**VILNIAUS UNIVERSITETAS**

**MATEMATIKOS IR INFORMATIKOS FAKULTETAS**

**DUOMENŲ MOKSLO STUDIJŲ PROGRAMA**

## Duomenų mokslo projektas - kursinis darbas

# SAULĖS ELEKTRINIŲ PAGAMINAMOS ELEKTROS KIEKIO PROGNOZAVIMAS

Darbo atliko: Matas Amšiejus,
Antanas Užpelkis
Darbo vadovė: Doc. Dr. Jurgita Markevičiūtė

**VILNIUS 2022**

# Turinys

# Lentelių sąrašas

# Iliustracijų sąrašas

# Saulės elektrinių pagaminamos elektros kiekio prognozavimas


**Santrauka**

Vienas iš geriausių atsinaujinančios energijos išteklių - saulės elektrinės. Norint išgauti optimalų kiekį energijos reikia žinoti, nuo ko priklauso jėgainių efektyvumas. Šiame darbe naudojame Šiaulių apskrityje esančių saulės elektrinių duomenis, kuriuos sujungiame su atitinkamų dienų orais. Priklausomas kintamasis buvo pagaminamos elektros kiekis, nepriklausomi - infraraudonieji spinduliai, saulės spinduliuotė, mėnuo, dienos ilgis (h), vidutinė paros temperatūra (°C), krituliai (mm), vidutinis vėjo greitis (km/h), vidutinis slėgis (hPa). Sudarėme keturis skirtingus modelius: modelis su temperatūra, krituliais, spinduliuote ir mėnesiais, modelis tik su meteorologiniais ir dienos ilgio kintamaisiais, modelis tik su kolektoriuje surinktais kintamaisiais ir mažiausiai multikolinearus modelis. Visi modeliai buvo didelio tikslumo ($R^2 > 89$). Nustatyta, kad svarbiausias modelio kintamasis buvo saulės spinduliuotė, o modelyje, kur nebuvo saulės spinduliuotės - dienos ilgis.



**Raktiniai žodžiai :** saulės kolektoriai; saulės elektrinės; energija; tiesinė regresija.


# Forecasting the amount of electricity produced by solar power plants


**Abstract**

One of the best sources of renewable energy is solar power plants. In order to produce the optimal amount of energy, it is necessary to know what the efficiency of power plants depends on. In this work, we use the data from solar power plants located in Šiauliai County, which we combined with the weather of the respective days. Dependent variable - amount of electricity produced, independent - infrared rays, solar irradiation, month, day length (h), average daily temperature (°C), precipitation (mm), average wind speed (km/h), average pressure (hPa). We made four different models: model with temperature, precipitation, irradiation and months, model with meteorological and day length variables only, model with collector variables only and least multicollinear model. All models performed well ($R^2 > 89$). The most important variable in the model was found to be solar irradiation, and in the model without irradiation - day length.



**Key words :** solar panels, solar power plant, linear regression.

# 1 Įvadas

Pasaulyje vyksta aktyvi klimato kaita [1], didėja elektros kaina [2]. Elektros energija taip gali būti priklausoma nuo geopolitinės situacijos. Šių problemų sprendimas galėtų būti atsinaujinanti energetika. Saulės (taip pat ir kitos) elektrinės nepriklauso nuo tarptautinių santykių ar esamos politinės situacijos bei yra tvarios, tai yra, prisideda prie klimato kaitos mažinimo. Taigi, atsinaujinanti energetika yra svarbi tiek klimato, tiek politiniame kontekste. Norint maksimizuoti saulės elektrinių pagaminamos elektros kiekį reikia atsižvelgti į meteorologines bei kitas sąlygas, kurios gali turėti įtakos elektros kiekio pagaminimui. Šiame darbe sudarysime matematinį modelį, kuris prognozuotų saulės elektrinių pagaminamos elektros kiekį. Taip pat ieškosime, kokie reiškiniai turi didžiausią įtaką elektros pagaminimui.

**Tikslas:** ištirti ir prognozuoti saulės elektrinių pagaminamos elektros kiekį, priklausomai nuo meteorologinių (bei kitų) duomenų, naudojant regresijos metodus.

**Uždaviniai:**

- Pradinis duomenų apdorojimas ir analizė;
- Regresijos modelių sudarymas;
- Modelių įvertinimas.

# 2 Literatūros apžvalga

Saulės jėgainių, atsinaujinančios energijos temos yra plačiai nagrinėjamos. Mus domino energijos modeliavimas, tam naudoti metodai ir pasirinkti kintamieji. T. Chuluunsaikhan, A. Nasridinov, W. S. Choi, D. B. Choi, S. H. Choi, ir Y. M. Kim straipsnyje [3] autoriai naudojo šešis skirtingus mašininio mokymosi metodus: tiesinę regresiją, kNN (artimiausio k kaimyno metodas), SVR (atraminių vektorių regresija), MLP (daugiasluoksnis perceptronas), RF (atsitiktinių miškų metodas), GB (padidinto gradiento metodas). Kintamieji buvo surinkti iš kolektoriaus (galios faktorius (pagamintos ir sunaudotos energijos santykis), nuokrypio kampas (kolektoriaus nukrypimas nuo horizontalaus paviršiaus), spinduliuotė patenkanti ant paviršiaus, kolektoriaus temperatūra), meteorologinius duomenis (drėgmė, giedrumas, spinduliuotė patenkanti ant žemės, debesuotumas, temperatūra) ir oro taršos duomenys (ozono, sieros rūgšties dujų, azoto dioksido, smalkių, smulkių dulkių, smulkių kietųjų dalelių kiekis). Visi metodai taikyti tyrime buvo tinkami (apie 95 % tikslumas), be to, oro taršos duomenys buvo nereikšmingi.

T. Verma, A. Tiwana, C. Reddy, V. Arora, ir P. Devanand straipsnyje [4] buvo atkreipta ir labiau įsigilinta į skirtingus regresijos modelius ir kintamųjų reikšmingumą (pagaminamos elektros kiekis, kolektoriaus temperatūra, debesuotumas, vėjo greitis, drėgmė, krituliai, saulės pakilimo ir azimuto kampai). Naudojant tiesinę regresiją vėjo greitis ir drėgmė buvo nereikšmingi,naudojant logaritminę regresiją debesuotumas buvo nereikšmingas, o naudojant polinominę regresiją visi pirmo ir trečio laipsnio kintamieji bei ketvirto laipsnio krituliai buvo nereikšmingi. Mažiausiai tiksli buvo logaritminė regresija, o tiesinė ir polinominė regresijos buvo panašaus tikslumo ($R^2$ $74,4$ % ir $75,1$ %). Tačiau tiksliausias buvo dirbtinių neuroninių tinklų metodas ($R^2 = 92$ %).

Y. S. Kim, H. Y. Joo, J. W. Kim, S. Y. Jeong, ir J. H. Moon straipsnyje [5] buvo kuriamas tiesinės regresijos modelis kiekvienam mėnesiui ir į jį buvo įtrauktas kintamasis dienos ilgis. Kiekvieno mėnesio modelių tikslumai buvo vidutiniškai didesni už bendrą (visų metų) modelį. Taip pat, reikšmingiausios kovariantės buvo spinduliuotė ir dienos ilgis.

Taigi, apibendrinus šaltinius, nusprendėme imti kolektoriaus duomenis (pagamintos elektros kiekis, saulės spinduliuotė ir infraraudonųjų spindulių spinduliuotė (angl. IR)), meteorologinius duomenis (vidutinė dienos temperatūra, krituliai, vidutinis vėjo greitis, vidutinis slėgis) bei dienos ilgumą. O modeliui kurti naudosime tiesinę regresiją.

# 3 Pirminė duomenų analizė

Tyrime naudojamus duomenis sudarėme iš trijų šaltinių:

- Saulės elektrinės, esančios Pakruojo rajone, duomenys;[1]
- Meteorologiniai duomenys (kadangi nepavyko gauti Pakruojo rajono duomenų, ėmėme Šiaulių apskrities duomenis)[2];
- Dienos ilgumas[3].

