji pretpostavke OLS-a izglednija. Ukoliko se OLS model ponaša tako da ne zadovoljava neku od pretpostavki, provodi se metoda regularizacije, čime se kontroliraju koeficijenti linearnog modela i smanjuje varijanca i greška pri primjeni modela na skupu podataka na kojem nije rađeno učenje (engl. *out of sample error*).

Prva pretpostavka OLS je da je regresijski model linearan u koeficijentima i u rezidualima, stoga je potrebno vizualizirati odnose reziduala i prilagođenih vrijednosti regresijskog modela, kako je prikazano na slici 3-32. Poželjno je da su reziduali pravilno raspoređeni duž nulte linije, te da rezultirajuća linija (plava linija na slici 3-32) u načelu prati nultu liniju.

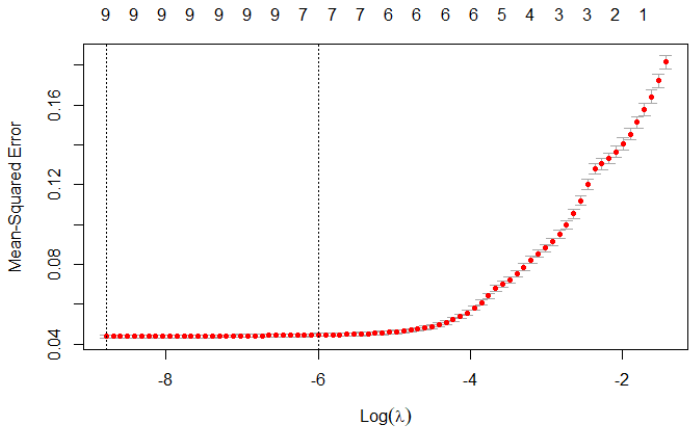


Slika 3-32 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za početni regresijski model za skup svih stanova

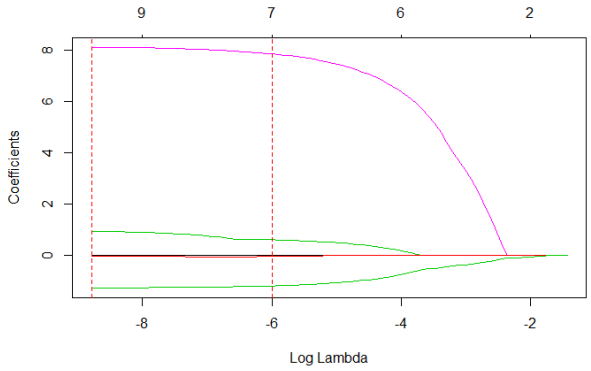
Analizom reziduala naspram prilagođenih vrijednosti, na ovom podatkovnom skupu se vidi uvjet heteroskedastičnosti nije u potpunosti zadovoljen (slika 3-32), te se u idućem koraku provodi regularizacija modela .

Regularizacija

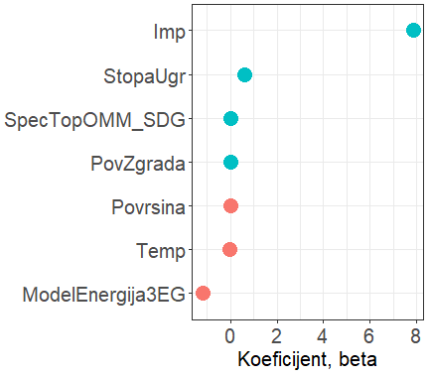
Vrijednost parametra λ, parametra za regularizaciju (opisano u poglavlju 2.5.2), određuju se unakrsnom provjerom. Prikaz 3-33 daje rezultate unakrsne provjere u 10 koraka provedene funkcijom *cv.glmnet* iz R paketa *glmnet* [97]. Zaključuje se da je vrijednost srednje kvadratne greške (MSE) gotovo konstantna za λ između -9 i -4, a da greška iznad vrijednosti -4 značajno raste. Brojevi na gornjem dijelu slika 3-33 i 3-34 predstavljaju broj varijabli u modelu. Također, vidi se da se s promjenom vrijednosti λ mijenja i broj varijabli, tj. da neke od varijabli dobivaju vrijednost nula te su izbačene iz modela.



Slika 3-33 Ovisnost log(λ) i MSE za unakrsnu provjeru u 10 koraka za skup svih stanova



Slika 3-34 Promjena vrijednosti regresijskih koeficijenata u ovisnosti o log λ za skup svih stanova



Slika 3-35 Rezultati *LASSO* regularizacije za skup svih stanova

Vertikalne osi na slici 3-34 označavaju vrijednost log(λ) za minimalnu MSE i vrijednost log(λ) za udaljenost od jedne standardne pogreške od minimalne vrijednosti MSE.

Rezultati *LASSO* regularizacije se vide na slici 3-35. Reducirano je početnih 11 zavisnih varijabli na 7. Iznosi i skup bitnih varijabli dobivenih *LASSO* regularizacijom vide se iz slike 3-35 i tablice 3-13, koja ukazuje da je dominatna varijabla *Imp*, zatim varijable *ModelEnergija*, *StopaUgr* i varijabla *SpecTopOMM_SDG*. Ostale varijable imaju neznatan utjecaj i vrlo su blizu vrijednosti nule.

Tablica 3-13 Regresijski parametri za regularizirani regresijski model svih stanova

	$\boldsymbol{\beta}$
(odsječak)	3,426
SpecTopOMM_SDG	0,011
StopaUgr	0,614
PovZgrada	0,000
Povrsina	-0,003
Imp	8,124
Temp	-0,028
ModelEnergija3EG	-1,192

Kako je pretpostavljeno, vidimo da je regularizacija eliminirala korelirane vrijednosti koje su prethodno detektirane kao nebitne. Tako više u modelu nemamo varijablu *KorekcijskiFaktor*, koja je visoko korelirana s varijablom *StopaUgr* koja je ostala u modelu. Također, iz modela je uklonjena i varijabla *SnagaKK*, koja visoko korelira s varijablom *Povrsina* koja je ostala u modelu.

Parametri modela su u različitim mjernim jedinicama. Regularizacija standardizira kontinuirane varijable prije provođenja višestruke linearne regresije. Na taj način se dobije model koji ima niže iznose, i apsolutne razlike između regresijskih parametra β, i može se koristiti u predviđanju, ali se regularizirani parametri ne mogu direktno interpretirati. S obzirom da je jedan od ciljeva analize intepretirati pojedine utjecajne parametre na potrošnju, regularizirani regresijski model se koristi za odabir utjecajnih parametara, a konačni regresijski model se određuje s varijablama koje nisu standardizirane, kako je opisano u nastavku.

Konačni model

Konačni model u prvom koraku sadrži varijable koje su rezultat modela dobivenog *LASSO* regularizacijom prema slici 3-35.

Regresijski model prikazan u tablici 3-14 dominatno označava varijablu *Imp*, zatim varijablu *ModelEnergija*, te *StopaUgr*, dok su ostale varijable (*SpecTopOMM_SDG*, *Povrsina* i *PovZgrada*) manje dominantne.

Tablica 3-14 Regresijski parametri i statistički pokazatelji konačnog regresijskog modela za skup svih stanova

	β	St. pogreška	t- vrijednost	p - vrijednost
(Intercept)	29,029	6,709	4,327	0,000
SpecTopOMM_SDG	0,955	0,008	112,555	0,000
StopaUgr	80,142	0,805	99,549	0,000
PovZgrada	0,002	0,000	37,793	0,000
Povrsina	-0,258	0,008	-30,398	0,000
Imp	639,666	5,179	123,501	0,000
Temp	-3,415	1,167	-2,927	0,003
ModelEnergija3EG	-123,126	0,734	-167,652	0,000

Redukcijom parametara provođenjem regularizacije, vrijednost R^2 na skupu za provjeru je ostala na istoj razini kao i u početnom modelu (tablica 3-12), dok je vrijednost RMSE na jednakoj razini i iznosi 16,48 kWh/m^2 (tablica 3-15).

Tablica 3-15 Pokazatelji točnosti početnog regresijskog modela

	Podskup za učenje	Podskup za provjeru
R^2	0,797	0,791
RMSE	16,13	16,48

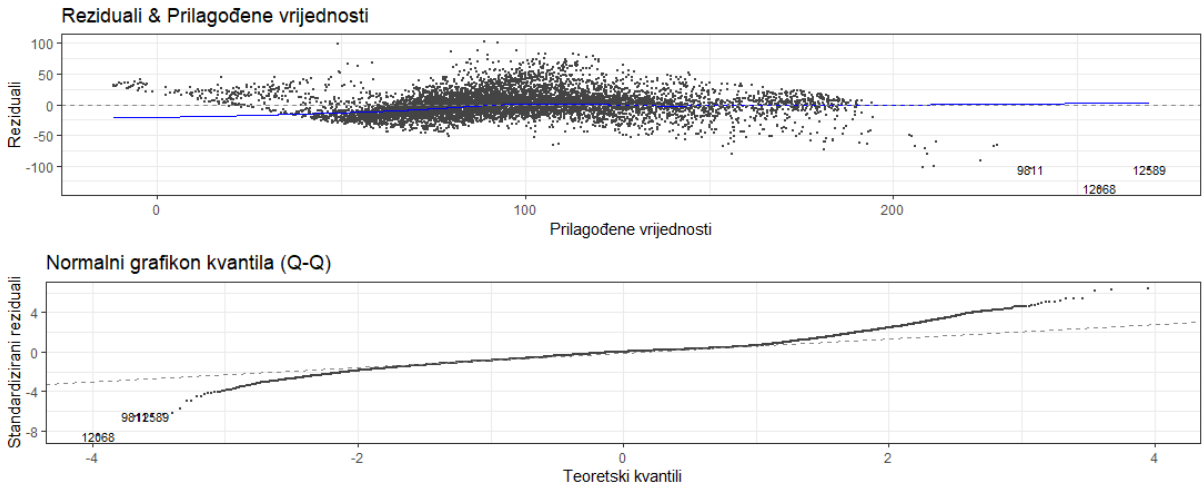
Pokazatelji točnosti ukazuju da je točnost finalnog, jednostavnijeg i intepretativnijeg regresijskog modela na jednakoj razini kao kompleksnijeg početnog modela. Ovime se potvrđuje korisnost regularizacije kod modela s velikim brojem parametara.

U dobivenom finalnom modelu vidi se dominacija regresijskog parametra uz varijablu *Imp* koja je i očekivana s obzirom da je:

$$Imp = \frac{ImpulsiKK}{ImpulsiOMM} \qquad (61)$$

Srednja vrijednost ove dvije varijable za podatkovni skup *psStan* prema tablici 3-3 jednaka je 0,007. Tako vrijednost umnoška regresijskog koeficijenta i omjera srednjih vrijednosti varijabli *ImpulsiKK* i *ImpulsiOMM* pridonosi predviđanju specifične potrošnje stana od približno 4 kWh/m^2 na godišnjoj razini, što je očekivani red veličine.

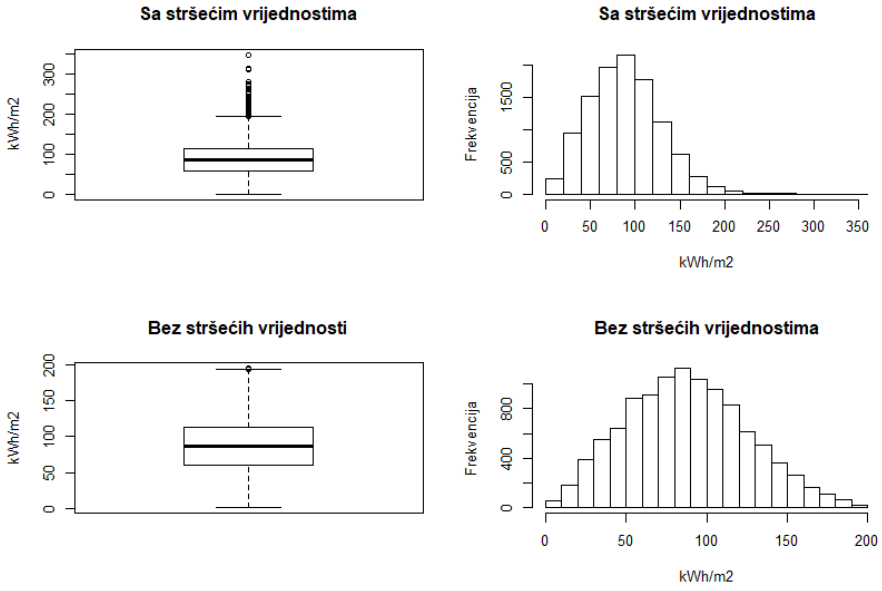
Prikazi vizualizacije modela dani su na slici 3-36 iz koje se može vidjeti da su reziduali regresijskog modela nakon regularizacije relativno jednoliko raspoređeni uzduž nulte horizontalne linije te se može reći da je većinom zadovoljen uvjet heteroskedastičnosti, uz blagu naznaku potencijalne nelinearnosti. Ukoliko nelinearnost postoji, pretpostavka je da će algoritmi koji će se koristiti u kasnijim analizama (regresijska stabla, slučajne šume, stroj potpornih vektora), imati višu točnost predviđanja od modela višestruke linearne regresije.



Slika 3-36 Dijagram reziduala i prilagođenih vrijednosti za konačni regresijski model za skup svih stanova

3.4.1.2. Regresijski model za stanove s ugrađenim RTTE

Posebno se analiziraju stanovi koji su ugradili individualna mjerila. U prvom koraku se podskup analizira s ciljem utvrđivanja postojanja stršećih vrijednosti. Ponovno nalazimo određeni broj stršećih vrijednosti koji čine oko 0,8 % svih opservacija sa srednjom vrijednošću od 220,9 kWh/m^2. Nakon uklanjanja stršećih vrijednosti srednja vrijednost svih opservacija blago pada, s 88,6 kWh/m^2 na 87,47 kWh/m^2. Primjećuje se da je srednja vrijednost specifične potrošnje topline za stanove s ugrađenim individualnim mjerenjem znatno niža nego što je srednja vrijednost čitavog podatkovnog skupa, kako je opisano u tablici 3-3. Utjecaj stršećih vrijednosti na ovaj podskup prikazan je na slici 3-37.



Slika 3-37 Utjecaj stršećih vrijednosti na analizirani skup za regresijski model na stanove s ugrađenim RTTE

Iz analiziranog podatkovnog skupa se uklanjaju stršeće vrijednosti te se daljnje analize provode na reduciranom skupu.

Početni model

Daljnje analize su provedene konzistentno analizi regresijskog modela za čitav podatkovni skup *psStan*. U prvom koraku se tako razvija početni regresijski model, dobivene vrijednosti se interpretiraju te se provodi postupak odabira linearnog modela postupkom regularizacije.

Dobiveni početni regresijski model korištenjem R paketa *stats* prepoznaje sve varijable iz podskupa značajnim (tablica 3-16), tj. svima je p-vrijednost manja od 0,05. Ponovno je dominantna varijabla *Imp*, zatim *StopaUgr* i *Temp*.

Tablica 3-16 Regresijski parametri i statistički pokazatelji početnog regresijskog modela za stanove koji su ugradili RTTE

	β	St. pogreška	t- vrijednost	p - vrijednost
(odsječak)	-100,061	10,569	-9,468	0,000
SpecTopOMM_SDG	0,650	0,013	48,566	0,000
SnagaKK	1,184	0,200	5,926	0,000
StopaUgr	150,648	5,718	26,345	0,000
PovZgrada	0,006	0,000	65,714	0,000
Povrsina	-1,029	0,029	-35,437	0,000
Imp	799,703	5,558	143,871	0,000
KorekcijskiFaktor	0,071	0,025	2,808	0,005
Temp	-10,582	1,711	-6,185	0,000

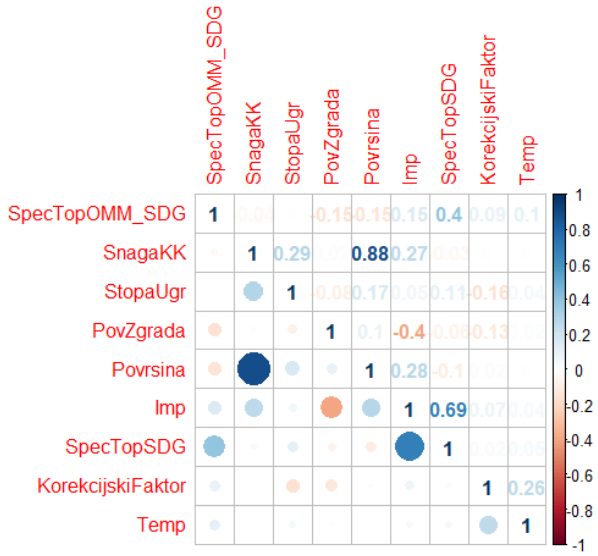
Prema tablici 3-17, pokazatelji točnosti su nešto slabiji nego za čitav podatkovni skup *psStan* što ukazuje na veću varijabilnost u opservacijama, koja se može pripisati izraženijim bihevioralnim utjecajima kod ovog skupa krajnjih korisnika.

Tablica 3-17 Pokazatelji točnosti početnog regresijskog modela stanove s RTTE

	Podskup za učenje	Podskup za provjeru
$\mathbf{R^2}$	0,762	0,760
RMSE	18,33	18,14

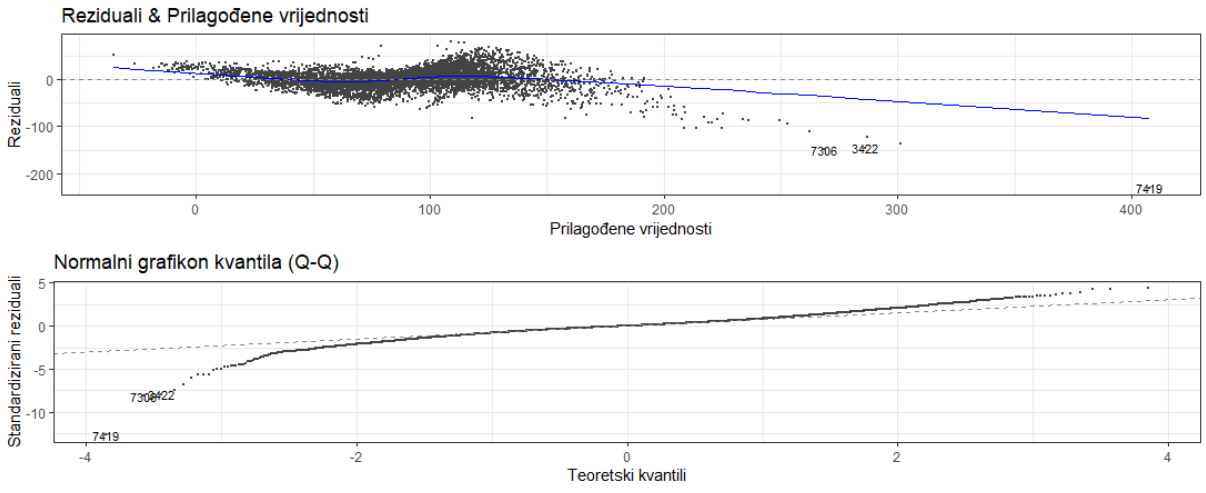
U modelima u kojima se nalazi visoka varijabilnost između regresijskih parametara, što je slučaj u modelu u tablici 3-16, postoji indikacija za potencijalnu kolinearnost među varijablama. Kolinearnost nije poželjna te je u parovima kolinearnih varijabli cilj izbaciti jednu od varijabli, kako bismo smanjili broj varijabli u modelu i izbjegli prenaučenost.

Iz slike 3-38 odnosi između varijabli su nešto drugačiji nego u slučaju čitavog podatkovnog skupa za stanove. Ipak, ponovno se visoka korelacija vidi između varijabli *SnagaKK* i *Povrsina*, što je očekivano jer je vrijednost *SnagaKK* (zakupljena toplinska snaga) vezana za grijanu površinu. Druge varijable nisu visoko korelirane, viša korelacija se vidi između same specifične potrošnje toplinske energije grijanog prostora, varijable *SpecTopSDG*, i omjera broja impulsa *Imp*. Među ovim varijablama se i očekuje visoka korelacija, posebice jer je je regresijski parametar uz varijablu *Imp* dominantan prema zavisnoj varijabli *SpecTopSDG*.



Slika 3-38 Korelacija između numeričkih varijabli u početnom regresijskom modelu za stanove koji su ugradili RTTE

Postupkom vizualizacije prikazujemo neželjene distribucije između reziduala i prilagođenih vrijednosti (slika 3-39). Distribucija normalnosti reziduala je u načelu zadovoljavajuća, osim u krajnjim vrijednostima. Regularizaciju provodimo s ciljem smanjenja varijance bez povećanja pristranosti modela.



Slika 3-39 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za početni regresijski model za stanove koji su ugradili RTTE

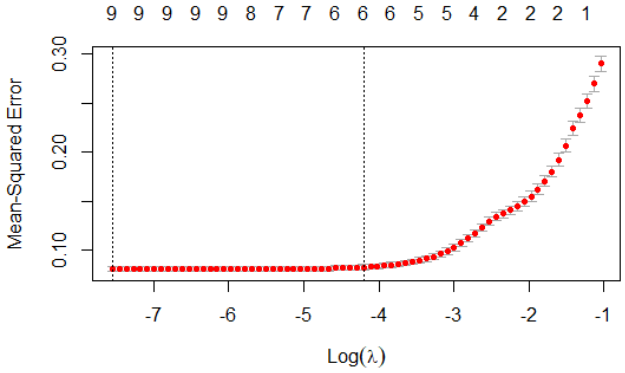
Regularizacija

Regularizacija se provodi istim postupkom kao u poglavlju 3.4.1.1. Unakrsnom provjerom u 10 koraka se određuje optimalna vrijednost parametra λ koji se dobiva za udaljenost jedne standardne pogreške od minimalne vrijednosti MSE. Na ovaj način reducira se broj varijabli, bez da se utječe na pogrešku, tj. broj varijabli se smanjuje s 9 na 6, kako se vidi na slici 3-40. Također, smanjenjem vrijednosti parametara, tj. odabirom takve vrijednosti λ koje minimiziraju broj koeficijenata, smanjujemo i apsolutni iznos koeficijenata (tablica 3-18), te se iznosi približavaju vrijednosti nula, prema slici 3-41. Teoretski bi broj varijabli mogli proizvoljno smanjiti, ali bi na taj način utjecali na točnost predviđanja. Ukoliko bi u modelu ostavili samo 2 varijable, utoliko bi točnost predviđanja bila duplo manja.

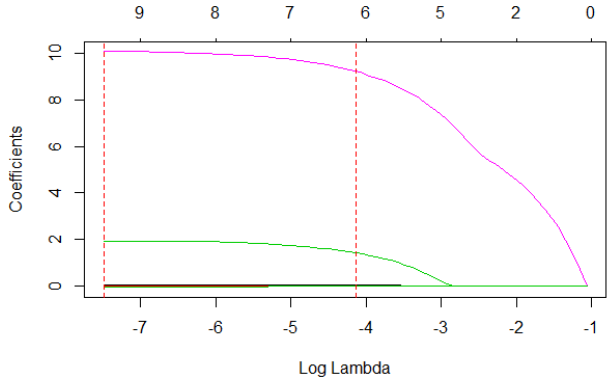
Tablica 3-18 Regresijski parametri za regularizirani regresijski model stanova s RTTE

	$\boldsymbol{\beta}$
(odsječak)	1.398
SpecTopOMM_SDG	0.013
StopaUgr	1.415
PovZgrada	0.000
Povrsina	-0.007
Imp	9.230
KorekcijskiFaktor	-0.001

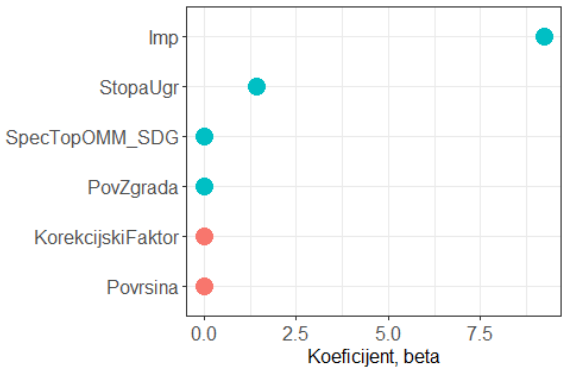
Konačan regularizirani *LASSO* regresijski model se vidi na slici 3-19, s izborom 6 značajnih varijabli za konačni model.



Slika 3-40 Ovisnost log(λ) i MSE za unakrsnu provjeru u 10 koraka za stanove sa ugrađenim RTTE



Slika 3-41 Promjena vrijednosti regresijskih koeficijenata u ovisnosti o log λ za stanove s RTTE



Slika 3-42 Rezultati *LASSO* regularizacije za stanove sa ugrađenim RTTE

Jednako kao i kod regularizacije za podatkovni skup svih stanova, ovdje regularizaciju koristimo da bi odredili utjecajne varijable za konačni model bez standardizacije varijabli.

Konačni model

Prema izboru varijabli, *LASSO* regularizacijom dobivamo model prikazan u tablici 3-19. Pri analizi korelacije prethodno je identificirana visoka korelacija između varijabli *KorekcijskiFaktor* i *StopaUgr*, što se u modelu i vidi iz toga što varijabla *KorekcijskiFaktor* ima p-vrijednost višu od 0,05. Iz modela se uklanja varijabla *KorekcijskiFaktor* i u drugom koraku se definira konačni model prema tablici 3-20.

