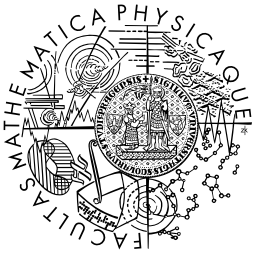


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Lukáš Kotlorz

Model pro krátkodobou predikci výroby elektrické energie z fotovoltaických zdrojů

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	prof. Ing. Emil Pelikán, CSc.
Studijní program:	Pravděpodobnost, matematická statistika a ekonometrie
Studijní obor:	Matematika

Praha 2017

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Rád bych poděkoval svému vedoucímu práce prof. Ing. Emilu Pelikánovi, CSc. a konzultantovi RNDr. Ing. Marku Brabcovi, PhD. za pomoc, kterou mi poskytli při tvorbě práce. Dále bych rád poděkoval i kolegům RNDr. Alexandrovi Černému, PhD. a Ing. Filipu Tichému, PhD. za cenné rady, které mi v úvodní fázi práce velmi pomohly. Poděkování rovněž patří i společnosti ČEZ Prodej, a.s. za poskytnutí dat, bez kterých by tuto práci nebylo možné napsat.

Název práce: Model pro krátkodobou predikci výroby elektrické energie z fotovoltaických zdrojů

Autor: Bc. Lukáš Kotlorz

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: prof. Ing. Emil Pelikán, CSc., Ústav informatiky Akademie věd České republiky

Abstrakt: V dnešní době nabírá výroba elektřiny z fotovoltaických elektráren na významu. Aby bylo možné přizpůsobit výrobu v ostatních elektrárnách, je potřeba predikovat výrobu elektřiny z těchto zdrojů. Práce je věnována zejména modelům pro krátkodobou predikci, která je založena na předpovědi počasí. Modely byly konstruovány zejména pomocí beta regrese a lineární regrese s transformovanou vysvětlovanou proměnnou. Součástí práce je i Clear sky model, který slouží k odhadu maximální možné výroby v danou hodinu.

Klíčová slova: krátkodobá predikce, fotovoltaická elektrárna, Clear sky model

Title: Model for short-term forecasting of photovoltaic energy production

Author: Bc. Lukáš Kotlorz

Department: Department of Probability and Mathematical Statistics

Supervisor: prof. Ing. Emil Pelikán, CSc., Institute of Computer Science Academy of Sciences of the Czech Republic

Abstract: Nowadays, electricity production from photovoltaics power plants is becoming important increasingly. In order to set production to other power plants, it is necessary to predict the generation of electricity from these sources. The thesis is mainly devoted to models for short-term prediction, which is based on weather forecast. The models were designated by beta regression and linear regression with transformed explanatory variable. One part of thesis is devoted to Clear sky model, which is used to estimated the maximum possible production at given hour.

Keywords: short-term prediction, photovoltaic power plant, Clear sky model

Obsah

Úvod

Elektrická energie je jednou ze základních potřeb lidstva. V době před 20 lety se elektřina vyráběla převážně v tepelných, jaderných a vodních elektrárnách. V dnešní době, kdy je kladen větší důraz na obnovitelné zdroje, začíná na významu nabývat i výroba elektřiny z větrných (dále VE) a fotovoltaických elektráren (dále FVE). Do budoucna je očekáváno, že tyto zdroje budou hrát hlavní roli v zásobování elektrickou energií. Hlavním problémem integrace solární a větrné energie do energetické sítě je nemožnost plánování její výroby a to z hlediska množstevního, kdy nemůžeme rozhodnout, kolik ji vyrobíme, tak i časového, kdy se nemůžeme rozhodnout, kdy ji vyrobíme. Teoreticky jedinou šanci jak ovlivnit výši výroby je nedodáním části této výroby do sítě. U ostatních elektráren jsme schopni s různou časovou prodlevou množství vyrobené elektřiny regulovat, tudíž zde problém se zvýšením nebo snížením výroby nemáme. Náklady spojené se změnou výroby jsou ovšem nepřímo úměrné s dobou, kterou máme na tuto změnu k dispozici.

Důvodem, proč potřebujeme množství vyráběné elektřiny měnit, je, na rozdíl od jiných zdrojů energie, nemožnost elektřinu efektivně skladovat. Možnosti pro skladování sice existují (např. přečerpávací elektrárny, baterie), ale skladovat elektřinu lze pouze v omezeném množství, navíc často za nemalých ztrát samotné energie, případně i za dodatečných finančních nákladů. Zejména z tohoto důvodu, který je navíc podpořen tím, že elektřina vyrobená z obnovitelných zdrojů má v ČR ze zákona při dodávkách do sítě přednost před elektřinou vyrobenou z ostatních zdrojů, je nutné předpovídat množství elektřiny, které bude v daném období v solárních a větrných elektrárnách vyrobeno, aby bylo možné efektivně naplánovat výrobu elektřiny na ostatních zdrojích. V budoucnosti je možné, že budou do sítě implementovány kapacitnější skladovací prostředky, jejichž výrobní náklady již nebudou neúměrně vysoké oproti finančnímu přínosu, ty nám ale umožní pouze větší variabilitu v plánování, a nikoliv možnost předpovědi vyrobené elektřiny z FVE a VE ignorovat.

Predikce výroby elektřiny z FVE se dělí na krátkodobou predikci, která zahrnuje období přibližně na 2 až 3 dny dopředu, a na dlouhodobou predikci, která je na vzdálenější období. V této diplomové práci se zaměříme na krátkodobou predikci výroby elektrické energie z FVE a to ze zdrojů, ze kterých vyrobenou elektřinu vykupuje společnost ČEZ Prodej, a.s. (dále ČEZ), na území celé ČR. Budeme se snažit najít takový model, který by z dostupných dat nejlépe předpovídal jednak výrobu z FVE v následujících hodinách aktuálního dne, stejně tak i model, který bude nejlépe předpovídat výrobu v následujícím dni. Tyto predikce bude možné použít i na odhad výroby FVE, ze kterých elektřinu vykupují jiní obchodníci, podmínkou ovšem je, že pro svou predikci budou mít k dispozici z území, kde se nachází jejich elektrárny, stejnou sadu dat, která budou použita ve výsledných modelech.

Veškeré statistické výpočty v této práci jsou prováděny pomocí 64-bitové verze statistického softwaru `R`, verze 3.4.0. Všechny použité zdrojové kódy jsou přílohou této práce, jejich seznam je uveden na konci práce v sekci Přílohy.

Na téma predikce výroby elektřiny z FVE již proběhlo v poslední době hodně výzkumů, tudíž existuje spousta možností, jak výrobu elektřiny z FVE predikovat,

stále ale neexistuje žádný jednoznačný model pro nejlepší predikci v libovolném místě na světě. Porovnávat výsledky jednotlivých studií lze jen velmi obtížné, neboť většina z nich je zaměřena na predikci pro konkrétní region, kdy tvorba modelů probíhala pouze na základě naměřených dat z tohoto regionu, navíc v každém jednotlivém případě máme k dispozici různý typ vstupních dat.

V první kapitole je představeno fungování trhu s elektrickou energií v ČR, které nám umožňuje si ujasnit, jakým způsobem a ve kterých časových intervalech, budeme předpovědi provádět. Dále je zde popsáno fungování odchylek. Na závěr kapitoly je popis, jakým způsobem probíhá výkup vyrobené elektřiny z FVE doplněný o způsoby finanční podpory této výroby.

Ve druhé kapitole zavádíme značení, které bude používáno v dalším průběhu práce. Dále se zaměříme na představení všech dostupných dat, které pro modelování máme k dispozici, provedení vizuální prohlídky dat pro odstranění zjevných chyb, dále kontroly dat na nepřípustné hodnoty a očištění nalezených chybných hodnot.

Ve třetí kapitole budou teoreticky představeny statistické metody, které budou při modelování používány v následujících kapitolách s odkazem na literaturu, kde se čtenář o těchto metodách může dozvědět více.

Ve čtvrté kapitole je představen Clear sky model.[1] Tento model odhaduje pro jednotlivé hodinové úseky v roce maximální potenciál výroby FVE, tj. potenciál za ideálních povětrnostních podmínek, který je pro každou hodinu v roce jiný. Tento model nám umožňuje místo predikování vyrobeného množství elektřiny z FVE predikovat procentuální množství vyrobené elektřiny z FVE z maximálního výkonu, které je jednodušší. Využití tohoto modelu je širší, neboť se z něj vychází při dlouhodobých predikcích výroby elektřiny z FVE.

V páté kapitole je popsán postup výstavby modelu pro predikci výroby. Kapitola je rozdělena na dvě části, kdy nejprve je vytvořen model pro predikci na následující den, který je pak následně vylepšován pro aktuální den na nejbližší budoucí hodiny.

[1] Do českého jazyka se dá přeložit jako model čisté oblohy, příp. model jasné oblohy. Vzhledem k tomu, že ani jeden z těchto výrazu není úplně výstižný, navíc se v energetice nepoužívají, bylo zvoleno ponechání anglického názvu pro tento model.

1. Obecné informace

V této kapitole nejprve popíšeme, jak funguje trh s elektřinou v ČR. Dále se podíváme na odchylky, kvůli kterým je mj. potřeba predikovat výrobu elektřiny z FVE. Následně popíšeme, jak v ČR probíhá podpora obnovitelných zdrojů elektřiny, zejména FVE, a kapitolu zakončíme krátkou úvahou o tom, pro které subjekty na trhu jsou krátkodobé predikce výroby elektřiny z FVE zajímavé.

1.1 Trh s elektřinou v ČR

Trhu s elektřinou se v ČR účastní výrobci, odběratelé, obchodníci, burza, provozovatelé distribuční soustavy, provozovatel přenosové soustavy, operátor trhu s elektřinou a Energetický regulační úřad. Jejich funkce a legislativní povinnosti lze nalézt např. na webu oEnergetice.cz[2]. Trh s elektřinou se v ČR dělí na trh organizovaný a neorganizovaný. Na neorganizovaném mohou dvě strany uzavřít libovolnou dohodu, kterou pouze musí nahlásit operátorovi trhu ještě před sjednaným obchodem. Organizovaný trh probíhá na burze a dělí se na dlouhodobý a krátkodobý trh.

1.1.1 Dlouhodobý trh

Dlouhodobý trh s elektřinou je v ČR organizován společností POWER EXCHANGE CENTRAL EUROPE, a.s. (dále PXE). Tato společnost byla založena 8. ledna 2007 pod názvem Energetická burza Praha a od té doby na trhu s elektrickou energií nabízí obchodování se standardizovanými produkty se zajištěním vypořádání. Předmětem obchodování na PXE je elektrická energie o hodinovém výkonu 1 MW ve všech hodinách, které zahrnuje příslušný produkt. V tomto případě se jedná o roční produkty, čtvrtletní produkty a měsíční produkty. Dále je možné tyto produkty nakoupit ve dvou pásmech a to buď pásmo Base Load, které zahrnuje všechny dny od 0:00 do 24:00, případně pásmo Peak Load, které zahrnuje hodiny od pondělí do pátku včetně od 8:00 do 20:00 bez ohledu na státní svátky. Pro představu, v roce 2016 byly na PXE zobchodovány české kontrakty s finančním vypořádáním o objemu 19,86 TWh.

1.1.2 Krátkodobý trh

Při vytváření krátkodobé predikce výroby elektřiny z FVE mají větší význam informace o krátkodobém trhu s elektřinou v ČR. Tyto trhy jsou v ČR organizovány společností OTE, a.s. Trh se dělí na blokový trh, denní trh a vyrovnávací trh, přičemž největší množství elektřiny se zobchoduje na denním trhu.

Na blokovém trhu je možné nakoupit samostatně pro každý den v roce elektrickou energii o hodinovém výkonu 1 MW a to buď blok Base Load, který zahrnuje všechny hodiny dne, blok Peak Load, který zahrnuje hodiny od 8:00 do 20:00 nebo Offpeak Load, který zahrnuje hodiny od 20:00 do 8:00. Obchodovat tyto

[2] `http://oenergetice.cz/elektrina/trh-s-elektrinou/trh-s-elektrinou/`

produkty je možné od pátého dne před začátkem dodávky od 9:30 až do 13. hodiny dne předcházejícího dodávce. V roce 2016 byly na tomto trhu zobchodovány kontrakty o objemu 61,9 GWh.

Nejzajímavější z těchto trhů je denní trh. Zde je možné koupit elektrickou energii na vybranou hodinu v roce o minimálním množství 1 MWh s přesností na desetiny. Tento trh neprobíhá klasickým způsobem, při kterém účastníci mohou přijmout nějakou ze zveřejněných nabídek a poptávek, ale formou dvoustranné aukce. Do doby uzávěrky, která je den před dodávkou v 11:00, má každý zájemce možnost podat na každou hodinu nabídku a poptávku elektřiny za jím požadovanou cenu. Po uzávěrce dojde k sesouhlasení nabídek, čímž se vytvoří výsledná cena a za ní dojde k uspokojení všech poptávek s vyšší cenou a nabídek za nižší cenu. Výsledky jsou zveřejněny 40 minut po uzávěrce. V roce 2016 bylo na tomto trhu zobchodováno 20,14 TWh elektrické energie, tj. cca 33 % tuzemské netto spotřeby.

Posledním trhem a zároveň trhem, kde se dá pořídit elektřina nejblíže termínu dodávky, je vnitrodenní trh. Ten se otevírá den před zahájením dodávky v 15 hodin, uzávěrka obchodování je 60 minut před začátkem dodávky. Zde je možné koupit elektrickou energii na vybranou hodinu v roce o minimálním množství 0,1 MWh s přesností na desetiny. Byť možnost nakoupit elektřinu na tomto trhu je již den před dodávkou, většina obchodů se zde uskuteční až v posledních hodinách před uzávěrkou. V roce 2016 bylo na tomto trhu zobchodováno 544,7 GWh elektrické energie.