Kadangi saulės elektrinės duomenys buvo nuo 2019-09-07 iki 2021-03-21, dėl to ir nusprendėme tirti šį laikotarpį. Viso turėjome informaciją apie 562 dienas. Kolektorių duomenys buvo pateikti kas valandą, tačiau kadangi mes negalėjome gauti kasvalandinių meteorologinių duomenų, dėl to iš valandinių sumuodami pakeitėme į duomenis parai. Pagamintos elektros kiekio ir spinduliuotės duomenys buvo suminiai, tai yra kaip skaitiklio duomenys, dėl to mes juos atstatėme į per dieną pagaminamos elektros kiekį rasdami skirtumus tarp dienų. Konvertuojant pastebėjome, kad kolektoriaus duomenyse buvo klaidų, pavyzdžiui neigiamas pagaminamos elektros kiekis per dieną, itin didelis pagaminamos elektros kiekis. Viso buvo aštuonios klaidos (4 klaidingos įvestys, bet kadangi duomenys buvo skirtumai, tai kita diena irgi buvo netiksli). Tad teko atsisakyti šių dienų - keturias iš jų paaiškinome žiemos laiko įvedimu (2019-10-27; 2019-10-28; 2020-10-25; 2020-10-26), o keturios liko nepaaiškintos (2020-05-29; 2020-05-30; 2020-06-01; 2020-06-02). Viso galutiniame duomenų rinkinyje turėjome 554-ių dienų duomenis.
Duomenų rinkinio kintamieji:

- saulės spinduliuotė (normuoti);
- IR (normuoti);
- pagaminamos elektros kiekis (normuoti);
- data (ymd formatu);
- vidutinė temperatūra (°C);
- krituliai (mm);
- vidutinis vėjo greitis (km/h);
- vidutinis slėgis (hPa);
- dienos ilgumas (h).

[1]Duomenys iš privačios įmonės.

[2]Nuoroda į meteorologinių duomenų šalitinį: `https://meteostat.net/en/place/lt/pakruojis?s=26524&t=2019-09-07/2021-03-21`

[3]Nuoroda į dienos ilgumo duomenis:`https://sunrise.maplogs.com/_iauli_miesto_savivaldyb_iauli_m_sav_lithuania.209371.html?year=2021`

1 lentelė: Kiekybinių kintamųjų aprašomoji statistika

| | vidurkis | s. nuok. | min | max | Q1 | Q2 | Q3 |
|---|---|---|---|---|---|---|---|
| Temp. | 6, 871 | 7, 606 | −19, 3 | 24, 2 | 2, 125 | 5, 5 | 12, 375 |
| Krituliai | 1, 27 | 2, 675 | 0 | 31, 2 | 0 | 0 | 1, 6 |
| Vėjo g. | 10, 22 | 4, 679 | 1, 6 | 26, 8 | 6, 7 | 10 | 13, 3 |
| Slėgis | 1013, 797 | 9, 373 | 982, 8 | 1042, 1 | 1008, 525 | 1014, 6 | 1020, 075 |
| D. ilgis | 11, 269 | 3, 398 | 6, 959 | 17, 618 | 8, 164 | 10, 716 | 13, 877 |
| Spind. | 0, 042 | 0, 042 | 0 | 0, 154 | 0, 007 | 0, 025 | 0, 071 |
| IR | 8, 865 | 2, 917 | 0, 667 | 16, 572 | 6, 678 | 8, 199 | 10, 829 |
| E. kiekis | 0, 042 | 0, 04 | 0 | 0, 142 | 0, 006 | 0, 028 | 0, 073 |

Iš 1 lentelės matome, kad stebimuoju laikotarpiu slėgis buvo normalus, vidutinė dienos temperatūra svyravo nuo -19 iki 24 laipsnių °C, taip pat įdomu pastebėti, kad nors daugiausia kritulių iškrito 31 mm, tačiau trečio kvartilio reikšmė rodo, kad 75 % visų dienų iškrito mažiau nei 1,6 mm kritulių, vadinasi stebimi metai nebuvo išskirtinai lietingi. Reikia pastebėti, kad dėl sukauptinių sumų atstatymo skirtumais, energijos kiekio ir spinduliuotės matavimo skalė smarkiai sumažėjo. Kadangi infraraudonoji spinduliuotė nebuvo sumuojama, tai kai kuriose vietose jos reikšmės dėl sumavimo pagal paras tapo didesnės už 1. Iš kintamųjų aprašomosios statistikos matome, kad stebimas laikotarpis buvo gan įprastas, nebuvo daug meteorologinių stichijų ar kokių kitų neįprastų reiškinių.

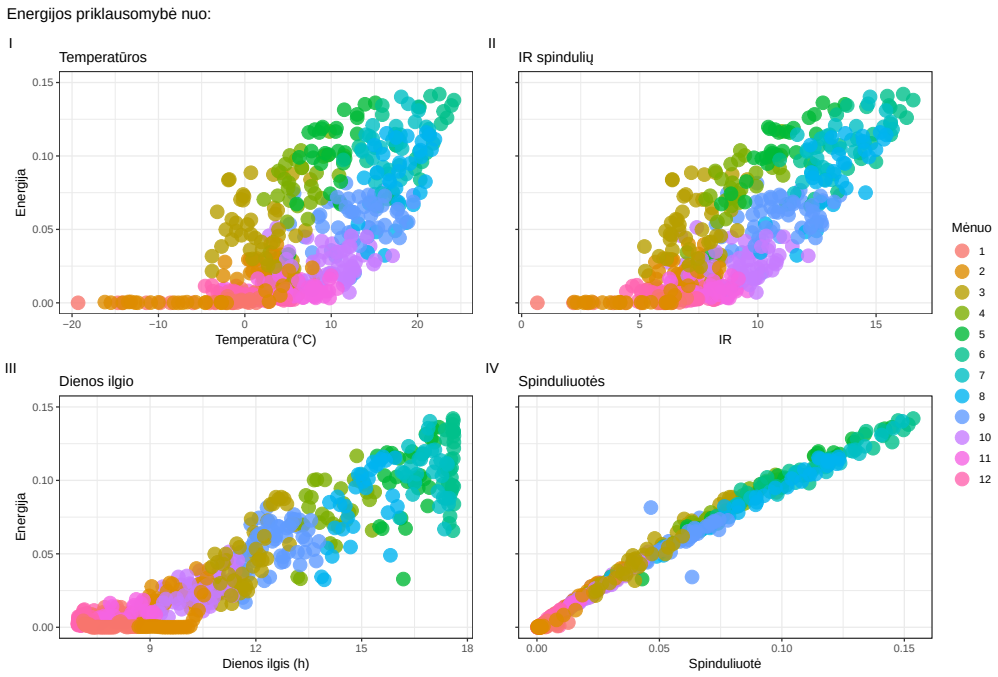
Norint nustatyti kintamųjų priklausomybę, paprasčiausias būdas yra


1 pav.: kiekybinių kintamųjų koreliacijų matrica

pasižiūrėti į jų koreliacijų matricą (1). Iš jos matome, kad pagaminamos elektros kiekis stipriai koreliuoja su saulės spinduliuote (0,99), dienos ilgiu (0,93), IR (0,82), vidutiniškai su su temperatūra (0,69). Taip pat matome, kad yra silpna neigiama koreliacija tarp pagaminamos elektros kiekio ir vėjo greičio, o koreliacija su slėgiu yra labai silpna, bei su krituliais jos visai nesimato. Be to, matome, kad krituliai su visais kitais kintamaisiais labai silpnai koreliuoja, stipriausia koreliacija yra -0,34 su slėgiu.

Žiūrint į koreliacijas taip pat įtariame, kad bus multikolinearumo problema, nes kintamieji, kurie stipriai koreliuoja su pagaminamos elektros kiekiu taip pat stipriai koreliuoja ir tarpusavyje: temperatūra su IR (0,96), dienos ilgis su spinduliuote (0,93), IR su spinduliuote (0,83), dienos ilgis su IR (0,81), temperatūra su spinduliuote (0,71), temperatūra su dienos ilgiu (0,7).