Tablica 3-19 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za stanove s RTTE nakon *LASSO* regularizacije

	β	St. pogreška	t- vrijednost	p - vrijednost
(odsječak)	-161,245	7,182	-22,452	0,000
SpecTopOMM_SDG	0,653	0,015	44,004	0,000
StopaUgr	156,543	6,262	24,998	0,000
PovZgrada	0,006	0,000	56,750	0,000
Povrsina	-0,863	0,016	-53,546	0,000
Imp	794,164	6,420	123,693	0,000
KorekcijskiFaktor	0,020	0,029	0,717	0,474

Testiranje točnosti modela provodi se na način da se ukloni svaka pojedinačna varijabla ili obje, koje imaju vrijednosti blizu nula (*PovZgrada* i *Povrsina*). U svim slučajevima primjećuje se značajan pad točnosti modela, tako da se sve varijable iz tablice 3-20 zadržavaju u modelu, kako bi se dobila zadovoljavajuća točnost prema tablici 3-21.

Tablica 3-20 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za stanove s RTTE nakon *LASSO* regularizacije nakon izbacivanja varijable *KorekcijskiFaktor*

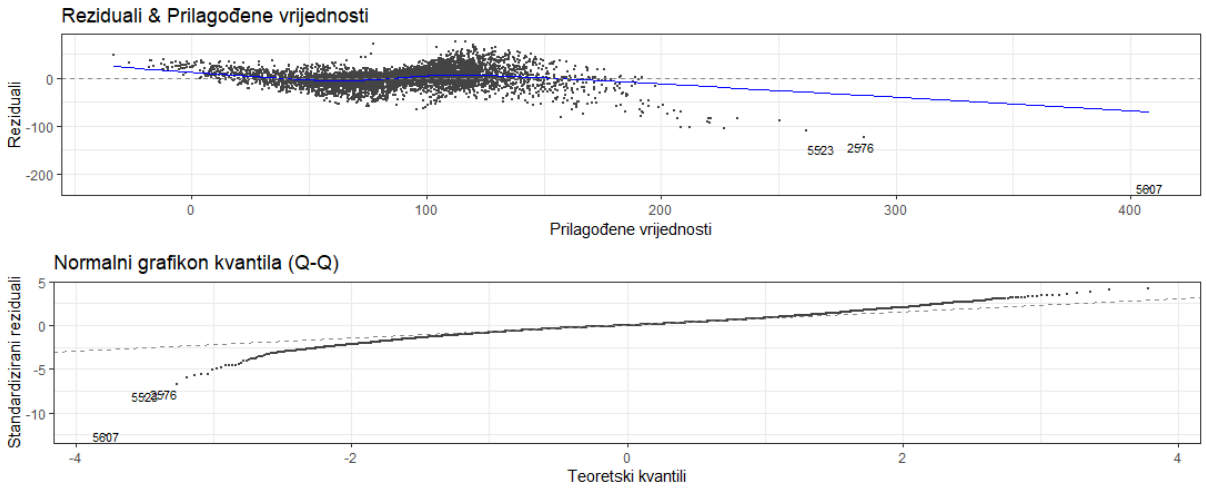
	β	**St. pogreška**	**t- vrijednost**	**p - vrijednost**
(odsječak)	-158,526	6,098	-25,996	0
SpecTopOMM_SDG	0,653	0,015	44,178	0
StopaUgr	155,749	6,163	25,270	0
PovZgrada	0,006	0,000	57,148	0
Povrsina	-0,863	0,016	-53,648	0
Imp	794,032	6,418	123,728	0

Tablica 3-21 Pokazatelji točnosti finalnog regresijskog modela za stanove s RTTE

	Podskup za učenje	**Podskup za provjeru**
R^2	0,759	0,759
RMSE	18,48	18,24

Ako želimo kvantificirati grešku predviđanja sa stajališta krajnjih potrošača i njihove očekivane potrošnje na godišnjoj razini, može se reći da ona iznosi +/- 18,24 kWh/m^2, a s MLR modelom se opisuje 76 % svih opservacija prema vrijednosti parametra točnosti R^2.

Prikazi vizualizacije finalnog modela su dani na slici 3-43 iz koje se može vidjeti da su nakon regularizacije reziduali relativno jednako raspoređeni uzduž nulte horizontalne linije, te se zaključuje da je većinom zadovoljen uvjet heteroskedastičnosti, uz dodatno zadovoljen uvjet normalnosti distribucije reziduala.



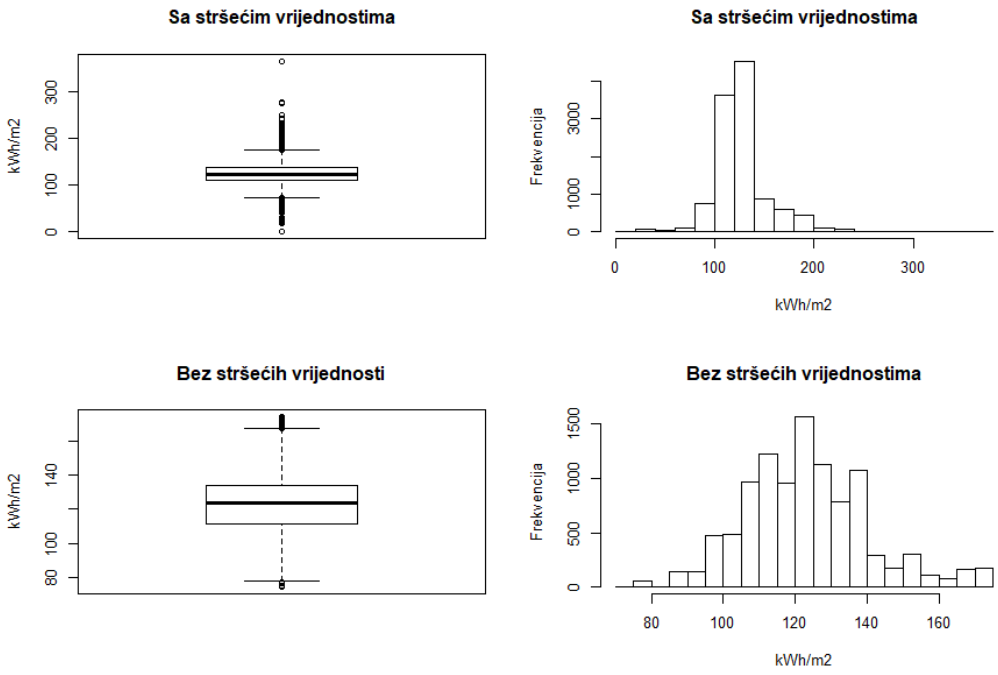
Slika 3-43 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za konačni regresijski model za psStan, model 3EG

3.4.1.3. Regresijski model za stanove bez ugrađenog mjerenja

Posljednji regresijski model za razinu stanova će biti razvijen za one stanove koji nisu ugradili individualna mjerila, prvenstveno kako bi se ocijenili utjecajni parametri na te stanove i potencijalni razlozi zašto krajnji korisnici iz ovog podskupa nisu motivirani da ugrade RTTE.

Pri analizi ovog podatkovnog skupa ponovno prepoznajemo stršeće vrijednosti. Pojavnost stršećih vrijednosti kod ove skupine potrošača može biti uslijed činjenice da se stanovi nalaze u tehnički lošijim zgradama, da se nalaze u zgradama u kojima je stopa ugradnje niska ili je jednaka nuli, ili se nalaze u zgradama u kojima je započeo postupak ugradnje ali nije postignuta razina od 100 % ugrađenosti.

Ukupno se prepoznaje 9,6 % stršećih vrijednosti, kojima je srednja vrijednost 166,05 kWh/m^2. Srednja vrijednost svih opservacija je 126,99 kWh/m^2 sa stršećim vrijednostima, te 123,23 kWh/m^2 nakon uklanjanja stršećih vrijednosti (slika 3-44).



Slika 3-44 Utjecaj stršećih vrijednosti na analizirani podskup skup za regresijski model za stanove bez RTTE

Stršeće vrijednosti se uklanjanju iz podatkovnog skupa, te se analiza nastavlja bez tih opservacija. U ovom podskupu podataka se očekuje jednostavniji i točniji model nego što je to slučaj s regresijskim modelom stanova koji su ugradili RTTE. Razlog ovoj pretpostavci je

činjenica da krajnji potrošači u ovoj skupini stanova ne mogu utjecati na svoju potrošnju, stoga nema utjecaja bihevioralnih parametara.

Početni model

Računski postupak izrade modela je jednak kao i za prethodna dva podatkovna skupa. Početni model (tablica 3-22) pokazuje da stanovima koji nemaju ugrađeno mjerenje stopa ugradnje nije relevantna varijabla za potrošnju, što je konzistentno s načinom alokacije topline za ovaj tip stanova, a prema Pravilniku [56].

Tablica 3-22 Regresijski parametri i statistički pokazatelji za početni regresijski model za stanove bez RTTE

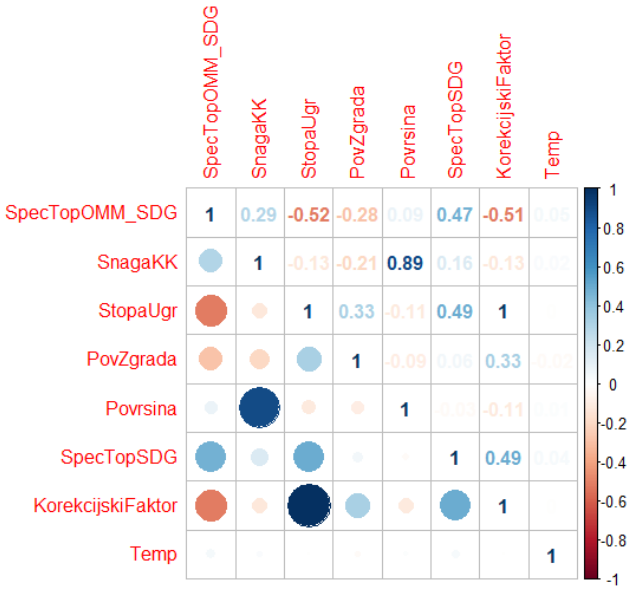
	β	St. pogreška	t- vrijednost	p - vrijednost
(odsječak)	5,605	1,798	3,117	0,002
SpecTopOMM_SDG	0,983	0,003	339,221	0,000
SnagaKK	-0,094	0,038	-2,497	0,013
StopaUgr	9,024	2,201	4,100	0,000
PovZgrada	0,000	0,000	4,460	0,000
Povrsina	0,016	0,005	2,967	0,003
KorekcijskiFaktor	0,718	0,021	34,883	0,000
Temp	-0,845	0,311	-2,717	0,007

Pokazatelji točnosti početnog regresijskog modela za 2EG, prema tablici 3-23, su očekivano visoki, ali postoji mogućnost visoke kolinearnosti među varijablama i pojave prenaučenosti.

Tablica 3-23 Pokazatelji točnosti početnog regresijskog modela za stanove bez RTTE

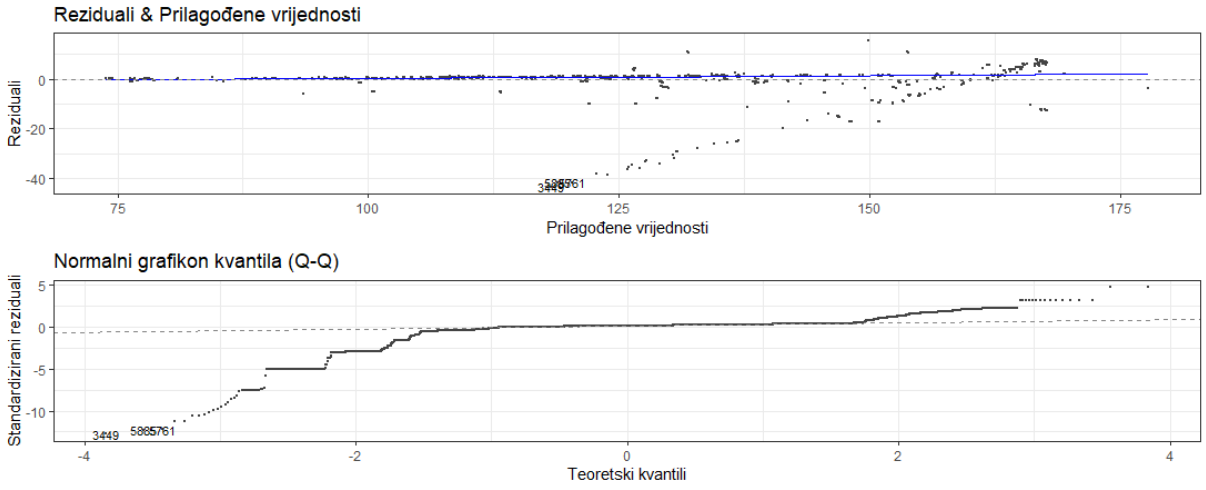
	Podskup za učenje	Podskup za provjeru
$\mathbf{R^2}$	0,963	0,955
RMSE	3,415	3,699

Kolinearnost među varijablama se ne razlikuje značajno od kolinearnosti u ostala dva regresijska modela za stanove. Na slici 3-45 ponovno vidimo visoku kolinearnost između *StopaUgr* i *KorekcijskiFaktor*, za ovaj podatkovni skup je ta korelacija savršena i iznosi 1, te korelaciju između varijabli *Povrsina* i *SnagaKK*, čija je tehnička povezanost objašnjena u poglavlju 3.4.1.2.



Slika 3-45 Korelacija između numeričkih varijabli početnog regresijskog modela za stanove bez RTTE

Vizualizacijom modela na slici 3-46 uočljive su nepravilnosti i značajan utjecaj na normalnost distribucije reziduala.



Slika 3-46 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za početni regresijski model za stanove bez RTTE

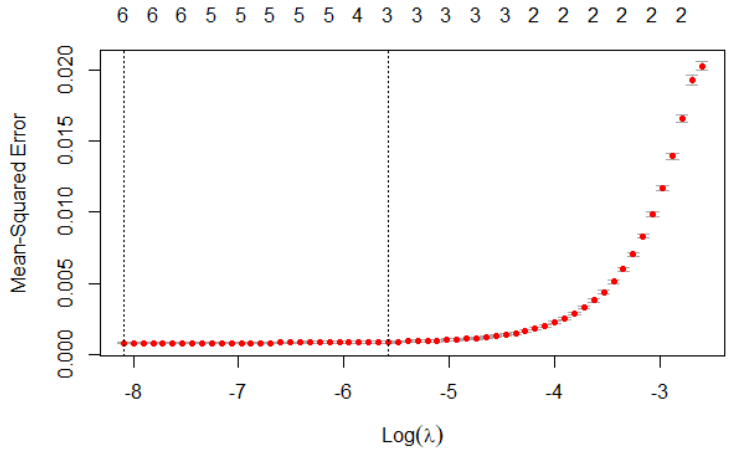
Regularizacija

Na skupu podataka provodi se *LASSO* regularizacija, s ciljem reduciranja varijabli. Regresijska analiza za stanove koji nisu ugradili individualno mjerenje je u početnom modelu pokazala svoju linearnost, a s obzirom na to da je način obračuna za ovakve stanove u potpunosti nezavisan o bihevioralnim parametrima, pretpostavlja se da se model može dodatno pojednostaviti.

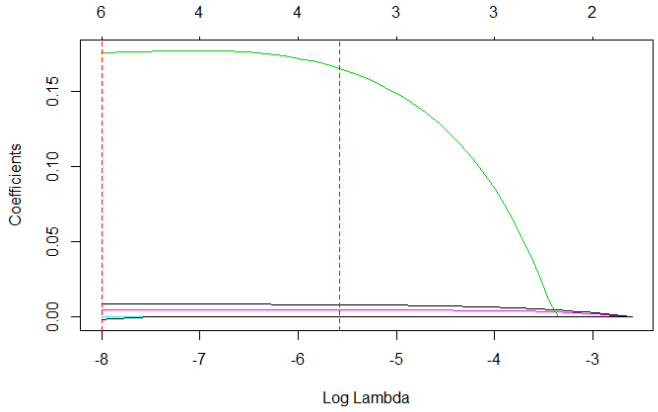
Tablica 3-24 Regresijski parametri za regularizirani regresijski model stanova bez RTTE

	β
(odsječak)	3,849
SpecTopOMM_SDG	0,008
StopaUgr	0,166
KorekcijskiFaktor	0,004

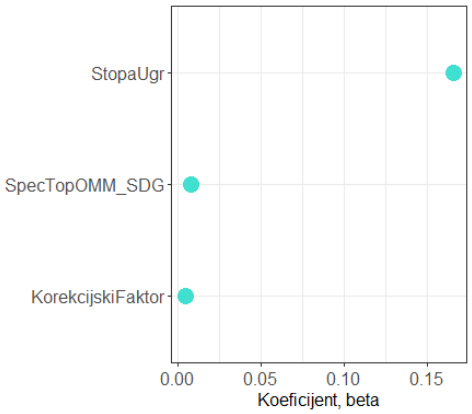
Regularizacija započinje sa 6 varijabli te se svede na 3 varijable pri log λ približno -5,5 (slike 3-47 i 3-48). Tri varijable koje nisu jednake nuli su *SpecTopOMM_SDG*, *StopaUgr* i *KorekcijskiFaktor*, s iznosima regresijskih parametara prema tablici 3-24 i slici 3-49.



Slika 3-47 Ovisnost log(λ) i MSE za unakrsnu provjeru u 10 koraka za stanove bez RTTE



Slika 3-48 Promjena vrijednosti regresijskih koeficijenata u ovisnosti o log λ za stanove bez RTTE



Slika 3-49 Rezultati LASSO regularizacije za stanove bez RTTE

Konačni model

Rezultati izbora varijabli za regresijski model su u skladu s očekivanim te se pokazuje da je model izabrao upravo one varijable koje i jesu najutjecajnije za alokacije topline u stanovima koji nemaju individualno mjerenje, a nalaze se u zgradama koje su provele ugradnju razdjelnika. Prema Pravilniku [56] toplina se ovim stanovima alocira zavisno o varijabli *SpecTopOMM_SDG* i *KorekcijskiFaktor*. Regresijski model, osim ove dvije varijable, prepoznaje i treću utjecajnu varijablu, a ta je *StopaUgr*. Tako se iz tablice 3-25 vidi da je varijabla *StopaUgr* dominantna i pozitivna, što ukazuje na to da se s povećanom stopom ugradnje unutar zgrade, alocirana toplina za one stanove koji nisu ugradili individualno mjerenje povećava.

Tablica 3-25 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za stanove bez RTTE

	β	St. pogreška	t- vrijednost	p - vrijednost
(odsječak)	1,993	0,348	5,718	0,000
SpecTopOMM_SDG	0,978	0,003	344,895	0,000
StopaUgr	7,508	2,529	2,968	0,003
KorekcijskiFaktor	0,732	0,024	31,018	0,000

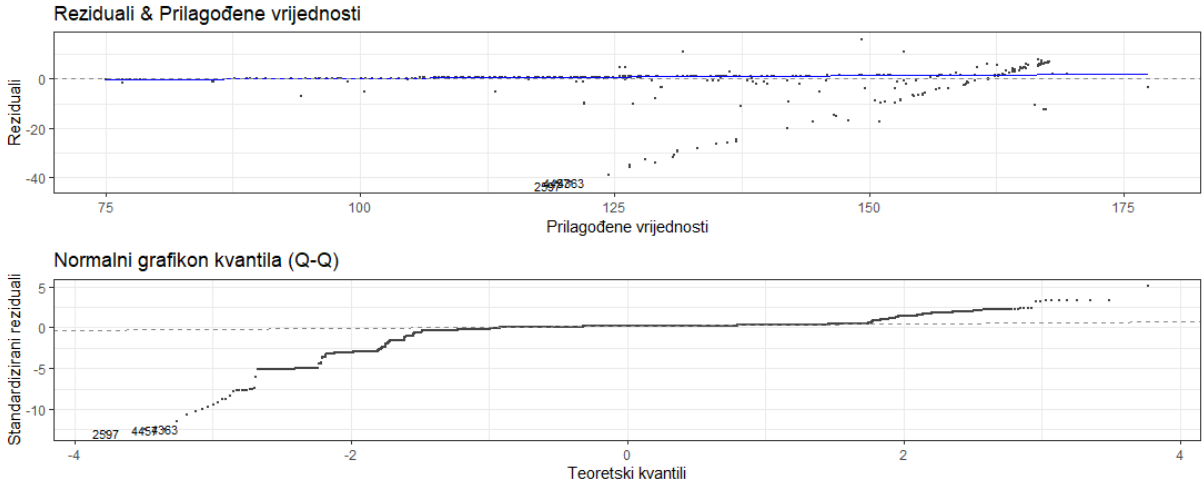
Pokazatelji točnosti za ovaj regularizirani regresijski model su vrlo visoki. Prema tablici 3-26 RMSE greške predviđanja su na razini od 3,63 kWh/m^2 godišnje, što je značajno manje od greške za podskup stanova koji imaju RTTE (tablica 3-21) ili točnosti koju daje regresijski model za čitav podatkovni skup stanova (tablica 3-15).

Tablica 3-26 Pokazatelji točnosti finalnog regresijskog modela za stanove bez RTTE

	Podskup za učenje	Podskup za provjeru
R^2	0,963	0,958
RMSE	3,38	3,63

Tako se može objasniti i činjenica da su pojedini stanovi koji su ugradili RTTE nezadovoljni novim sustavom na način da u komparaciji sa stanovima koji nemaju ugrađene RTTE potencijalno plaćaju isto ili možda čak i više, ali imaju veću varijabilnost potrošnje i manju točnost predviđanja.

Dobiveni regresijski model, prema slici 3-50, pokazuje ravnomjerni raspored reziduala i prilagođenih vrijednosti, i ukazuje na linearnost modela. Pozitivne vrijednosti reziduala indiciraju da je model podcijenio rezultate, dok negativne da ih je precijenio. U prosjeku, model predviđa na vrlo zadovoljavajućoj razini.



Slika 3-50 Dijagram reziduala konačni regresijski model za stanove bez RTTE

3.4.2. Regresijska stabla

Regresijska stabla se razvijaju korištenjem R paketa *rpart* [99] i *randomForest* [100]., dok su vizualizacije napravljene s R paketima *rpart.plot* [101] i *caret* [102].

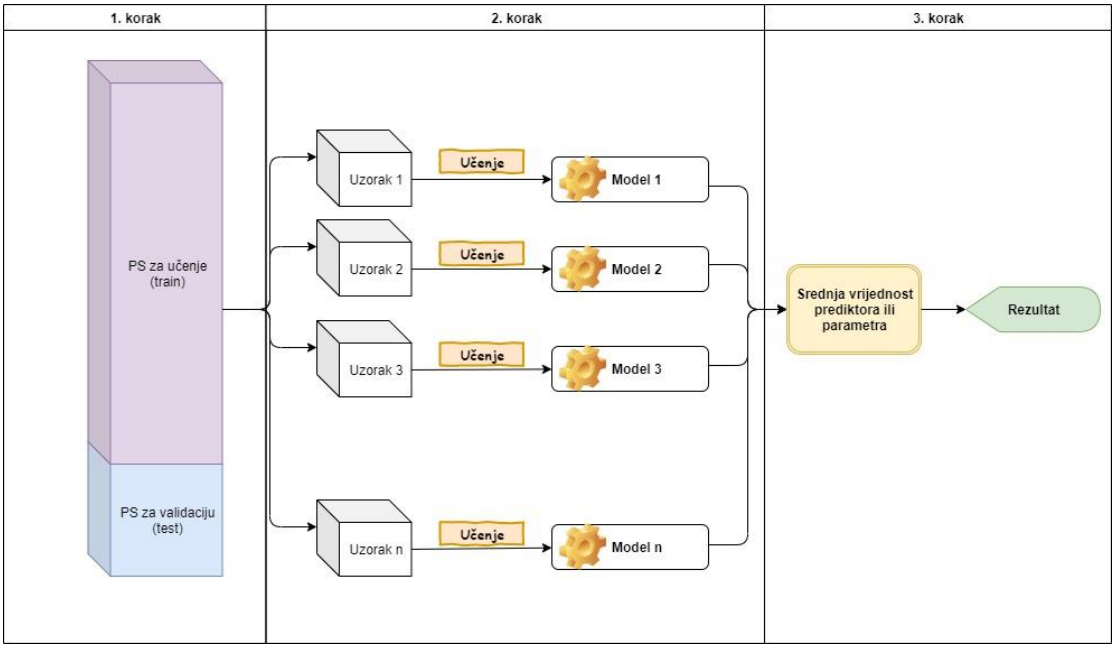
Kako je opisano u poglavlju 2.6.2, modeliranje metodom regresijskih stabala temelji se na optimizaciji između broja čvorova grananja, tj. složenosti stabla, i točnosti predviđanja korištenjem optimizacijskog parametra α. Pojednostavljeno, cilj je pronaći „najmanje“ stablo koje daje najmanju grešku predviđanja na način da prvo izgradimo „najveće“ moguće stablo na kojem provodimo „rezidbu“.

Postupak rezidbe, ili skraćivanja, regresijskog stabla opisan je u nastavku:

1. Rekurzivnim binarnim dijeljenjem i na temelju ulaznih podataka promatranja, dobiva se veliko stablo. Grananje stabla staje kada svaki krajnji čvor sadržava dovoljno mali broj promatranja.
2. Primjenom metode smanjenja složenosti na veliko stablo, dobiva se niz najboljih podstabala u ovisnosti o funkciji α.
3. *K*-grupnom unakrsnom provjerom određuje se α. Odnosno, ulazna promatranja dijele se na *K* grupa. Za svaki $k = 1, \dots, K$:
    a. Koraci 1. i 2. ponavljaju se sa svim ulaznim podacima, osim s *k*-tim podatkom.
    b. Procjenjuje se srednja kvadratna pogreška prediktora u funkciji od α, na temelju podatka u *k*-toj izostavljenoj grupi. Vrijednosti, dobivene za svaki α, usrednjuju se i zatim se odabire α kako bi se smanjila srednja pogreška.
4. Postupak se vraća na podstablo iz 2. koraka koje odgovara odabranoj vrijednosti α.