Dále existuje vyrovnávací trh, který řídí provozovatel přenosové soustavy ČEPS, a.s. Ten zde má možnost až do doby 30 minut před začátkem dodávky nakoupit regulační energii. V roce 2016 bylo na tomto trhu zobchodováno 21,2 GWh elektrické energie.

1.1.3 Maloobchod s elektřinou

Obchodníci s elektřinou svoji nakoupenou elektřinu prodávají zákazníkům (domácnosti, firmy) na maloobchodním trhu, kde mezi sebou soutěží zejména cenou, případně jinými dodatečnými službami. Vzhledem k tomu, že tento obchod nijak nesouvisí s FVE, nebudeme jej dále rozebírat.

1.2 Odchylky

Jak jsme viděli výše, tak obchodování s elektřinou má mnoho podob. Základem tohoto obchodování ovšem je vyrobit tolik, kolik jsem prodal a odebrat tolik, kolik jsem si koupil. Pokud se to nepovede, tak jakákoliv odchylka pro obchodníka většinou představuje pokutu. Každý subjekt, který s elektřinou obchoduje, musí být zodpovědný za odchylku, kterou svým obchodováním v síti způsobí. Tato odpovědnost může být přenesená. Příkladem je spotřebitel typu domácnost, který využívá elektřinu tak, jak ji sám potřebuje. Odpovědnost za odchylku za něj nese obchodník, se kterým má uzavřenou smlouvu.

Cena za odchylku se odvíjí od skutečnosti, zdali byla systémová odchylka kladná nebo záporná. Přesný postup pro vypočtení ceny je dán v Příloze č. 8 k vyhlášce č. 408/2015 Sb. Pro subjekt zúčtování platí, že má-li odchylku ve shodě se systémem, pak platí pokutu, je-li v protiodchylce, dostává od operátora

trhu příspěvek, který je ale významně nižší než pokuta. Pro subjekty zúčtování je místo nákladů na odchylky přínosnější počítat tzv. vícenáklady na odchylku, které v sobě zahrnují náklady na pořízení přebytečné elektřiny a úsporu za nepořízenou spotřebovanou elektřinu. Tato elektřina se oceňuje cenou denního trhu. Tyto vícenáklady přesněji odrážejí skutečné náklady, které subjektu zúčtování při odchylce vznikly.

1.3 Podpora obnovitelných zdrojů na výrobu elektřiny v ČR[3]

Forma podpory výroby elektřiny z obnovitelných zdrojů v ČR je dána zákonem č. 165/2012 Sb. o podporovaných zdrojích energie. Jednou z možností podpory je podpora elektřiny formou výkupních cen, druhou možností je podpora formou zelených bonusů. Pro výrobce elektřiny z FVE platí, že mají-li jeho zdroje instalovaný výkon do 100 kW včetně, může si z obou forem podpory vybrat tu, která mu více vyhovuje. Výrobce s instalovaným výkonem nad 100 kW má právo pouze na podporu formou zelených bonusů.

1.3.1 Podpora formou povinného výkupu

Pokud si výrobce zvolí podporu formou výkupních cen, potom má povinně vykupující povinnost vykoupit veškerý objem vyrobené elektřiny naměřený v předávacím místě. Povinně vykupujícím je dodavatel poslední instance v příslušném území. Elektřinu vykupuje za cenu stanovenou aktuálním cenovým rozhodnutím Energetického regulačního úřadu (dále ERÚ). V současné době jsou dodavateli poslední instance společnosti ČEZ Prodej, a.s., E.ON Energie, a.s. a Pražská energetika. Cena je v tomto případě stejná ve všech hodinách daného kalendářního roku a je na každý rok upravována novým Cenovým rozhodnutím ERÚ. Odpovědnost za odchylku v předávacím místě výroby nese povinně vykupující.

1.3.2 Podpora formou zeleného bonusu

Jestliže má výrobce podporu formou zeleného bonusu, musí si sám najít svého vykupujícího a s ním si dohodnout výkupní cenu. V případě, že dochází k přetokům vyrobené elektřiny do elektrizační soustavy a výrobce nemá sjednanou smlouvu o dodávce, jedná se o neoprávněnou dodávku bez nároku na podporu. Ke sjednané ceně za výkup od vykupujícího dostane navíc bonus, který je stanoven Cenovým rozhodnutím ERÚ. Podpora je realizována buď v hodinovém režimu, pro zdroje s instalovaným výkonem nad 100 kW nebo ročním režimu pro zdroje s instalovaným výkonem do 100 kW. Rozdíl mezi těmito režimy je dán ve stanovení výše bonusu. U ročního režimu je cena stejná ve všech hodinách daného kalendářního roku, u hodinového je vypočtena na základě vzorce z přílohy č. 22 vyhlášky č. 408/2015 Sb. Zelený bonus má pro výrobce zpravidla vyšší výnos, ten je ale vyvážen vyšším rizikem. Odpovědnost za odchylku nese vykupující s výjimkou případů, kdy výrobce je sám subjektem zúčtování nebo přenesl tuto odpovědnost na jiný subjekt zúčtování.

[3]Zdrojem pro tuto část byly internetové stránky ERÚ, viz [1]

1.4 Pro koho mají predikce význam?

Na závěr se ještě podíváme, pro které subjekty mají jaké predikce význam. Pokud výrobce elektřiny z FVE není subjektem zúčtování nebo si neponechal odpovědnost za odchylku, pak je pro něj samotného krátkodobá predikce vyrobené elektřiny nezajímavá, neboť jediné, co mu může přinést, je odhad jeho příjmů v nejbližších hodinách a dnech. Vzhledem k tomu, že množství vyrobené elektřiny, a tudíž ani svůj příjem nemůže nijak ovlivnit, je tato predikce pro něj nejvýše informativního charakteru. Pro vykupujícího je naopak krátkodobá predikce výroby elektřiny důležitá, neboť se mu takto nakoupena elektřina přičte k elektřině zakoupené na burze, tudíž se promítne do odchylky.

Čerpá-li výrobce podporu formou zeleného bonusu, je pro něj samotného zajímavější dlouhodobá predikce, pomocí které může odhadnout cenu, za kterou by měl jím vyrobenou elektřinu prodat. Pro vykupujícího je dlouhodobá predikce rovněž zajímavá, protože pomocí ní si stanovuje cenu, za kterou je sám ochoten elektřinu odkoupit, dále mu slouží i při plánování dlouhodobého nákupu elektřiny, protože při tomto nákupu musí počítat i s elektřinou vyrobenou z obnovitelných zdrojů.

2. Data

2.1 Obecné informace o datech

V této kapitole představíme data, která budou pro modelování použita, zavedeme značení, provedeme základní kontrolu a případné očištění dat. Všechna data, která máme k dispozici, jsou v hodinových intervalech od 1. ledna 2012 do 31. května 2017. Protože budeme modelovat výrobu elektřiny z fotovoltaických zdrojů, která má závislost pouze na fyzikálních proměnných, pro zjednodušení práce převedeme všechna data, která jsou v SELČ (Středoevropský letní čas) na SEČ (Středoevropský čas, tzv. zimní). Tímto přechodem nedojde k ovlivnění výsledku predikčního modelu, výrazně se tím však zjednoduší práce s používaným statistickým softwarem R, neboť nebude potřeba přechody času zapracovávat do modelů. Komerční predikční softwary, které jsou v energetice používány, mají většinou přechody mezi letním a zimním časem zpracovány automaticky, takže není potřeba modely složitě upravovat. Dále pro zjednodušení modelování vynecháme z dat 29. únor 2012 a 29. únor 2016, abychom měli v každém roce právě 365 dní a pořadová čísla kalendářních dní v roce si ve všech letech odpovídala.

V celé práci s výjimkou kapitoly 3 zavedeme indexy Y označující rok (z angl. year), index D označující pořadové číslo dne v roce (z angl. day of year) a index H označující pořadové číslo hodiny ve dni (z angl. hour of day). Tyto indexy zejména v souvislosti s jednotlivými proměnnými budou identifikovat, o která pozorování se jedná. Pokud nebude uvedeno jinak, bude index Y nabývat hodnot od 2012 do 2017, index D bude nabývat hodnot od 1 do 365, a index H od 1 do 24 s tím, že první hodina dne je v časovém rozmezí od 0:00 do 1:00, druhá hodina dne od 1:00 do 2:00 atd. Kombinací těchto indexů můžeme docílit výběr konkrétního pozorování, např. $Y,D,H = 2015, 35, 15$ značí 15. hodinu dne 4. února 2015. Dále pro tyto tři indexy platí, že příslušný index v kombinaci se znaménkem plus nebo mínus a číslem, označuje posun o uvedený počet roků/dní/hodin od uvažovaného roku/dne/hodiny, kterému budeme říkat nultý rok/den/hodina. Tedy např. $D-1$ značí předchozí den od nultého dne D, $H-3$ posun o 3 hodiny zpět od nulté hodiny H, $D-1,H-1 = H-25$ značí předchozí hodinu předchozího dne, tedy je-li např. nultý den 14. července 2016 a nultá hodina 15., potom $D-1,H-1$ znamená 14. hodinu dne 13. července 2016. V případě všech takových posunů je vždy aplikována kalendářní konvence, tzn. je-li nultý den 1. leden, pak $D-1$ je 31. prosinec předcházejícího roku, je-li nultá hodina 23. hodina, pak $H+3$ je 2. hodina následujícího dne apod. V žádném případě se nebude jednat o posun mimo hranice např. na 26. hodinu, který by odkazoval na neexistující data.

Dále definujme sjednocený časový index i, který bude značit pořadové číslo hodiny od počátku dat. Index i může nabývat hodnot od 1 do 47 448, není-li časový index jinak omezen, kde $i = 1$ odpovídá $(Y,D,H) = (2012, 1, 1)$ a $i = 47\,448$ odpovídá $(Y,D,H) = (2017, 152, 24)$.

Další index, tentokráte již ne časový, ale místní, používaný pro odlišení jednotlivých proměnných bude R pro region. Data pocházející od společnosti ČEZ, která souvisejí se samotnými elektrárnami (měření, instalovaný výkon atd.), máme k dispozici v 6 regionech, které územně odpovídají působnosti původních regionálních energetik v ČR (viz Obrázek 2.1). Seznam regionů a jejich zkratky

Regiony ČEZ		Regiony ČHMÚ		
Zkr.	*Název regionu*	*Zkr.*	*Název kraje*	*Region*
ZC	Západní Čechy	CB	Jihočeský kraj	JC
SC	Severní Čechy	HK	Královéhradecký kraj	VC
ST	Střední Čechy	JM	Jihomoravský kraj	JM
VC	Východní Čechy	KV	Karlovarský kraj	ZC
SM	Severní Morava	LB	Liberecký kraj	SC
JC	Jižní Čechy	MS	Moravskoslezský kraj	SM
JM	Jižní Morava	OL	Olomoucký kraj	SM
AB	Praha	PH	Praha, hlavní město	AB
		PL	Plzeňský kraj	ZC
OS	Ostatní (JM, JC)	PU	Pardubický kraj	VC
CZ	Celá ČR	SK	Středočeský kraj	ST
		US	Ústecký kraj	SC
		VY	Vysočina	OS
		ZL	Zlínský kraj	SM

Tabulka 2.1: Zkratky jednotlivých krajů a regionů ČR.

najdeme v tabulce 2.1. Regiony Jižní Čechy a Jižní Morava jsou v datech sloučeny v jeden region Ostatní z důvodu malého instalovaného výkonu FVE v těchto dvou regionech, od nichž elektrickou energii vykupuje společnost ČEZ. Data z regionu Praha nemáme k dispozici, ale vzhledem k zanedbatelnému instalovanému výkonu v tomto regionu jej můžeme bez výrazné ztráty informace včlenit do regionu Střední Čechy.

Dalším indexem pro místní odlišení jednotlivých proměnných bude index K pro kraj. Předpovědi modelu Aladin pocházející od Českého hydrometeorologického ústavu (dále ČHMÚ) máme k dispozici po jednotlivých krajích ČR. Zkratky pro jednotlivé kraje najdeme rovněž v tabulce 2.1.