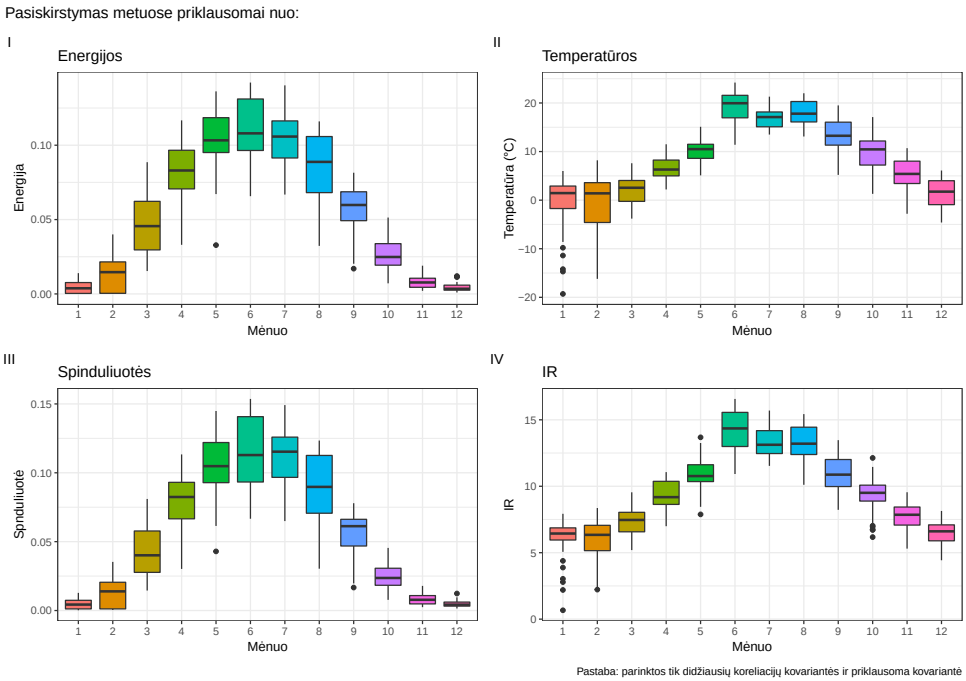
Kintamųjų priklausomybes patikriname vizualiai, pavaizdavus kintamųjų sklaidos diagramas:




2 pav.: pagaminamos elektros kiekio priklausomybė.

Taigi, tiek iš koreliacijų matricos 1, tiek iš sklaidos diagramų 2 įtariame, kad regresijos modeliuose reikšminga bus temperatūra, IR, dienos ilgis ir spinduliuotė, o krituliai (4), slėgis (5) ir vėjo greitis (6) - nereikšmingi.

Norint įsitikinti, kad daugiausiai energijos pagaminama šiltuoju metų laikotarpiu, nusibrėžiame kiekvieno mėnesio stačiakampes diagramas kiekvienam iš dominančių kintamųjų:




3 pav.: kintamųjų pasiskirstymas metuose

Taigi, iš grafikų 3 patvirtiname hipotezę, kad daugiausiai elektros sugeneruojama šiltuoju metų laikotarpiu (nuo gegužės iki liepos) (7). Šiuo laikotarpiu dienos yra ilgesnės, suintensyvėja spinduliuotė ir IR, kas lemia ir didesnę temperatūrą.

# 4 Modelių sudarymas

Šiame tyrime modeliavimui naudosime tiesinės regresijos modelius. Prieš tai atlikta duomenų analizė patvirtino, kad egzistuoja tiesiniai sąryšiai tarp pagaminamos energijos kiekio ir dienos ilgumo, temperatūros, infraraudonųjų spindulių ir saulės spinduliuotės. Taigi, pradiniame regresijos modelyje priklausoma kovariantė ($Y$) bus pagaminamos energijos kiekis, o nepriklausomos kovariantės ($X_i$) bus temperatūra, krituliai, vėjo greitis, slėgis, šviesiojo paros meto ilgumas, spinduliuotė, patenkanti ant kolektoriaus, infraraudonųjų spindulių kiekis bei mėnuo. Atitinkamai, elektrinių pagaminamos energijos kiekio įvertis bus suskaičiuojamas pagal formulę:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + e_i,$$

kur $k$ - duomenų dimensija, $i$ - i-tasis duomenų įrašas, $e_i$ - nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai, $e_i \sim N(0, \sigma^2)$. $\beta$ koeficientai galės parodyti, kokią įtaką kiekvienas faktorius turi priklausomam kintamajam (teigiamą ar neigiamą), o jį standartizavus galėsime nuspręsti, kurie faktoriai labiausiai keitė pagaminamos elektros kiekį. Tyrime naudosime reikšmingumo lygmenį $\alpha = 0,05$.

Pirma sudarome modelį įtraukdami visas kovariantes. Patikrinus standartizuotų paklaidų ir Kuko mato grafikus (9, 10) pastebėjome, kad turime dvi labiau išsiskiriančias reikšmes, kurios gali iškraipyti modelio įvertinimą, tad jas šaliname (vieną dieną pagamino išskirtinai nedaug elektros, nors, lyginant su panašiais parametrais, energijos turėjo būti pagaminta daugiau. Kita diena automatiškai tapo išskirtis dėl sukauptųjų duomenų atstatymo).

Toliau, tikriname ar liekanos yra pasiskirsčiusios pagal normalųjį skirstinį naudojant Šapiro-Vilko testą:

$$\begin{cases} H_0 : \text{liekanos pasiskirsčiusios pagal normalųjį skirstinį.} \\ H_1 : \text{liekanos nėra pasiskirsčiusios pagal normalųjį skirstinį.} \end{cases}$$

Kadangi gauname, kad p-reikšmė $= 3,216 \cdot 10^{-10} < 0,05 = \alpha$, dėl to $H_0$ atmetama, vadinasi sąlyga nėra tenkinama. Tą patį patvirtina ir kvantilių grafikai (8). Problemai išspręsti naudojame Bokso-Kokso transformaciją priklausomam kintamajam:

$$Y^* = \frac{Y^{\lambda-1}}{\lambda}, \qquad \text{čia } \lambda = 0,74.$$

Taip pat transformuojame kovariantę IR ištraukdami šaknį.

Dar kartą atlikus anksčiau naudotą testą, modeliui su transformuotais kintamaisiais, gauname, kad p-reikšmė $= 0,0592 > 0,05 = \alpha$, dėl to neatmetame $H_0$, vadinasi, liekanos yra pasiskirsčiusios normaliai (tai patvirtina

ir kvantilių grafikas (11)).

Tiesinės regresijos modeliuose liekanos $e_i$ turi tenkinti homoskedastiškumo prielaidą - jų dispersijos turi būti vienodos. Naudodami Breušo ir Pagano kriterijų nustatėme, kad paklaidos yra heteroskedastiškos. Dėl to tolimesniame tyrime naudosime koreguotas standartines paklaidas HC1 metodu.

Toliau atliekame nereikšmingų kovariančių šalinimą iš modelio. Atlikus pažingsninę regresiją gauname, kad nereikšmingos yra slėgio ir vėjo greičio kovariantės. Gauto modelio determinacijos koeficientas $R^2 = 0,9862$.

Nors pastarasis modelis turi didelę $R^2$ reikšmę, jo interpretuoti nederėtų, nes jame yra kelios multikolinearios kovariantės. Labiausiai tai pasireiškė tarp dienos šviesos ilgio, temperatūros, spinduliuotės, $\sqrt{IR}$ ir mėnesio. Nors matėme, kad pagaminamos energijos kiekis ir infraraudonieji spinduliai turi teigiamą sąryšį, mūsų modelyje koeficientas prie šios kovariantės tapo neigiamas. Taip nutiko dėl multikolinearumo, tad tą sutvarkėme sukurdami kelis modelius pretendentus:

- M1 - modelis su temperatūra, krituliais, spinduliuote ir mėnesiais. Šiame modelyje atmetėme $\sqrt{IR}$ ir dienos ilgį, kadangi šios kovariantės buvo labiausiai multikolinearios. 2 lentelėje matome šio ir kitų modelių koeficientus. Taip pat tik šiame modelyje buvo įtraukta mėnesio kovariantė (3 lentelė). Tačiau matome, kad koeficientai neatitinka pradinės duomenų analizės (vasarą mažiau didėja pagaminamos energijos skaičius nei, pavyzdžiui, kovą). Tai indikuoja potencialų multikolinearumą, todėl tolimesniuose modeliuose mėnesio kovariantės nebenaudosime. Taip pat iš 4 lentelės[4] matome, kad reikšmingiausia kovariantė yra patenkantis šviesos kiekis ant kolektoriaus. Toliau seka temperatūra ir krituliai, kurie yra žymiai mažiau reikšmingi. Reikėtų pastebėti, kad regresijos modeliuose gauname neigiamą koeficientą prie kritulių kovariantės, nors iš koreliacijų matricos turėjome teigiamą sąryšį tarp šios kovariantės ir pagaminamos elektros kiekio. Tačiau taip gali būti, nes patikrinus, ši koreliacija nebuvo statistiškai reikšminga;

- M2 - modelis su temperatūra, krituliais ir dienos ilgiu (kovariantės, kurių reikšmes galima lengvai sužinoti pagal orų prognozes arba kalendorių). Šis modelis būtų realus pretendentas bandyti nuspėti, kiek bus pagaminama energijos per artimiausias dienas. Iš 4 lentelės matome, kad reikšmingiausia kovariantė yra dienos ilgis, toliau seka krituliai ir vidutinė dienos temperatūra. Dienos ilgiui ir temperatūrai didėjant, didėja ir pagaminamos elektros kiekis. Lyjant - mažėja;

- M3 - modelis su spinduliuote ir $\sqrt{\text{IR}}$ (kovariantės, kurių reikšmės gaunamos iš kolektoriaus). 4 lentelėje gauname, kad spinduliuotė yra svarbiausia kovariantė, o $\sqrt{IR}$ - kur kas mažiau svarbi;

[4]standartizuoti koeficientai randami pagal formulę: $\beta^* = \frac{S_x}{S_y}\beta$, čia $S_x$-priklausomo kintamojo standartinis nuokrypis, $S_y$-nepriklausomo kintamojo standartinis nuokrypis

- M4 - modelis su temperatūra, krituliais ir spinduliuote (mažiausiai multikolinearus modelis). Šis modelis yra labiausiai tinkamas koeficientų interpretacijai. Pagal M4, svarbiausia kovariantė yra spinduliuotė, tada temperatūra ir krituliai. Pastarieji yra labai nežymūs lyginant su spinduliuote.

2 lentelė: Skaitinių kovariančių koeficientai

| Modelis | Laisv. narys | Temp. | Krituliai | D. ilgis | Spind. | $\sqrt{IR}$ |
|---|---|---|---|---|---|---|
| M1 | -1,3382 | 0,0006 | -0,0004 | * | 1,9231 | * |
| M2 | -1,5068 | 0,0012 | -0,0035 | 0,0239 | * | * |
| M3 | -1,3444 | * | * | * | 2,1367 | 0,0068 |
| M4 | -1,3274 | 0,0005 | -0,0003 | * | 2,1448 | * |

3 lentelė: M1 modelio mėnesio pseudokintamųjų koeficientai (lyginant su sausiu)

| Vas. | Kov. | Bal. | Geg. | Birž. | Liep. | Rugpj. | Rugs. | Spal. | Lapkr. | Gruod. |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.014 | 0.0413 | 0.0422 | 0.0316 | 0.0212 | 0.0187 | 0.0234 | 0.032 | 0.024 | 0.0063 | 0.0008 |

4 lentelė: Skaitinių kovariančių standartizuoti koeficientai

| Modelis | Temp. | Krituliai | D. ilgis | Spind. | $\sqrt{IR}$ |
|---|---|---|---|---|---|
| M1 | 0,0501 | -0,0110 | * | 0,8594 | * |
| M2 | 0,0984 | -0,1050 | 0,8840 | * | * |
| M3 | * | * | * | 0,9549 | 0,0370 |
| M4 | 0,0377 | -0,0086 | * | 0,9585 | * |

Taigi visuose modeliuose, kuriuose yra spinduliuotės kovariantė, ji yra reikšmingiausia. Tuo tarpu modeliai, kuriuose yra kritulių kovariantė, ji turi mažiausią įtaką prognozei. Visuose modeliuose spinduliuotė yra bent 80 kartų svarbesnė už kritulius ir 35 kartus - už temperatūrą. M2 modelyje dienos ilgis buvo apie 9 kartus svarbesnis už kritulius ir temperatūrą.

# 5 Modelių vertinimas

Modeliams vertinti naudojame išlaikymo (angl. hold-out) metodą. Išlaikymo metodu duomenys yra skirstomi į mokymo ir testavimo aibes, kur mokymo aibė yra naudojama sudaryti modeliui, o testavimo - jam įvertinti. Mokymo aibę sudaro 80 % visų duomenų (443 stebėjimai), o testavimo - 20 % (111 stebėjimai). Modelius vertiname naudodami penkis tikslumo matus:

- MAE - vidutinė absoliutinė paklaida (angl. *mean square error*):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,;$$

- RMSE - šaknis iš vidutinės kvadratinės paklaidos (angl. *root mean square error*):

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2};$$

- NMAE - normuota MAE (angl. *normalized MAE*):

$$NMAE = \frac{MAE}{\max(y) - \min(y)};$$

- NRMSE - normuota RMSE (angl. *normalized RMSE*):

$$NRMSE = \frac{RMSE}{\max(y) - \min(y)};$$

- $R^2$ - determinacijos koeficientas (pastaba: šiame darbe jis randamas naudojant mokymo duomenis):

$$R^2 = 1 - \frac{S_e^2}{S_y^2}$$

Čia $n$ - duomenų aibės dydis, $y_i$ - tikroji (stebėtoji) reikšmė, $\hat{y}_i$ - prognozuota reikšmė, $S_e^2$ - liekanų dispersija, $S_y^2$ - priklausomo kintamojo (mūsų atveju pagaminamos elektros kiekio) dispersija.

$R^2$ nurodo kiek procentų priklausomo kintamojo elgesio nusako nepriklausomi kintamieji, taigi, kuo ši reikšmė didesnė (didžiausia galima reikšmė 1, mažiausia 0) tuo modelis yra tikslesnis. Kiti tikslumo matai nurodo paklaidas, vadinasi atvirkščiai nei $R^2$, kuo reikšmės mažesnis tuo modelis yra geresnis. Kadangi duomenyse pagaminamos elektros kiekis buvo transformuotas, dėl to MAE ir RMSE interpretacija yra sunkesnė. Dėl šios priežasties naudojame ir normuotas MAE ir RMSE, kurios nusako ne konkrečią paklaidą,

bet paklaidos procentinę išraišką, tai yra, keliais procentais prognozė suklydo.

Kiekvieną sudarytą modelį vertiname šiais metodais ir gauname rezultatų lentelę:

5 lentelė: Modelių tikslumo rezultatai

| Modelis | MAE | RMSE | NMAE | NRMSE | $R^2$ |
|---|---|---|---|---|---|
| M1 | 0,0095 | 0,0118 | 0,0299 | 0,037 | 0,9855 |
| M2 | 0,0225 | 0,0288 | 0,0707 | 0,0905 | 0,8955 |
| M3 | 0,0149 | 0,0174 | 0,0468 | 0,0547 | 0,9692 |
| M4 | 0,0147 | 0,0173 | 0,0463 | 0,0544 | 0,9694 |