Kod modela pojedinačnog stabla javlja se visoka varijanca. Rezidba stabla pomaže pri smanjenju varijance, no postoje metode koje iskorištavaju varijabilnost pojedinačnog stabla tako da se značajno poboljšaju njegove performanse. *Bagging* kombinira i usrednjuje više modela. Usrednjenje više stabala smanjuje varijabilnost bilo kojeg stabla i sprječava preodređivanje i poboljšava točnost predviđanja. *Bagging* se provodi u tri jednostavna koraka, kako slijedi i kako je prikazano na slici 3-51:

1. Najprije se stvaraju *m bootstrap* uzorka iz skupa za učenje. Uzorci na kojima je proveden *bootstrap* omogućuju stvaranje više različitih skupova podataka, ali s distribucijom kao i skup za učenje.
2. Učenje se provodi za svaki *bootstrap* uzorak i izrađuje se regresijsko stablo.
3. Pojedini prediktori iz svakog stabla usrednjuju se kako bi se stvorila sveukupna srednja vrijednost predviđanja.



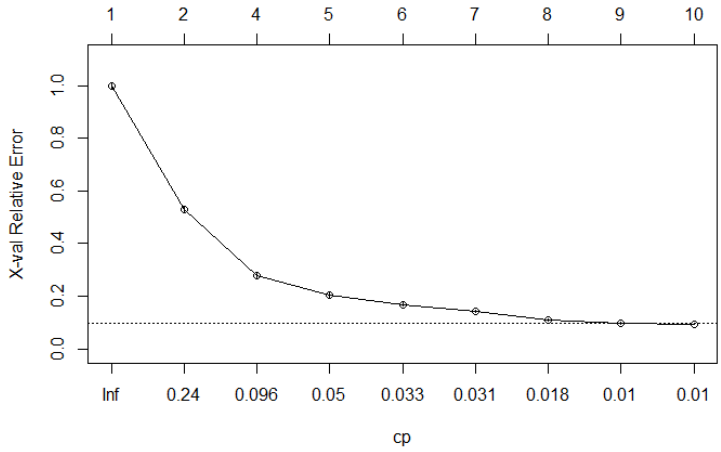
Slika 3-51 Postupak provođenja *bagging* metode za regresijska stabla

Ovaj postupak može se primijeniti za poboljšanje svakog regresijskog ili klasifikacijskog modela koji ima visoku varijancu. Primjena ovog postupka na stabilnim parametarskim modelima, kao što je linearna regresija, ne rezultira značajnim poboljšanjem. Prednost *bagging*-a je to da će *bootstrap* uzorak u prosjeku sadržavati 65 % podataka za učenje, a ostalih 35 % podataka ostaje izvan *bootstrap* uzorka. Taj preostali uzorak naziva se *out-of-bag* (OOB) uzorak. OOB promatranja mogu se iskoristiti za procjenu točnosti modela prilikom koje se provodi unakrsna provjera.

3.4.2.1. Model za čitav podatkovni skup stanova

Za podatkovni skup odabire se optimalno stablo postupkom skraćivanja stabla. R paket *rpart* [99] traži optimalnu vrijednost parametra ugađanja α unakrsnom provjerom u 10 koraka na skupu za provjeru. Rezultati unakrsne provjere su prikazani na slici 3-52. Vidimo da se greška unakrsne provjere na skupu za provjeru (y-os) stabilizira za $|T| = 10$ terminalnih čvorova

(vrijednosti na gornjoj x-osi označavaju $|T|$) i $\alpha = 0{,}1$ (vrijednost označena s cp na donjoj x-osi je α).

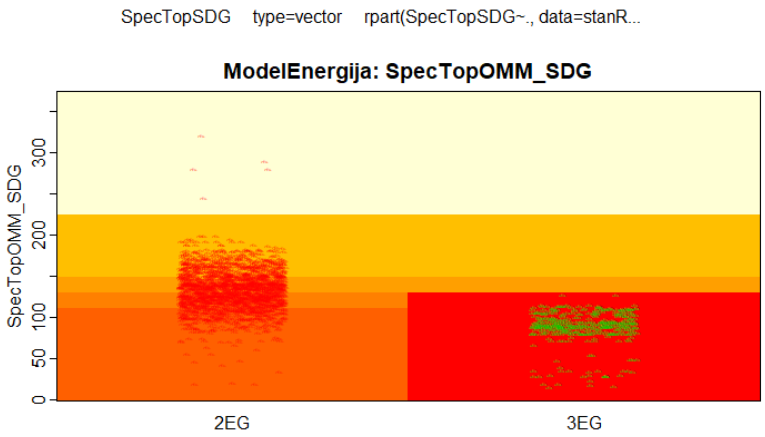


Slika 3-52 Rezultati unakrsne provjere za učenje regresijskim stablima za sve stanove

Funkcija *rpart* dodatno optimizira regresijska stabla na način da za najmanju grešku unakrsne provjere računa vrijednosti najmanjeg broja opservacija (engl. *minsplit*) potrebnih da se provodi grananje prije nego se točka grananja proglasi terminalnim čvorom. Smanjivanjem ove vrijednosti dobiva se stablo s više listova. Drugi optimizacijski parametar određuje veličinu stabla, tj. maksimalni broj grananja (engl. *maxdepth*) između korjena i terminalnog čvora.

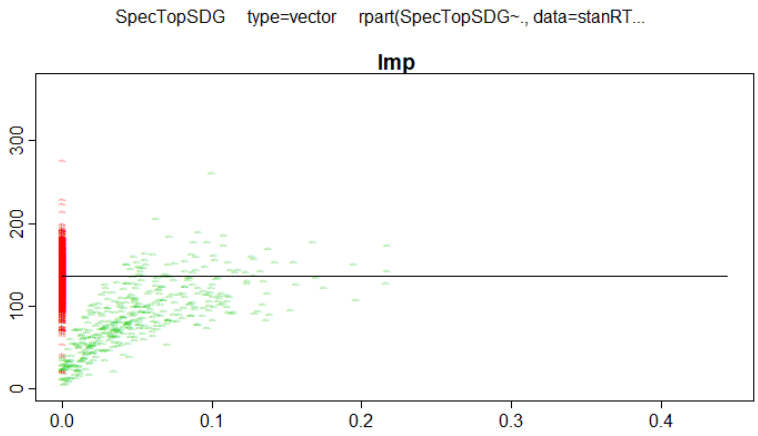
Za ovaj podatkovni skup optimalni broj opservacija *minsplit* je 10, a maksimalni broj grananja *maxdepth* je 20. Izgrađeno regresijsko stablo je prikazano na slici 4-1, a na slici 3-53 analizira se stablo prema varijabli *ModelEnergija* i prema pravokutnim područjima R_J.

Izgrađeno regresijsko stablo provodi dijeljenja prema 5 varijabli, a od interesa je pokazati razliku između modela 3EG (stanovi s ugrađenim RTTE) i modela 2EG (stanovi bez ugrađenih RTTE). Potrošnja stanova koji su ugradili RTTE (crveni pravokutnik, zelene točkice) se grananjem grupirala u pravokutnik u kojem su maksimalni iznosi potrošnje zgrade do 130 kWh/m^2, što ukazuje da se ugradnjom RTTE smanjuje potrošnja na razini zgrada (slika 3-53). Na ovaj način odredilo se korijensko grananje te su stanovi grupirani u dvije grane prema potrošnji zgrade i modelu obračuna energije, što je kasnije prikazano u rezultatima (slika 4-1).



Slika 3-53 Prikaz grananja regresijskog stabla prema varijablama i definirana pravokutna područja prema modelu obračuna energije i potrošnji na razini zgrade

Kod stanova koji su ugradili RTTE svi bihevioralni parametri se očituju u broju impulsa zabilježenih u jednom stanu. Prema slici 3-54, ukoliko se u stanu odbroji manje od približno 4 % impulsa može se očekivati potrošnja do 100 kWh/m^2. Upravo u ovu kategoriju spadaju i stanovi u listu s najmanjom potrošnjom prema slici 4-1. Stanovi koji nisu ugradili RTTE (označeni crvenim točkicama) ne mogu ostvarivati niže potrošnje od 70 kWh/m^2, što stanovi s ugrađenim RTTE s malim brojem impulsa mogu (zelene točkice). Raspon specifičnih potrošnji topline za stanove koji jesu i koji nisu ugradili RTTE prikazan je u tablici 3-7.



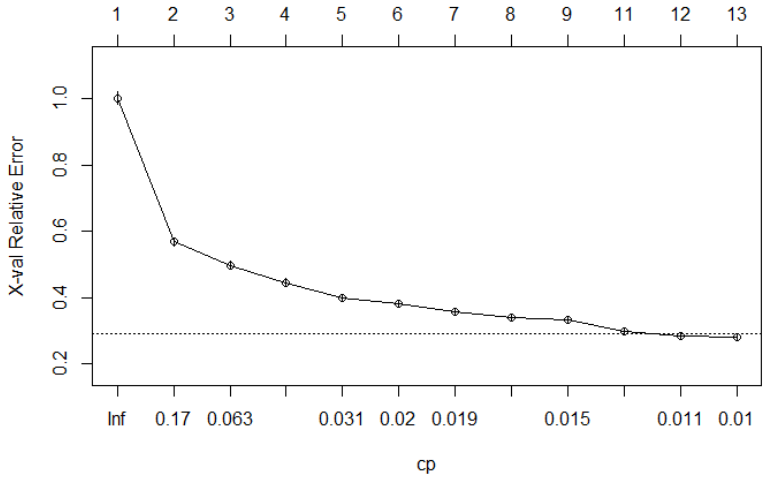
Slika 3-54 Prikaz ovisnosti izbrojanih impulsa u stanu sa specifičnom potrošnjom (y-os)

Točnost predviđanja regresijskih stabla za čitav podatkovni skup za razinu stana za OOB RSME iznosi 11,78 kWh/m^2.

3.4.2.2. Model za podatkovni skup stanova sa ugrađenim individualnim mjerenjem

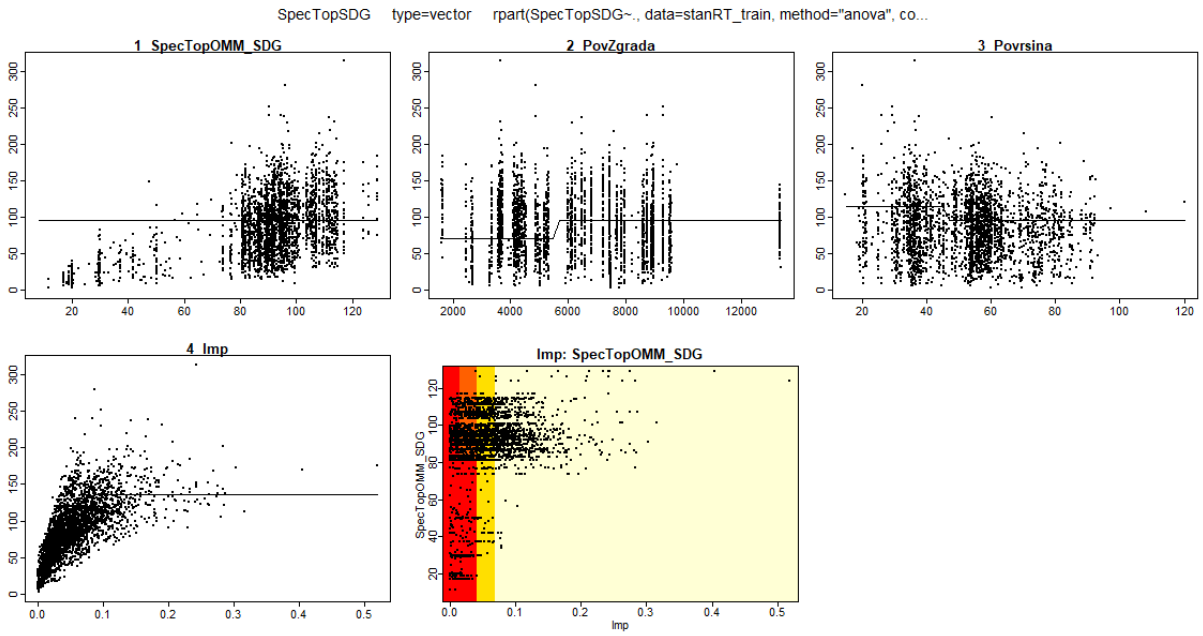
Rezultati unakrsne provjere u 10 koraka za skup stanova s ugrađenim RTTE daju vrijednost $|T|$ = 13, α = 0,1 za grešku unakrsne provjere od 0,3, kako je prikazano na slici 3-55 (y-os predstavlja grešku unakrsne provjere, vrijednosti na gornjoj x-osi označavaju $|T|$, vrijednost označena s cp na donjoj x-osi je α).

Za ovaj podatkovni skup optimalni broj opservacija *minsplit* je 20, a maksimalni broj grananja *maxdepth* je 11. Izgrađeno regresijsko stablo je prikazano na slici 4-2, a na slici 3-56 analizira se stablo prema varijablama dijeljenja i prema pravokutnim područjima R_J.



Slika 3-55 Rezultati unakrsne provjere za učenje regresijskim stablima za stanove s RTTE

Izgrađeno regresijsko stablo za stanove koji su ugradili RTTE provodi grananje prema 4 varijable. Prepoznaje se veza između potrošnje pojedinog stana s ugrađenim RTTE i površine zgrade prema prvom grafu na slici 3-56. Što se tiče odnosa između površine zgrade i potrošnje stanova ne vidi se značajnija ovisnost, a dijeljenje stabla se provodi iznad površine zgrade od približno 5000 m^2, s obzirom da se iznad tih površina pojavljuju stršeće vrijednosti nižih iznosa. Slična je situacija i kod površine stanova u odnosu na specifičnu potrošnju, manje stršećih vrijednosti se pojavljuje iznad površina od približno 50 m^2, te se na toj vrijednosti provodi grananje.

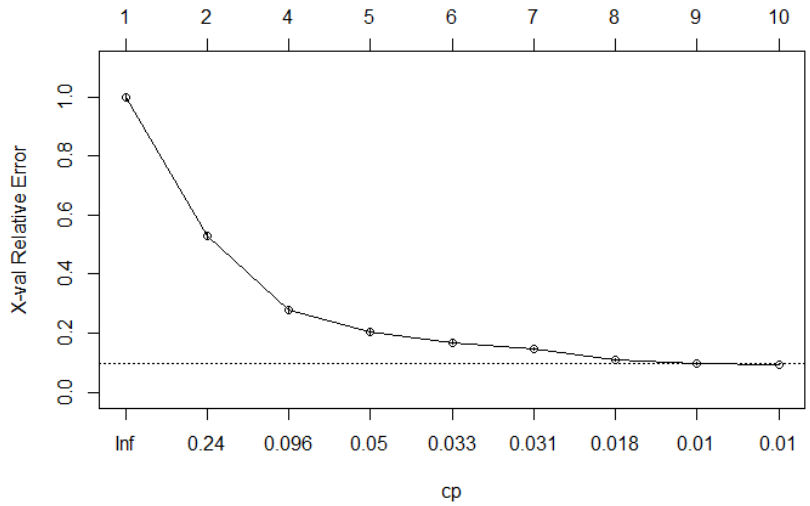


Slika 3-56 Prikaz grananja regresijskog stabla prema varijablama i definirana pravokutna područja prema modelu obračuna energije

Točnost predviđanja regresijskih stabala za podatkovni skup stanova koji imaju ugrađena individualna mjerila je niža nego za čitav podatkovni skup stanova te OOB RSME iznosi 19,31 kWh/m^2.

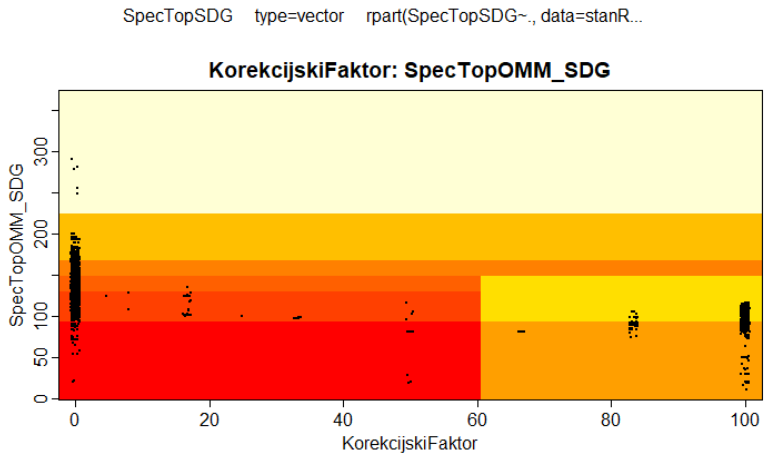
3.4.2.3. Model za podatkovni za skup stanova bez ugrađenog individualnog mjerenja

Rezultati unakrsne provjere u 10 koraka za skup stanova bez ugrađenih RTTE daju vrijednost $|T|$ = 10, α = 0,1 za grešku unakrsne provjere od 0,3, kako je prikazano na slici 3-57 (y-os predstavlja grešku unakrsne provjere, vrijednosti na gornjoj x-osi označavaju $|T|$, vrijednost označena s cp na donjoj x-osi je α). Za ovaj podatkovni skup optimalni broj opservacija *minsplit* je 10, a maksimalni broj grananja *maxdepth* je 12. Izgrađeno regresijsko stablo prikazano je na slici 4-3, a na slici 3-58 analizira se stablo prema varijablama dijeljenja *KorekcijskiFaktor* i *SpecTopOMM_SDG* i prema pravokutnim područjima R_J.



Slika 3-57 Rezultati unakrsne provjere za učenje regresijskim stablima za stanove bez RTTE

Dominantna varijabla koja utječe na potrošnju stanova u ovoj skupini je *KorekcijskiFaktor*, ali na slici 3-58 primjećuje se tendencija smanjenja potrošnje s povećanjem ove varijable. Razlog leži u tome što u slučajevima iznosa varijable *KorekcijskiFaktora* blizu 100% stanovi nalaze u zgradama koje su većinom ugradile RTTE i u svojim obračunima reflektiraju postignute uštede u zgradi.



Slika 3-58 Prikaz grananja regresijskog stabla prema varijablama specifične potrošnje na razini zgrade i korekcijskom faktoru

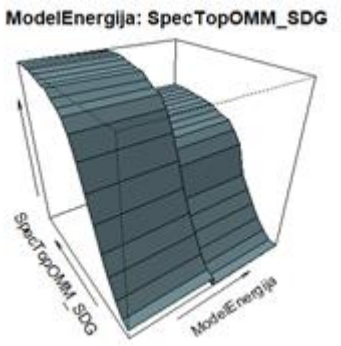
Točnost predviđanja regresijskih stabla za podatkovni skup stanova koji nemaju ugrađena individualna mjerila je viša nego za čitav podatkovni skup stanova te OOB RSME iznosi 6,53 kWh/m^2. Viša točnost predviđanja je djelomično i očekivana jer je način alokacije energije kod ovakvih stanova najjednostavniji i na njega bihevioralni parametri nemaju utjecaja.

3.4.3. Slučajne šume

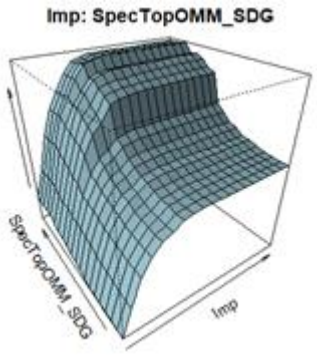
Za analize metodom slučajnih šuma korišten je R paket *ranger* [103], prema opisima analiza u nastavku.

3.4.3.1. Model za čitav podatkovni skup stanova

Vrijednost MSE za OOB predviđanje iznosi 15, tj. RMSE iznosi 3,87 kWh/m^2 za godišnje predviđanje potrošnje stanova ukoliko se predviđanje radi na skupu za učenje. Ovim rezultatima je objašnjeno 98,52 % opservacija. Iako za slučajne šume nije potrebno provoditi učenje i provjeru na posebnim skupovima (vidjeti poglavlje 2.7), što je slučaj pri implementaciji drugih analiziranih algoritama, isto se može učiniti kako bi se dodatno procijenila točnost. Modeliranje slučajnim šumama se radi tako da se traži veći broj manjih stabala razmatrajući sve varijable, nakon čega se izrađena stabla uklapaju u šumu. U načelu se radi višestruka analiza regresijskih stabala ali u višim dimenzijama. Trodimenzijski prikazi modeliranja potrošnje stanova u ovisnosti o parovima varijabli dani su na slikama 3-59 i 3-60.



Slika 3-59 Ovisnost zavisne varijable o modelu za obračun energije i potrošnji zgrade



Slika 3-60 Ovisnost zavisne varijable o broju impulsa i potrošnji zgrade

Ako se točnost predviđanja želi izračunati, a ne temeljiti na OOB točnosti, učenje se provodi na 60 % opservacija, a točnost predviđenih RMSE se evaluira na skupu za provjeru kojeg čine ostalih 40 % opservacija. Ovako izračunate greške predviđanja su manje, te RMSE iznosi 4,27 kWh/m^2, što nije značajna razlika.

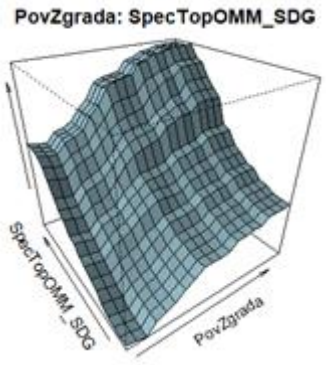
Dobiveni rezultati se mogu interpretirati na način da metoda slučajne šume, na temelju prethodno opisanih ulaznih podataka, točno predviđa potrošnju uz grešku od +/- 4,27 kWh/m^2 na godišnjoj razini.

3.4.3.2. Model za podatkovni skup stanova sa ugrađenim individualnim mjerenjem

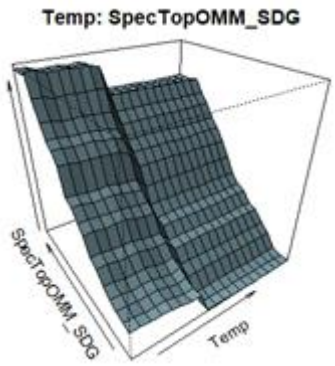
Za podskup stanova koji imaju ugrađene RTTE, očekuju se nešto manje točni rezultati predviđanja u usporedbi s točnošću predviđanja na podatkovnim skupovima za sve stanove ili stanove bez ugrađenih RTTE, kao što je to bilo u slučaju analiza višestrukom linearnom regresijom i regresijskim stablima.

Vrijednost MSE za OOB predviđanje iznosi 54,55, tj. RMSE iznosi 7,39 kWh/m^2 za godišnje predviđanje potrošnje stanova ukoliko se predviđanje radi na skupu za učenje. Ovim rezultatima je objašnjeno 96,65 % opservacija. Dobivene greške ukazuju da je izdvojeni model za predviđanje na stanovima koji su ugradili RTTE manje točan nego što je model za čitav skup stanova.

Trodimenzijski prikazi modeliranja potrošnje stanova koji su ugradili RTTE u ovisnosti o parovima varijabli dani su na slikama 3-61 i 3-62.



Slika 3-61 Ovisnost zavisne varijable površini i potrošnji zgrade



Slika 3-62 Ovisnost zavisne varijable o temperaturi i potrošnji zgrade

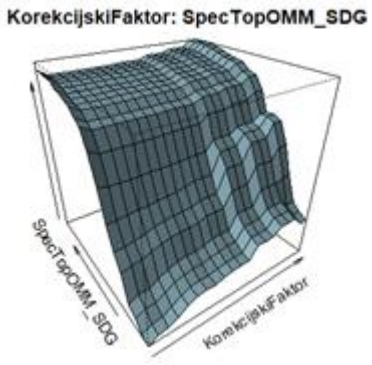
Ponovno se zasebno provodi učenje koji se sastoji od 60 % opservacija, a točnost predviđenih RMSE od 7,48 kWh/m^2 se evaluira na skupu za testiranje kojeg čine ostalih 40 % opservacija. Ovako izračunate greške predviđanja su manje, te RMSE iznosi 7,25 kWh/m^2, što nije značajna razlika.

Dobiveni rezultati se mogu interpretirati na način da metoda slučajne šume, na temelju prethodno opisanih ulaznih podataka, točno predviđa potrošnju uz grešku od +/- 7,25 kWh/m^2 na godišnjoj razini.

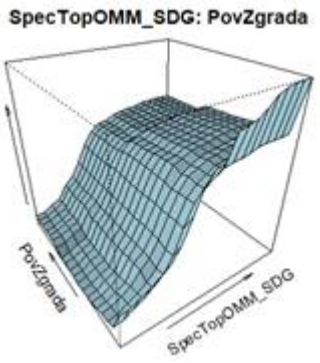
3.4.3.3. Model za podatkovni za skup stanova bez ugrađenog individualnog mjerenja

Dosad provedene analize su pokazale da je model za skupinu stanova koji nisu ugradili RTTE najtočniji u predviđanju, pa se takvi rezultati očekuju i za predviđanje algoritmom slučajnih šuma. Vrijednost MSE za OOB predviđanje iznosi 6,51, RMSE iznosi 2,55 kWh/m^2 za predviđanje potrošnje stanova na godišnjoj razini. Ovim rezultatima je objašnjeno 98,77 % opservacija. Dobivene greške ukazuju da je izdvojeni model za predviđanje na stanovima koji su nisu ugradili RTTE višestruko točniji od ostala dva modela.