Posledním indexem pro místní odlišení jednotlivých proměnných bude index S pro meteorologickou stanici ČHMÚ, ze kterých máme k dispozici měření skutečných hodnot meteorologických jevů. Seznam stanic, jejich zkratky a lokaci v rámci kraje a regionu najdeme v tabulce 2.2

2.2 Data ČEZ

Nyní se podíváme na časové řady o FVE, od kterých je vyrobená elektrická energie vykupována společností ČEZ. Hlavními časovými řadami, bez kterých by nebylo možné cokoliv modelovat, jsou naměřené skutečné hodinové výroby FVE v příslušném regionu. Ty jsou udávány v kWh s přesností na celá čísla. Tyto řady budeme značit P_R. Dále máme pro každý region k dispozici instalovaný výkon všech FVE v daném regionu. Výkon je udáván v kW s přesností na celá čísla. Tyto řady budeme značit V_R. Protože zveřejnění těchto dat by mohlo pro společnost ČEZ znamenat obchodní riziko, bylo poskytnutí dat podmíněno převodem těchto časových řad na časové řady, které budou udávat procentuální množství vyrobené elektřiny z maximálního výkonu. Tyto řady budeme značit p_R a pro jejich výpočet platí $p_{R,Y,D,H} = P_{R,Y,D,H}/V_{R,Y,D,H}$. Predikci výkonu FVE v příslušném

Meteorologické stanice ČHMÚ			
Zkr.	*Název stanice*	*Kraj*	*Region*
CH	Cheb	KV	ZC
KY	Karlovy Vary	KV	ZC
PM	Plzeň - Mikulka	PL	ZC
UL	Ústí nad Labem	US	SC
LI	Liberec	LB	SC
TE	Temelín	CB	JC
LS	Praha - Libuš	PH	AB
UO	Ústí nad Orlicí	PU	VC
PB	Přibyslav	VY	VC
BR	Brno - Tuřany	JM	JM
HO	Holešov	ZL	SM
MO	Ostrava - Mošnov	MS	SM

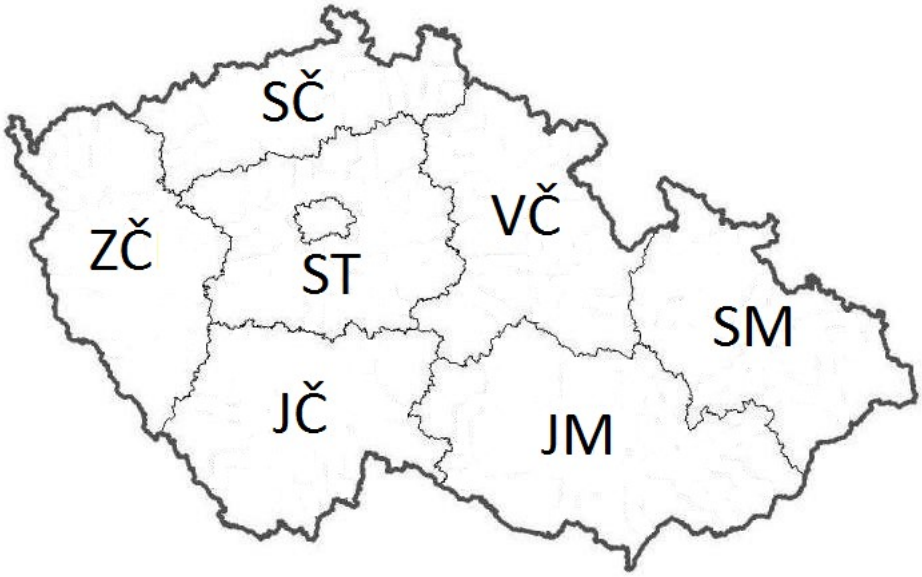
Tabulka 2.2: Seznam stanic ČHMÚ, jejich zkratky a umístění.

regionu potom můžeme z predikce procentuálního výkonu získat snadným vynásobením instalovaným výkonem jako $\hat{P}_{R,Y,D,H} = \hat{p}_{R,Y,D,H} \cdot V_{R,Y,D,H}$. Instalovaný výkon je firmě vykupující elektřinu znám, proto není nutné jej predikovat. Vzhledem k tomu, že výkon většiny elektráren je stahován jednou měsíčně, jsou tato data o naměřeném výkonu, které poskytuje společnost ČEPS, a.s., k dispozici nejdříve následující den. Pro přibližnou představu, na jakém desetinném místě je již z pohledu nákupu elektřiny změna nevýznamná, alespoň prozradíme, že instalovaný výkonu v jednotlivých regionech se řádově pohybuje v nižších stovkách MW.

Aby bylo možné modelovat predikci i na základě znalosti aktuálnějšího měření výkonu a nikoliv pouze na měření z předchozího dne, máme k dispozici i částečné hodinové „online“ měření výkonu spolu s informací o instalovaném výkonu elektráren, ze kterých je toto měření k dispozici. Tyto řady budeme značit Q_R a W_R. Hodnoty jsou opět v kWh resp. kW s přesností na celá čísla. Ze stejných důvodů uvedených výše provedeme i u „online“ měření převod časových řad do časových řad, které budou udávat procentuální množství vyrobené elektřiny z , instalovaného výkonu. Tyto řady budeme značit q_R a pro výpočet opět platí $q_{R,Y,D,H} = Q_{R,Y,D,H}/W_{R,Y,D,H}$. Dále zavedeme řady označené s_R, které budou udávat procento instalovaného výkonu „online“ měření z celkového instalovaného výkonu. Pro tuto řadu platí $s_{R,Y,D,H} = Q_{R,Y,D,H}/P_{R,Y,D,H}$. Vzhledem k malému instalovanému výkonu elektráren v regionu Ostatní, ve kterých by „online“ měření bylo k dispozici, nejsou řady Q_{OS} a W_{OS} a z nich odvozené časové řady k dispozici.

2.3 Data ČHMÚ

Druhou sadu dat, kterou budeme pro modelování používat, jsou časové řady od ČHMÚ. Tyto řady rozdělíme do dvou skupin. První skupinou budou předpovědi meteorologických jevů modelu Aladin, druhou naměřená meteorologická data ze stanic ČHMÚ.



Obrázek 2.1: Mapa ČR s hranicemi regionů původních energetik.

2.3.1 Data z modelu Aladin

První skupinou dat od ČHMÚ jsou predikce modelu Aladin. Zde máme z každého kraje v ČR k dispozici predikci teploty vzduchu ve dvou metrech nad zemí ve °C. Tyto řady budeme označovat $\widehat{t}_K$. Dále máme k dispozici celkovou, nízkou, střední a vysokou oblačnost. Ta je udávaná v procentech, přičemž 0 znamená jasnou oblohu a 1 zataženou oblohu. Tyto řady budeme postupně značit $\widehat{Nc}_K$, $\widehat{Nl}_K$, $\widehat{Nm}_K$, $\widehat{Nh}_K$. Poslední datovou řadou, kterou máme k dispozici, je intenzita globálního slunečního záření ve W/m^2. Tuto řadu budeme označovat $\widehat{G}_K$. Všechny tyto řady jsou označovány symbolem se stříškou, aby bylo zdůrazněno, že se nejedná o naměřené hodnoty, ale o predikce, navíc i z důvodu jejich odlišení od měření stejných jevů. Predikce meteorologických jevů sami provádět nebudeme, proto k záměně symboliky v tomto případě dojít nemůže. U těchto predikcí modelu Aladin předpokládejme, že nejsou systematicky vychýlené a že se jedná o nejlepší možné predikce, tudíž odchylky od skutečnosti jsou minimální možné. Všechny predikce máme k dispozici přibližně na následujících 48 hodin s tím, že jsou aktualizovány jednou za 6 hodin. Tzn. v den D máme k dipozici vždy predikci na celý den D+1 a částečně i na den D+2. Bohužel systém ukládání dat byl navržen tak, že se při každém nahrávání aktualizované predikce původní hodnoty přepíšou. Proto pro každou hodinu máme k dispozici pouze nejaktuálnější predikci z modelu Aladin, která ale při modelování výroby na následující dny nemusí korespondovat s predikcí z modelu Aladin, kterou jsme měli v daný den k dispozici při předpovědi na dny $D + 1$ a $D + 2$.

2.3.2 Data naměřená na stanicích ČHMÚ

Druhou skupinou dat od ČHMÚ je měření vybraných meteorologických jevů ze 12 meteorologických stanic ČHMÚ. Zde máme k dipozici z každé stanice měření teploty vzduchu ve dvou metrech nad zemí ve °C. Tyto řady budeme označovat t_S. Dále máme k dispozici měření celkové a vyšší vrstvy oblačnosti. Tato měření nabývají devíti hodnot numericky značených $0, 1, ..., 8$, kdy 0 znamená jasno a 8 zataženo. Tyto řady budeme značit Nc_S a Nh_S. Dále máme k dispozici měření slunečního svitu. Toto měření nabývá jedenácti hodnot numericky značených $0, 1, ..., 10$, kdy 0 znamená nulovou intenzitu slunečního svitu, zatímco 10 maximální intenzitu slunečního svitu. Všechna tato měření jsou pro hodinu H k dispozici v průběhu následující hodiny $H + 1$. Přesný čas se liší a závisí na zpracování těchto údajů v ČHMÚ, jejich odeslání a zpracování v ČEZ.

Z jednotlivých stanic máme k dispozici měření srážek. To je ale bohužel k dispozici až v nočních hodinách následujícího dne, což je pro predikci výroby pozdě, proto tyto časové řady nebudeme v modelech používat a ani pro ně zavádět nějaké označení.

2.4 Kontrola a očištění dat

Před tím, než začneme s daty pracovat, provedeme jejich kontrolu a očištění.

2.4.1 Data ČEZ

Nejprve si prohlédneme data od ČEZ. Zde jsme již při převodu řad na řady p_R a q_R provedli základní očištění. Hodnoty menší než 0,000 1 byly nahrazeny 0. V tomto případě se jednalo o očištění o nízké naměřené hodnoty v nočních hodinách, kdy se jedná nejspíše o chybu. Při výpočtu řady s_R jsme v případě, kdy $q_{R,i} = 0$ položili $s_{R,i} = 0$. Následně jsme z dat ručně nahradili několik nenulových hodnot řady p_R v nočních hodinách 0, v tomto případě se jednalo o chybu, která nebyla očištěna již při převodu. Při načtení dat do softwaru R a použití funkce summary zjistíme, že data neobsahují žádná chybějící pozorování a že se v datech nevyskytují žádná pozorování mimo interval $[0\,;1]$. V datech ČEZ se nenacházejí žádné zjevné chyby. Pokud se nějaké vyskytují, narazíme na ně nejspíše až při vytváření modelů v případě jejich nezvyklého chování na určité skupině dat.

Jediné omezení, které z dat můžeme vypozorovat je, že v regionech SC a SM jsou $\forall\, d \in D \wedge \forall\, h \in H$ řady $q_{SC,2012,d,h} = q_{SM,2012,d,h} = 0$. V těchto dvou regionech nejspíše „online“ měření bylo k dispozici až od roku 2013, s čímž bychom měli při vytváření modelů počítat.

2.4.2 Data ČHMÚ

Nyní provedeme kontrolu dat od ČHMÚ. Kontrolu opět provedeme ve dvou skupinách. Při načtení a kontrole predikcí modelu Aladin v softwaru R zjistíme, že predikce neobsahují chybějící hodnoty, pro oblačnosti se všechny hodnoty pohybují v intervalu $[0\,;1]$, u slunečního záření se všechny hodnoty pohybují v intervalu $[0\,;1\,000]$ a teploty v intervalu $[-20\,;40]$, což v obou případech odpovídá

fyzikálním předpokladům. V predikcích modelu Aladin nenajdeme žádné zjevné chyby.

U časových řad představujících měření meteorologických jevů se nejprve podíváme na validitu jednotlivých hodnot. Teploty se pohybují v intervalu $[-25\,;40]$, což odpovídá fyzikálním předpokladům, měření slunečního svitu a oblačností nenabývají jiných než povolených hodnot, takže z tohoto hlediska jsou data v pořádku. Bohužel v tomto případě již nemáme data kompletní, ve všech řadách se objevují chybějící pozorování. V případě teplot se jedná o náhodné výskyty, které byly nejspíše způsobené poruchou měřících přístrojů. V případě měření slunečního svitu je u každé stanice přibližně 11 800 chybějících hodnot, ale pokud z dat vynecháme pozorování, kdy $H \leq 4$ nebo $H \geq 20$, zjistíme v tomto výřezu již minimum chybějících hodnot. Chybějící hodnoty, tedy až na drobné výjimky, připadají na noční hodiny, kdy slunce nesvítí, tudíž nám chybějící hodnoty nemusí vadit. V případě měření oblačnosti je počet chybějících hodnot přibližně 20 000 s výjimkou stanic Brno - Tuřany, Ostrava - Mošnov a Karlovy Vary, kde chybí přibližně 1 000 hodnot u celkové oblačnosti a přibližně 4 000 hodnot u vyšší vrstvy oblačnosti. Nižší počet chybějících pozorování je dán tím, že tyto tři stanice se nacházejí na mezinárodních letištích, kde měření oblačnosti probíhá i v noci. Při vizuálním prohlédnutí dat z ostatních stanic můžeme stejně jako u slunečního svitu konstatovat, že většina chybějících pozorování se objevuje v nočních hodinách, kdy přístroje nejsou schopny oblačnost měřit, pro predikce tyto chybějící hodnoty nebudou podstatné. Bohužel je ale nutné podotknout, že se zde objevují ve větší míře chybějící hodnoty i v době slunečního svitu.

2.5 Volba dat v regionech

Protože data ČHMÚ máme k dispozici podle krajů případně z jednotlivých stanic ČHMÚ, bude potřeba přiřadit k jednotlivým regionům R vhodná data z krajů K a stanic S. Protože predikování výroby v regionu R na základě meteorologických dat ze stanic ČHMÚ nebo krajů, které v daném regionu neleží, nedává žádný logický význam, zaměříme se pouze na volbu dat ze stanic ČHMÚ a krajů, které územně spadají do daného regionu. Určit, do jakého regionu data spadají, není náročné. Po rozdělení dat nastává problém v tom, že se nám v jednotlivých regionech objevila data, která jsou závislá a v případě zařazení obou časových řad do modelu budeme mít problém s multikolinearitou. Vypočteme pro všechny časové řady, které spadají do stejného regionu a které se týkají stejného meteorologického jevu, Pearsonovy korelační koeficienty $\rho_{X,Z} = \mathrm{cov}(X,Z)/(\sigma_X \cdot \sigma_Z)$.

Pro časové řady týkající se teploty dostáváme jak u předpovědí modelu Aladin, tak i u měření všechny korelační koeficienty větší než 0,96, které potvrzují předpoklad závislosti. Podobné korelační koeficienty pozorujeme i u predikce intenzity globálního slunečního záření. V případě měření intenzity slunečního svitu se korelační koeficienty pohybují okolo hodnot 0,8, což je nejspíše způsobeno tím, že měření nabývá pouze 11 hodnot. U oblačností se již jednotlivé korelační koeficienty pohybují v rozmezí hodnot 0,65 až 0,98 podle regionů a typu oblačností, takže i tady můžeme očekávat (zejména u predikcí) lineární závislost. Jedinou výjimkou je $\rho_{Nh_{TE},Nh_{BR}} = 0{,}495$, zde je ale nižší korelační koeficient způsoben velkou vzdáleností obou stanic ČHMÚ.