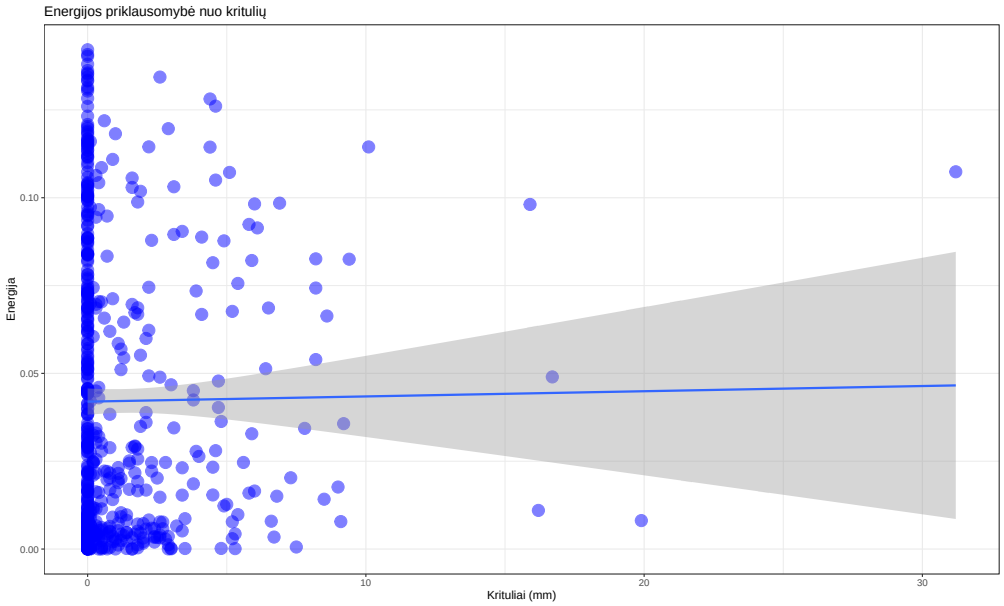
Taigi iš 5 lentelės, matome, kad visi modeliai yra labai geri (net prasčiausio modelio $R^2 = 0,8955$). Geriausias modelis (mažiausios paklaidos ir didžiausias $R^2$) yra M1, tai yra su temperatūra, krituliais, spinduliuote ir mėnesiais. Prasčiausias modelis yra M2, tai yra su temperatūra, krituliais ir dienos ilgiu. Modelis M3 ir M4 yra labai panašūs ir prognozavimo tikslumo matai skiriasi tik per ketvirtą skaičių po kablelio.

# 6 Išvados bei rekomendacijos

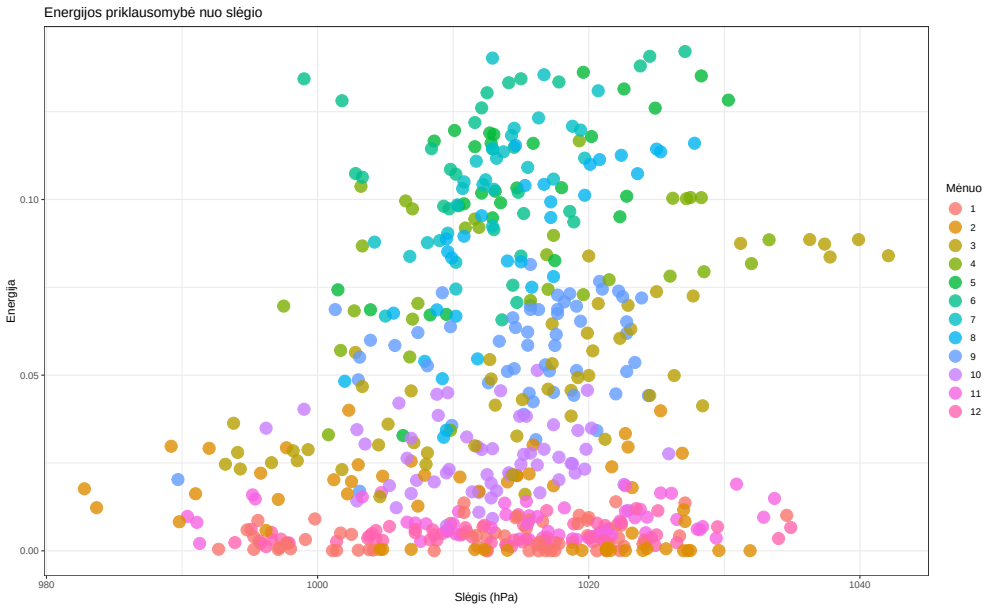
Saulės elektrinių pagaminamos elektros kiekiui prognozuoti sudarėme keturis tiesinės regresijos modelius naudodami duomenis, surinktus iš kolektoriaus (pagaminamos elektros kiekis, saulės spinduliuotė, infraraudonoji spinduliuotė, data) ir duomenis, rastus kituose šaltiniuose (slėgis, krituliai, temperatūra, vėjo greitis, dienos ilgumas). Slėgis ir vėjo greitis visuose modeliuose buvo nereikšmingi. Siekiant patenkinti modelių prielaidas buvo naudojama Bokso-Kokso transformacija energijos kiekiui ir ištraukta šaknis iš infraraudonosios spinduliuotės. Modelių liekanos buvo heteroskedastiškos, dėl to buvo naudojamas HC1 paklaidų korekcijos metodas. Visi sudaryti modeliai buvo tikslūs ($R^2 \geq 0,8955$). Geriausius rezultatus parodė modelis su temperatūra, krituliais, spinduliuote ir mėnesiais ($R^2 = 0,9855, MAE = 0,0095, RMSE = 0,0118$), o prasčiausias modelis buvo sudarytas iš temperatūros, kritulių ir dienos ilgio (duomenis, kuriuos galima žinoti iš anksto) ($R^2 = 0,8955, MAE = 0,0225, RMSE = 0,0288$). Reikšmingiausia modelio kovariantė buvo spinduliuotė, o modelyje be šios kovarinatės - dienos ilgis.

Nors visuose gautuose modeliuose $R^2$ reikšmės buvo palyginus didelės ($> 85\%$), patys modeliai vos tenkino tiesinės regresijos prielaidas. Duomenis reikėjo transformuoti, taip pat paklaidos netenkino homoskedastiškumo prielaidos. Šios problemos nulėmė faktą, kad modelio interpretacija gali būti nepastovi. Ateityje, tiriant saulės elektrinių pagaminamos elektros kiekį būtų galima įtraukti naują dienų iki paskutinio lietaus kovariantę, kuri indikuotų, kada paskutinį kartą buvo lyta (ir taip potencialiai nuplautos dulkės, nešvarumai nuo kolektorių). Taip pat orų kovariantės, turėjusios gana mažą reikšmingumą galėjo būti svarbesnės, jei jos būtų stebėtos arčiau nei mūsų tyrime (atstumas pas mus apie 40km). Taip pat potencialiai prasminga kovariantė galėtų būti didžiausia paros temperatūra vietoje vidutinės.