Trodimenzijski prikazi modeliranja potrošnje stanova koji nisu ugradili RTTE u ovisnosti o parovima varijabli dani su na slikama 3-63 i 3-64 .



Slika 3-63 Prikaz ovisnosti zavisne varijable o potrošnji zgrade i korekcijskom faktoru



Slika 3-64 Prikaz ovisnosti zavisne varijable o potrošnji i površini zgrade

Ponovno se zasebno provodi učenje koji se sastoji od 60 % opservacija, a točnost predviđenih RMSE od 2,71 kWh/m^2 se evaluira na skupu za testiranje kojeg čine ostalih 40 % opservacija. Ovako izračunate greške predviđanja su manje, te RMSE iznosi 2,73 kWh/m^2, što nije značajna razlika.

Dobiveni rezultati se mogu interpretirati na način da metoda slučajne šume, na temelju prethodno opisanih ulaznih podataka, točno predviđa potrošnju uz grešku od +/- 2,73 kWh/m^2 na godišnjoj razini.

3.4.4. Stroj potpornih vektora

Analize metodom stroja potpornih vektora se razvijaju korištenjem R paketa *e1071* [104] i *kernlab* [105].

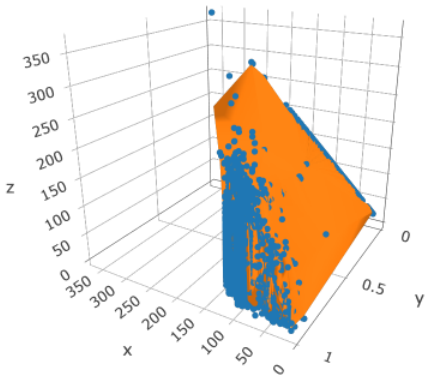
Općenito, stroj potpornih vektora se pokazuje učinkovitim kada imamo neregularnost podataka, kada nam nije poznata distribucija ili bilo koja druga karakteristika podataka. Iz toga razloga je ova metoda dobra za nelinearne probleme i ne zahtjeva prethodno znanje o obliku i vrsti funkcijske ovisnosti. Udaljenosti između podataka se određuju najvećom marginom te je stoga model robustan i ne smetaju mu stršeće vrijednosti ili bilo kakva pristranost u podacima. Najveći nedostatak ove metode je u podešavanju, jer je moguće dobiti veliku razliku u točnosti u zavisnosti o različitim mjerama podešavanja.

Točnost predviđanja za čitav podatkovni skup regresijskim strojem potpornih vektora, korištenjem radijalnog kernela, iznosi RMSE +/- 5,96 kWh/m^2.

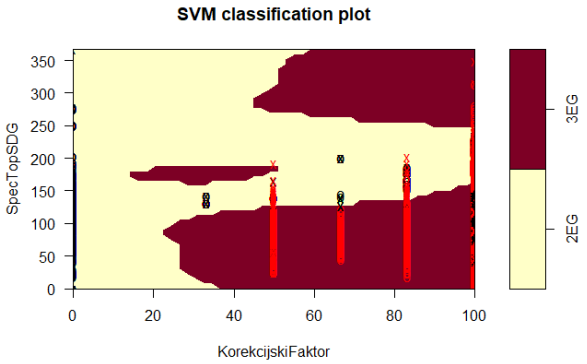
Ovdje je zanimljivo istražiti utjecaj korekcijskog faktora u zgradi na alociranu potrošnju po stanovima. S obzirom da je cilj korekcijskog faktora da penalizira one stanove koji nisu ugradili RTTE, za očekivati bi bilo da u jednoj skupini stanova (npr. stanovi s istom površinom), stanovi koji nisu ugradili RTTE budu penalizirani na način da uvijek imaju alociranu višu energiju od onih stanova koji jesu ugradili RTTE i provode mjere energetske učinkovitosti.

Iz slike 3-66 vidi se da ovaj učinak nije postignut te da se potrošnje topline grupiraju u tri skupine. Iako je točno da je najniža potrošnja za stanove koji su ugradili RTTE kod svih iznosa korekcijskih faktora, iznad nje se zatim prostire potrošnja onih stanova koji nisu ugradili RTTE, ali onda opet postoji skupina stanova koji jesu ugradili RTTE a imaju višu potrošnju od onih koji nisu ugradili RTTE. Ova pojavnost je demotivirajuća za ugradnju RTTE i trebalo bi poduzeti korake u modifikaciji sustava alokacije kako se ne bi pojavljivala.

Razina točnosti predviđanja za podatkovni skup stanova koji su ugradili RTTE i onih koji nisu je na sličnoj razini kao i predviđanje za čitav podatkovni skup stanova.. Tako predviđanje potrošnje za one stanove koji jesu ugradili ima točnost RMSE od +/- 5,45 kWh/m^2, a onih koji nisu + / - 5,38 kWh/m^2.

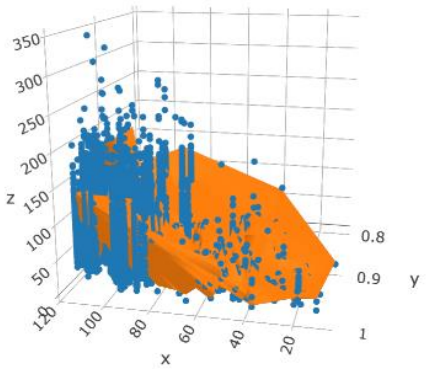


Slika 3-65 Predviđanje na razini svih stanova. X= SpecTopOMM_SDG, Y=StopaUgr, Z=SpecTopSDG

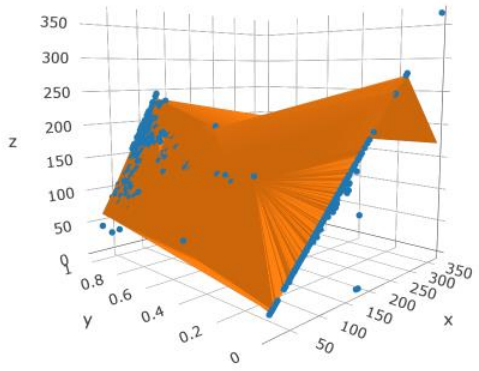


Slika 3-66 Klasifikacija strojem potpornih vektora ovisnosti potrošnje o modelu obračuna energije i korekcijskom faktoru.

Narančasta površina na slikama 3-65, 3-67 i 3-68 pokazuje ravninu predviđanja strojem potpornih vektora za svaki podatkovni skup, a dobro predviđanje je kad su predviđene vrijednosti, označene plavim točkicama, blizu ravnine. Vidi se da metoda stroja potpornih vektora dobro provodi predviđanja za svaki podatkovni skup. Sva tri prikaza daju ovisnosti varijabli *SpecTopOMM, StopaUgr* i *SpecTopSDG*. Može se primijetiti kako se narančasta površina predviđanja mijenja u obliku zavisno o skupu podataka te najviše linearnih karakteristika pokazuje za skup stanova koji nisu ugradili RTTE (slika 3-68), najmanje za skup onih koji jesu (slika 3-67), dok u zajedničkom skupu površina predviđanja predstavlja kombinaciju (slika 3-65).



Slika 3-67 Predviđanje na razini stanova s ugrađenim RTTE. X= SpecTopOMM_SDG, Y=StopaUgr, Z=SpecTopSDG



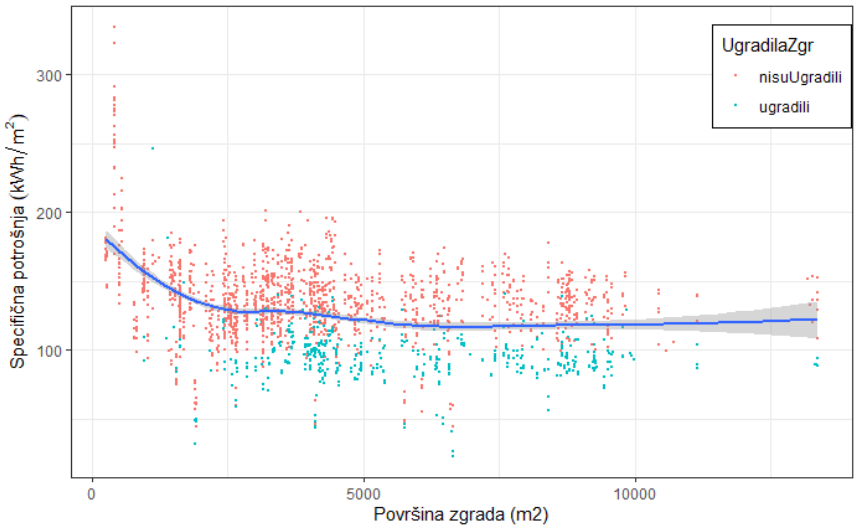
Slika 3-68 Predviđanje na razini stanova bez ugrađenih RTTE. X= SpecTopOMM_SDG, Y=StopaUgr, Z=SpecTopSDG

3.5. Strojno učenje na razini zgrade

Predviđajući model na razini zgrade ima za cilj odgovoriti na pitanje da li se metodama strojnog učenja može, i na kojoj razini točnosti, predvidjeti potrošnja zgrade spojene na sustav daljinskog grijanja, a s obzirom na postojanje ili ne postojanje individualnog mjerenja. Ukoliko bi razvijeni model rezultirao zadovoljavajućom točnošću, bio bi pri primjenjiv pri ocjeni učinka ugradnje individualnog mjerenja na razini zgrade.

U modelu su razmatrane samo zgrade koje su ugradile RTTE u određenoj stopi ugradnje. Analize su provedene na podatkovnom skupu opisanom u poglavlju 3.1.3.

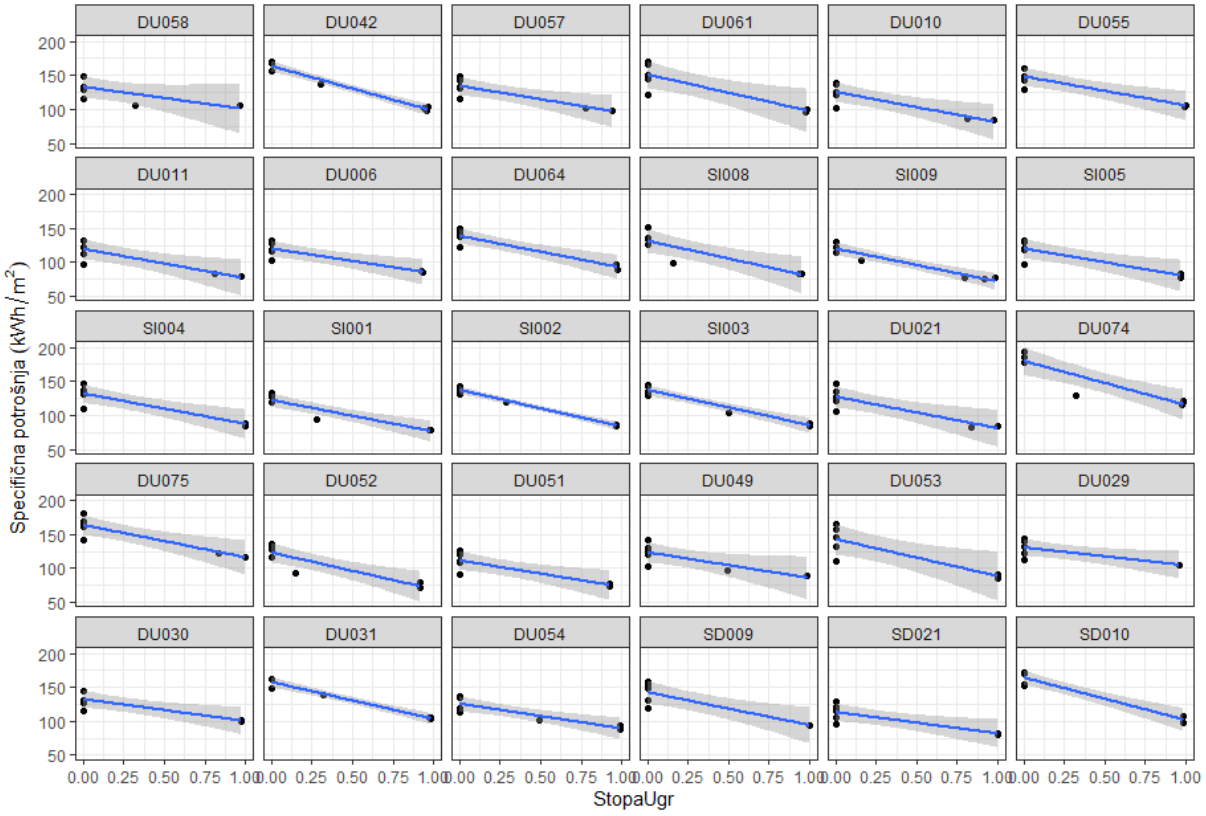
Iz dijagrama rasipanja, slika 3-69, vidi se da nema jasnog grupiranja između onih zgrada koje jesu i onih koje nisu ugradile RTTE. Ono što se može vidjeti da se one zgrade koje jesu ugradile nalaze oko regresijskog pravca, te da maksimalne, ali i neke minimalne vrijednosti, bilježimo kod zgrada koje nisu ugradile individualno mjerenje. Iz tog razloga se analiza neće provoditi na čitavom skupu zgrada, nego samo na podskupu zgrada koje jesu ugradile RTTE – kako bi se dobila ocjena učinka te mjere na potrošnju topline na razini zgrade.



Slika 3-69 Dijagram rasipanja specifične godišnje potrošnje topline u zgradama prema ugrađenosti RTTE

Analizom odabranog skupa zgrada (ukupno 161 zgrada) je zaključeno da su sve zgrade koje su ugradile individualna mjerila bez iznimke postigle uštede na razini zgrade, kako se vidi na slici 3-70 na kojoj je prikazan dio analiziranih zgrada.

Usporedbom s rezultatima dobivenim na razini stana, može se vidjeti da učinci na zgradi nisu prenosivi na učinke na pojedine stanove. Pod pretpostavkom da su u zgradi ispunjeni svi tehnički uvjeti za pravilno funkcioniranje sustava, činjenica da zgrada postiže uštede, a krajnji korisnici koji su uveli individualno mjerenje ne postižu uštede, se može pripisati bihevioralnim karakteristikama svakog krajnjeg korisnika.



Slika 3-70 Utjecaj ugradnje RTTE na potrošnju na razini zgrade za odabrani skup zgrada

3.5.1. Višestruka linearna regresija

Regresijskom analizom želi se naći funkcijska ovisnost između potrošnje topline na razini zgrade u ovisnosti o nezavisnim varijablama od koje je jedina *StopaUgr* koja ilustrira da li je u zgradi uvedeno individualno mjerenje i u kojem udjelu. Regresijski model sadrži značajno manje relevantnih varijabli, tj. sadrži samo one varijable koje su dostupne iz skupa za naplatu, a smatra ih se relevantnim za potrošnju zgrade (tablica 3-27).

Rezultati dobiveni korištenjem R paketa *stats* [66] daju sljedeću regresijsku funkciju:

Tablica 3-27 Regresijski parametri i statistički pokazatelji regresijskog modela za zgrade

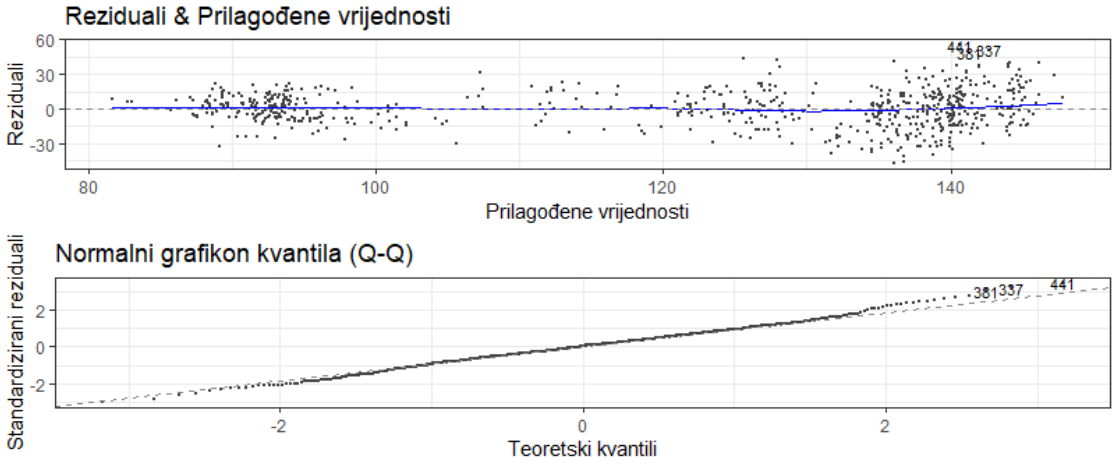
	β	St. pogreška	t- vrijednost	p - vrijednost
(odsječak)	172,887	4,068	42,498	0
Temp	-5,210	0,648	-8,040	0
StopaUgr	-46,160	1,458	-31,654	0
PovZgrada	-0,001	0,000	-4,269	0

Analiza korelacija pokazuje da među ovim varijablama nema korelacija te da je ista jednaka 0. Pokazatelji točnosti predviđanja dani su u tablici 3-28.

Tablica 3-28 Pokazatelji točnosti regresijskog modela z arazinu zgrade

	Podskup za učenje	Podskup za provjeru
R^2	0,627	0,633
RMSE	15,97	16,559

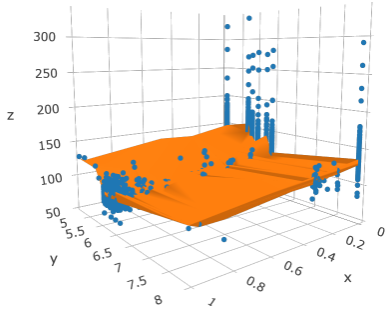
Regresijski model daje vrijednosti RMSE na razini točnosti iste metode na razini stana (tablica 3-21). Linearnost modela je veća, što se može vidjeti iz 3-71.



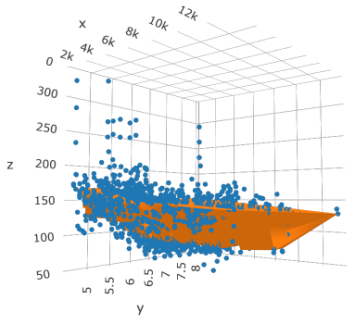
Slika 3-71 Dijagram reziduala i normalni Q-Q dijagram za razinu zgrade

3.5.2. Nelinearni modeli strojnog učenja

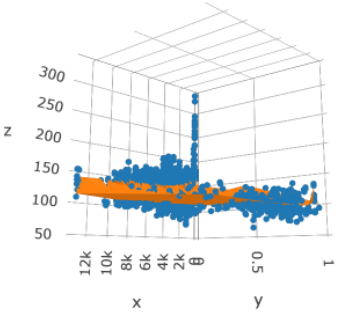
Model slučajnih šuma je oblikovan R paketom *ranger* [103]. Predviđanje je na razini točnosti metode regresijskih stabala, u modelu slučajnih šuma s 500 definiranih stabala. OOB greška predviđanja varijable RMSE je 13,24 kWh/m^2. Grafički prikazi točnosti predviđanja dani su na slikama 3-72, 3-73 i 3-74. Može se vidjeti da model dobro predviđa osim u slučajevima zgrada koje nisu ugradile RTTE. U tim slučajevima model ne može predvidjeti potrošnju (jako udaljene plave točke od narančaste ravnine predviđanja).



Slika 3-72 Predviđanje na razini zgrada za X=StopaUgr, Y=Temp, Z=SpecTopOMM_SDG

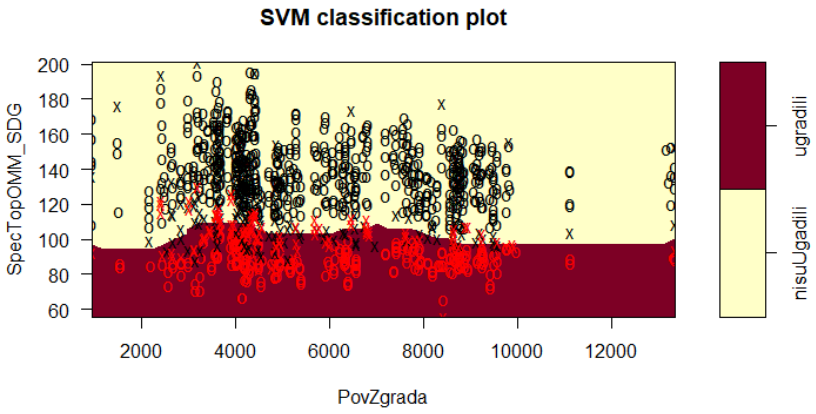


Slika 3-73 Predviđanje na razini zgrada za X=PovZgrada, Y=Temp, Z=SpecTopOMM_SDG



Slika 3-74 Predviđanje na razini zgrada za X=PovZgrada, Y=StopaUgr, Z=SpecTopOMM_SDG

Stroj potpornih vektora ima nešto manju točnost predviđanja u usporedbi sa slučajnim šumama i regresijskim stablima, ali nam daje mogućnost da prema slici 3-75 evaluiramo koliko stopa ugradnje utječe na specifičnu potrošnju topline u pojedinim zgradama. S obzirom da su sve zgrade u ovom podatkovnom skupu ugradile RTTE u nekoj stopi ugradnje, na slici su prikazane iste zgrade (površina zgrade nije promjenjiva, prije i poslije ugradnje). Vidi se da sve analizirane zgrade nakon ugradnje imaju potrošnju do približno 100 kWh/m^2 na godišnjoj razini, dok su prije ugradnje imale znatno više (u žutoj površini na slici).



Slika 3-75 Klasifikacijski prikaz stroja potpornih vektora koji pokazuje vezu između StopeUgr i SpecTopOMM_SDG

OOB greška predviđanja prema vrijednosti RMSE ovom metodom iznosi +/- 13,38 kWh/m^2 na godišnjoj razini.

3.6. Strojno učenje na višim razinama

Na višim razinama je od interesa usporediti potrošnju na nekom distribucijskom području ili gradu, tj. na nacionalnoj razini. S obzirom na karakteristike sektora daljinskog grijanja u Hrvatskoj, opisane u poglavlju 1.1.1.2, broj distribucijskih područja je relativno mali s obzirom na broj gradova i općina i iznosi ukupno 11 područja. Dodatno, obaveza ugradnje RTTE je nastala od 31.12.2016. [2], tako da je i vremenski niz u kojem postoje podaci za ova područja relativno malen i čini ukupno 7 godina, od koje su dvije pune godine nakon ugradnje RTTE. U ovom radu su analize napravljene na dva najveća distribucijska područja po broju zgrada i stanova, ali za analize na višoj razini podatkovni skup je relativno malen. Stoga je statističko učenje na višim razinama moguće provesti jedino metodom višestruke regresije, što je i učinjeno.

Regresijski modeli se određuju bez regularizacije i predstavljaju regresijski pravac u sedmogodišnjem nizu. Od dominantnih varijabli se ponovno prepoznaje stopa ugradnje. Drugim riječima, i na višim razinama predviđanja je ova varijabla od značaja u smislu predviđanja potrošnje, što se vidi u tablici 3-29 koja daje regresijske parametre za razinu distribucijskog sustava. Regresijski model za nacionalnu razinu je u sadržaju varijabli jednak, uz dodatnu kategorijsku varijablu *DistPodrucje* (tablica 3-30).