Máme tři možnosti, jak se se závislostí potenciálních vysvětlujících proměnných vypořádat:

1. pro každý region si zvolíme jeden kraj a jednu stanici, ze které budeme data používat a ostatní zanedbáme,
2. vytvoříme novou proměnnou, která vznikne váženým průměrem příslušných proměnných,
3. použijeme všechny proměnné a budeme zkoumat, zdali nemá model problém s multikolinearitou (tuto možnost má smysl uvažovat pouze v případě, kdy Pearsonův korelační koeficient není blízký jedné).

Konkrétní volbu provedeme až při predikování a to pouze u těch časových řad, které se do modelu rozhodneme zahrnout.

2.6 Přehled dat

Na závěr ještě přiložíme tabulku 2.3 s přehledem značení jednotlivých časových řad. V této tabulce jsme označili typ řady následovně: E = řady související s výkonem FVE, P = predikce modelu Aladin, M = měření ze stanic ČHMÚ.

Typ	*Značení*	*Název řady*	*Jednotka*
E	P_R	naměřená výroba FVE	kWh
E	V_R	instalovaný výkon FVE	kW
E	p_R	procentuální množství naměřeného výkonu FVE z instalovaného výkonu	%
E	Q_R	naměřená „online“ výroba vybraných FVE	kWh
E	W_R	instalovaný výkon vybraných FVE	kW
E	q_R	procentuální množství naměřeného „online“ výkonu vybraných FVE z instalovaného výkonu	%
E	s_R	procentuální množství instalovaného výkonu vybraných FVE oproti instalovanému výkonu všech FVE	%
P	$\widehat{t}_K$	teplota vzduchu ve 2 metrech nad zemí	°C
P	$\widehat{Nc}_K$	celková oblačnost	%
P	$\widehat{Nl}_K$	nízká oblačnost	%
P	$\widehat{Nm}_K$	střední oblačnost	%
P	$\widehat{Nh}_K$	vysoká oblačnost	%
P	$\widehat{G}_K$	intenzita globálního slunečního záření	W/m^2
M	t_S	teplota vzduchu ve 2 metrech nad zemí	°C
M	Nc_S	celková oblačnost	%
M	Nh_S	vysoká oblačnost	%
M	G_S	intenzita slunečního svitu	%

Tabulka 2.3: Přehled všech časových řad včetně značení.

3. Statistické metody

V této kapitole budou stručně popsány statistické metody, které mohou být některým čtenářům neznámé a budou použity při vytváření predikcí.

3.1 Kvantilová regrese [4]

Obvyklou statistickou úlohou je modelovat závislost vysvětlované náhodné veličiny Y v závislosti na vysvětlujících náhodných veličinách $X_1, ..., X_n$. Na základě znalosti naměřených hodnot $\mathbf{X} = (X_1, ...X_n)^\top$ chceme provést predikci hodnot závislé proměnné Y, tedy se snažíme najít nějakou vhodnou funkci $g(\mathbf{X})$ takovou, která závislou proměnnou Y dobře aproximuje. Kvalitu predikce posuzujeme podle ztrátové funkce $L(u)$.

V klasické regresi se volí kvadratická ztrátová funkce $L(u) = u^2$. Za g nejčastěji volíme lineární funkci $g(\mathbf{X}) = X_1\beta_1 + ... + X_n\beta_n$. Výsledek měření bývá v praxi vždy zatížen náhodnou chybou $\varepsilon_1, ..., \varepsilon_n$. Získané hodnoty můžeme sestavit do maticové rovnice $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Nyní je třeba odhadnout $\boldsymbol{\beta}$ za podmínky, že ztrátové funkce $L(\varepsilon_i)$ jsou minimální. Tedy dostáváme odhad

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{k} L(\varepsilon_i) = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\varepsilon})$$

a při použití kvadratické ztrátové funkce dostaneme

$$S(\boldsymbol{\varepsilon}) = \sum_{i=1}^{k} \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon},$$

což vede k podmíněné střední hodnotě. V tomto případě můžeme odhad $\boldsymbol{\beta}$ vyjádřit vztahem $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, který se nazývá odhad metodou nejmenších čtverců.

Střední hodnota obvykle není jediná charakteristika, kterou chceme o náhodné veličině vědět, neboť nám to nestačí pro určení rozdělení. Znalost všech momentů je pro charakteristiku rozdělení náhodné veličiny dostačující. V tomto případě můžeme vypočítat všechny kvantily rozdělení.

Definice 1. *Nechť $F(x) = P(X \leq x)$ je distribuční funkce daného rozdělení pravděpodobnosti náhodné veličiny X a $\alpha \in (0,1)$. Potom funkce*

$$Q(\alpha) = F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$$

se nazývá kvantilová funkce *a číslo $y_\alpha = Q(\alpha)$ se nazývá* α-kvantil *rozdělení s distribuční funkcí $F(x)$.*

Z definice je zřejmé, že α-kvantil y_α rozděluje definiční obor náhodné veličiny X na dvě části tak, že platí

$$P(X \leq y_\alpha) = \alpha \text{ a } P(X \geq y_\alpha) = 1 - \alpha$$

[4] Někdy též známá pod pojmem expektilová regrese

Další možností je nalézt kvantily rozdělení přímo. Kvantilová regrese byla poprvé představena v práci Koenker a Bassett [2]. Regresní kvantil, jak je prezentováno v [2], je lineární regrese s n vysvětlujícími proměnnými

$$\widehat{Q}(\alpha, X_t) = \widehat{\beta}_1(\alpha)x_{1,t} + ... + \widehat{\beta}_n(\alpha)x_{n,t} = X^\top \widehat{\boldsymbol{\beta}}(\alpha), \tag{3.1}$$

kde $\widehat{Q}(\alpha)$ je α-kvantil, X_1 obvykle bývá vektor jedniček, tedy β_1 je intercept. Položíme-li za ztrátovou funkci

$$L(\varepsilon) = \rho_\alpha(\varepsilon) = \begin{cases} \alpha\varepsilon, & \varepsilon \geq 0, \\ (1-\alpha)\varepsilon, & \varepsilon < 0, \end{cases}$$

kde ε je residuum, tj. $\varepsilon_i = y_i - \widehat{Q}(\alpha, X_i)$. Odhad $\boldsymbol{\beta}$ získáme minimalizací $\sum_i \rho_\alpha(\varepsilon_i)$ vzhledem k $\boldsymbol{\beta}$, tudíž dostáváme odhad

$$\widehat{\boldsymbol{\beta}}(\alpha) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{k} \rho_\alpha(\varepsilon_i) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{k} \rho_\alpha(y_i - X^\top \boldsymbol{\beta}).$$

Tato úloha se řeší pomocí metod lineárního programování. Detaily minimalizace lze nalézt např. v Møller [3]. V případě vážené kvantilové regrese, kdy jednotlivá pozorování mají různé váhy ω_i, je nutné odhad doplnit na

$$\widehat{\boldsymbol{\beta}}(\alpha) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{k} \omega_i \rho_\alpha(\varepsilon_i) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{k} \omega_i \rho_\alpha(y_i - X^\top \boldsymbol{\beta}). \tag{3.2}$$

V softwaru `R` můžeme kvantilovou regresi použít funkcí `qr`, kterou najdeme v balíčku `quantreg`, viz článek Koenker [9].

3.2 Beta regrese

Jedná se o regresní model, který byl představen v článku Ferrariová a Cribari-Neto [6]. Jejich cílem bylo navrhnout regresní model, který je vhodný pro situaci, kdy závislá proměnná Y je spojitá a pro její hodnoty platí $0 < Y < 1$. Model je založen na předpokladu, že závislá proměnná Y má beta rozdělení. Hustota beta rozdělení je obvykle dána vztahem

$$f(y; p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \;\; 0 < y < 1,$$

kde $p, q > 0$ a $\Gamma(\cdot)$ je gamma funkce. Ferrariová a Cribari-Neto navrhli jinou parametrizaci, kde $\mu = p/(p+q)$ a $\phi = p+q$:

$$f(y; \mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \;\; 0 < y < 1,$$

kde $0 < \mu < 1$ a $\phi > 0$. V této parametrizaci platí $\mathsf{E}\, Y = \mu$ a $\mathrm{var}(Y) = \mu(1-\mu)/(1+\phi)$. Parametr ϕ je nazýván parametr přesnosti, protože při pevném μ, čím větší je ϕ, tím menší je rozptyl. ϕ^{-1} je rušivý parametr.

Nechť $y_1, ..., y_n$ je náhodný výběr takový, že $y_i \sim \mathcal{B}(\mu_i,\phi)$, $y_i = 1, ..., n$. Model je získán za předpokladu, že střední hodnotu y_i lze zapsat vztahem

$$g_1(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j = X_i^\top \boldsymbol{\beta}, \tag{3.3}$$

kde $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^\top$ je vektor neznámých regresních parametrů ($\boldsymbol{\beta} \in \mathbb{R}^k$), $X_i = (x_{i1}, ..., x_{ik})^\top$ je vektor k vysvětlujících proměnných ($k < n$), obvykle $x_{i1} = 1$ pro všechna i, tudíž model má intercept. $g_1(\cdot) : (0,1) \mapsto \mathbb{R}$ je linková funkce, která je striktně rostoucí a dvakrát diferencovatelná. Motivace pro zavedení linkové funkce je dvojí. Za prvé, obě strany regresní rovnice předpokládají hodnoty na celé reálné ose, když je linková funkce aplikována na μ_i. Za druhé je zde přidána dodatečná flexibilita, kdy si uživatel může zvolit funkci, která nejlépe odhaduje model.

Vhodné linkové funkce jsou např. logit $g(\mu) = \log(\mu/(1-\mu))$; probit $g(\mu) = \Phi^{-1}(\mu)$, kde $\Phi(\cdot)$ je distribuční funkce standardního normálního rozdělení; komplementární log-log linková funkce $g(\mu) = \log[-\log(1-\mu)]$ a log-log linková funkce $g(\mu) = -\log[-log(\mu)]$ a další. Obzvláště užitečnou linkovou funkcí je logit, kdy v tomto případě dostaneme

$$\mu_i = \frac{e^{X_i^\top \boldsymbol{\beta}}}{1 + e^{X_i^\top \boldsymbol{\beta}}}.$$

Všimněme si, že rozptyl Y je funkcí μ, což činí z regresního modelu založeného na této parametrizaci heteroskedastický model. Odhad parametru $\boldsymbol{\beta}$ dostaneme metodou maximální věrohodnosti. Detaily lze nalézt v článku Ferrariová a Cribari-Neto [6].

V článku Simas a kol. [7] bylo představeno rozšíření beta regresního modelu. V tomto modelu není parametr přesnosti konstantou pro všechna pozorování, ale je modelován podobným způsobem jako parametr střední hodnoty. Nechť $y_1, ..., y_n$ je náhodný výběr takový, že $y_i \sim \mathcal{B}(\mu_i,\phi_i)$, $y_i = 1, ..., n$. Pro tento model máme rovnice

$$\begin{aligned} g_1(\mu_i) &= \sum_{j=1}^{k} x_{ij}\beta_j = X_i^\top \boldsymbol{\beta}, \\ g_2(\phi_i) &= \sum_{j=1}^{h} z_{ij}\gamma_j = Z_i^\top \boldsymbol{\gamma}, \end{aligned} \tag{3.4}$$

kde $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^\top$ a $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_h)^\top$ jsou vektory neznámých regresních parametrů ($\boldsymbol{\beta} \in \mathbb{R}^k$, $\boldsymbol{\gamma} \in \mathbb{R}^h$), $X_i = (x_{i1}, ..., x_{ik})^\top$ a $Z_i = (x_{i1}, ..., x_{ih})^\top$ jsou vektory k, resp. h vysvětlujících proměnných ($k + h < n$). $g_1(\cdot) : (0,1) \mapsto \mathbb{R}$ je linková funkce, která je striktně rostoucí a dvakrát diferencovatelná, $g_2(\cdot) : (0,\infty) \mapsto \mathbb{R}$ je linková funkce, která je rovněž striktně rostoucí a dvakrát diferencovatelná. Vhodné volby linkové funkce g_2 jsou např. logaritmus $g_2(\phi) = \log\phi$, druhá odmocnina $g_2(\phi) = \sqrt{\phi}$ nebo identická funkce $g_2(\phi) = \phi$ (zde si ale musíme dát pozor, aby odhad byl kladný).

V softwaru `R` je pro model beta regrese možno použít funkci `betareg`, kterou najdeme v balíčku `betareg`. Je-li zadána formule typu `y` $\sim$ `x1` $+$ `x2`, y_i a X_i poskytují odhad střední hodnoty beta regresního modelu. Zde je ϕ_i konstanta, $z_i = 1$ a g_2 je identická linková funkce. Formule odpovídá modelu vyjádřenému vztahem 3.3. Je-li zadána formule typu `y` $\sim$ `x1` $+$ `x2` $\mid$ `z1` $+$ `z2` $+$ `z3`, pak máme stejnou rovnici pro odhad střední hodnoty jako v předchozím případě, parametr přesnosti bude odhadován z části `z1 + z2 + z3`. Tato formule odpovídá modelu vyjádřeného vztahy 3.4. Další parametry funkce `betareg` a ostatní funkce stejnojmenného balíčku nalezneme v článku Zeilers [10].