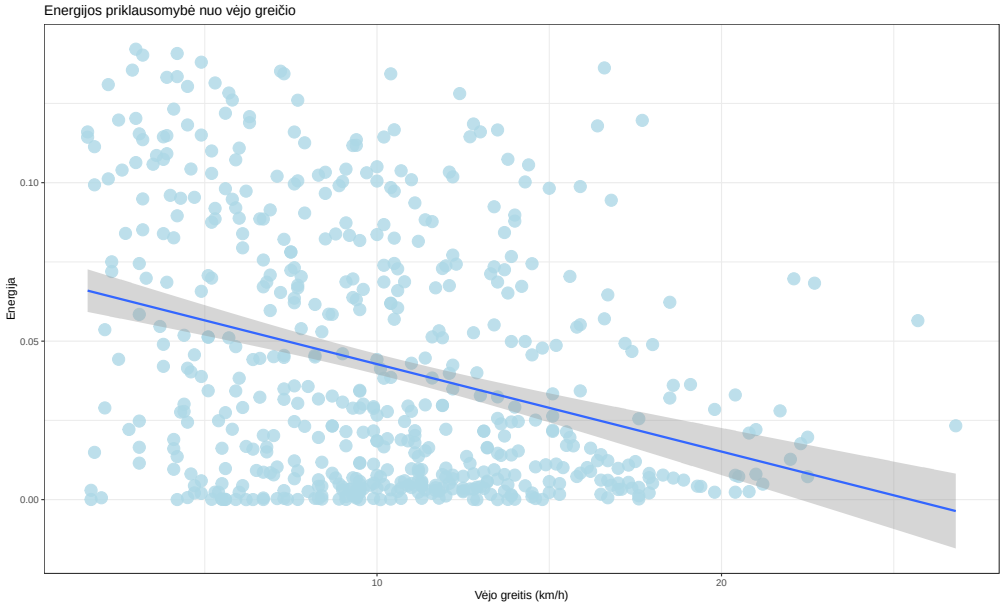
# 7 Pirmas priedas



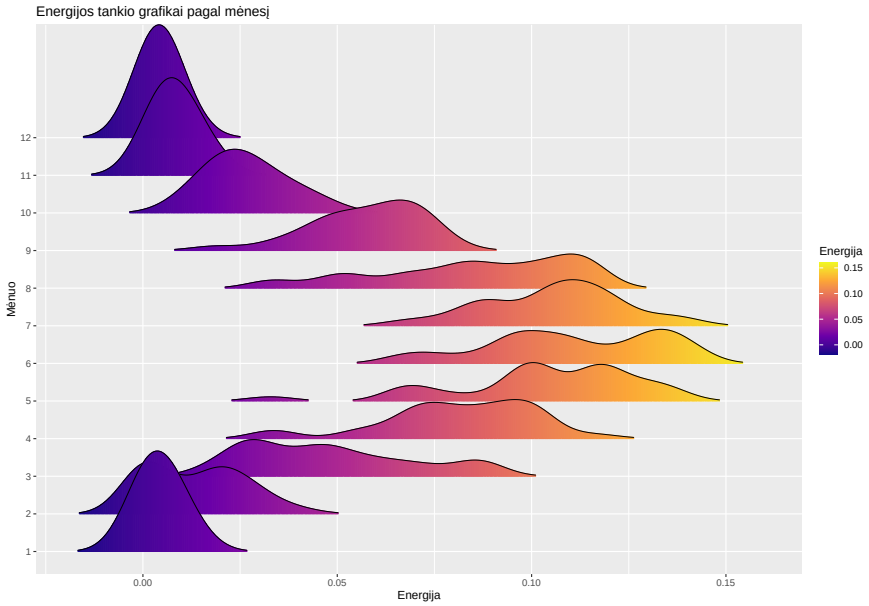

4 pav.: elektros kiekio priklausomybė nuo kritulių



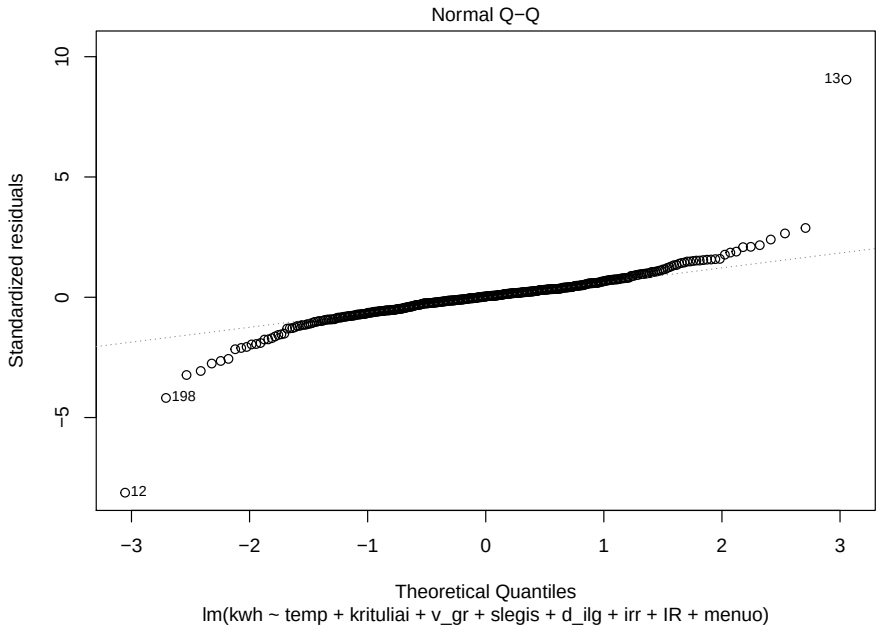

5 pav.: elektros kiekio priklausomybė nuo slėgio



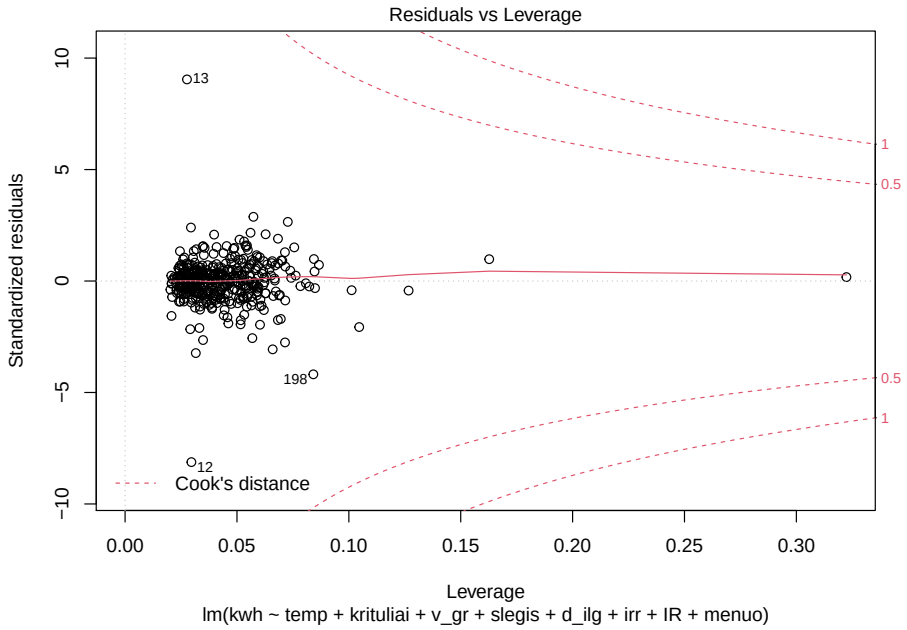

6 pav.: elektros kiekio priklausomybė nuo vėjo greičio



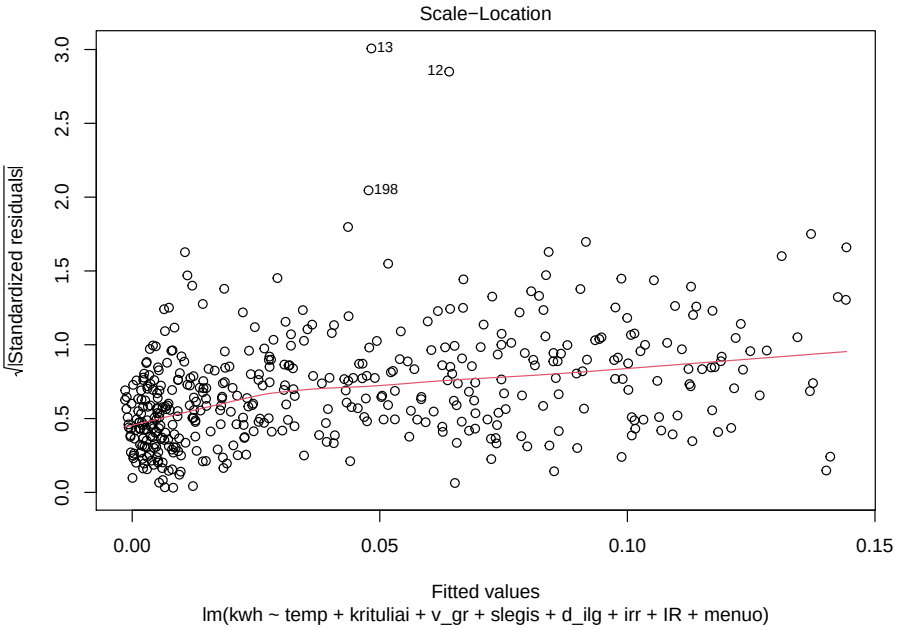

7 pav.: pagaminamos elektros kiekio pasiskirstymas metuose.