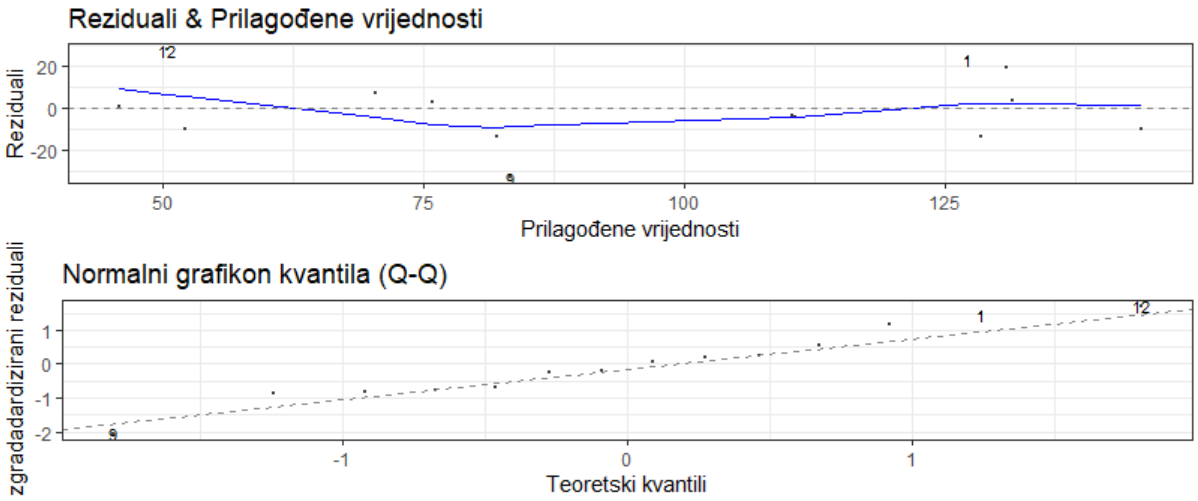
Tablica 3-29 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za razinu distribucijskog područja

	β	St. pogreška	t- vrijednost	p - vrijednost
(odsječak)	-333,343	83,226	-4,005	0,002
StopaUgr	-43,997	14,511	-3,032	0,013
PovZgrada	0,095	0,019	5,103	0,000
Temp	5,297	5,328	0,994	0,344

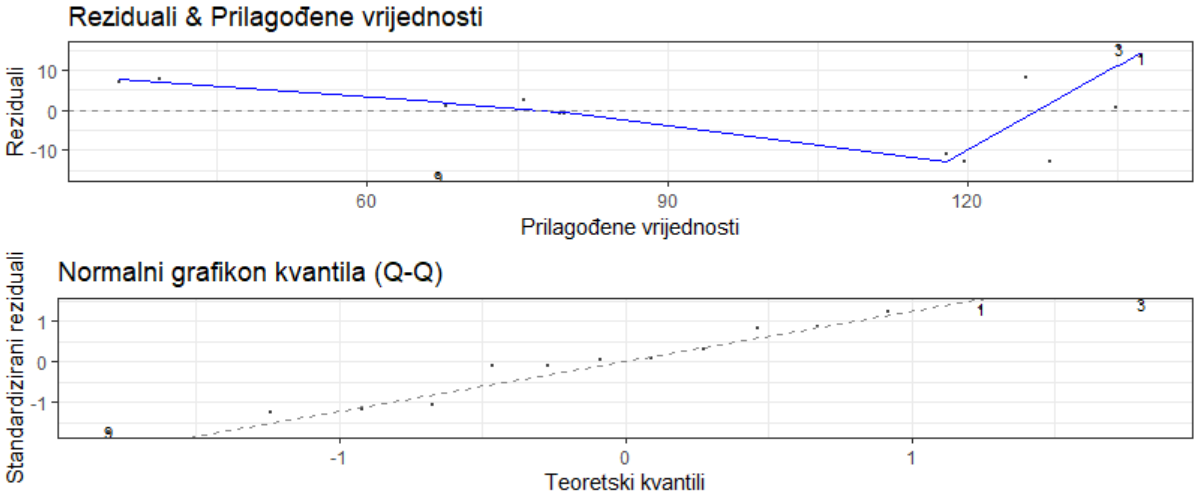
Tablica 3-30 Regresijski parametri i statistički pokazatelji finalnog regresijskog modela za nacionalnu

	β	St. pogreška	t- vrijednost	p - vrijednost
(odsječak)	731,074	286,122	2,555	0,031
DistPodrucjeDP2	-128,831	33,998	-3,789	0,004
StopaUgr	-21,597	11,184	-1,931	0,086
PovZgrada	-0,126	0,059	-2,116	0,063
Temp	-3,038	4,122	-0,737	0,480

Dijagrami kvantil-kvantil na slikama 3-76 i 3-77 pokazuje normalnost distribucije reziduala predviđanja, a dijagram rasporeda reziduala i prilagođenih vrijednosti daje nesavršene krivulje što je razlog manjeg skupa podataka. Ipak, predviđanje modela je zadovoljavajuće na skupu s obzirom da su iznosi standardnih pogreški u tablicama 3-29 i 3-30 u granicama koje ne upućuju na veliku stopu rasipanja.



Slika 3-76 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za model na razini distribucijskog područja



Slika 3-77 Dijagram reziduala i prilagođenih vrijednosti i normalni dijagram Q-Q za model na nacionalnoj razini

4. PRIKAZ REZULTATA

Sustavi daljinskog grijanja imaju određenih posebnosti u odnosu na druge umrežene energetske sustave (električna energija, prirodni plin), što se posebno očituje u statusu krajnjeg kupca, načinu mjerenja ili alokacije potrošene energije i međusobnim odnosima krajnjih kupaca u višestambenim zgradama. Prava potrošača u sustavima daljinskog grijanja dolaze u fokus tek nedavno, u širem kontekstu politika povećanja energetske učinkovitosti. Do stupanja na snagu Direktive o energetskoj učinkovitosti [1] individualno mjerenje u sustavima daljinskog grijanja bilo je rijetko, pogotovo na području južne i istočne Europe [55]. Alokacija potrošene topline u Republici Hrvatskoj do 2016. velikim se dijelom provodila paušalno, prema udjelima u površini višestambene zgrade, te se ni na koji način nije temeljila na karakteristikama potrošnje u pojedinom stanu. Takav model nije poticao krajnje korisnike na upravljanje svojom potrošnjom, niti su se postigle energetske uštede i konzekventno smanjenje računa za toplinu.

Transpozicijom odredbi Direktive o energetskoj učinkovitosti u hrvatsko zakonodavstvo započinje se s ugradnjom RTTE u stanove spojene za daljinsko grijanje, ali kod određenog broja krajnjih korisnika dolazi do nezadovoljstva. Manji dio korisnika nakon ugradnje RTTE, a time i provođenja sekundarnih mjera energetske učinkovitosti, bilježi negativan učinak i ne ostvaruje uštede. Kod jednog dijela stanova to se može pripisati činjenici da sustavi grijanja unutar zgrada nisu bili pripremljeni za rad u režimu promjenjivog toplinskog opterećenja, a u stanovima u zgradama koje su poduzele sve potrebne tehničke prilagodbe u sustavu takva se pojavnost može pripisati bihevioralnim čimbenicima krajnjih potrošača. Dodatnu prepreku korektnoj alokaciji po pojedinim stanovima koji su ugradili RTTE predstavlja i činjenica da u većini zgrada nisu ugrađeni u 100 % grijanih prostora te se zbog toga alokacija djelomično provodila paušalno, a djelomično prema propisanoj formuli za alokaciju [56], što nije obuhvatilo stvarnu potrošnju u svim stanovima. Model razvijen u ovom radu objašnjava koje su utjecajne varijable i određuje stupanj utjecaja svake utjecajne varijable na potrošnju topline u stanovima.

S druge strane, analiza je pokazala da se ugradnjom RTTE u zgradi, bez obzira na to što možda nije riječ o ugradnji u 100 % stanova, potrošnja na razini zgrade smanjila u svih 350 analiziranih zgrada. Jednako tako, potrošnja topline pala je i na razini dvaju analiziranih distribucijskih sustava te je pad potrošnje prepoznat i na nacionalnoj razini, što se može vidjeti i u godišnjim statističkim izvještajima [37].

Učinke ugradnje individualnog mjerenja tako treba gledati na svim razinama. Na najvišim razinama, razini distribucijskog područja (grada) i nacionalnoj razini – oni su najočitiji.

Na nižim razinama – razini zgrade i stana – oni se prepoznaju, ali nisu jasno predvidivi te su odnosi između utjecajnih parametara na uštede u zgradi i uštede u stanovima u toj zgradi, kao i njihove interakcije, kompleksni i teško ih je objasniti klasičnim simulacijskim metodama.

Stoga su predloženi modeli za predviđanje na svakoj razini, koji se temelje na metodama strojnog učenja.

Predviđajući modeli razvijeni u ovom radu daju indikaciju kako pristupiti modeliranju potrošnje u sustavima daljinskog grijanja s ciljevima predviđanja potrošnje topline i ocjene učinka primjene mjera energetske učinkovitosti. Modeli su razvijeni na svim razinama potrošnje – razini stana, razini zgrade, razini distribucijskog sustava i nacionalnoj razini.

4.1. Predviđajući model za razinu stana

Razvoj modela na razini stanova proveden je za tri podatkovna skupa, kako slijedi:

1. skup svih stanova,
2. skup stanova koji su ugradili RTTE i
3. skup stanova koji nisu ugradili RTTE, ali se nalaze u zgradama u kojima su djelomično ugrađeni RTTE.

Prvi model daje predviđanja u slučajevima kada je nakana predvidjeti potrošnju stana bez prethodnog grupiranja stanova prema ugradnji RTTE.

Finalni regularizirani regresijski model za ovu skupinu stanova prikazan je jednadžbom:

$$\begin{aligned} \boldsymbol{SpecTopSDG} = \mathbf{29,03} - \mathbf{123,13} \cdot \boldsymbol{ModelEnergija3EG} + \mathbf{639,67} \cdot \boldsymbol{Imp} + \\ + \mathbf{0,96} \cdot \boldsymbol{SpecTopOMM_{SDG}} + \mathbf{80,14} \cdot \boldsymbol{StopaUgr} - \mathbf{3,42} \cdot \boldsymbol{Temp} - \\ - \mathbf{0,258} \cdot \boldsymbol{Povrsina} + \mathbf{0,002} \cdot \boldsymbol{PovZgrada} \end{aligned} \quad (62)$$

Prema ovom modelu, dominantna je varijabla *Imp*, koja predstavlja omjer odbrojenih impulsa u pojedinom stanu i sume svih odbrojenih impulsa u zgradi. Iznos regresijskog faktora β indicira značaj svake varijable u regresijskom modelu. Idući parametar je *ModelEnergija3EG*. Ovdje je riječ o kategorijskoj varijabli koja je procesom kodiranja dobila kontinuirane karakteristike i uključena je u regresijski model. Iznos ove varijable može biti 0 ili 1, s tim da vrijednost 0 ima kada stan nije ugradio RTTE, a vrijednost 1 kada stan jest ugradio RTTE.

Vrijednost pokazatelja točnosti predviđanja R^2 za sve stanove iznosi 0,791, tj. 80 % opservacija opisano je ovim modelom.

Osim predviđanja za čitav skup stanova, od interesa je odrediti i odvojene modele za stanove koji imaju ugrađen RTTE i one koji nemaju. Finalni regularizirani regresijski model za stanove koji imaju ugrađen RTTE prikazan je jednadžbom:

$$\begin{aligned} \boldsymbol{SpecTopSDG} = -\mathbf{158,53} + \mathbf{794,03} \cdot \boldsymbol{Imp} + \mathbf{0,65} \cdot \boldsymbol{SpecTopOMM_SDG} + \\ + \mathbf{155,75} \cdot \boldsymbol{StopaUgr} - \mathbf{0,863} \cdot \boldsymbol{Povrsina} + \mathbf{0,006} \cdot \boldsymbol{PovZgrada} \end{aligned} \quad (63)$$

Varijable su slične kao u modelu za sve stanove, ali bez varijable *ModelEnergija*, jer je ona u ovom slučaju singularitet. Vrijednost pokazatelja točnosti predviđanja R^2 za sve stanove s ugrađenim RTTE očekivano je malo niža, što je posljedica veće varijabilnosti u podacima zbog bihevioralnih utjecaja te iznosi 0,759, tj. 76 % opservacija opisano je ovim modelom.

Model predviđanja za stanove bez ugrađenih RTTE, a koji se nalaze u zgradama koje su započele ugradnju te postoji određena stopa ugradnje, prikazan je jednadžbom *(64)*. Očekivano je ovaj model u strukturi najjednostavniji te ima najvišu točnost predviđanja R^2 od 0,955, tj. čak 95 % opservacija stanova koji nisu ugradili RTTE može se opisati ovim modelom.

$$\boldsymbol{SpecTopSDG = 1,987 + 0,978 \cdot SpecTopOMM_{SDG} + 7,508 \cdot StopaUgr + \\ +0,732 \cdot KorekcijskiFaktor} \qquad (64)$$

Iz jednadžbe *(64)* vidljiv je utjecaj varijable *KorekcijskiFaktor*. Osnovni cilj uvođenja ove varijable u zakonodavstvo bila je svojevrsna penalizacija onih stanova koji nisu ugradili RTTE te je iznos korekcijskog faktora kroz godine mijenjan od 50 % do 100 %. Ukratko, ako bi korekcijski faktor bio 100 %, alokacija se radi tako da se površina stana koji nema ugrađen RTTE poveća za 100 % te se zatim alocira ona količina ukupno izmjerene toplinske energije u zgradi (varijabla *ToplinaOMM*, poglavlje 3.1.1) pomnožena s omjerom uvećane površine stana (tj., $2 \cdot Povrsina$) i ukupne površine zgrade *PovZgrada*.

Takva bi penalizacija prema finalnom regresijskom modelu najviše iznosila kada je *KorekcijskiFaktor* jednak 100 % i uvećala godišnju potrošnju za cca 7,2 kWh/m^2. Stopa ugradnje za stanove koji nisu gradili RTTE može utjecati na alociranu potrošnju najviše do iznosa od približno 7,5 kWh/m^2. Ako zbrojimo utjecaje varijabli *StopaUgr* i *KorekcijskiFaktor*, maksimalni mogući apsolutni iznos za koji ove dvije varijable povećavaju potrošnju stana bez ugrađenih RTTE prema prosječnoj specifičnoj potrošnji zgrade iznosi 14,7 kWh/m^2. Ako se apsolutni iznos ove penalizacije usporedi s prosječnom specifičnom potrošnjom topline svih analiziranih stanova u tablici 3-3, ona iznosi 11,3 %.

Iz predviđajućih modela u jednadžbama (*62*), (*63*) i (*64*) možemo zaključiti sljedeće:

1. U modelu za čitav skup zgrada, jednadžba (*63*), regresijski parametar β ispred varijable *ModelEnergija* umanjuje potrošnju za -123,13 za stanove koji su ugradili RTTE. S druge strane, na prvi pogled neočekivano, regresijski parametar ispred varijable

StopaUgr ima pozitivan predznak i iznosi +80,14. Rezultirajuća razlika između ove dvije varijable za stanove koji su ugradili RTTE iznosi -53, tj. ovaj iznos je minimalna apsolutna razlika specifične potrošnje stanova koji jesu i koji nisu ugradili RTTE, sve dok je broj impulsa u stanu koji je ugradio RTTE manji od 20 % svih u zgradi. Ako je nekom stanu alocirano više od 20 % impulsa u zgradi, taj stan imat će alociranu potrošnju veću nego da nije ugradio RTTE.

2. Usporedi li se alocirana potrošnja za dva stana jednake površine u istoj zgradi, od kojih je jedan ugradio RTTE, a drugi nije, možemo odrediti kolika je ušteda u stanovima koji su ugradili RTTE. Pretpostavke su kako slijedi:
    - stan koji je ugradio RTTE izbrojio je 4 % impulsa u cijeloj zgradi,
    - specifična potrošnja topline u zgradi iznosi 100 kWh/m^2,
    - stopa ugradnje je 95 %,
    - srednja vanjska temperatura u sezoni grijanja je 5,8 °C,
    - površina stanova koje uspoređujemo je 50 m^2,
    - površina zgrade u kojoj se nalaze stanovi je 4000 m^2.

    Specifična godišnja potrošnja, prema modelu za sve stanove kojem je vrijednost točnosti predviđanja prema RMSE-u 16,48 kWh/m^2, iznosi:

    - za stan koji je ugradio RTTE 83,68 kWh/m^2,
    - za stan koji nije ugradio RTTE 181,22 kWh/m^2.

    Ako pretpostavimo najvišu pozitivnu grešku za stan koji je ugradio RMSE +16,48 kWh/m^2 i najvišu negativnu grešku za stan koji nije ugradio RTTE od -16,48 kWh/m^2, dobivamo predviđanje godišnje specifične potrošnje za ova dva stana:

    - za stan koji je ugradio RTTE od 100,16 kWh/m^2,
    - za stan koji nije ugradio RTTE od 164,74 kWh/m^2.

    Uz navedene pretpostavke, ušteda u usporedbi ovih dvaju stanova je minimalno 40 % u korist stana koji je ugradio RTTE.

3. Ako se predviđanje provodi za stanove prema pretpostavkama u točki 2. i po modelima samo za taj tip stanova po jednadžbama (*64*) i (*65*), rezultati predviđanja specifične godišnje potrošnje toplinske energije su kako slijedi:

- za stan koji je ugradio RTTE predviđena specifična potrošnja je 67,04 kWh/m^2, a vrijednost točnosti predviđanja prema RMSE-u 16,481 kWh/m^2 i
- za stan koji nije ugradio RTTE i uz *KorekcijskiFaktor* jednak 100 % predviđa potrošnju od 180,13 kWh/m^2.

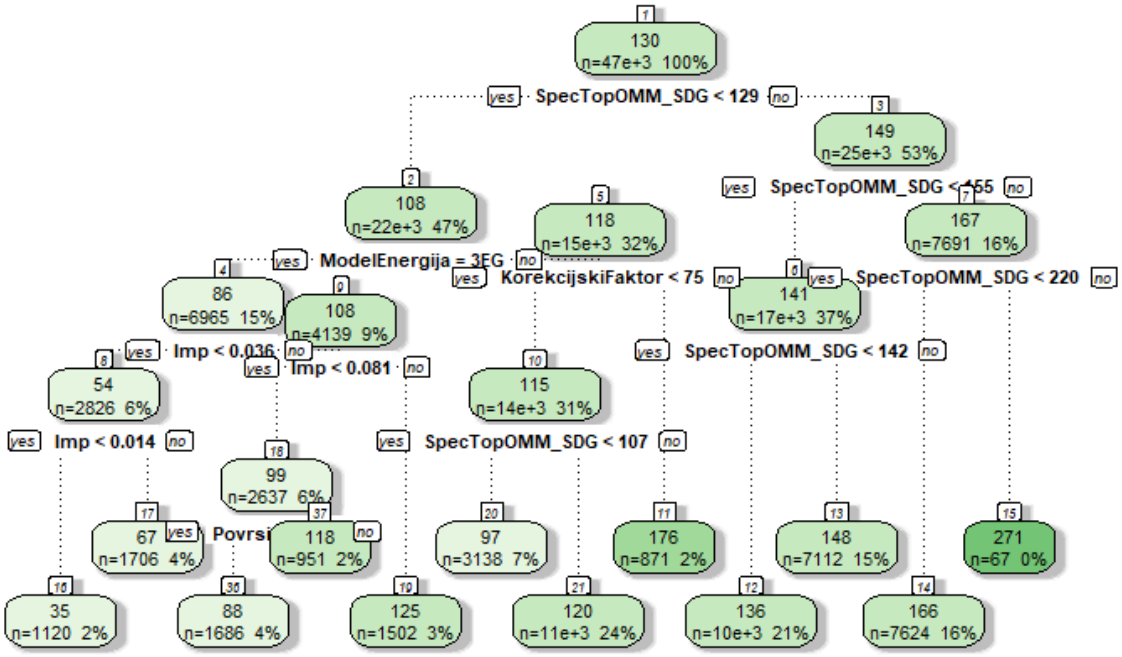
Vidi se da sva tri modela imaju visoku točnost i daju rezultate u skladu s regresijskom analizom. Regresijski modeli daju točnije rezultate kada se primjenjuju specifični modeli za stanove koji su ugradili i stanove koji nisu ugradili RTTE nego kada se primjenjuje model opisan jednadžbom (*63*) koji obuhvaća i jedan i drugi tip stanova. Komparativno, manju točnost ima model za predviđanje za stanove koji su ugradili RTTE, što se objašnjava većom nesigurnosti modeliranja zbog utjecaja bihevioralnih čimbenika.

Za dodatnu interpretaciju koristimo modele regresijskih stabala kao metodu strojnog učenja koju, kao i metodu linearne regresije, karakterizira visoka razina intepretativnosti.

Iz analize regresijskih stabala za čitav podatkovni skup za razinu stana na slici 4-1 vidi se da je varijabla dijeljenja u korijenu *SpecTopOMM_SDG,* a vrijednost dijeljenja je 129 kWh/m^2, što se interpretira tako da je najutjecajnija varijabla na potrošnju pojedinog stana specifična potrošnja zgrade u kojoj se stan nalazi.

Ako je potrošnja manja od tog iznosa, drugi unutarnji čvor radi grananja prema činjenici ima li stan ili nema ugrađene RTTE, što varijablu *ModelEnergija* čini drugom utjecajnom varijablom kod stanova koji se nalaze u zgradi čija je specifična potrošnja manja od 129 kWh/m^2. Ako stan ima ugrađene RTTE, daljnja grananja provode se ovisno o vrijednosti varijable *Imp*, što je i očekivano, jer će se za one stanove koji imaju manji relativni broj impulsa u obračunskom razdoblju prema ukupnom broju impulsa alocirati i manja količina toplinske energije. Godišnje specifične potrošnje za stanove koji imaju ugrađena individualna mjerila i nalaze se u zgradama s godišnjom specifičnom potrošnjom manjom od 129 kWh/m^2 u rasponu su od 35 do 125 kWh/m^2, s time da najveći broj stanova spada u razred srednje godišnje potrošnje od 99 kWh/m^2. Ako se nakon drugog unutarnjeg čvora u zgradama koje imaju specifičnu potrošnju manju od 129 kWh/m^2 radi o stanovima koji nisu ugradili RTTE, potrošnja tih stanova ovisit će o varijabli *KorekcijskiFaktor* i raspon potrošnje tih stanova bit će od 97do 176 kWh/m^2. Možemo pretpostaviti da se u korijenskom dijeljenju koje je manje od 129 kWh/m^2 radi o zgradama koje su započele ugradnju RTTE.

Ako pogledamo drugu granu nakon dijeljenja u korijenu, s vrijednošću specifične potrošnje zgrade višom od 129 kWh/m^2 očekivano imamo stanove koji imaju višu specifičnu godišnju potrošnju, od 136 do 271 kWh/m^2 te se daljnja interna grananja do listova vrše samo prema varijabli *SpecTopOMM_SDG*, što upućuje na to da je ovdje najvjerojatnije riječ o stanovima koji se nalaze u zgradama čija je stopa ugradnje RTTE mala ili je uopće nema. Vjerojatno se radi o stanovima koji energiju obračunavaju prema modelu 2EG, ali ne može se isključiti da se na ovoj grani nalaze i neki stanovi s modelom obračuna 3EG.

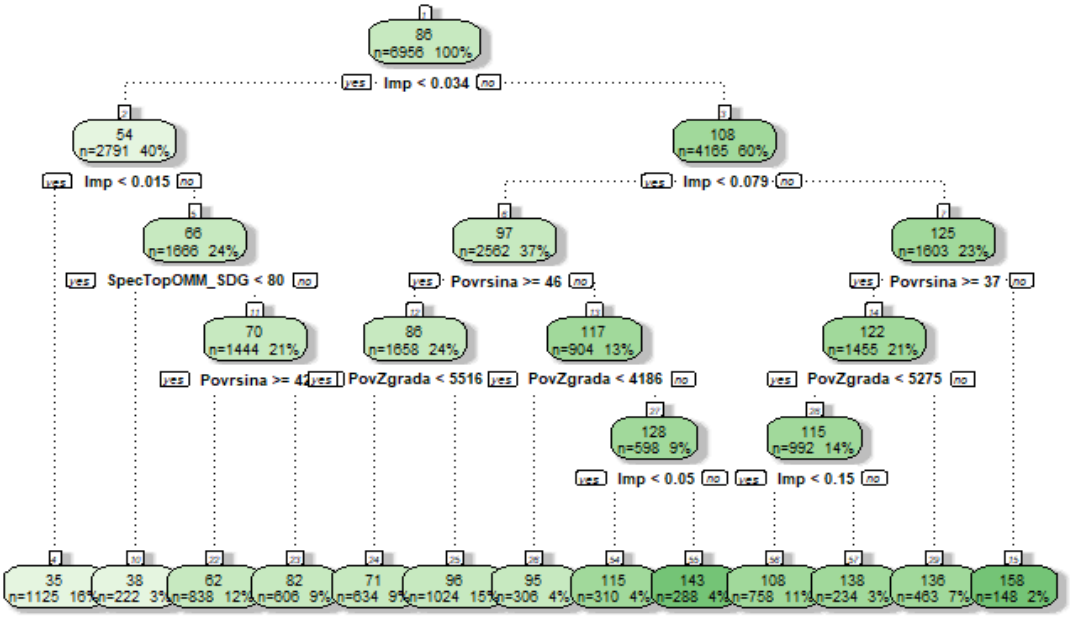


Slika 4-1 Regresijsko stablo za razinu stanova za čitav podatkovni skup

Analiza metodom regresijskih stabala na podatkovnom skupu koji se sastoji samo od onih stanova koji su ugradili individualna mjerila pokazuje da su dominantni parametri po kojima se radi grananje varijable *Imp, Povrsina, PovZgrada* i *SpecTopOMM_SDG*. Korijensko dijeljenje provodi se tako da se u nižu granu potrošnje kategoriziraju stanovi koji u obračunskom razdoblju izmjere manje od 3,4 % impulsa u cijeloj zgradi. Raspon potrošnje za takve stanove kreće se od 35 do 82 kWh/m^2, a najvišu potrošnju od 82 kWh/m^2 imaju stanovi površine manje od 42 m^2.

Za stanove koji u obračunskom razdoblju izmjere više od 3,4 % impulsa u cijeloj zgradi raspon potrošnje kreće se od 71 do 158 kWh/m^2, s time da najvišu potrošnju od 158 kWh/m^2 imaju stanovi manje površine koji mjere više od 8 % svih impulsa u zgradi u obračunskom razdoblju. Za stanove koji na godišnjoj razini izmjere više od 3,4 % svih impulsa u gradi, vrše se daljnja

grananja prema broju impulsa i površinama. Idući interni čvor dijeli stanove na one koji na godišnjoj razini imaju više ili manje od 7,9 % impulsa. Stanovi koji imaju manje od ovog postotka impulsa dalje se dijele prema površini stana, površini zgrade u kojoj se nalaze i mogu očekivati godišnju specifičnu potrošnju od 71 do 143 kWh/m^2. Najvišu potrošnju u ovoj grani imaju stanovi koji imaju malu površinu, nalaze se u većoj zgradi i imaju izmjerenih više od 5 % godišnjih impulsa zgrade. Stanovi koji godišnje izmjere više od 7,9 % impulsa, očekivano, imaju i najviši raspon potrošnje, od 108 do 158 kWh/m^2. Niže razine potrošnje imaju stanovi koji se nalaze u manjim zgradama i imaju izmjereno manje od 15 % izmjerenih impulsa, dok najviše potrošnje imaju mali stanovi, s manje od 37 m^2, koji mjere više od 7,9 % impulsa u cijeloj zgradi.