4. Clear sky model

Časové řady procentuálního množství vyrobené elektřiny z maximálního výkonu $p_R = \{p_{R,i}, i = 1, ..., I\}$, kde $I = 47\,448$, mohou být chápany jako realizace náhodných procesů s diskrétním časem. Abychom mohli použít obvyklé metody pro analýzu časových řad, je potřeba prozkoumat charakteristiky těchto řad. Většina obvyklých metod předpokládá stacionaritu. Časové řady $\{p_R\}$ ji nesplňují, neboť např. střední hodnota je závislá na hodině a dni realizace pozorování. Z toho vyplývá, že n-dimenzionální distribuční funkce náhodného procesu není invariantní na změny v čase a procesy nejsou stacionární. Z tohoto důvodu bude vhodné transformovat časové řady $\{p_R\}$ tak, abychom obdrželi stacionární procesy.

Časové řady $\{p_R\}$ rozdělíme na deterministickou a stochastickou část

$$p_{R,i} = cs_R(i) \cdot \tau_{R,i}, \tag{4.1}$$

kde $\tau_{R,i}$ je náhodný proces a $cs_R(i)$ je deterministická funkce, která udává procentuální množství vyrobené elektřiny z maximálního výkonu za předpokladu jasné oblohy (v angl. clear sky) v čase i. Jinými slovy $cs_R(i)$ představuje maximální možný procentuální dosažitelný výkon fotovoltaických elektráren v čase i za ideálních povětrnostních podmínek.

Protože poloha slunce na obloze je deterministickou funkcí času a je periodická s periodou 1 rok můžeme $cs_R(i)$ modelovat jako $cs_R(D,H)$, jelikož pro $cs_R(i)$ platí implikace

$$i_1 = i_2 + k \cdot 8\,760,\ k \in \mathbb{Z} \Rightarrow cs_R(i_1) = cs_R(i_2).$$

Transformaci procesu $\{p_{R,i}\}$ na proces $\{\tau_{R,i}\}$ získáme vztahem

$$\tau_{R,i} = \frac{p_{R,i}}{cs_R(i)}. \tag{4.2}$$

Abychom mohli transformaci provést, je nutné najít metodu, jak získat $cs_R(i)$. V případě jasné oblohy, kdy platí

$$\tau_{R,i} = 1 \Rightarrow cs_R(i) = p_{R,i},$$

můžeme $cs_R(i)$ pozorovat přímo z $\{p_{R,i}\}$, ale bohužel takové povětrnostní podmínky nejsou po celý rok. Je potřeba najít metodu pro odhad $cs_R(i)$ v případě, kdy $\{\tau_{R,i}\}$ jsou různé od 1.

Je nutné poznamenat, že v případě, kdy $p_{R,i} = 0$, tj. v době, kdy v příslušném regionu na zemi nedopadá žádné sluneční záření, není možné $\tau_{R,i}$ pozorovat. Dále v období úsvitu a soumraku, kdy $cs_R(i) \longrightarrow 0$ mohou mít šum v $p_{R,i}$ a chyba v odhadu $cs_R(i)$ významný vliv na $\tau_{R,i}$.

4.1 Odhad Clear sky modelu

Metod, jak získat Clear sky model, je více. Nejjednodušší je najít jasné dny v měření a poté jednoduchou interpolací mezi těmito dny dostaneme odhad Clear sky modelu. Další možností je vypočítat Clear sky model teoreticky. V případě

znalosti zeměpisné délky a šířky umístění dané elektrárny, jejího směru a úhlu sklonu k zemskému povrchu lze teoreticky vypočítat úhel dopadajícího slunečního záření na povrch elektrárny a odsud určit maximální dosažitelný výkon v danou denní a roční dobu.

Další možností jak Clear sky model získat dostaneme, když si uvědomíme, že 100% kvantil, případně kvantil blízký 100 %, podmíněného rozdělení $p_{R,i}$ ve dni D a hodině H je hledaná hodnota Clear sky modelu. Můžeme tedy použít kvantilovou regresi pro odhad kvantilu podmíněného rozdělení $p_{R,i}$. Protože pro dané D a H máme málo pozorování, navíc není zaručeno, že pro každou dvojici (D,H) nastaly pro alespoň jedno pozorování ideální podmínky (jasná obloha), využijeme toho, že pozorování v sousedních dnech a hodinách jsou závislá. Inspirací pro použití kvantilové regrese byla práce Bachera [5], kapitola 4.

Mějme dán 3-dimenzionální prostor (a,b,Z), kde Z je náhodný proces a (a,b) systém souřadnic. Při odhadu $cs_R(D,H)$ máme $Z = \{p_{R,i}\}$, $a = D$ a $b = H$. Cílem je najít odhad podmíněného rozdělení daného hustotou $f_Z(a,b)$, v našem případě potřebujeme pouze znát příslušný kvantil, který najdeme pomocí kvantilové regrese. Pro odhad kvantilu pro danou dvojici (a,b) použijeme všechna pozorování náhodného procesu Z s tím, že jim přiřadíme váhy podle toho jak moc jsou „daleko“ od dané dvojice (a,b). Toto je možné udělat, protože existuje lokální závislost mezi $f_Z(a,b)$ a $f_Z(a + x_a,b + x_b)$, která roste s tím jak x_a a x_b klesají.

Příslušné váhy získáme pomocí 2-dimenzionálního jádrového odhadu. Za jádrovou funkci zvolíme normované normální rozdělení, tedy pro bod a_i dostaneme

$$\omega(a,a_j,l) = f_N\left(\frac{|a - a_j|}{l}\right),$$

kde f_N je hustota normovaného normálního rozdělení. Toto použijeme v 2-dimenzionální jádrové funkci

$$k(a_j,b_j) = \frac{\omega(a,a_j,l_a) \cdot \omega(b,b_j,l_b)}{\sum_{i=1}^{n} \omega(a,a_j,l_a) \cdot \omega(b,b_j,l_b)}, \tag{4.3}$$

kde l_a a l_b jsou vyhlazovací parametry postupně pro dimenzi a a b. Vyhlazovací parametry řídí váhu mezi vychýlením a rozptylem v modelu. V případě nízkých hodnot bude model málo vychýlený, ale bude mít větší rozptyl a bude „nevyhlazený“. Budou-li vyhlazovací parametry velké, bude model vychýlený a zejména v obdobích s menším potenciálním výkonem budou výsledky nadhodnocené. Při odhadu modelu se omezíme pouze na data za období let 2012 až 2016 a to z důvodu nekompletního roku 2017. Pokud bychom neměli pro všechny hodiny v roce stejný počet pozorování, byl by jmenovatel ve výrazu 4.3 pro každý den D různý, což by vedlo v průběhu roku k různým váhám i v případě shodné hodnoty v čitateli výrazu 4.3, což není úplně žádoucí. Navíc v okolí dne 31. května by měly v sumě větší váhu pozorování z předchozích dní než ze stejného počtu dní následujících.

4.1.1 Vážená kvantilová regrese

Pro odhad α-kvantilu $f_Z(a,b)$ použijeme váženou kvantilovou regresi s váhami z jádrového odhadu. V tomto případě máme úlohu (3.1) zjednodušenou na odhad

konstanty

$$\widehat{Q}(\alpha, Z) = \widehat{\beta}(\alpha).$$

Odhadem kvantilu je dle (3.2)

$$\widehat{\beta}(\alpha) = \arg\min_{\beta} \sum_{j=1}^{I} k(a_j,b_j)\rho_\alpha(\varepsilon_i) = \arg\min_{\beta} \sum_{j=1}^{I} k(a_j,b_j)\rho_\alpha(z_j - \beta),$$

kde příslušné váhy $k(a_j,b_j)$ jsou jádrovou funkcí.

4.1.2 Volba parametrů modelu

Neméně důležité při odhadu $cs_R(D,H)$ je volba parametrů (α, l_a, l_b), jejichž volba bude mít velký vliv na to, jak kvalitní odhad získáme. Bohužel v tomto případě je těžké porovnat kvalitu odhadu, neboť skutečné hodnoty $cs_R(D,H)$ neznáme (pokud bychom znali, pak bychom je vůbec nemuseli odhadovat). Pro částečné porovnání můžeme použít jednoduchý odhad, kdy pro každý den a hodinu použijeme nejvyšší naměřený výkon v daném dni a hodině za celé období, tedy

$$\widehat{cs}_{max,R}(D,H) = \max_{y \in Y} p_{R,y,D,H}.$$

Tento odhad nazveme maximálním modelem. Porovnání maximálního modelu s Clear sky modelem můžeme vidět na obrázku 4.1.

Vzhledem k tomu, že máme data za 5 let, má smysl volit parametr α blízko jedné a parametry l_a a l_b blíže k 0. Protože v období úsvitu s rostoucím H a v době soumraku s klesajícím H hodnoty $cs_R(D,H)$ velmi rychle rostou, vede volba $l_b > 0{,}05$ k výrazně nadhodnocenému odhadu $cs_R(D,H)$ v těchto hodinách. Protože při volbě $l_b \leq 0{,}05$ jsou váhy u pozorování $p_{R,Y,D,h}$, kde $h \neq H$ velmi malé, odhad nezhoršíme, položíme-li v dimenzi b

$$\omega(b,b_j,l_b) = \begin{cases} 1, & b_j = b, \\ 0, & b_j \neq b. \end{cases}$$

Toto povede k tomu, že celý proces kvantilové regrese zjednodušíme tím, že bude stačit z náhodného procesu $\{p_{R,i}\} = \{p_{R,y,d,h}\}$ vybrat ta pozorování, kde $h = H$. Tímto přejdeme do 2-dimenzionálního prostoru (a,Z), kde máme $Z = \{p_{R,y,d,H}\}$ a $a = D$, jádrová funkce se zjednoduší na

$$k(a_j) = \frac{\omega(a,a_j,l_a)}{\sum_{j=1}^{n} \omega(a,a_j,l_a)}$$

a počet parametrů snížíme na dva.

Pro omezení škály vhodných parametrů (α, l_a) je vhodné nejprve použít vizuální porovnání odhadu kvantilovou regresí $\widehat{cs}_{rq,R}(\alpha, l_a)$ s maximálním odhadem $\widehat{cs}_{max,R}$. Nejprve se zaměříme na volbu parametru α, tedy jaký kvantil v regresi zvolit. Při volbě kvantilu $\alpha < 0{,}97$ bez ohledu na volbu l_a již dochází k tomu, že odhad Clear sky modelu kvantilovou regresí je ve slunečných dnech v průběhu celého roku podhodnocen, proto se při volbě parametru l_a zaměříme pouze na hodnoty $\alpha \in [0{,}97\,;0{,}995]$. Pro hodnoty α velmi blízko 1 je model v zimních měsících výrazně nadhodnocen.

Nyní se podíváme, jaké hodnoty parametru l_a je možné uvažovat. Zvolíme-li $l_a > 10$, pak je opět odhad kvantilovou regresí v zimním období ve všech hodinách nadhodnocen, zejména v ranních a večerních hodinách je již odhad nedostatečný. Při volbě $l_a < 4$ vypadá při vizuálním porovnání odhad kvantilovou regresí jako kvalitní, ale bohužel odhad působí „nevyhlazeně“ a schodovitě, v ročním průběhu je možné pozorovat několik lokálních minim a maxim, které by se v případě teoretického výpočtu Clear sky modelu neobjevily.

4.1.3 Pomocná kritéria pro nalezení nejlepšího modelu

Pomocí vizuálního porovnání modelů $\widehat{cs}_{rq,R}(\alpha, l_a)$ a $\widehat{cs}_{max,R}$ jsme se dostali k omezení parametrů (α, l_a) na kartézský součin intervalů $[0{,}97\,;0{,}995] \times [4\,;10]$. Vzhledem k tomu, že mezi dvojicemi parametrů z výše uvedeného kartézského součinu již lze vizuálně jen obtížně pozorovat rozdíly v kvalitě modelů $\widehat{cs}_{rq,R}(\alpha, l_a)$, zavedeme pět pomocných kritérií, které nám mohou pomoci určit nejvhodnější volbu parametrů. Jedná se o těchto pět kritérií:

1. počet dní, kdy je odhad $\widehat{cs}_{rq,R}(\alpha, l_a)$ menší než $\widehat{cs}_{max,R}$,
2. maximální rozdíl mezi odhady $\widehat{cs}_{max,R}$ a $\widehat{cs}_{rq,R}(\alpha, l_a)$ a to pouze ve dnech, kdy je první z odhadů větší,
3. minimální rozdíl mezi odhady $\widehat{cs}_{max,R}$ a $\widehat{cs}_{rq,R}(\alpha, l_a)$ v libovolném souvislém dvacetidenním intervalu za podmínky, že je první odhad ve všech dnech v tomto intervalu menší než druhý odhad,
4. v období prvních 100 a posledních 120 kalendářních dní by se neměly vyskytovat žádné lokální extrémy[5],
5. pro každý den by měl odhad do 12. hodiny včetně růst, od 13. hodiny včetně klesat.

První kritérium má za cíl korigovat, aby odhad kvantilovou regresí nebyl systematicky nižší, než skutečnost. Druhé kritérium hlídá, aby ve dnech, kdy je odhad nižší, nebyl rozdíl oproti skutečnosti velký. Třetí kritérium naopak hlídá, aby nebyl odhad kvantilovou regresí v některém období výrazně vyšší než je skutečnost. Dvacetidenní interval byl zvolen z důvodu, že se v datech vyskytuje období 18 po sobě jdoucích dní, ve kterém v letech 2012 až 2016 nebyl zaznamenán žádný jasný den. Čtvrté kritérium odpovídá fyzikální skutečnosti, kdy v tomto období je skutečný Clear sky model do dne zimního slunovratu striktně klesající a ode dne zimního slunovratu striktně rostoucí. Poslední kritérium kontroluje průběh v jednotlivých dnech a kontroluje výskyt nežádoucích lokálních extrémů.