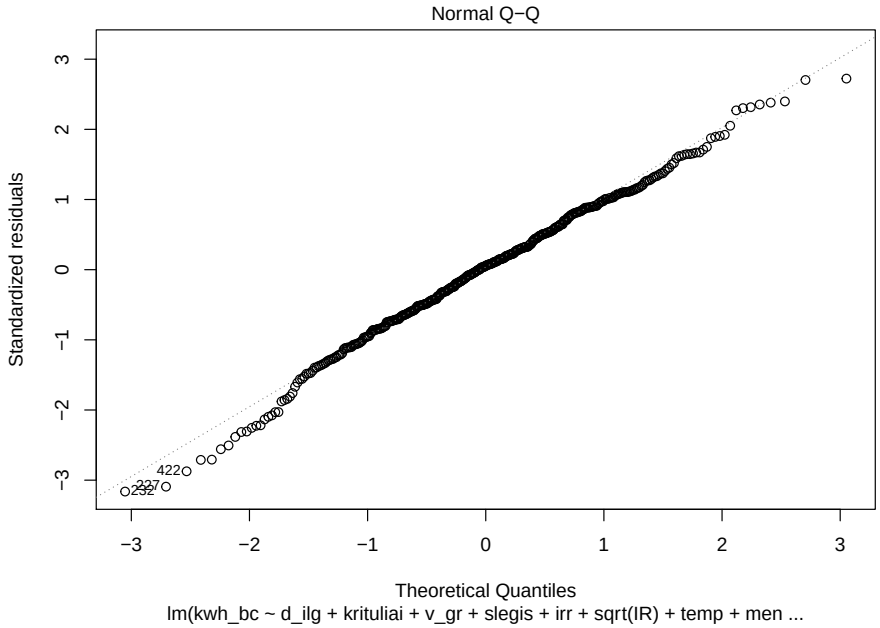

8 pav.: kvantilių grafikas (pradinis modelis su visomis kovariantėmis ir išskirtimis)




9 pav.: Kuko mato grafikas (pradinis modelis su visomis kovariantėmis ir išskirtimis)




10 pav.: standartizuotų liekanų grafikas (pradinis modelis su visomis kovariantėmis ir išskirtimis)




11 pav.: kvantilių grafikas(transformuotų duomenų modelis su visomis kovariantėmis)

# 8 Antras priedas

## 8.1 Pradinė duomenų analizė

```
Sys.setlocale("LC_ALL","Lithuanian") # lietuviškos raidės
# naudojamos bibliotekos
library(data.table)
library(lubridate)
library(tidyverse)
library(corrplot)
library(psych)
library(patchwork)
library(ggridges)

# nuskaitomi sutvarkyti duomenys
duom<-as.data.frame(fread("galDuom.csv"))
duom$menuo <- as.factor(month(ymd(duom$data)))

##############################################
# APRAŠOMOJI STATISTIKA

A <- describe(duom[,2:9], skew = F, quant = c(0.25,0.5, 0.75))
A <- A[,c(3:7,9,10, 11)]
options(scipen = 999)
htmlTable::htmlTable(format(round(A,3)))

##############################################
# TAŠKINĖS DIAGRAMOS

# DUOMENYS SUSIJĘ SU SAULE
(p1 <- ggplot(duom, aes(x=temp, y=kwh, color = menuo)) +
  geom_point(size = 5, alpha = 0.8) +
  theme_bw() + ggtitle("Temperaturos") +
  xlab("Temperatura (°C)") + ylab("Energija") +
  scale_colour_discrete("Menuo") + theme(plot.title = element_text(size=12)))

(p2 <- ggplot(duom, aes(x=IR, y=kwh, color = menuo)) +
  geom_point(size = 5, alpha = 0.8) +
  theme_bw()+ ggtitle("IR spinduliu") +
  xlab("IR") + ylab("") +
  scale_colour_discrete("Menuo") +
  theme(plot.title = element_text(size=12), axis.text.y=element_blank()))

(p3 <- ggplot(duom, aes(x=d_ilg, y=kwh, color = menuo)) +
  geom_point(size = 5, alpha = 0.8) +
```

```
  theme_bw() + ggtitle("Dienos ilgio") +
  xlab("Dienos ilgis (h)") + ylab("Energija") +
  scale_colour_discrete("Menuo") + theme(plot.title = element_text(size=12)))

(p4 <- ggplot(duom, aes(x=irr, y=kwh, color = menuo)) +
  geom_point(size = 5, alpha = 0.8) +
  theme_bw() + ggtitle("Spinduliuotes") +
  xlab("Spinduliuote") + ylab("") +
  scale_colour_discrete("Menuo") +
    theme(plot.title = element_text(size=12), axis.text.y=element_blank()))

p1234 <- p1 + p2 + p3 + p4
p1234 + plot_layout(ncol = 2, guides = "collect") +
  plot_annotation(title = "Energijos priklausomybe nuo:", tag_levels = 'I') &
  scale_y_continuous(limits = c(0, 0.15)) &
  theme(plot.tag = element_text(size = 12))

ggplot(duom[-c(14,15),], aes(x=irr, y=kwh, color = menuo)) +
  geom_point(size = 5, alpha = 0.8) +
  theme_bw() + ggtitle("Energijos priklausomybe nuo spinduliuotes") +
  xlab("Spinduliuote") + ylab("Energija") +
  scale_colour_discrete("Menuo")

# VĖJAS
ggplot(duom, aes(x=v_gr, y=kwh)) +
  geom_point(size = 5, alpha = 0.8, color = "light blue") +
  theme_bw()+
  geom_smooth(method=lm) +
  ggtitle("Energijos priklausomybe nuo vejo greicio") +
  xlab("Vejo greitis (km/h)") + ylab("Energija")

# KRITULIAI
ggplot(duom, aes(x=krituliai, y=kwh)) +
  geom_point(color = 'blue', size = 5, alpha = 0.5) +
  theme_bw()+
  geom_smooth(method=lm) +
  ggtitle("Energijos priklausomybe nuo krituliu") +
  xlab("Krituliai (mm)") + ylab("Energija")

# SLĖGIS
ggplot(duom, aes(x=slegis, y=kwh, color = menuo)) +
  geom_point(size = 5, alpha = 0.8) +
  theme_bw() +
  ggtitle("Energijos priklausomybe nuo slegio") +
  xlab("Slegis (hPa)") + ylab("Energija") +
```

```
  scale_colour_discrete("Menuo")

##########
# STAČIAKAMPĖS DIAGRAMOS

# MĖNUO
(p21 <- ggplot(duom, aes(x=menuo, y=kwh, fill = menuo)) +
  theme_bw()+
  geom_boxplot() + theme(legend.position="none") +
  ggtitle("Energijos") +
  xlab("Menuo") + ylab("Energija"))

(p22 <- ggplot(duom, aes(x=menuo, y=temp, fill = menuo)) +
  theme_bw()+
  geom_boxplot() + theme(legend.position="none") +
  ggtitle("Temperaturos") +
  xlab("Menuo") + ylab("Temperatura (°C)"))

(p23 <- ggplot(duom, aes(x=menuo, y=irr, fill = menuo)) +
  theme_bw()+
  geom_boxplot() + theme(legend.position="none") +
  ggtitle("Spinduliuotes") +
  xlab("Menuo") + ylab("Spnduliuote"))

(p24 <- ggplot(duom, aes(x=menuo, y=IR, fill = menuo)) +
  theme_bw()+
  geom_boxplot() + theme(legend.position="none") +
  ggtitle("IR") +
  xlab("Menuo") + ylab("IR"))

(p25 <- p21 + p22 + p23 + p24  +
  plot_annotation(title = "Pasiskirstymas metuose priklausomai nuo:",
  tag_levels = 'I',
caption="Pastaba: parinktos tik didziausiu koreliaciju kovariantes
ir priklausoma kovariante")&
  theme(plot.tag = element_text(size = 12)))

##########################################
# TANKIS

ggplot(duom, aes(x = kwh, y = menuo, fill = stat(x))) +
  geom_density_ridges_gradient(scale = 3, size = 0.3,
  rel_min_height = 0.01) +
  scale_fill_viridis_c(name = "Energija", option = "C") +
  labs(title = 'Energijos tankio grafikai pagal menesi') +
```

```
  xlab("Energija") + ylab("Menuo")

##############################################
# HISTOGRAMOS

duom20 <- duom[as.factor(year(ymd(duom$data)))=='2020',]

ggplot(duom20, aes(x=kwh, fill = menuo)) +
  geom_histogram(bins = 15, alpha=0.9, color = "black") +
  theme_bw() +
  xlab("Energija") + ylab("Skaicius") +
  ggtitle("Pagaminamos energijos pasiskirstymas") +
  scale_fill_discrete(name = "Menuo")

##############################################
# KORELIACIJŲ MATRICA

M<-cor(duom[,2:9])
colnames(M) <- c("Temp.", "Krituliai", "Vejo g.",
                    "Slegis", "D. ilgis", "Spind.",
                    "IR", "E. kiekis")
corrplot.mixed(M, upper = 'ellipse', title = 'Kintamuju koreliacijos',
               mar=c(0,0,2,0), tl.col = 'black', tl.pos = 'd')
```