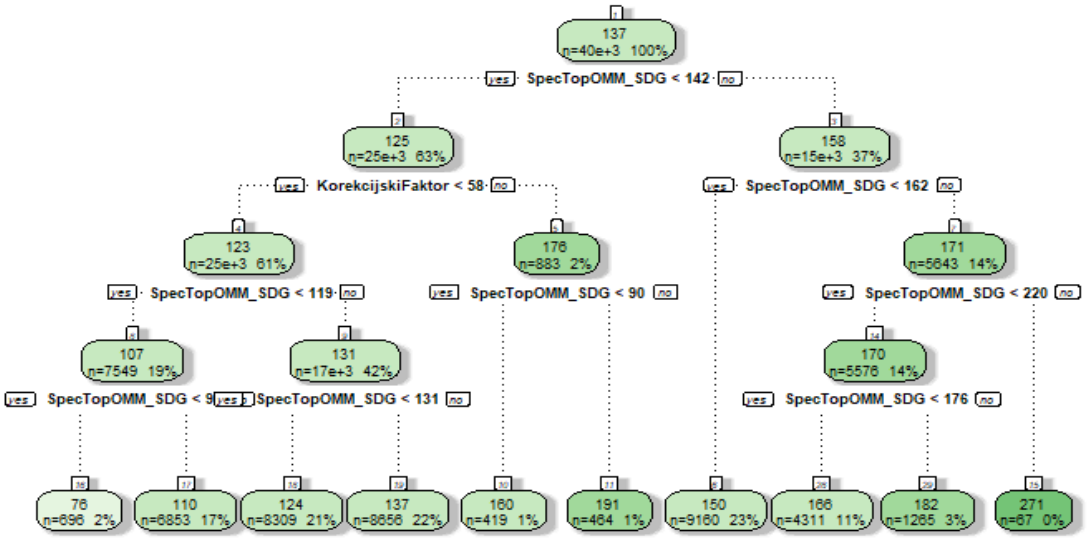


Slika 4-2 Regresijsko stablo za razinu stanova s ugrađenim individualnim mjerenjem

Analiza podatkovnog podskupa stanova koji nisu ugradili individualna mjerila i obračunavaju energiju prema modelu 2EG, pokazuje da je raspon potrošnje ovih stanova znatno viši nego kod stanova koji imaju ugrađena individualna mjerila i kreće se od 76 do 271 kWh/m^2. Korijensko grananje provodi se prema vrijednosti specifične potrošnje zgrade, varijabla *SpecTopOMM_SDG* prema iznosu od 142 kWh/m^2. S obzirom na to da je grananje u idućem unutarnjem čvoru nakon korijena za vrijednosti manju od 142 kWh/m^2, prema varijabli *KorekcijskiFaktor* koja ima vrijednost samo za one stanove koji se nalaze u zgradama u kojima je u nekim stanovima ugrađeno individualno mjerenje, može se zaključiti da i stanovi koji nemaju individualno mjerenje, a nalaze se u zgradama koje djelomično imaju ugrađeno

mjerenje, imaju manju specifičnu potrošnju od stanova koji nemaju ugrađeno mjerenje na razini čitave zgrade. Iz ovoga se vidi pozitivan učinak na smanjenje potrošnje energije čak i kod djelomične ugradnje.

Ostala grananja za ovaj podatkovni podskup provode se isključivo prema vrijednosti specifične potrošnje zgrade, što je i očekivano jer se pri obračunu prema modelu 2EG alokacija topline za stanove provodi jedino prema površini stana u odnosu prema površini zgrade.



Slika 4-3 Regresijsko stablo za razinu stanova bez ugrađenog individualnog mjerenja

Provedene analize odredile su modele predviđanja potrošnje za stanove u sustavima daljinskog grijanja primjenom metoda strojnog učenja. Kako je u tijeku ovoga rada navedeno, svaka od provedenih metoda ima pripadajuću razinu interpretativnosti. Metode korištene u ovom radu, kao što su višestruka linearna regresija i regresijska stabla, imaju višu razinu interpretativnosti od modela slučajnih šuma i stroja potpornih vektora. S druge strane, tzv. *black-box* metode, kao što su slučajne šume i stroj potpornih vektora, imaju višu razinu točnosti i jednostavniju primjenu te su zanimljive za primjenu u slučajevima kada je cilj postići što veću točnost predviđanja. Takve metode zahtijevaju visoku razinu ekspertize modelara o pojedinom problemu, pripremu modela i rad na podešenju optimalnih parametara, a kao rezultat daju samo predviđanja u smislu kvantifikacije zavisne varijable bez dodatne interpretativnosti.

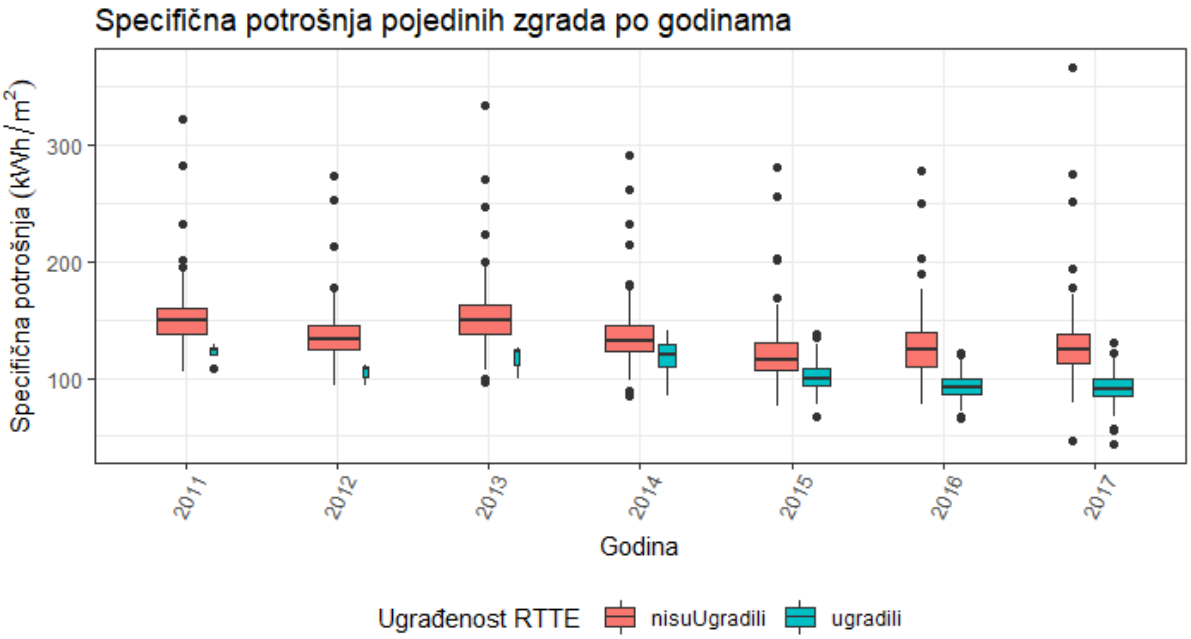
4.2. Predviđajući model na razini zgrada

Razvoj regresijskog modela za ovu razinu napravljen je na skupu 350 analiziranih zgrada. Osim deskriptivne statistike iz koje je vidljiva razlika u potrošnji energije na razini zgrade prije i poslije ugradnje RTTE, u predviđajućim modelima od glavnog je interesa kako stopa ugradnje, varijabla *StopaUgr*, utječe na potrošnju zgrade.

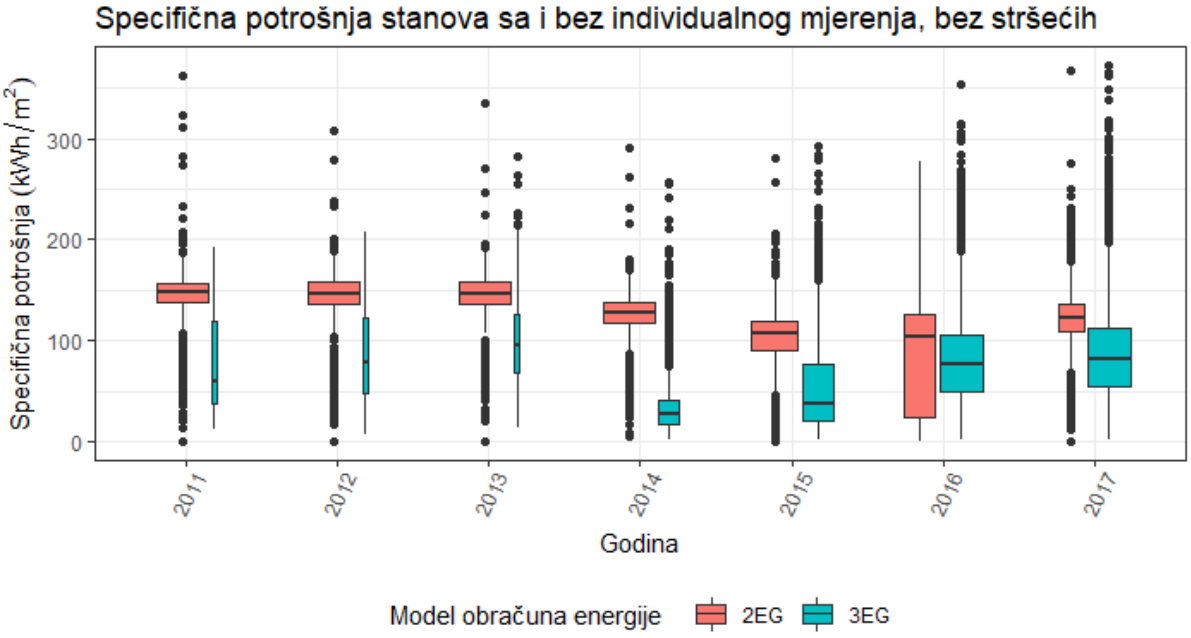
Treba ponovno napomenuti da ocjena uštede na razini zgrade zahvaljujući ugradnji RTTE nije dovoljna da bi se moglo ocijeniti da su i svi stanovi koje se nalaze u zgradi ostvarili uštede. S obzirom na to da zakonodavni okvir nije propisao da svi stanovi u jednoj zgradi moraju ugraditi individualna mjerenja, Pravilnikom [56] određeno je na koji način se alocira potrošnja za svaki stan u ovisnosti je li ili nije ugradio RTTE. Ovaj model alokacije potpuno je temeljen na matematičkoj formuli za izračun, bez ikakvih fizikalnih ili tehničkih temelja, čime se u nekim posebnim slučajevima otvara mogućnost da stan koji ima ugrađeno individualno mjerenje (a time i mogućnost upravljanja svojim troškovima i uštede) ima veću potrošnju od stanova koji nemaju mogućnost mjerenja.

Kako se može vidjeti iz slike 4-4, srednja vrijednost specifične potrošnje za zgrade koje nisu ugradile RTTE bez iznimke su veće od vrijednosti za zgrade koje jesu. Istovremeno se vidi da za godine 2014., 2015. i 2016. postoji mogućnost da neke zgrade koje nisu ugradile RTTE, a smještene su u donjem kvartilu, imaju niže vrijednosti nego zgrade su ugradile RTTE. Također se vidi da u 2017. više nema takvih pojavnosti i da su gotovo sigurno sve zgrade koje su ugradile individualno mjerenje imale manju potrošnju od zgrada koje to nisu učinile. Iskustva iz prakse pokazuju da je potreban određeni period učenja u kojem se krajnji korisnici privikavaju na novi način rada sustava grijanja, što obično traje oko 3 godine.

Ako usporedimo slike 4-4 i 4-5 za 2017. godinu, vide se naznake uzroka nezadovoljstva nekih krajnjih kupaca. Iako su zgrade (slika 4-4) postigle uštede te se specifična potrošnja zgrade u gornjem kvartilu kreće oko 100 kWh/m^2, i iako je gornji kvartil stanova koji su ugradili RTTE na razini gornjeg kvartila zgrada (slika 4-5), kod stanova koji su ugradili RTTE postoji velika pojavnost stršećih vrijednosti potrošnje u odnosu na onu koja je u statističkom prosjeku (između donjeg i gornjeg kvartila). Stršeće vrijednosti za stanove su greške mjerenja ili alokacije te je potrebno poduzeti radnje kako bi se te greške otklonile. Ova pojavnost izrazito je negativna, potrošači u stanovima koji bilježe stršeće vrijednosti zasigurno su nezadovoljni i žele ukinuti svoj priključak na sustav daljinskog grijanja.



Slika 4-4 Komparativni prikaz potrošnje pojedinih zgrada s obzirom na način alokacije topline



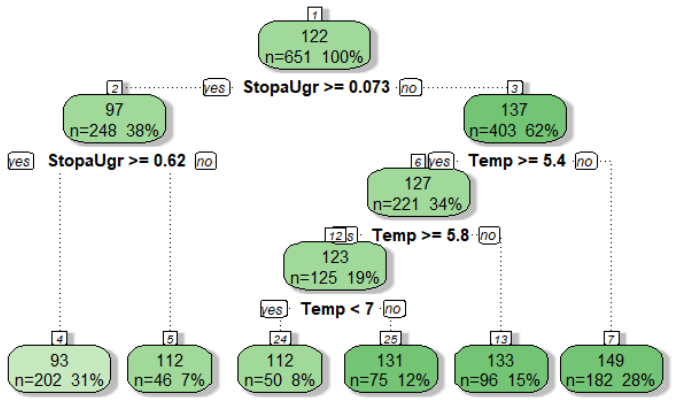
Slika 4-5 Komparativni prikaz potrošnje stanova s obzirom na način alokacije topline

Regresijski model dobiven je bez regularizacije i prikazan je u jednadžbi:

$$\boldsymbol{SpecTopOMM_{SDG} = 172,89 - 5,21 \cdot Temp - 46,16 \cdot StopaUgr - \ 0,001 \cdot PovZgrada} \qquad (65)$$

Ovim modelom obuhvaćene su samo one zgrade koje su provele postupak ugradnje RTTE. Dominantna utjecajna varijabla je *StopaUgr*, čime se dodatno naglašava važnost 100-postotne ugradnje u stanovima u pojedinim zgradama kako bi se postigle maksimalne uštede na razini zgrade. Također, to je jedina varijabla u regresijskom modelu na koju se može utjecati, ima negativan predznak te s njezinim povećanjem pada specifična potrošnja zgrade. Ilustrativno se može reći da ako usporedimo dvije zgrade jednake površine, s time da jedna ima stopu ugradnje 0 %, a druga ima stopu ugradnje 100 % - zgrada koja ima ugrađenost 100% ostvaruje manju potrošnju od približno 46 kWh/m^2 na godišnjoj razini.

Metoda regresijskih stabala omogućuje nam da interpretiramo što pojedine zgrade mogu očekivati nakon ugradnje RTTE. Rezultati ove metode identificiraju stopu ugradnje kao najutjecajniju varijablu koja utječe na potrošnju topline unutar pojedine zgrade (slika 4-6). Ovime se potvrđuje činjenica da ugradnja individualnog mjerenja dovodi do ušteda na razini zgrade. U skupu analiziranih zgrada, zgrade koje imaju stopu ugradnje nižu od približno 7 % mogu očekivati godišnju specifičnu potrošnju u rasponu od 112 do 149 kWh/m^2, što ovisi o srednjoj vanjskoj temperaturi u sezoni grijanja. S druge strane, zgrade koje su ugradile individualna mjerenja mogu očekivati srednju godišnju specifičnu potrošnju topline između 93 i 112 kWh/m^2. Potrošnja je na donjoj granici ako je stopa viša od 62 %.



Slika 4-6 Regresijsko stablo za razinu zgrade

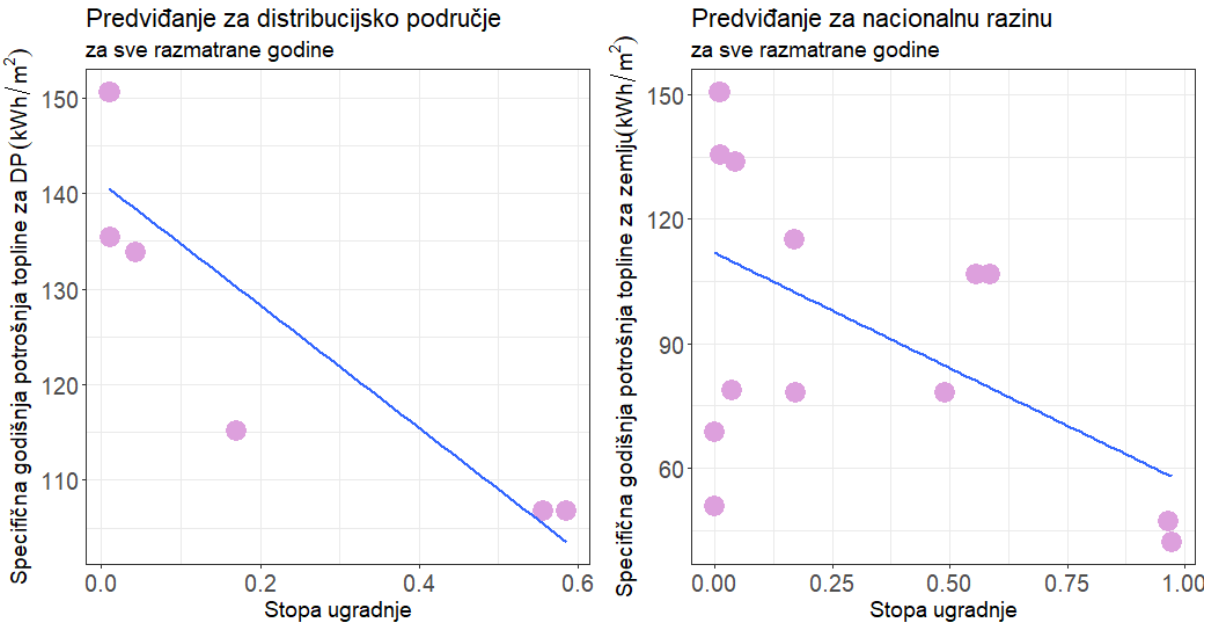
4.3. Predviđajući model na višim razinama

Regresijski model za predviđanje potrošnje topline u distribucijskom području razvijen je prema dostupnim podacima. U ovoj analizi dostupni su bili podaci za dva najveća distribucijska područja u smislu broja zgrada i stanova. Karakteristike regresijskog pravca za najveće distribucijsko područje i za nacionalnu razinu prikazane su na slici 4-7. Konačni regresijski model za predviđanje potrošnje na razini distribucijskog područja glasi:

$$\boldsymbol{SpecTopDP = -333,33 - 44,0 \cdot StopaUgr + 0,095 \cdot Sr.V(PovZgrada) + 5,3 \cdot Temp} \quad (66)$$

Konačni model za predviđanje potrošnje na nacionalnoj razini glasi:

$$\boldsymbol{SpecTopDP = 731,074 - 128.8 * DP2 - 21,6 \cdot StopaUgr - 0,13\, Sr.Vr(PovZgrada) - 3,04 \cdot Temp} \quad (67)$$



(a) *(b)*

Slika 4-7 Dijagram rasipanja i regresijski pravac za (a) najveće analizirano distribucijsko područje i (b) nacionalnu razinu

4.4. Vrednovanje algoritama strojnog učenja

Vrednovanje algoritama strojnog učenja radi se prema točnosti predviđanja i razini interpretativnosti. U tablici 4-1 dan je sumarni pregled točnosti po svakom razmatranom algoritmu strojnog učenja za razinu stana.

Tablica 4-1 Vrednovanje RMSE točnosti predviđanja razmatranih algoritama strojnog učenja na razini stana

Metoda strojnog učenja	Model razvijen za razinu	RMSE
Višestruka linearna regresija	Svi stanovi	16,44 kWh/m^2
	Stanovi s ugrađenim RTTE	18,20 kWh/m^2
	Stanovi bez ugrađenog RTTE	3,70 kWh/m^2
Regresijska stabla	Svi stanovi	11,78 kWh/m^2
	Stanovi s ugrađenim RTTE	19,31 kWh/m^2
	Stanovi bez ugrađenog RTTE	6,53 kWh/m^2
Slučajne šume	Svi stanovi	4,27 kWh/m^2
	Stanovi s ugrađenim RTTE	7,25 kWh/m^2
	Stanovi bez ugrađenog RTTE	2,73 kWh/m^2
Stroj potpornih vektora	Svi stanovi	5,96 kWh/m^2
	Stanovi s ugrađenim RTTE	5,45 kWh/m^2
	Stanovi bez ugrađenog RTTE	5,38 kWh/m^2

U tablici 4-2 dan je sumarni pregled točnosti po svakom razmatranom algoritmu strojnog učenja za razinu zgrade. Testiranja između pojedinih modela nije moguće napraviti u ovom istraživanju, jer analize pojedinim metodama nisu rađene na istom skupu podataka, npr. za višestruku regresiju konačni modeli se regularizirani.

Tablica 4-2 Vrednovanje RMSE točnosti predviđanja razmatranih algoritama strojnog učenja na razini zgrade

Metoda strojnog učenja	RMSE
Višestruka linearna regresija	16,59 kWh/m^2
Regresijska stabla	13,38 kWh/m^2
Slučajne šume	13,24 kWh/m^2
Stroj potpornih vektora	13,65 kWh/m^2

Najvišu interpretativnost imaju metode višestruke linearne regresije i regresijskih stabala, s tim da višu točnost ima model slučajnih šuma. Gledajući samo točnost predviđanja modela, najvišu točnost ima model slučajnih šuma, ali i vrlo slabu intepretativnost.

4.5. Određivanje stupnja utjecaja pojedinih faktora na potrošnju toplinske energije

4.5.1. Utjecaj na razini stanova

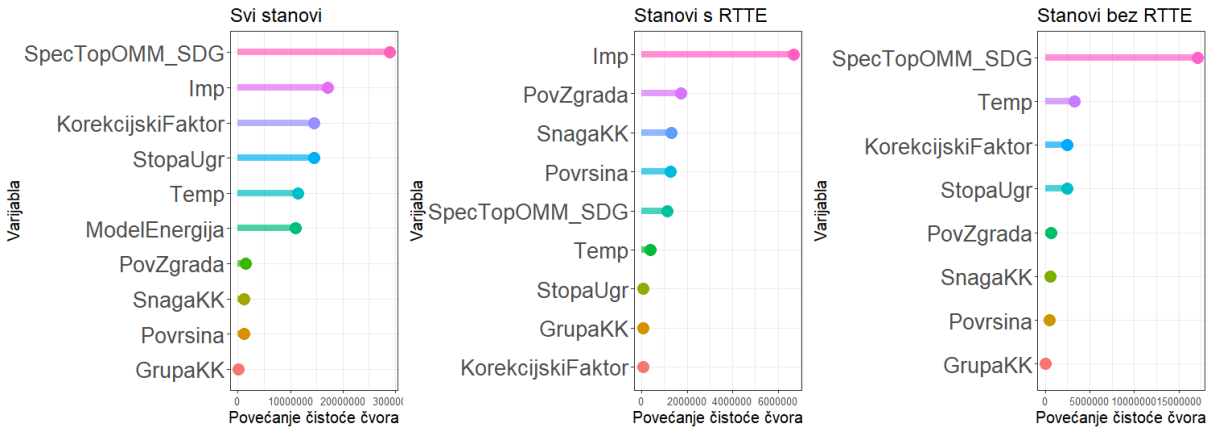
Određivanje stupnja utjecaja na potrošnju bitno je s aspekta energetske učinkovitosti kako bi se mogli identificirati parametri od većeg utjecaja te osmisliti metode i politike kojima se može utjecati na prepoznate parametre radi postizanja što veće uštede energije na troškovno učinkovit način.