Protože pro každou hodinu H provádíme odhad na jiných datech, je možné mít pro každou hodinu různou volbu parametrů. V tomto případě by ale bylo vhodné, kdyby pro různé hodiny H byla volba parametrů řízená nějakou funkcí závislou na H. Z výsledků pro jednotlivé parametry je možné vyčíst, že pro ranní a večerní hodiny je vhodnější volba parametru α blíže dolní hranici, neboť při vysokých hodnotách je v zimním období odhad nadhodnocený, pro hodiny okolo

[5]Lokální minimum, které se vyskytuje okolo 21. prosince a je zároveň globálním minimem mezi tyto extrémy počítat nebudeme

poledne je vhodnější volit parametr α vyšší, neboť při nižších hodnotách je naopak v letním období odhad podhodnocený. Důvodem je skutečnost, že v ranních a večerních hodinách, kdy je maximální možný výkon elektráren nižší, je mezi dnem s jasnou a dnem se zataženou oblohou, kdy se např. výkon pohybuje okolo 30 % maximálního dosažitelného výkonu v danou dobu, menší absolutní rozdíl ve výkonu, zatímco v poledních hodinách, kdy je maximální dosažitelný výkon elektráren v rámci dne nejvyšší, je mezi absolutní rozdíl ve výkonu mezi dnem s jasnou a dnem se zataženou oblohou větší. Z tohoto důvodu výsledné hodnoty parametru α určíme za pomoci vhodné funkce hodiny H.

Dále je možné z výsledků jednotlivých odhadů vyčíst, že volba l_a blízko dolní hranice vede k nesplnění 4. kritéria, kdy se v poledních hodinách objevují v odhadu lokální minima. Při volbě l_a blízko horní hranice dostáváme při nízkých hodnotách parametru α u 2. kritéria hodnoty, které svědčí o nepříliš vhodné aproximaci, při vysokých hodnotách parametru α se naopak vyskytují vysoké hodnoty u 3. kritéria, což svědčí o výrazně nadhodnoceném odhadu v zimním období.

4.1.4 Závěrečná volba parametrů

Parametr α budeme volit pomocí funkce

$$\alpha(H) = \begin{cases} 0{,}97, & H \in \{1,2,3,4,5,6\}, \\ 0{,}97 + (H-6)/5 \cdot (\alpha_{max} - 0{,}97), & H \in \{7,8,9,10\}, \\ \alpha_{max}(l_a), & H \in \{11,12,13,14\}, \\ 0{,}97 + (19-H)/5 \cdot (\alpha_{max} - 0{,}97), & H \in \{15,16,17,18\}, \\ 0{,}97, & H \in \{19,20,21,22,23,24\}, \end{cases}$$

Parametr α postupně z konstantní hodnoty 0,97 v nočních hodinách lineárně roste během dopoledne až k hodnotě $\alpha_{max}(l_a)$, které dosahuje v 11. až 14. hodině. Poté v odpoledních hodinách lineárně klesá až k hodnotě 0,97, na které zůstane až do 24. hodiny. Maximální hodnotu $\alpha_{max}(l_a)$ volíme tak, aby hodnoty v 1. kritériu nebyly ve všech případech rovny 0, protože v tomto případě by Clear sky model byl systematicky, byť mírně, nadhodnocen. Hodnota $\alpha_{max}(l_a)$ se při různých volbách parametru l_a liší.

Pro parametr l_a je obtížné jednoznačně zvolit nejlepší hodnotu, neboť pro hodnoty v intervalu $[5\,;7{,}5]$ je model i podle všech kritérií dostatečně kvalitní. Přesná volba parametru l_a již závisí na tom, zdali upřednostníme model méně „vyhlazený“ a více schodovitý, ale s mírně menšími hodnotami ve 2. a 3. kritériu, nebo model, který je „vyhlazenější“, méně schodovitý, ale s mírně vyššími hodnotami ve 2. a 3. kritériu.

Pro výsledný Clear sky model jsem se rozhodl vybrat hodnotu $l_a = 6$, které odpovídá hodnota $\alpha_{max}(6) = 0{,}986\,5$. Pro region Střední Čechy najdeme hodnoty jednotlivých kritérií modelu $\widehat{cs}_{rq,ST}(\alpha(H), 6)$ v tabulce 4.1

4.2 Vyhlazení výsledného odhadu Clear sky modelu

Poslední možnou úpravou výsledného modelu bude porovnání odhadu Clear sky modelu kvantilovou regresí $\widehat{cs}_{rq,R}(\alpha(H), 6)$ s odhadem $\widehat{cs}_{sm,R}(\alpha(H), 6)$, který se od původního bude lišit tím, že pro každou hodnotu provedeme roční vyhlazení. Pro zvolený krok délky $\kappa \in \mathbb{N}$ položíme:

$$\widehat{cs}_{sm,R}(D,H) = \frac{\sum_{j=-\kappa}^{\kappa} \widehat{cs}_{rq,R}(D+j,H) \cdot 2^{|j|}}{\sum_{j=-\kappa}^{\kappa} 2^{-|j|}}.$$

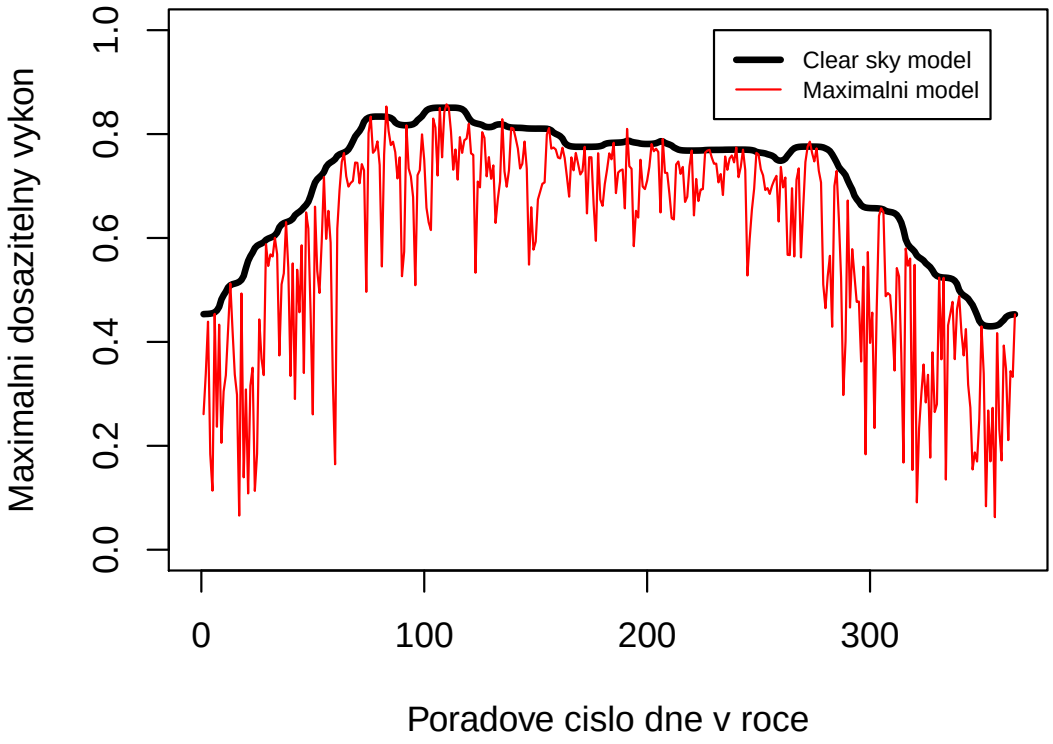
Pro výsledný vyhlazený Clear sky model $\widehat{cs}_{sm,R}$ jsem se rozhodl zvolit $\kappa = 5$. U získaného „hladšího“ odhadu $\widehat{cs}_{sm,R}$ se rovněž podíváme na hodnoty pomocných kritérií a porovnáme je s původním „nevyhlazeným“ odhadem. Hodnoty kritérií najdeme v tabulce 4.1. Z ní můžeme vidět, že cenou za „vyhlazení“ je mírné zhoršení kvality modelu zejména v prvním kritériu, ale u výsledného modelu se v zimních měsících ztratil schodovitý průběh, což lze brát jako adekvátní přínos za cenu mírného zhoršení kritérií.

4.3 Grafické vyobrazení výsledného odhadu Clear sky modelu

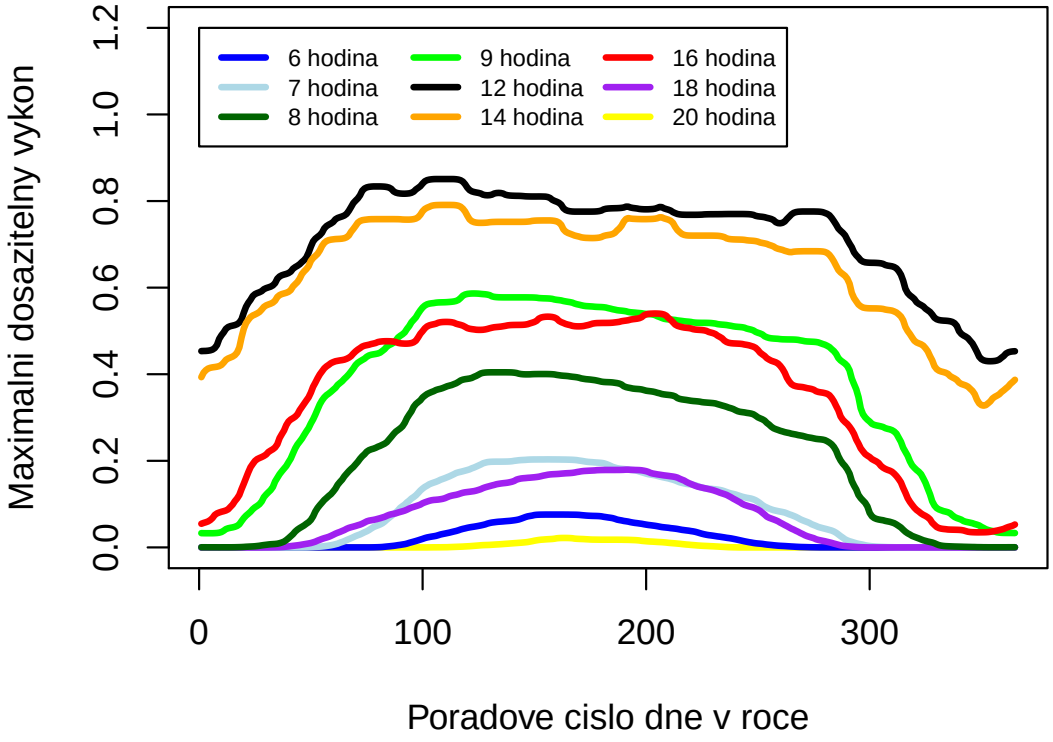
Na závěr této kapitoly dodáme grafické vyobrazení výsledného Clear sky modelu z regionu Středních Čech. V dalších regionech je výsledný Clears sky model podobný, v grafickém zobrazení nerozlišitelný. Nejprve máme na na obrázku 4.1 porovnání maximálního modelu $\widehat{cs}_{max,ST}$ s Clear sky modelem $\widehat{cs}_{sm,ST}(0{,}986\,5, 6)$ ve 12. hodině ve všech dnech v roce. Z obrázku můžeme vidět, že odhad Clear sky modelu respektuje maximální naměřené hodnoty a jeho roční průběh odpovídá očekáváním. Na obrázku 4.2 jsou pro vybrané hodiny vidět denní průběhy Clear sky modelu $\widehat{cs}_{sm,ST}(\alpha(H), 6)$. Postupný nárůst směrem k poledním hodinám a poté pokles směrem k večerním opět odpovídá skutečnosti. Na obrázku 4.3 jsou pro vybrané dny vidět hodinové průběhy Clear sky modelu $\widehat{cs}_{sm,ST}(\alpha(H), 6)$. Grafy pro jednotlivé hodiny mají očekávaný „kopcovitý“ průběh, výše maximálního výkonu v jednotlivých dnech opět neodporuje představám. Na obrázku 4.4 je 3D model Clear sky modelu $\widehat{cs}_{sm,ST}(\alpha(H), 6)$ a nakonec na obrázku 4.5 je jeho konturový diagram. Vizuálním pohledem se i na těchto obrázcích můžeme přesvědčit, že výsledný Clear sky model odpovídá představám, které jsme o něm měli před začátkem modelování, akorát by možná bylo vhodné lokálně provést drobné ruční opravy pro „hladší“ tvar modelu.