## 8.2 Modeliavimas

```
Sys.setlocale("LC_ALL","Lithuanian") # lietuviškos raidės
# Naudojamos bibliotekos
library(data.table)
library(lubridate)
library(tidyverse)
library(MASS)#stepAIC
library(lmtest)#bptest
library(car)#vif
library(sandwich)
library(Metrics)#RMSE ir MAE
library(caTools)

# nuskaitomi sutvarkyti duomenys
duom<-as.data.frame(fread("galDuom.csv"))
duom$menuo <- as.factor(month(ymd(duom$data)))
# padalinami duomenys į mokymo ir testavimo
set.seed(67)
split = sample.split(duom$kwh, SplitRatio = 0.8)
```

```
sum(split)
sum(!split)
train = subset(duom, split == TRUE)
test = subset(duom, split == FALSE)

# Atstatome numeravimą
train <- train %>% as.data.frame(row.names = 1:nrow(.))
test <- test %>% as.data.frame(row.names = 1:nrow(.))

#modelis su visom kovariantėm
names(train)
m0 <- lm(kwh ~ temp+krituliai+v_gr+slegis+d_ilg+irr+IR+menuo, data = train)
summary(m0)

# grafikai
plot(m0)
plot(m0$residuals)
plot(cooks.distance(m0))

# Saliname isskirtis:
# didžiausios išskirtys yra 12 ir 13 reikšmės
df01 <- train[-c(12, 13),]
m00 <- lm(kwh ~ temp+krituliai+v_gr+slegis+d_ilg+irr+IR+menuo, data = df01)
summary(m00)
plot(m00)
shapiro.test(m00$residuals)

# boxcox transformacija
lambda <- 0.74
df01$kwh_bc <- (df01$kwh**lambda-1)/lambda
m01 <- lm(kwh_bc ~ d_ilg+krituliai+v_gr+slegis+irr+sqrt(IR)+temp+menuo,
data = df01)
plot(m01)
summary(m01)

# Patikriname liekanų normalumą ir heteroskedatiškumą.
shapiro.test(m01$residuals)
bptest(m01)

#Naudojame HC modelio korekciją
coeftest(m01, vcov = vcovHC, save = T)

# a)
# Nereikšmingas kovariantes šaliname po vieną, pradėdami nuo slėgio:
m02 <- lm(kwh_bc ~ temp+krituliai+v_gr+d_ilg+irr+sqrt(IR)+menuo, data = df01)
```

```
coeftest(m02, vcov = vcovHC(m02))
# Šaliname vėjo greitį:
m03 <- lm(kwh_bc ~ temp+krituliai+d_ilg+irr+sqrt(IR)+menuo, data = df01)
coeftest(m03, vcov = vcovHC(m03))

# grafikai
plot(m03)
plot(m03$residuals)
plot(cooks.distance(m03))

summary(m03)

# b)
# Atliekame pažingsninę regresiją naudodami stepAIC
stepAIC(m01, direction = "both")
# Gauname, kad modelis parenkamas toks pats, kaip ir
# atmetus po vieną kovariantę (m03).

##########################################################
#       M U L T I K O L I N E A R U M A S
# Tikriname, ar turime multikolinearumo problema:
vif(m03)
# Tikriname, kurios kovariantės mažiausiai reikšmingos, kad jas pašalintume
df02<-df01
df02$sqrt_IR <- sqrt(df02$IR)

b <- summary(m03)$coef[-c(1,7:17), 1]
sx <- df02[-c(4,5,8,9,11)] %>% summarise_if(is.numeric, sd)
sy <- sd(df02$kwh_bc)
(beta <- b * sx/sy)

# Sudarome modelius pretendentus:
# Be sqrt(IR) ir d_ilg (nes labiausiai multikolinearu)
m1 <- lm(kwh_bc ~ temp+krituliai+irr+menuo, data = df01)
vif(m1)
summary(m1)


# Be irr ir IR (spėti pagal iš anksto randamus kintamuosius (mėnuo multikol.))
m2 <- lm(kwh_bc ~ temp+krituliai + d_ilg, data = df01)
vif(m2)
summary(m2)

# Tik kolektoriaus duomenys
m3 <- lm(kwh_bc ~ irr+sqrt(IR), data = df01)
```

```
vif(m3)
summary(m3) # Adjusted R-squared:  0.969

# Be mėnesio, sqrt(IR), d_ilg (mažiausiai multikolinearus modelis)
m4 <- lm(kwh_bc ~ temp+krituliai+irr, data = df01)
vif(m4)
summary(m4)

cor.test(df01$krituliai, df01$kwh_bc)

####################################
# Modeliu vertinimas
# transformuojami testavimo duomenys boxcox
lambda <- 0.74
test$kwh_bc <- (test$kwh**lambda-1)/lambda

# Modelis m1 (geras R-square, didelis vif)
prediction1 <- predict(m1, newdata=test[-c(1,9)])
# Pakladios
round(rmse(test$kwh_bc, prediction1),4)
round(mae(test$kwh_bc, prediction1),4)
round(rmse(test$kwh_bc, prediction1)/(max(test$kwh_bc)-min(test$kwh_bc)),4)
round(mae(test$kwh_bc, prediction1)/(max(test$kwh_bc)-min(test$kwh_bc)),4)

# Modelis m2 (be irr ir IR (speti pagal is anksto randamus kintamuosius))
prediction2 <- predict(m2, newdata=test[-c(1,9)])
# Pakladios
round(rmse(test$kwh_bc, prediction2),4)
round(mae(test$kwh_bc, prediction2),4)
round(rmse(test$kwh_bc, prediction2)/(max(test$kwh_bc)-min(test$kwh_bc)),4)
round(mae(test$kwh_bc, prediction2)/(max(test$kwh_bc)-min(test$kwh_bc)),4)

# Modelis m3 (Tik kolektoriaus duomenys)
prediction3 <- predict(m3, newdata=test[-c(1,9)])
# Pakladios
round(rmse(test$kwh_bc, prediction3),4)
round(mae(test$kwh_bc, prediction3),4)
round(rmse(test$kwh_bc, prediction3)/(max(test$kwh_bc)-min(test$kwh_bc)),4)
round(mae(test$kwh_bc, prediction3)/(max(test$kwh_bc)-min(test$kwh_bc)),4)

# Modelis m4 (Be menesio, sqrt(IR), d_ilg (maziausiai multikolinearus modelis))
prediction4 <- predict(m4, newdata=test[-c(1,9)])
# Pakladios
round(rmse(test$kwh_bc, prediction4),4)
round(mae(test$kwh_bc, prediction4),4)
```

```
round(rmse(test$kwh_bc, prediction4)/(max(test$kwh_bc)-min(test$kwh_bc)),4)
round(mae(test$kwh_bc, prediction4)/(max(test$kwh_bc)-min(test$kwh_bc)),4)
```

# Literatūra


[1] Rebecca Lindsey ir LuAnn Dahlman. "Climate change: Global temperature". In: *Climate. gov* 16 (2020 m.).

[2] VERT. *Elektros energijos kainos*. `https://www.regula.lt/elektra/Puslapiai/tarifai/elektros-energijos-kainos.aspx`. 2021 m.

[3] Tserenpurev Chuluunsaikhan ir kt. "Predicting the Power Output of Solar Panels based on Weather and Air Pollution Features using Machine Learning". In: *Journal of Korea Multimedia Society* 24.2 (2021 m.), p. 222–232.

[4] Tushar Verma ir kt. "Data analysis to generate models based on neural network and regression for solar power generation forecasting". In: *2016 7th international conference on intelligent systems, modelling and simulation (ISMS)*. IEEE. 2016 m., p. 97–100.

[5] Young Seo Kim ir kt. "Use of a big data analysis in regression of solar power generation on meteorological variables for a Korean solar power plant". In: *Applied Sciences* 11.4 (2021 m.), p. 1776.