Indikaciju utjecaja regresijskih modela iz jednadžbi (*63*), (*64*) i (*65*) možemo sumirati kako slijedi:

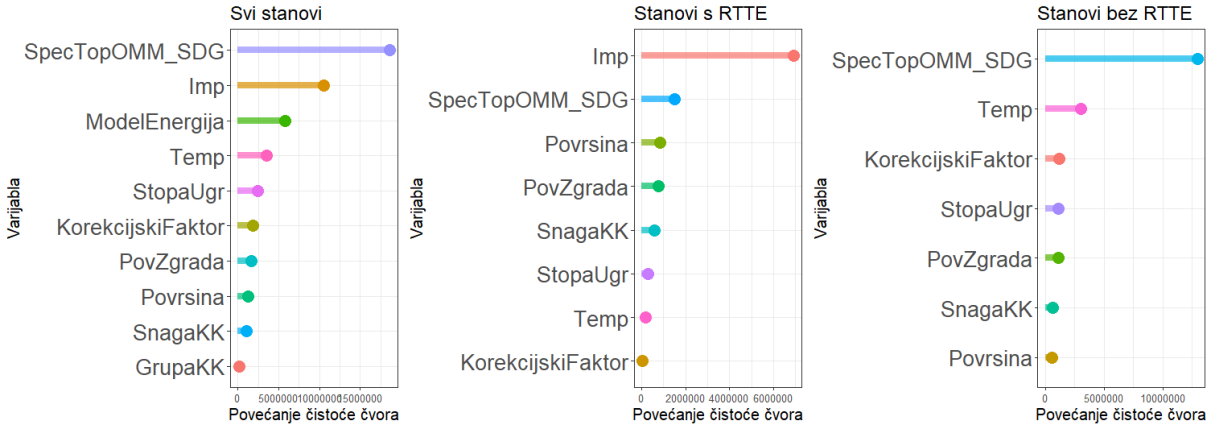
- Dominantan utjecaj na potrošnju stanova ima ugradnja individualnog mjerenja (varijabla *ModelEnergija*), što je u skladu s pretpostavkama i ciljevima Direktive o energetskoj učinkovitosti [1].
- Drugi parametar je očitanje impulsa s RTTE u onim stanovima u kojima su oni ugrađeni, što je u skladu s pretpostavkama i ciljevima Direktive o energetskoj učinkovitosti [1].
- Treći parametar je stopa ugrađenosti na razini zgrade. Regresijski faktor β ispred varijable *StopaUgr* je +80, koji daje indikaciju da ako se stopa ugradnje poveća na 100 %, da će to povećati potrošnju. Ali u tom slučaju bi i regresijski član ($\beta \cdot ModelEnergija$) iznosio -123 za sve stanove, tako da bi pri 100-postotnoj stopi ugradnje ukupan benifit za stanove koji su ugradili RTTE bio oko 53 kWh/m^{2}·na godišnjoj razini.
- Zadnji utjecajan parametar na koji se može djelovati specifična je potrošnja zgrade (varijabla *SpecTopOMM_SDG*). Na nju se može djelovati primjenom mjera energetske učinkovitosti kao što je poboljšanje tehničkih karakteristika zgrade i na taj način direktno utjecati i na potrošnju svih stanova u zgradi.
- Za stanove koji imaju ugrađene RTTE dodatno se kao utjecajan parametar pojavljuju površine stanova (varijabla *Povrsina*) i površina zgrade (varijabla *PovZgrada)*.
- Za stanove koji nemaju ugrađene RTTE kao utjecajan parametar pojavljuje se korekcijski faktor (varijabla *KorekcijskiFaktor*), što je u skladu s Pravilnikom [56], jer se putem ove varijable korigira grijana površina stana u smislu penalizacije.

Utjecaj bihevioralnih parametara prepoznaje se na razini stanova i reflektira se u broju impulsa u pojedinom stanu, a opisan je u poglavlju 3.3.

Utjecajne varijable koje su identificirane metodama strojnog učenja regresijskih stabala (4-8), slučajne šume i stroja potpornih vektora (4-9) prikazane su kroz modeliranje metodom *VarImp*, s time da su rezultati za metodu stroja potpornih vektora identični rezultatima metode regresijskih stabala. Na čitavom podatkovnom skupu najutjecajniji parametar prema metodi regresijskih stabala je *SpecTopOMM_SDG*, specifična potrošnja topline na razini zgrade, što je za očekivati s obzirom na to da stanovi s oba modela obračuna u alokacijskim formulama koriste ovu vrijednost. Slijede varijable: *Imp*, koje se odnose prije svega na stanove s ugrađenim mjerenjem pa *KorekcijskiFaktor* koji se odnosi na stanove bez ugrađenih mjerila. Zatim po utjecaju slijede parametri *StopaUgr* i *Temp* te varijabla *ModelEnergija*.



Slika 4-8 Utjecajne varijable za svaki podskup na razini stanova za metodu regresijski stabala



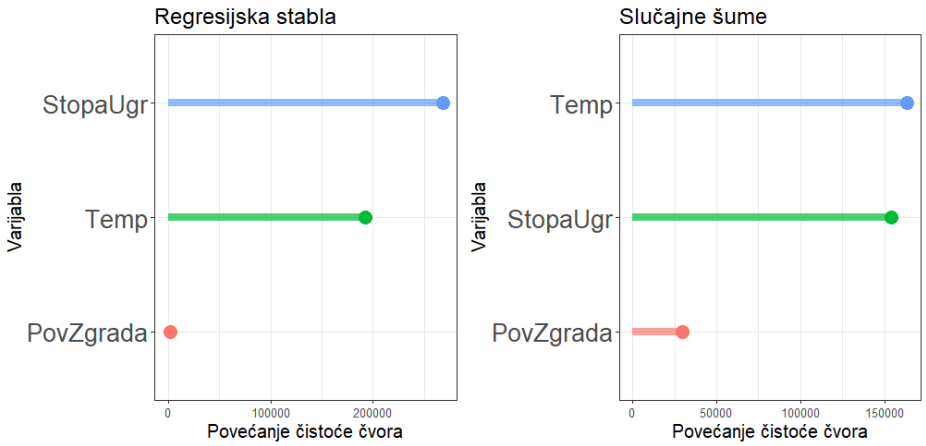
Slika 4-9 Utjecajne varijable za svaki podskup na razini stanova za metodu slučajnih šuma

Algoritmom slučajnih šuma dobivamo visoku točnost predviđanja, ali bez dodatne interpretativnosti rezultata, osim analizom utjecajnih varijabli, kako je prikazano na slici 4-9. Utjecaj svake varijable mjeri se tako da se mjeri pojavnost smanjenja MSE svaki put kad se varijabla koristi kao kriterij za grananje u čvoru stabla. Nakon svakog grananja dobivamo novu grešku predviđanja koja se naziva nečistoćom čvora (engl. *node impurity*). Varijable koje smanjuju nečistoću čvora smatraju se utjecajnijim i važnijim od onih koje nemaju takav učinak. Prilikom provedbe analize metodom slučajnih šuma učinak smanjenja MSE za svaku se pojedinu varijablu pribraja, i one varijable koje imaju najveći učinak na smanjenje MSE smatraju se utjecajnijima. Vrijednost učinka je relativna i u načelu označava sumu svih smanjenja MSE relativno prema ostalim varijablama te nema mjernu jedinicu. Prepoznaje se da je u modelu slučajnih šuma faktor koji dominantno utječe na potrošnju pojedinog stana u načelu potrošnja zgrade u kojoj se stan nalazi, što je i očekivano s obzirom na to da se specifična toplina stanova zapravo alocira iz izmjerene topline u toplinskoj podstanici na koju je zgrada spojena. Iduća utjecajna varijabla je *Imp*, koja predstavlja kvocijent očitanih impulsa sa svih razdjelnika u pojedinom stanu i sume svih očitanih razdjelnika u jednoj zgradi. Ova varijabla također je utjecajna, jer zapravo predstavlja mjerenje koje je alocirano pojedinom stanu. Treća utjecajna varijabla je *ModelEnergija,* i ovdje se vidi da znatan utjecaj na potrošnju stana ima način obračuna te da stanovi koji su ugradili individualna mjerenja mogu očekivati manju specifičnu potrošnju. Kao četvrtu utjecajnu varijablu model prepoznaje vanjsku temperaturu koja je direktno povezana s toplinskim opterećenjem. Također se vidi znatan utjecaj varijable *StopaUgr*. Zgrade koje imaju višu stopu ugradnje imaju stanove s manjim specifičnim potrošnjama topline, te zbog toga treba poduzimati sve radnje da stopa ugrađenosti bude 100 % ako se želi postići maksimalna ušteda svih stanova (i zgrada).

Algoritam slučajnih šuma kao utjecajne varijable za podatkovni skup stanova koji su ugradili RTTE konzistentno prepoznaje jednake varijable kao drugi analizirani algoritmi i kako je prikazano na slici 4-9. Tako je ponovno dominantna varijabla *Imp*, s obzirom na to da je to omjer koji zajedno s varijablama *Povrsina* i *PovStana* utječe na način na koji se alocira izmjerena toplinska energija na razini zgrade, *SpecTopOMM_SDG*. Prvih 6 utjecajnih varijabli konzistentno je s identificiranim utjecajnim varijablama za algoritam regresijskih stabala (slika 4-8). Također, 5 varijabli iz regresijskog modela za stanove na 3EG, jednadžba (*64*), nalazi se u prvih 6 utjecajnih varijabli prema algoritmu slučajnih šuma, što pokazuje konzistentnost u stupnju utjecaja pojedinih nezavisnih varijabli na zavisnu varijablu.

4.5.2. Utjecaj na razini zgrada

Utjecajni faktori određeni su samo za one zgrade koje su provele postupak ugradnje RTTE prema slici 4-10.



Slika 4-10 Utjecajne varijable na razini zgrada za metodu regresijskih stabala i slučajnih šuma

Najutjecajniji parametar koji se prepoznaje u analizama svih metoda strojnog učenja za razinu zgrade je stopa ugradnje RTTE, što je ujedno jedini parametar na koji se može utjecati i koji ponovno pokazuje da je od iznimne važnosti što prije provesti 100-postotnu ugradnju u zgradama kako bi se postigla maksimalna ušteda na razini zgrade. Od ostalih utjecajnih varijabli prepoznaju se vanjska temperatura i površina zgrade.

4.5.3. Utjecaj na razini distribucijskog sustava i nacionalna razina

Utjecaj na razini distribucijskog sustava i na nacionalnoj razini jednak je utjecaju na razini zgrade te zapravo predstavlja agregirani utjecaj svih zgrada. Dominantna utjecajna varijabla je stopa ugradnje, iz čega se dolazi do zaključka da je nužno provesti sve mjere potrebne da stopa ugradnje na nacionalnoj razini dođe na 100 % kako bi se postigle maksimalne uštede. Prema analizi u ovom radu koja je provedena na određenom broju godina nakon ugradnje RTTE, očekivano prosječno smanjenje potrošnje po svakom grijanom kvadratnom metru u sustavima daljinskog grijanja u Hrvatskoj može se očekivati na godišnjoj razini od 44 kWh/m^2 .

5. ZAKLJUČAK

U sklopu ovog rada analiziran je stvarni učinak ugradnje razdjelnika troškova toplinske energije (RTTE) u sustavima daljinskog grijanja i definirani su utjecajni parametri na potrošnju na razini stana, zgrade i višim razinama, razvojem predviđajućih modela i primjenom metoda strojnog učenja. Razvijeni modeli imaju visoku točnost predviđanja, posebno u usporedbi s klasičnim simulacijskim metodama i softverima, koji nerijetko rade znatne pogreške u predviđanju stvarne potrošnje toplinske energije u zgradama [52].

Razvijeni regresijski modeli na razini stanova pokazuju da stanovi s ugrađenim RTTE mogu očekivati 40 % manju potrošnju toplinske energije od stanova koji ih nisu ugradili, u slučaju da na godišnjoj razini ne broje više od 4 % svih impulsa u zgradi. Dodatna interpretativnost dobiva se metodom regresijskih stabala, koja pokazuje da najmanju srednju godišnju specifičnu potrošnju imaju oni stanovi koji su ugradili RTTE i koji godišnje broje manje od 3,6 % svih impulsa u zgradi. Takvih stanova ima ukupno 10 % u analiziranom podatkovnom skupu, a njihova se specifična potrošnja topline kreće od 35 do 88 kWh/m^2. Najvišu potrošnju u skupu stanova koji su ugradili RTTE imaju stanovi koji na godišnjoj razini broje više od 8,1 % svih impulsa i njihova srednja potrošnja kreće se oko 125 kWh/m^2. Takvih je stanova 3 % i, prema provedenoj analizi, predstavljaju nezadovoljne krajnje kupce koji su dobili negativan učinak nakon ugradnje RTTE. Dodatni razlog za nezadovoljstvo prepoznaje se u tome što su specifične potrošnje onih stanova koji nisu ugradili RTTE, a nalaze se u zgradama u kojima je provedena ugradnja, u rasponu od 97 do 176 kWh/m^2. Dakle, postojeći način obračuna omogućava da u istoj zgradi neki od stanova koji nisu ugradili RTTE plaćaju manje nego stanovi koji su ih ugradili.

Na razini zgrade, sve analizirane zgrade imaju više od 70 stanova i ostvarile su uštede nakon ugradnje RTTE. Dobiveni regresijski model daje indikaciju da se 100-postotnom ugradnjom RTTE u zgradi može očekivati apsolutna specifična ušteda na godišnjoj razini od 46 kWh/m^2. Model regresijskih stabala i na razini zgrada daje dodatnu intepretativnost te se tako može reći da na razini zgrade najnižu potrošnju (93 kWh/m^2) imaju one zgrade koje imaju stopu ugradnje RTTE prema grijanoj površini višu od 62 %.

Točnost predviđanja svakog od četiri modela razvijena metodama strojnog učenja (višestruka linearna regresija, regresijska stabla, slučajne šume i stroj potpornih vektora) na razini stana i razini zgrade visoka je u usporedbi s klasičnim simulacijskim metodama. Najbolje rezultate

vrednovanja točnosti predviđanja prema RMSE dobiva se metodom slučajnih šuma (stanovi +/- 4,27 kWh/m^2, zgrade 13,24 kWh/m^2), dok su najlošiji rezultati kod modeliranja metodom višestruke linearne regresije (stanovi +/- 16,44 kWh/m^2, zgrade 16,59 kWh/m^2).

Pri odabiru metoda treba naći kompromis između točnosti i interpretativnosti. Ako je želja modeliranja i predviđanja dobiti što točniji model s određenom razinom interpretativnosti, na način da se vrednuje utjecaj varijabli na potrošnju toplinske energije, predlaže se koristiti model regresijskih stabala. S druge strane, ako nam je jedini cilj predviđanja dobiti što veću točnost, npr. želi se kvantificirati utjecaj mjere ugradnje RTTE, predlaže se koristiti metodu strojnog učenja slučajnim šumama.

Razmatrajući stupnjeve utjecaja pojedinih varijabli zaključuje se da su kod stanova dominantne varijable koje su vezane za izmjerenu energiju (omjer izbrojenih impulsa u stanu i ukupno izbrojenih impulsa u zgradi, izmjerena toplinska energija na obračunskom mjerilu) zajedno s varijablama koje opisuju je li stan ugradio RTTE i kolika je stopa ugradnje u zgradi. Ako se želi smanjiti potrošnja toplinske energije u što većem broju stanova, trebalo bi djelovati na te varijable. Preporuka je da se u sve stanove ugradi RTTE (mjere vezane u implementaciju sekundarne mjere energetske učinkovitosti), da se provedu sve troškovno optimalne mjere na tehničkim sustavima u zgradama (mjere vezane za implementaciju primarnih mjera energetske učinkovitosti) te da se provedu mjere edukacije krajnjih korisnika kako bi smanjili postotak izbrojenih impulsa u svom stanu u odnosu na sve izmjerene impulse u zgradi (mjere kojima se utječe na bihevioralne parametre potrošnje).

Pozitivni učinci ugradnje RTTE očigledni su na razini zgrade, distribucijskih područja i u širim nacionalnim okvirima, ali na razini stanova dolazi do velikog rasipanja rezultata i pojava stršećih vrijednosti zbog više utjecajnih bihevioralnih parametara identificiranih u ovom radu. Fokus u sustavima daljinskog grijanja treba usmjeriti na pravilnu raspodjelu troškova za toplinsku energiju temeljenu na izmjerenoj potrošnji.

Modeli razvijeni u ovom radu mogu služiti kao osnova za predviđanje potrošnje u sustavima daljinskog grijanja s visokom razinom točnosti i mogu se koristiti kao alat za ocjenu učinka provedbe mjera energetske učinkovitosti.

POPIS LITERATURE

[1] European Parliament and Council. Directive 2010/31/EU of the European Parliament and of the Council of 19May 2010 on the energy performance of buildings. Official Journal of the European Union 2010;L153:13–35.

[2] Zakon o tržištu toplinske energije (Narodne novine, br. 80/13, 14/14, 86/19)

[3] Zakon o energetskoj učinkovitosti(Narodne novine, br. 127/14, 116/18)

[4] Razdjelnici troškova za utvrđivanje potrošnje radijatora za grijanje prostorija – Uređaji s napajanjem električnom energijom (EN 834:2013 i HRN EN 834)

[5] Zakon o gradnji (Narodne novine , br. 153/13, 20/17, 39/19, 125/19)

[6] Mapping and analyses of the current and future (2020 - 2030) heating/cooling fuel deployment (fossil/renewables)" ENER/C2/2014-641, Fraunhofer and alia, 2015.

[7] L. De Boeck, S. Verbeke, A. Audenaert, L. De Mesmaeker, Improving the energy performance of residential buildings: A literature review, Renewable and Sustainable Energy Reviews, Volume 52, December 2015, Pages 960-975.

[8] Catalina, J. Virgone, E. Blanco, Development and validation of regression models to predict monthly heating demand for residential buildings, Energy and Buildings 40 (10) (2008) 1825–1832.

[9] L. Magnier, F. Haghighat, Multiobjective optimization of building design using TRNSYS simulations, genetic algorithm, and Artificial Neural Network, Building and Environment 45 (2010) 739–746

[10] F.F. Al-ajmi, V.I. Hanby, Simulation of energy consumption for Kuwaiti domestic buildings, Energy and Buildings 40 (6) (2008) 1101–1109.

[11] International Energy Agency, Mapping the energy future: energy modelling and climate change policy, Energy and environment policy analysis series. Paris, France; International Energy Agency/Organisation for Economic Cooperation and Development; 1998.

[12] Kristoffer S. Andersen, Lars B. Termansen, Maurizio Gargiulo, Brian P. O Gallachoirc, Bridging the gap using energy services: Demonstrating a novel, framework for soft linking top-down and bottom-up models, Energy 169 (2019) 277-293, 2019.

[13] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, M. Djurovic-Petrovic, A review of bottom-up building stock models for energy consumption in the residential sector, Building and Environment, Volume 45, Issue 7, July 2010, Pages 1683–1697

[14] Liu Yang, Haiyan Yan, Joseph C. Lam, Thermal comfort and building energy consumption implications – A review, Applied Energy, Volume 115, 15 February 2014, Pages 164–173

[15] Tuan Anh Nguyen, , Marco Aiello, Energy intelligent buildings based on user activity: A survey, Energy and Buildings, Volume 56, January 2013, Pages 244–257

[16] Olofsson T, Andersson S, Sjogren JU. Building energy parameter investigations based on multivariate analysis, Energy and Buildings, Volume 41, Issue 1, January 2009, Pages 71–80

[17] Rishee K. Jain, Kevin M. Smith, Patricia J. Culligan, John E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Applied Energy, Volume 123, 15 June 2014, Pages 168–178

[18] H. Lund, B. Möller, B.V. Mathiesen, A. Dyrelund, The role of district heating in future renewable energy systems, Energy Volume 35, Issue 3, March 2010, Pages 1381–1390.

[19] Thellufsen, J.Z. ,Lund, H., Energy saving synergies in national energy systems, Energy Conversion and Management, Volume 103, 4 July 2015, Article number 7294, Pages 259-265.

[20] Communication from the Commision to the European Parlament, the Council, The European Economic and Social Committee and the Committee of the Regions, An EU Strategy on Heating and Cooling, {SWD(2016) 24 final}, Brussels, 16.2.2016.

[21] European Commission, 2012, DIRECTIVE 2012/27/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC, Off. J. Eur. Communities 2012.

[22] S. Siggelsten, Reallocation of heating costs due to heat transfer between adjacent apartments, Energy and Buildings, 75 (2014), pp. 256–263.

[23] Lanbin Liu, Lin Fu, Yi Jiang, Shan Guo, Major issues and solutions in the heat-metering in China, Renewable and Sustainable Energy Reviews 15 (2011) 673–680

[24] H. Burak Gunay, William O'Brien, Ian Beausoleil-Morrison, Andrea Perna, On the behavioral effects of residential electricity submetering in a heating season, Building and Environment 81 (2014) 396-403

[25] Søren Andersen, Rune Korsholm Andersen, Bjarne W. Olesen, Influence of heat cost allocation on occupants' control of indoor environment in 56 apartments: Studied with measurements, interviews and questionnaires, Building and Environment 101 (2016) 1-8.

[26] Tomasz Cholewa, , Alicja Siuta-Olcha, Long term experimental evaluation of the influence of heat cost allocators on energy consumption in a multifamily building, Energy and Buildings, Volume 104, 1 October 2015, Pages 122–130

[27] Milan Protić, Shahaboddin Shamshirband, Dalibor Petković, ,Almas Abbasi, Miss Laiha Mat Kiah, Jawed Akhtar Unar, Ljiljana Živković, Miomir Raos, Forecasting of consumers heat load in district heating systems using the support vector machine with a discrete wavelet transform algorithm, Energy, Volume 87, 1 July 2015, Pages 343–351

[28] B. Yildiz, J.I. Bilbao, A.B. Sproul, A review and analysis of regression and machine learning models on commercial building electricity load forecasting, Renewable and Sustainable Energy Reviews, Volume 73, June 2017, Pages 1104–1122

[29] J. Sánchez-Oroa, A. Duartea, S. Salcedo-Sanzb, Robust total energy demand estimation with a hybrid Variable Neighborhood Search – Extreme Learning Machine

algorithm, Energy Conversion and Management, Volume 123, 1 September 2016, Pages 445–452

[30] P.Woods, J.Overgaard, Historical development of district heating and characteristics of a modern district heating system, Advanced District Heating and Cooling (DHC) Systems, Woodhead Publishing Series in Energy, 2016, Pages 3-15.

[31] R. F. Babus'Haq, S. D. Probert, M. J. Shilston, The total energy approach: Evolution of combined heat and power for district heating and/or cooling, Applied Energy, Volume 25, Issue 2, 1986, Pages 97-166.

[32] Henrik Lund, Sven Werner, Robin Wiltshire, Svend Svendsen, Jan Eric Thorsen, Frede Hvelplund, Brian Vad Mathiesen, 4th Generation District Heating (4GDH): Integrating smart thermal grids into future sustainable energy systems, Energy, Volume 68, 15 April 2014, Pages 1-11.

[33] Henrik Lund, Neven Duic, Poul Alberg Østergaard, Brian Vad Mathiesen, Future district heating systems and technologies: On the role of smart energy systems and 4th generation district heating, Energy, Volume 165, Part A, 15 December 2018, Pages 614-619.

[34] S. Frederiksen, S. Werner, District Heating and Cooling, Studentlitteratur, ISBN 978-91-44-08530-2, 2014.

[35] European Commission, An EU Strategy on Heating and Cooling, COM(2016) 51 final, 2016.

[36] Hrvatska energetska regulatorna agencija, Godišnje izvješće za 2018. godinu, 2019., https://www.hera.hr/hr/docs/HERA_izvjesce_2018.pdf

[37] Branko Vuk et el, Energija u Hrvatskoj 2017 – Godišnji energetski pregled, Ministarstvo zaštite okoliša i energetike Republike Hrvatske, 2018., ISSN 847-0602.

[38] Connolly, David; Mathiesen, Brian Vad; Lund, Henrik; Møller, Bernd; Hvelplund, Frede Kloster; Sperling, Karl; Østergaard, Poul Alberg; Werner, Sven, Heat Roadmap Europe, European Commission, 2015.

[39] O. Guerra Santin, Actual Energy Consumption in Dwellings: the Effect of Energy Performance Regulations and Occupant Behaviour, Delft University Press, Delft, 2010.

[40] Guerra-Santin Olivia, Tweed Aidan Christopher, In-use monitoring of buildings: An overview and classification of evaluation methods, Energy and Buildings 86 (2015) 176–189

[41] Laure Itard, Tasos Ioannou, Arjen Meijer, Arash Rasooli, Wim Kornaat, Devoplopment of improved models for the accurate predicition of energy consumption in dwellings, MONICAIR project, Delft University Press, Delft, 2016.

[42] Simon Siggelsten, Stefan Olander, Individual metering and charging of heat and hot water in Swedish housing cooperatives, Energy Policy, Volume 61, October 2013, Pages 874-880.

[43] J.Terés-Zubiag, E.Pérez-Iribarren, I.González-Pino, J.M.Sala, Effects of individual metering and charging of heating and domestic hot water on energy consumption of buildings in temperate climates, Energy Conversion and Management, Volume 171, 1 September 2018, Pages 491-506.

[44] Kaiyu Sun, Tianzhen Hong, A simulation approach to estimate energy savings potential of occupant behavior measures, Energy and Buildings, Volume 136, 1 February 2017, Pages 43-62.

[45] Olivia Guerra Santin, Laure Itard, Henk Visscher, The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock, Energy and Buildings, Volume 41, Issue 11, November 2009, Pages 1223-1232.

[46] Elisha R. Frederiks, Karen Stenner, Elizabeth V. Hobman, Household energy use: Applying behavioural economics to understand consumer decision-making and behaviour, Renewable and Sustainable Energy Reviews, Volume 41, January 2015, Pages 1385-1394.

[47] Baoping Xu, Ang Huang, Lin Fu, Hongfa Di, Simulation and analysis on control effectiveness of TRVs in district heating systems, Energy and Buildings 43 (2011) 1169–1174

[48] Simona D'Oca, Valentina Fabi, Stefano P. Corgnati, Rune Korsholm Andersen, Effect of thermostat and window opening occupant behavior models on energy use in homes, Building Simulation (2014) 7: 683–694, DOI 10.1007/s12273-014-0191-6

[49] J. Toftum, R.V. Andersen, K.L. Jensen, Occupant performance and building energy consumption with different philosophies of determining acceptable thermal conditions, Building and Environment 44 (2009) 2009–2016

[50] Jared Langevin, Jin Wen, Patrick L. Gurian, Including occupants in building performance simulation: integration of an agent-based occupant behavior algorithm with EnergyPlus, 2014 ASHRAE/IBPSA-USA, 2014.