	Nevyhlazený model			**Vyhlazený model**		
Hod.	*Krit. 1*	*Krit. 2*	*Krit. 3*	*Krit. 1*	*Krit. 2*	*Krit. 3*
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	1	0,0001	0	1	0,0001	0
5	6	0,0021	0,0016	8	0,002	0,0016
6	7	0,0073	0,0057	11	0,0073	0,005
7	4	0,0035	0,0142	5	0,0036	0,0132
8	9	0,0113	0,0324	16	0,01	0,0354
9	12	0,016	0,0355	20	0,0162	0,0339
10	10	0,0126	0,0216	17	0,0126	0,0265
11	8	0,0123	0,029	12	0,0123	0,0265
12	6	0,0194	0,0182	15	0,0228	0,0175
13	7	0,0208	0,0179	15	0,0223	0,0193
14	7	0,014	0,0295	13	0,014	0,0239
15	13	0,0346	0,0328	20	0,0354	0,0247
16	9	0,0334	0,0326	17	0,0331	0,034
17	10	0,0103	0,021	18	0,0101	0,0231
18	7	0,0066	0,016	9	0,0073	0,0154
19	9	0,0121	0,0033	10	0,0122	0,0038
20	5	0,0094	0,0015	8	0,009	0,0016
21	5	0,001	0	7	0,001	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	0	0	0	0	0	0
Krit. 4	0			0		
Krit. 5	0			0		

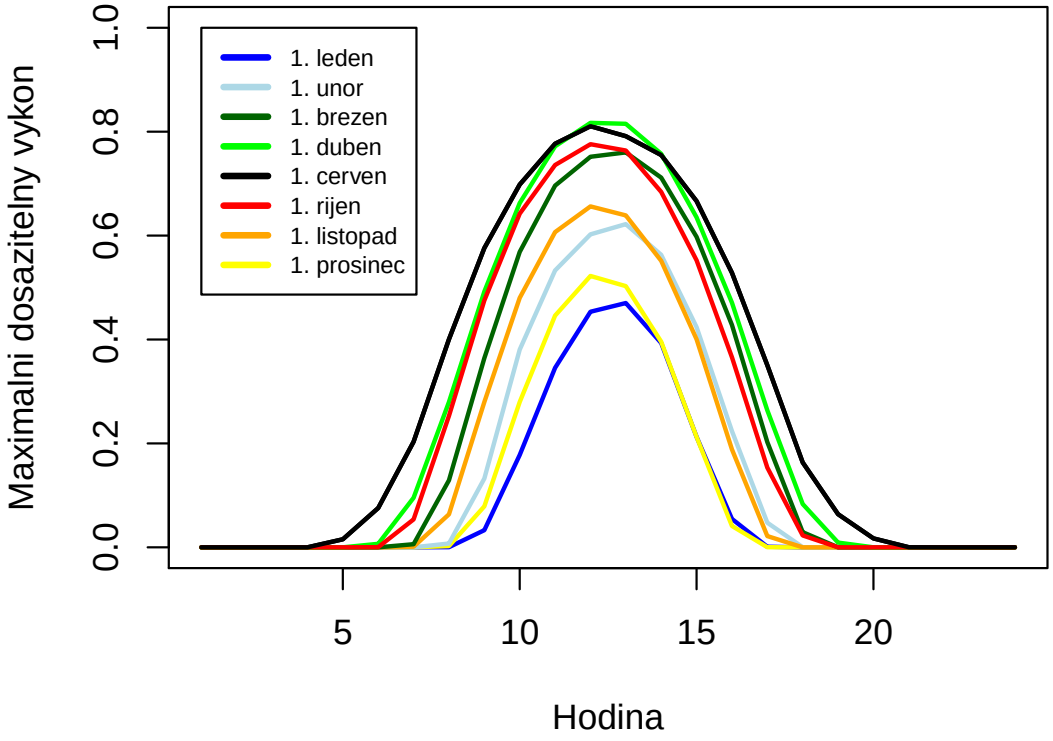
Tabulka 4.1: Porovnání hodnot všech kritérií u Clear sky modelů $\widehat{cs}_{rq,ST}(0{,}986\,5, 6)$ a $\widehat{cs}_{sm,ST}(0{,}986\,5, 6)$.



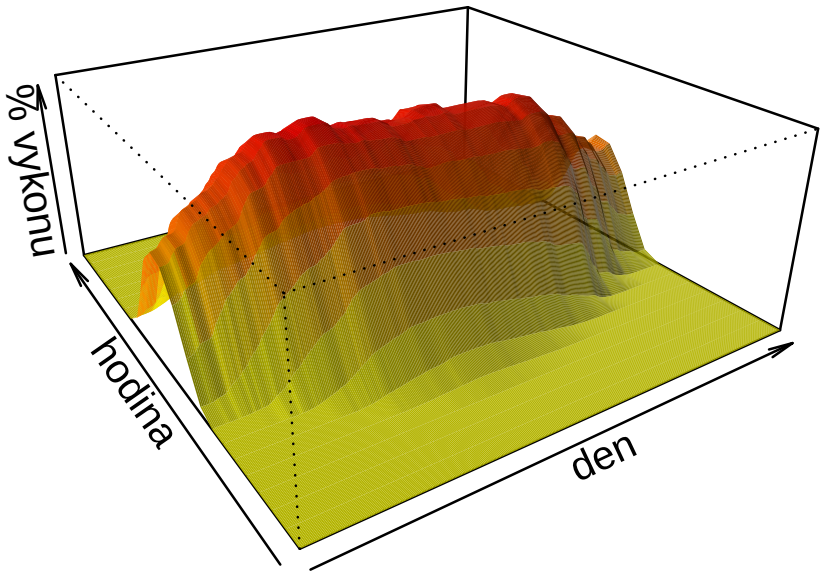
Obrázek 4.1: Porovnání Clear sky modelu $\widehat{cs}_{sm,ST}(0{,}986\,5, 6)$ s maximálním modelem $\widehat{cs}_{max,ST}$ pro $H = 12$.



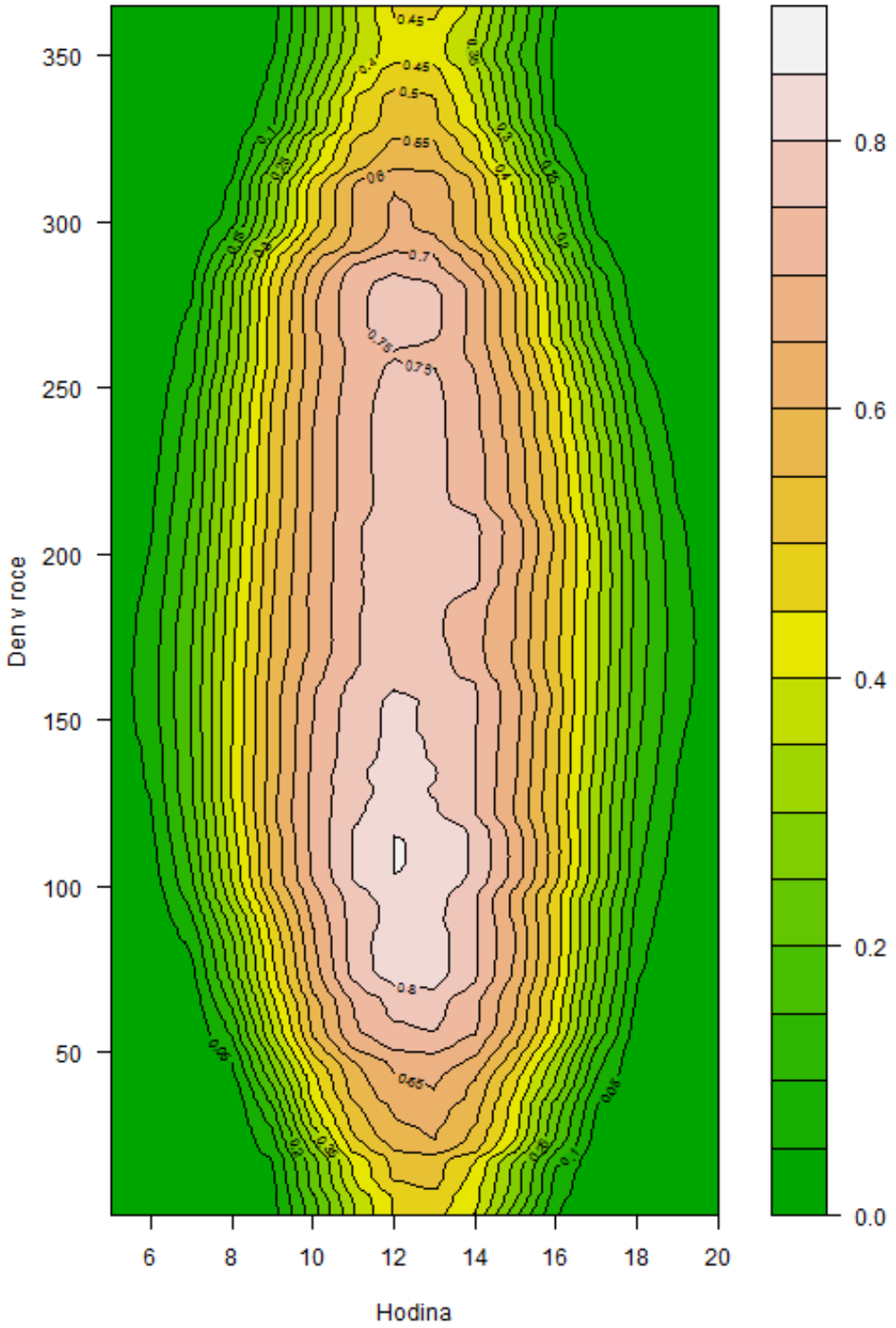
Obrázek 4.2: Denní průběh Clear sky modelu $\widehat{cs}_{sm,ST}(\alpha(H), 6)$ ve vybraných hodinách.



Obrázek 4.3: Hodinový průběh Clear sky modelu $\widehat{cs}_{sm,ST}(\alpha(H), 6)$ ve vybraných dnech.



Obrázek 4.4: 3D model Clear sky modelu $\widehat{cs}_{sm,ST}(\alpha(H), 6)$



Obrázek 4.5: Konturový diagram Clear sky modelu $\widehat{cs}_{sm,ST}(\alpha(H), 6)$

5. Predikce

V této kapitole se zaměříme na modely, které již budou předpovídat výrobu elektřiny z FVE. Nejdříve začneme s modelem na den $D+1$, poté bude následovat model na hodiny $H+2, H+3$ a další ve dni D. Než začneme modely vytvářet, bude potřeba zavést kritéria, jakými budeme potenciální modely mezi sebou porovnávat, jinak bychom těžko mohli posoudit, který z nich je vhodnější.

5.1 Vyhodnocování modelů

Veškeré modely budou vyhodnocovány na základě měření chyb popsaných v této části. Při kontrole správnosti vybraného modelu můžeme pro jednotlivé předpovědi $\widehat{x}_{i+k|i}$ v čase i o kroku k při znalosti skutečné hodnoty x_{i+k} spočítat chybu jako

$$e_{i+k|i} = x_{i+k} - \widehat{x}_{i+k|i}.$$

V případě, že jsme vybrali správný model a zároveň jsme správně odhadli jeho parametry, by měla tato chyba zahrnovat pouze reziduální složku. Chceme-li různé modely souhrnně porovnat, můžeme pro porovnání použít jednu z následujících metod měření chyb. V tomto případě J značí počet uskutečněných předpovědí a index j probíhá postupně přes všechny časy i, ve kterých jsme předpovědi uskutečnili. Můžeme použít buď střední čtvercovou chybu (ozn. MSE)

$$MSE_k = \frac{1}{J}\sum_{j=1}^{J}\left(x_{j+k} - \widehat{x}_{j+k|j}\right)^2 = \frac{1}{J}\sum_{j=1}^{J} e_{j+k|j}^2,$$

odmocninovou střední čtvercovou chybu (ozn. RMSE), která na rozdíl od střední čtvercové chyby vrací výsledky v jednotkách, v jakých jsou předvídané hodnoty

$$RMSE_k = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(x_{j+k} - \widehat{x}_{j+k|j}\right)^2} = \sqrt{\frac{1}{J}\sum_{j=1}^{J} e_{j+k|j}^2},$$

případně střední absolutní chybu (ozn. MAE)

$$MAE_k = \frac{1}{J}\sum_{j=1}^{J}\left|x_{j+k} - \widehat{x}_{j+k|j}\right| = \frac{1}{J}\sum_{j=1}^{J}|e_{j+k|j}|.$$

Dále pro posouzení správnosti modelu použijeme i vychýlení (ozn. BIAS), které ukazuje systematickou chybu modelu

$$BIAS_k = \frac{1}{J}\sum_{j=1}^{J}\left(x_{j+k} - \widehat{x}_{j+k|j}\right) = \frac{1}{J}\sum_{j=1}^{J} e_{j+k|j},$$

kdy kladné hodnoty indikují, že model podhodnocuje skutečnost, záporné hodnoty indikují nadhodnocení skutečnosti. Místo vychýlení můžeme použít procentuální vychýlení (ozn. PBIAS)

$$PBIAS_k = 100 \cdot \frac{\sum_{j=1}^{J}\widehat{x}_{j+k|j} - x_{j+k}}{\sum_{j=1}^{J} x_{j+k}} = 100 \cdot \frac{\sum_{j=1}^{J} -e_{j+k|j}}{\sum_{j=1}^{J} x_{j+k}},$$

které se od vychýlení liší pouze o konstantu, ale umožňuje vzájemně porovnat vychýlení u modelů v případě, že se pozorování, pro která máme předpovědi, liší. V případě PBIAS díky změně znamének v čitateli indikují kladné hodnoty, že model nadhodnocuje skutečnost, záporné hodnoty indikují podhodnocení skutečnosti.

Všimněme si, že veškeré metody počítající chyby modelů počítáme pouze pro předpovědi o stejném kroku k. Toto dává smysl, protože každý model se může se změnou kroku k chovat jinak. Délku kroku k při stavbě modelu známe předem.

V našem případě se jako vhodnější metoda pro porovnávání jeví odmocninová střední čtvercová chyba (dále RMSE), která zachovává jednotky a více penalizuje větší chyby predikce. Současně budeme u modelu kontrolovat i střední absolutní chybu (dále MAE), která odpovídá průměrné hodinové velikosti odchylky. Současně s RMSE a MAE budeme používat i procentuální vychýlení (dále PBIAS), kterým budeme kontrolovat systematickou chybu modelu. V softwaru `R` najdeme funkce `rmse`, `mae` a `pbias` v balíčku `hydroGOF`, viz článek Zambrano-Bigiarini [11].