[51] Valentina Fabi, Rune Vinther Andersen, Stefano P. Corgnati, Bjarne W. Olesen, A methodology for modelling energy-related human behaviour: Application to window opening behaviour in residential buildings, Building Simulation, DOI 10.1007/s12273-013-0119-6.

[52] Elham Delzendeh, Song Wu, Angela Lee, Ying Zhou, The impact of occupants' behaviours on building energy analysis: A research review, Renewable and Sustainable Energy Reviews 80 (2017) 1061–1071

[53] Jelena Ziemele, Ieva Pakere, Dagnija Blumberga, Gatis Zogla, Economy of heat cost allocation in apartment buildings, Energy Procedia 72 (2015) 87 – 94, 2015.

[54] UNI 10200:2015, Norma Italiana, Impianti termici centralizzati di climatizzatione internale, estiva e produzione di acqua calda sanitaria – Criteri di ripartizione delle spese di climatizzazione internale, estiva e produzione di acqua calda sanitaria.

[55] Roberto Colombo, Fiorenzo Zerbetto, La contabilizzazione del calore negli edifici con riscaldamento centralizzato, Maggioli Editore, 2015.

[56] Pravilnik o načinu raspodjele i obračunu troškova za isporučenu toplinsku energiju (Narodne novine, br. 99/14, 27/15, 124/15)

[57] Mehmood, M. U., Chun, D., Zeeshan, Han, H., Jeon, G., Chen, K., A review of the applications of artificial intelligence and big data to buildings for energy-efficiency

and a comfortable indoor living environment, Energy and Buildings, Volume 202, 1 November 2019, 109383, https://doi.org/10.1016/j.enbuild.2019.109383

[58] Rishee K. Jain, Kevin M. Smith, Patricia J. Culligan, John E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Applied Energy, Volume 123, 15 June 2014, Pages 168–178

[59] Smola, A. J., Vishwanathan, S.V.N., and Le, Q. V., Bundle Methods for Machine Learning, Advances in Neural Information Processing Systems 20, J.C. Platt and D. Koller and Y. Singer and S. Roweis (Eds.), MIT Press, Cambridge, MA 2008.

[60] Hai Wang, Haiying Wang, Zhou Haijian, Tong Zhu, Optimization modeling for smart operation of multi-source district heating with distributed variable-speed pumps, Energy 138 (2017) 1247-1262, https://doi.org/10.1016/j.energy.2017.08.009

[61] Zhong, H., Wang, J., Jia, H., Mu, Y., Lv, S., Vector field-based support vector regression for building energy consumption prediction, Applied Energy, Volume 242, 15 May 2019, Pages 403-414, https://doi.org/10.1016/j.apenergy.2019.03.078

[62] Francesco Guarino, Maurizio Cellura, Sonia Longo, Teresa Gulotta, Marina, Mistrettab Giovanni Tumminia, Marco Ferraro, Vincenzo Antonucci, Integration of building simulation and life cycle assessment: a TRNSYS application, Energy Procedia 101 (2016) 360 – 367, https://doi.org/10.1016/j.egypro.2016.11.046.

[63] Federico Bava, Simon Furbo, Development and validation of a detailed TRNSYS-Matlab model for large solar collector fields for district heating applications, Energy 135 (2017) 698-708, https://doi.org/10.1016/j.energy.2017.06.146

[64] Saleh Seyedzadeha, Farzad Pour Rahimian, Parag Rastogi, Ivan Glesk, Tuning machine learning models for prediction of building energy loads, Sustainable Cities and Society 47 (2019) 101484, https://doi.org/10.1016/j.scs.2019.101484

[65] Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org

[66] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[67] Ciulla, G., D'Amico, A. , Lo Brano, V., Traverso, M., Application of optimized artificial intelligence algorithm to evaluate the heating energy demand of non-residential buildings at European level, Energy, Volume 176, 1 June 2019, Pages 380-391, https://doi.org/10.1016/j.energy.2019.03.168

[68] Al-Shammari, E.T., Keivani, A., Shamshirband, S., Mostafaeipour, A., Yee, P.L., Petković, D., Sudheer Ch, Prediction of heat load in district heating systems by Support Vector Machine with Firefly searching algorithm, Energy, Volume 95, 15 January 2016, Pages 266-273

[69] Ahmad, M. W., Mourshed, M., Rezgui. Y., Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy and Buildings, Volume 147, 15 July 2017, Pages 77-89, https://doi.org/10.1016/j.enbuild.2017.04.038

[70] Chuan Li, Ying Tao, Wengang Ao, Shuai Yang, Yun Bai, Improving forecasting accuracy of daily enterprise electricity consumption using a random forest based on ensemble empirical mode decomposition, Energy 165 (2018) 1220-1227, 2018.

[71] Geysen, D., De Somer, O., Johansson, C., Brage, J., Vanhoudt, D., Operational thermal load forecasting in district heating networks using machine learning and expert advice, Energy and Buildings, Volume 162, 1 March 2018, Pages 144-153, https://doi.org/10.1016/j.enbuild.2017.12.042

[72] Saloux, E., Candanedo, J. A., Forecasting District Heating Demand using Machine Learning Algorithms, Energy Procedia, Volume 149, September 2018, Pages 59-68, https://doi.org/10.1016/j.egypro.2018.08.169

[73] Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., Ahrentzen, S., Random Forest based hourly building energy prediction, Energy and Buildings, Volume 171, 15 July 2018, Pages 11-25, https://doi.org/10.1016/j.enbuild.2018.04.008

[74] RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

[75] Dave Olsthoorn, Fariborz Haghighat, Parham A. Mirzaei, Integration of storage and renewable energy into district heating systems: A review of modelling and optimization, Solar Energy 136 (2016) 49–64.

[76] Markridakis, Wheelwright, Hyndman, Forecasting – Methods and Application, 3rd edition, Wiley, 1998.

[77] Alpaydin, Introduction to Machine Learning, 3rd edition, MIT Press, 2014.

[78] Mongomery, Runger, Applied Statistics and Probability for Engineers, Wiley, 2014.

[79] James, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R, 7th edition, 2017.

[80] Harrel, Regression Modeling Strategies – With Application to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd edition, Springer, 2015.

[81] Kuhn, Johnson, Applied Predictive Modelling, Springer, 2013.

[82] Trevor Hastie, Robert Tibshirani, Jarome Friedman. The Elements of Statistical Learning, Seventh Edition, Springer. 2017.

[83] Jiawei Han and Michelle Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[84] Smola A (1996). “Regression Estimation with Support Vector Learning Machines.” Master’s thesis, Technische Universitet Munchen

[85] Drucker H, Burges C, Kaufman L, Smola A, Vapnik V (1997). “Support Vector Regression Machines.” Advances in Neural Information Processing Systems, pp. 155–161.

[86] Pravilnik o metodologiji za praćenje, mjerenje i verifikaciju uštede energije u neposrednoj potrošnji (NN 77/12)

[87] Frost Jim, 2014, Regression Analysis: An Intuitive Guide, https://statisticsbyjim.com/

[88] Kreyszig Erwin, Advanced Engineering Mathematics, Wiley, 2011.

[89] Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. B, 2001, 63:411–423., https://doi.org/10.1111/1467-9868.00293

[90] Gulanbaier Tuerhong, Seoung Bum Kim, Gower distance-based multivariate control charts for a mixture of continuous and categorical variables, Expert Systems with Applications 41 (2014) 1701–1707, 2014.

[91] Laurens van der Maaten, Geoffrey Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9, 2008.

[92] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[93] Alboukadel Kassambara and Fabian Mundt (2019). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.6. https://CRAN.R-project.org/package=factoextra

[94] EC Guidelines on good practice in cost-effective cost allocation and billing of individual consumption of heating, cooling and domestic hot water in multiapartment and multi-purpose buildings, MBIC (ENER/C3/2013-977), 2016.

[95] Jesse H. Krijthe (2015). Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation, URL: https://github.com/jkrijthe/Rtsne

[96] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[97] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

[98] Breiman, Friedman, Olshen, Stone, Classification and Regression Treesm, dsworth Statistics/Probability, 1st Edition, 1984.

[99] Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

[100] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[101] Stephen Milborrow (2019). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.8. https://CRAN.R-project.org/package=rpart.plot

[102] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret

[103] Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1-17. doi:10.18637/jss.v077.i01

[104] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-2. https://CRAN.R-project.org/package=e1071.

[105] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. URL http://www.jstatsoft.org/v11/i09/

ŽIVOTOPIS

Danica Maljković rođena je u Beogradu 1980. godine. Diplomirala je na Procesno-energetskom smjeru Fakulteta strojarstva i brodogradnje Sveučilišta u Zagrebu 2005. godine s temom rada „Analiza tehničkih potencijala primjene mjera energetske učinkovitosti u vrtićima“. Dobitnica je Rektorove nagrade 2003. i Posebnog priznanja Dekana za izuzetne doprinose za promicanje ugleda Fakulteta 2005. god. Tijekom studija bila dobitnica stipendije Ministarstva znanosti za nadarene studente i stipendije ALSTOM Hrvatska za najbolje studente FSB. Tijekom studija bila je studentska predstavnica Povjerenstva za odnose s javnošću FSB. Godine 2003. je jedna od inicijatora ponovne aktivacije udruge Hrvatska studentska asocijacija strojarskih fakulteta (HSA-SF) u kojoj je obnašala funkciju tajnice studentske udruge od 2003. do 2005. godine.

Godine 2018. završava poslijediplomski specijalistički studij MBA Ekonomije energetskog sektora na Ekonomskom fakultetu Sveučilišta u Rijeci, te je dobitnica nagrade MBA Ekonomija energetskog sektora i HRO CIGRE za najbolji MBA EES završni rad u ak.god. 2018./2019. na temu „Analiza troškova i koristi implementacije individualnog mjerenja u sustavima daljinskog grijanja u Hrvatskoj“.

Od 2005. godine je zaposlena u Odjelu za proizvodnju i transformaciju energije Energetskog instituta Hrvoje Požar u Zagrebu kao istraživač, a od 2010. godine je zamjenica voditeljice Odjela za organizaciju i gospodarenje u energetici u istom Institutu. Sudjelovala i vodila preko 70 projekata u energetskom sektoru za javne i privatne naručitelje, s fokusom na projekte u sustavima daljinskog grijanja, obnovljivih izvora energije i razvoja institucionalnih okvira, kao i projekte financirane iz međunarodnih financijske institucija i fondova EU.

Članica je Hrvatskog energetskog društva.

Radovi objavljeni u časopisima:

Maljkovic, Danica, Modelling Influential Factors of Consumption in Buildings Connected to District Heating Systems, Energies (1996-1073) 12 (2019), 4; https://doi.org/10.3390/en12040586

Balić, Dražen; Maljković, Danica; Lončar, Dražen: Multi-criteria analysis of district heating system operation strategy. Energy Conversion and Management 2017; 144; 414–428; doi: dx.doi.org/10.1016/j.enconman.2017.04.072

Radovi u zbornicima skupova:

Maljković, Danica; Balen, Igor: Modelling the impact of installation of heat cost allocators in DH in Croatia. Skup Healthy Buildings 2017 Europe. Zbornik radova. Paper ID 0275 ISBN: 978-83-7947-232-1.

Maljković, Danica; Budim, Ružica: Forecasting of energy consumption based on simulation and statistical methods. 25. MEĐUNARODNI SIMPOZIJ O GRIJANJU, HLAĐENJU I KLIMATIZACIJI / 25th INTERNATIONAL SYMPOSIUM ON HEATING, REFRIGERATING AND AIR CONDITIONING. Interklima 2019 / Dović, D ; Soldo, V ; Mudrinić, S. (ur.). - Zagreb : Fakultet strojarstva i brodogradnje, Sveučilište u Zagrebu, Proceedings of International Conference Interklima ISSN 2671-0609, 2019.

Balić, Dražen; Maljković, Danica; Lončar, Dražen; Schneider, Daniel Rolph: Selection of optimal generation set-up considering the socio economic impact. Digital Proceedings of 11th Conference on Sustainable Development of Energy, Water and Environment Systems - SDEWES / Ban, Marko et al. (ur.). - Zagreb : International Centre for Sustainable Development of Energy, Water and Environment Systems – SDEWES Centre , 2016. 0642-1-0642-20.

PRILOZI

Prilog A - Kratki pregled osnovnih statističkih metoda

Raspon je razlika između maksimalne i minimalne vrijednosti skupa, te je izražen donjih izrazom:

$$Raspon(x_i) = x_n - x_1$$

U uređenom skupu, 25 % podataka je manje ili jednako gornjoj granici raspona prvog kvartila ili Q1. Za kontinuiranu raspodjelu kvartil je skup vrijednosti od 0 % do 25 % dobivenih iz kumulativne raspodjele vrijednosti ili funkcije. Ako je Q2 medijan čitavog skupa podataka, Q1 je medijan vrijednosti od minimalne do i uključujući vrijednost za Q2.

U uređenom skupu 75 % podataka je manje ili jednako gornjoj granici raspona, treći kvartil ili Q3. Za kontinuiranu distribuciju Q3 je skup vrijednosti od 75 % do 100 % vrijednosti dobivenih iz kumulativne raspodjele ili funkcije.

Razlika između donje i gornje kvartilne vrijednosti, dakle pokriva srednjih 50 % raspodjele. Raspon između kvartila ili međukvartilni raspon (engl. *interquartile range* – IQR) može se dobiti uzimanjem medijana skupa podataka, zatim pronalaženjem medijana gornje i donje polovice skupa. IQR je u tom slučaju razlika između ova dva sekundarna medijana. IQR = UQ-LQ = Q3-Q1. IQR je neparametarska mjera rasipanja i nije pod utjecajem ekstremnih vrijednosti u gornjem i donjem repu uzorka.

Varijanca je prosječna kvadratna razlika vrijednosti u skupu podataka odnjihove populacijske sredine, μ, ili iz uzorka srednje vrijednosti (također poznata kao varijacija uzorka gdje su podaci iz uzorak veće populacije).

$$\sigma^2 = \sum_i \frac{(\mu - x_i)}{N}$$

Varijanca uzorka na temelju uzorka od n elemenata, n-1 označava stupnjeve slobode pridružene procjeni varijance, računa se prema izrazu:

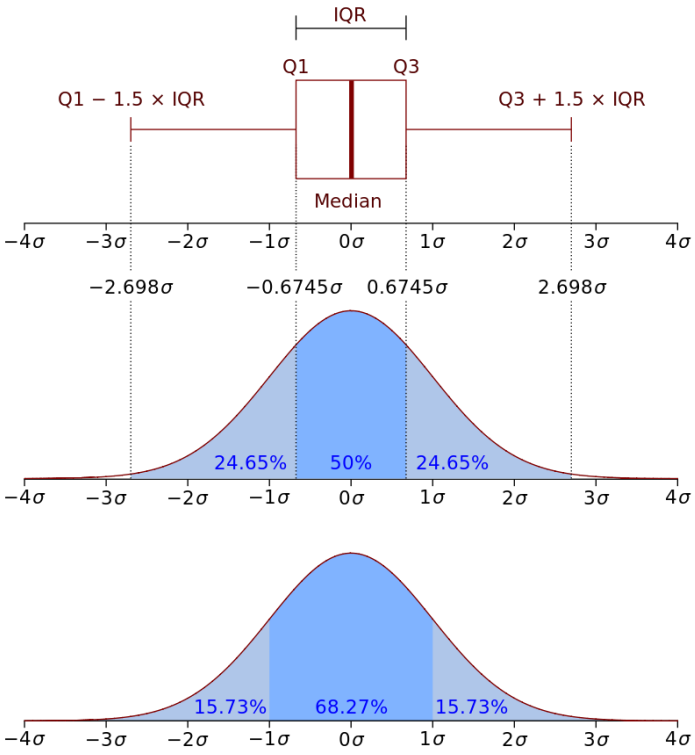
$$s^2 = \frac{\sum_i (\overline{x} - x_i)}{n - 1}$$

Ova mjera je jedna od najšire korištenih mjera disperzije i često se opisuje kao srednje kvadratno odstupanje (engl. *mean squared deviation - MSD*). Iznos varijance može biti izvan opsega izvornih podataka i obično se prilagođava uzimanjem kvadratnog korijena kako bi se rezultirajuća vrijednost svela na mjere podatkovnog skupa, dajući korijen srednjeg standardnog odstupanja (engl. *root mean squared deviation - RMSD*) ili standardnu devijaciju koja je jednaka:

$$\sigma = \sqrt{s^2} = \sqrt{\frac{\sum_i (\overline{x} - x_i)}{n - 1}}$$

Histogrami prikazuju jednodimenzionalnu raspodjelu dijeljenjem na intervalu i broje broj opažanja u svakom od njih. Dok prethodno razmatrane *sumarne mjere* - srednja vrijednost, medijana, standardna devijacija, asimetrija - opisuju samo jedan aspekt numeričke varijable, histogram daje cjelovitu sliku ilustrirajući središte distribucije, varijabilnost, asimetričnost i druge aspekte u jednom prikladnom grafikon.

Pravokuktni dijagrami (engl. *box and whiskers plot* ili *box plot*) su alternativni način prikaza oblika distribucije varijable i sažeti su način ilustriranja standardnih kvantila, oblika i stršećih vrijednosti. Kao što pokazuje generički dijagram, pravokutnik se proteže, s lijeva na desno, od 1. kvartila do 3. kvartila. To znači da sadrži srednju polovicu podataka. Linija unutar kutije je postavljena na medijan. Linije koji izlaze s obje strane kutije protežu se do 1,5 interkvartelnih raspona (IQR) od četvrtine. To obično uključuje većinu podataka izvan kutije. Vrijednosti koje se nalazi izvan *izdanaka, brkova*, predstavljaju stršeće vrijednosti i označene su zasebno pojedinačnim točkama.



Slika 0-1 Pravokutni dijagram i distribucija varijable (Izvor: By Jhguch at en.wikipedia, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=14524285)

Stršeće vrijednosti u podacima mogu *iskriviti* rezultate predviđanja i utjecati na njihovu točnost. Stoga je važno razumjeti jesu li prisutne stršeće vrijednosti i koja su ta opažanja. Stršeće vrijednosti su one vrijednosti koje su veće od Q3 za 1,5*IQR (max gornje = Q3 + 1,5IQR), tj. one koje su manje od vrijednosti Q1 za 1,5*IQR (min donje = Q1 - 1,5IQR).

Ako se povezanosti između X i Y može zapisati kao $Y = f(X) + \epsilon$, gdje je f nepoznata funkcija, a ϵ pogreška, onda se uz pretpostavku da f ima linearni oblik, može zapisati:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Ovdje je β_0 odsječak pravca (odnosno vrijednost Y kada je X=0), a β_1 nagib pravca (prosječni porast Y s jediničnim porastom X). Pogreška se odnosi na sve pogreške koje se javljaju zbog aproksimacije stvarnog modela linearnim: veza X i Y vjerojatno nije linearna, te možda postoje druge varijable koje utječu na Y i možda koeficijenti imaju krivo izračunati iznos. Pretpostavljamo da je pogreška neovisna o X.

U stvarnosti se populacijska regresijska linija i linija najmanjih kvadrata ne poklapaju (čak i kada koristimo isti ulazni uzorak). Regresijska linija opisuje stvarno kretanje, a linija najmanjih kvadrata samo pretpostavlja oblik linije. Usprkos tome, linija najmanjih kvadrata dobro opisuje stvarno kretanje jer uzima uzorak iz određene skupine podataka.

Pri određivanju populacijska srednja vrijednost μ neke varijable *Y*, ta se vrijednost procjenjuje uz pomoć *n* promatranja *Y* ($y_1,\ldots,y_n$). Procjena $\hat{\mu} = \hat{y}$, gdje je $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, predstavlja srednju vrijednost uzorka. Isto tako, nepoznati koeficijenti β_0 i β_1 u linearnoj regresiji definiraju populacijsku regresijsku liniju, i te koeficijente aproksimirano s dobivenih $\hat{\beta}_0$ i $\hat{\beta}_1$ koji definiraju liniju najmanjih kvadrata.

Procjena, donesena na temelju određenog uzorka, predstavlja odstupanja od stvarnog modela linearne regresije. Ako koristimo vrijednost $\hat{\mu}$, dobivenu na uzorku, umjesto stvarne μ očekujemo da u prosjeku neće biti odstupanja između tih vrijednosti, ali $\hat{\mu}$ nekad može precijeniti, a nekad i podcijeniti stvarnu vrijednost μ. Ako pak pri proračunu uzmemo srednju vrijednost $\hat{\mu}$ velikog broja promatranja, onda će ona biti jednaka μ, čemu i težimo. Isto vrijedi i za određivanje β_0 i β_1 uz pomoć srednje vrijednosti velikog broja promatranja.

Kako je naglašeno, $\hat{\mu}$ dobivena na uzorku, može precijeniti i podcijeniti stvarnu vrijednost μ, a nekada, ako u obzir uzmemo veliki broj promatranja, može vrlo točno procijeniti stvarnu μ. Koliko će ta $\hat{\mu}$ odstupati od stvarne μ, pokazuje nam standardna pogreška $\hat{\mu}$ ($SE(\hat{\mu})$):

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n} \qquad \textbf{(68)}$$

Gdje je σ standardna devijacija svakog y_i od *Y*. Standardna pogreška pokazuje nam prosječnu vrijednost kojom se $\hat{\mu}$ razlikuje od μ. Iz jednadžbe (68) vidljivo je da porastom broja promatranja *n*, opada i standardna pogreška $\hat{\mu}$. Kako bismo odredili standardna odstupanja (standardnu pogrešku, engl. *standard error* - SE) $\hat{\beta}_0$ i $\hat{\beta}_1$ od β_0 i β_1 koristimo:

$$SE(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \qquad \textbf{(69)}$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Gdje je $\sigma^2 = Var(\epsilon)$. Da bismo koristili ove formule moramo pretpostaviti da greške ϵ_i svakog pojedinog promatranja nisu povezane sa zajedničkom varijancom σ^2. Standardna pogreška, $SE(\hat{\beta}_0)$, opada što su vrijednosti x_i međusobno više udaljene, dok je $SE(\hat{\beta}_1)$ je jednaka $SP(\hat{\mu})$ kad je $\bar{x}$ jednak 0 (u tom slučaju $\hat{\beta}_0$ bio bi jednak $\bar{y}$). Vrijednost σ^2 nije poznata te se mora odrediti iz danih podataka, a sama procjena σ^2 zove se preostala standardna pogreška (eng *residual standard error* RSE) i dana je jednadžbom $RSE = \sqrt{RSS/(n-2)}$.

Prilog B

		Min	D.kv.	Median	Sr.vr	G.kv.	Maks.
	Jedinica						
SpecTopSDG	**kWh/m^2**	16,96	18,9	19,81	20,03	20,94	23,53
Kat		0					
VrijemeBoravka		veći dio dana					
ToplinskaUgodnost		toplo					
Grupa	**1**						
SpecTopSDG	**kWh/m^2**	16,84	17,88	19,27	18,97	20,36	20,49
Kat		1					
VrijemeBoravka		veći dio dana					
ToplinskaUgodnost		umjereno					
Grupa	**2**						
SpecTopSDG	**kWh/m^2**	01,11,1990	17,4	24,13	22,17	26,72	30,12
Kat		0	1				
VrijemeBoravka		cijeli dan					
ToplinskaUgodnost		toplo	umjereno				
Grupa	**3**						
SpecTopSDG	**kWh/m^2**	7.511	8.774	17.807	14,695	18,463	20,917
Kat		1					
VrijemeBoravka		veći dio dana					
ToplinskaUgodnost		toplo					
Grupa	**4**						
SpecTopSDG	**kWh/m^2**	01,10,1953	18,07,2020	19,85	21,57	25,42	39,28
Kat		3					
VrijemeBoravka		veći dio dana					
ToplinskaUgodnost		toplo					
Grupa	**5**	5	5	5	5	5	5
SpecTopSDG	**kWh/m^2**	13,6	13,77	17,18	18,02	21,44	24,13
Kat		4					
VrijemeBoravka		veći dio dana					
ToplinskaUgodnost		toplo					
Grupa	**6**	6	6	6	6	6	6
SpecTopSDG	**kWh/m^2**	8.045	10.404	12.239	12,153	13,988	16,089
Kat		3					
VrijemeBoravka		rijetko					
ToplinskaUgodnost		umjereno					
Grupa	**7**	7	7	7	7	7	7
SpecTopSDG	**kWh/m^2**	12,10,2020	01,12,2029	18,42	17,24	19,39	24,03
Kat		4					

VrijemeBoravka		veći dio dana	cijeli dan				
ToplinskaUgodnost		umjereno					
Grupa	**8**	8	8	8	8	8	8
SpecTopSDG	**kWh/m^2**	23,27	24,88	25,43	25,45	25,99	27,67
Kat		0					
VrijemeBoravka		veći dio dana					
ToplinskaUgodnost		toplo					
Grupa	**9**	9	9	9	9	9	9
SpecTopSDG	**kWh/m^2**	01,11,1963	13,5	14,84	14,09	15,01	15,45
Kat		4					
VrijemeBoravka		rijetko					
ToplinskaUgodnost		umjereno					
Grupa	**10**	10	10	10	10	10	10
SpecTopSDG	**kWh/m^2**	3,741	7,788	12.344	11,149	15,706	16,168
Kat		2					
VrijemeBoravka		veći dio dana					
ToplinskaUgodnost		umjereno					
Grupa	**11**						
SpecTopSDG	**kWh/m^2**	01,11,2018	13,29	16,95	17,85	21,51	26,33
Kat		3					
VrijemeBoravka		veći dio dana					
ToplinskaUgodnost		umjereno					
Grupa	**12**	12	12	12	12	12	12