Pro vzájemné porovnání dvou modelů zavedeme hodnotu IMPR, která bude udávat míru vylepšení (z angl. improvment) zkoumaného modelu oproti vybranému referenčnímu modelu, tedy

$$IMPR_{CH,k} = 100 \cdot \frac{CH_{ref,k} - CH_{zk,k}}{CH_{ref,k}},$$

kde CH je vybraná metoda měření chyb, ref odkazuje na referenční model a zk na zkoumaný model. Budeme-li chtít mezi sebou modely nejen pomocí míry vylepšení porovnávat, bude potřeba určit, na základě jaké časové řady budeme chyby predikcí vyhodnocovat. Model pro predikci výroby elektrické energie může predikce poskytovat buď přímo, tj. predikovat hodnoty $\widehat{p}_{R,j+k|j}$ případně může predikovat hodnoty $\widehat{\tau}_{R,j+k|j}$ a hodnoty $\widehat{p}_{R,j+k|j}$ potom získáme pomocí rovnice 4.1. Pokud bychom porovnávali modely, přičemž jeden z nich predikuje $\widehat{p}_{R,j+k|j}$ a druhý $\widehat{\tau}_{R,j+k|j}$, dostaneme nic neříkající porovnání, protože z něj nelze o vhodnosti modelů vyvodit žádný závěr. Veškeré chyby tudíž budeme vyhodnocovat na základě rozdílů $\widehat{p}_{R,j+k|j} - p_{R,j+k|j}$, které na rozdíl od rozdílů $\widehat{\tau}_{R,j+k|k} - \tau_{R,j+k|j}$ vyjadřují skutečnou výši odchylky.

Modely můžeme mezi sebou porovnat i na různých homogenních podskupinách času i, např. pro $H = 12$. Tato porovnání nám umožní posoudit, zdali není některý model lepší na určité skupině dat nebo naopak.

5.2 Referenční model

Před tím než začneme vytvářet modely, je zapotřebí vytvořit model, který nastaví referenční úroveň pro porovnání chyb u pokročilejších modelů. Jako nejjednodušší model, který budeme považovat za referenční, vezmeme model, kdy pro danou hodinu H vezmeme za predikci poslední známé procentuální množství naměřeného výkonu ze stejné hodiny H, tedy

$$\widehat{p}_{R,j+k|j} = p_{R,j+k-24m}, \tag{5.1}$$

kde m je nejmenší přirozené číslo takové, že $p_{R,j+k-24m}$ již známe. Ve většině případů bude m nabývat hodnoty 1 nebo 2, která bude závislá na tom, jak velký krok k máme.

5.3 Model na den D+1

Byť by mohlo někomu přijít logické začít nejdříve s modelem, který bude předpovídat výrobu na nejbližší hodiny téhož dne, v tomto případě to bude naopak, protože začneme s modelem, který bude předpovídat výrobu elektřiny na jeden den dopředu, neboť tento model je jednodušší. V tomto případě nás nejvíce zajímá předpověď v čase přibližně 1,5 až 2 hodiny před dobou uzávěrky denního trhu, která je v 11 hodin. Tudíž se jedná o predikci v průběhu 9. hodiny SEČ ve dnech platnosti zimního času a o predikci v průběhu 8. hodiny SEČ ve dnech platnosti letního času. Podíváme-li se na data, která v tomto čase máme k dispozici pro vytvoření predikce na den $D + 1$, zjistíme, že z aktuálních měření se jedná o hodnoty, které jsou od doby, kdy budeme výrobu predikovat, vzdálené minimálně o 20 hodin v případě predikce brzkých ranních hodin dne $D+1$, u pozdějších hodin dne $D + 1$ jsou vzdálená o více než 24 hodin. Vzhledem k tomu, že počasí není až tak stabilní, aby se za takovou dobu nemohlo významně změnit (zejména to platí pro všechny typy oblačnosti a dopadající sluneční záření), je nesmyslné do modelů zařazovat jako vysvětlující proměnné jednak měření ze stanic ČHMÚ, tak i „online“ měření výkonu z vybraných FVE. Jako vysvětlují proměnné modelu pro den $D + 1$ budeme uvažovat pouze data z predikcí modelu Aladin. Toto je právě důvod, proč je predikce na den dopředu jednodušší než predikce na nejbližší hodiny.

Vzhledem ke způsobu ukládání dat, kdy máme k dispozici pouze nejaktuálnější predikce modelu Aladin, budeme v průběhu celého dne model na den $D + 1$ vytvářet na základě stejných vysvětlujících proměnných. Z tohoto důvodu je nesmyslné vytvářet model na predikci výroby ve dne $D + 1$ v každé hodině $h \in H$, protože i při zmenšujícím se kroku predikce k se tyto modely nebudou na základě dostupných dat nijak lišit a dospěli bychom ke stejnému výsledku. Přejdeme proto ke zjednodušení označení, kdy časový index i označující čas, ve kterém predikci provádíme, a krok k z označení proměnné vynecháme, neboť v modelech budeme používat pouze data z příslušného dne a hodiny, tudíž nemůže dojít k záměně.

Podrobněji si popíšeme pouze tvorbu modelu ve Středočeském regionu. Tento region byl vybrán díky tomu, že se na jeho území nachází pouze jeden kraj, proto díky tomu nemusíme uvažovat nad tím, z jakého kraje proměnné vybírat. V části 2.5 jsme si již ukázali, že zařazení více proměnných stejného typu ze sousedních krajů způsobí v modelu multikolinearitu. Tvorba modelu u ostatních regionů by probíhala úplně stejným způsobem jako tvorba modelu ve Středočeském regionu. Protože se nedá očekávat, že by v každém regionu měl být jako nejlepší úplně jiný model, stačí pak v ostatních regionech provést pouze odhad regresních parametrů na výsledném modelu pro Středočeský region spolu s výběrem dat z příslušných krajů z daného regionu.

5.3.1 Tvorba modelu na den D+1

Protože je vhodné při ověřování kvality modelu jej otestovat i na datech, která jsme pro jeho tvorbu nepoužili, rozdělíme vstupní data na trénovací a testovací množinu. Do trénovací množiny zařadíme data z let 2012 až 2016, tedy data, kdy $Y \leq 2016$, do testovací množiny zařadíme data z roku 2017.

Při tvorbě modelu na den $D+1$ máme možnost predikovat buď přímo hodnoty

$\widehat{p}_{ST}$, případně můžeme predikovat $\widehat{\tau}_{ST}$ a poté výsledné predikce pomoci vztahu 4.2 převést na $\widehat{p}_{ST}$. Vzhledem k tomu, že se v obou případech jedná o proměnnou vyjádřující procentuální šanci, tudíž jsou hodnoty omezené na interval $[0\,;1]$, bude nutné, aby se predikce z výsledného modelu rovněž nacházely ve stejném intervalu. V tomto případě můžeme použít buď klasickou lineární regresi s tím, že příslušnou vysvětlovanou proměnnou vhodně transformujeme tak, aby nabývala hodnot na celé reálné ose, případně můžeme použít beta regresi. Z těchto úvah vycházejí čtyři možnosti, jak model vytvořit a to buď beta regresí na τ, beta regresí na p, lineární regresí na transformované τ a lineární regresí na transformované p, které postupně prozkoumáme.

Protože tyto metody vyžadují, aby vysvětlovaná proměnná byla na otevřeném intervalu $(0\,;1)$ a nikoli na uzavřeném, provedeme transformaci řad p a τ podle Smithsona a Verkuilena [8], kdy jednotlivá pozorování transformujeme vztahem $x_i = (x_i \cdot (n-1) + 0{,}5)/n$, kde n je velikost celého transformovaného výběru.

Než začneme, podívejme se ještě na hodnoty RMSE a MAE pro referenční model. Při predikci na následující den máme poslední známé hodnoty p z předchozího dne, proto v rovnici 5.1 položíme $l = 2$. Na trénovacích datech dostáváme hodnoty $RMSE_{ref,tr} = 0{,}126\,334$ a $MAE_{ref,tr} = 0{,}058\,09$, na testovacích datech máme hodnoty $RMSE_{ref,test} = 0{,}130\,82$ a $MAE_{ref,test} = 0{,}062\,239$.

5.3.2 Beta regrese na τ

Prvním modelem, který vyzkoušíme, bude beta regresní model s vysvětlovanou proměnnou τ_{ST}. Vzhledem k tomu, že máme pouze 6 vysvětlujících proměnných, nebude chybou, když začneme s úplným modelem, což je model, ve kterém použijeme všechny tyto proměnné. Vzhledem k tomu, že budeme provádět pouze bodové predikce, předpokládejme, že $\phi_i = \phi$ pro všechna i a střední hodnotu lze zapsat vztahem 3.3. Jako linkovou funkci zvolíme logit. Pro úplný model dostaneme

```
 > M1_full <- betareg(tau ~ G + T + Nl + Nc + Nm + Nh, data_ST, link="logit")
> coeftest(M1_full)

z test of coefficients:

               Estimate  Std. Error   z value  Pr(>|z|)
(Intercept) -2.4106e+00  1.8062e-02 -133.4625 < 2.2e-16 ***
G            9.9458e-03  5.8762e-05  169.2558 < 2.2e-16 ***
T            1.5319e-02  7.5889e-04   20.1855 < 2.2e-16 ***
Nl          -1.0690e-01  2.5211e-02   -4.2402 2.233e-05 ***
Nc           2.4373e-01  4.2756e-02    5.7006 1.194e-08 ***
Nm          -2.2091e-02  2.7399e-02   -0.8063  0.420085
Nh          -7.0269e-02  2.6320e-02   -2.6698  0.007589 **
(phi)        1.4008e+00  1.0996e-02  127.3964 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M1_full)*data_ST$CS, data_ST$p)
[1] 0.0746733
> rmse(predict(M1_full, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.08244758
```

Z modelu nejprve vyřadíme proměnnou $\widehat{Nm}_{ST}$, která je podle Waldova testu nevýznamná. Následně zkusíme z modelu vyřadit i proměnnou $\widehat{Nh}_{ST}$, která je nevýznamná na hladině 0,001. Po tomto vyřazení se hodnota RMSE u nového modelu zvětšila jen nepatrně. Následně z odhadů regresních parametrů zjistíme, že u proměnné $\widehat{Nc}_{ST}$ je odhad regresního parametru kladný, což odporuje našim představám, neboť kladná hodnota by říkala, že při jasné obloze je výroba elektřiny z FVE menší než při zatažené obloze. Při vyřazení proměnné $\widehat{Nc}_{ST}$ vyjde opět podle Waldova testu nevýznamná proměnná $\widehat{Nl}_{ST}$ a dostaneme již pouze model

```
 > M1 <- betareg(tau ~ G + T, data_ST, link="logit")
> coeftest(M1)

z test of coefficients:

               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -2.3164e+00  1.0655e-02 -217.392 < 2.2e-16 ***
G            9.8935e-03  5.6864e-05  173.984 < 2.2e-16 ***
T            1.5583e-02  7.1764e-04   21.714 < 2.2e-16 ***
(phi)        1.4012e+00  1.1005e-02  127.329 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M1)*data_ST$CS, data_ST$p)
[1] 0.07498194
> rmse(predict(M1, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.08273891
```

Tento model má všechny vysvětlující proměnné významné a od úplného modelu se podle hodnot RMSE významně nezhoršil. Protože nám z modelu vypadly veškeré oblačnosti, o kterých víme, že mají vliv na množství dopadajícího slunečního záření, zatímco zde zůstala teplota, o které víme, že má ve skutečnosti na výrobu elektřiny z FVE marginální vliv, zkusíme ještě model sestavit tak, že z úplného modelu nejprve vyřadíme proměnou $\widehat{t}_{ST}$, jejíž přítomnost může potlačit významnost jiných proměnných. Z modelu bez teploty následně vyřadíme podle Waldova testu nevýznamné proměnné $\widehat{Nm}_{ST}$ a $\widehat{Nh}_{ST}$ a následně i proměnnou $\widehat{Nc}_{ST}$ kvůli kladnému odhadu regresního parametru. Nyní máme model

```
 > M1 <- betareg(tau ~ G + Nl, data_ST, link="logit")
> coeftest(M1)

z test of coefficients:

               Estimate  Std. Error   z value  Pr(>|z|)
(Intercept) -2.1737e+00  1.1042e-02 -196.8636 < 2.2e-16 ***
G            1.0386e-02  5.2267e-05  198.7033 < 2.2e-16 ***
Nl          -7.9662e-02  1.8002e-02   -4.4253 9.632e-06 ***
(phi)        1.3891e+00  1.0918e-02  127.2267 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M1)*data_ST$CS, data_ST$p)
[1] 0.074995
> rmse(predict(M1, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.08152935
```

Porovnáním hodnot RMSE je možné usoudit, že model, kde jsme teplotu $\widehat{t}_{ST}$ nahradili nízkou oblačností $\widehat{Nl}_{ST}$, je malinko lepší. Vyzkoušíme ještě model, kdy nízkou oblačnost $\widehat{Nl}_{ST}$ nahradíme celkovou oblačností $\widehat{Nc}_{ST}$, protože kladný odhad regresního parametru u této proměnné mohl být způsobený přítomností proměnné $\widehat{Nl}_{ST}$. Protože i v tomto modelu zůstal kladný odhad regresního parametru u proměnné $\widehat{Nc}_{ST}$, tento model zamítneme.

Zkusme se dále podívat (nyní již trochu rychleji) na model, kde k odhadu regresních parametrů nepoužijeme všechna data, ale pouze data ze dní a hodin, kdy $\widehat{cs}_{sm,ST}(D,H) > 0{,}2$. Tímto z odhadování odstraníme hodiny, ve kterých je potenciální výroba malá, což povede k tomu, že model budeme odhadovat pouze na hodinách s významnějším potenciálem výroby. Modely, které budeme predikovat pouze na této množině dat nazývejme dále cut modely. Z úplného modelu postupně odstraníme podle Waldova testu nevýznamnou proměnnou $\widehat{Nh}_{ST}$ a následně proměnnou $\widehat{Nm}_{ST}$ díky kladnému odhadu regresního parametru. Poté zkusíme odstranit teplotu $\widehat{t}_{ST}$. Po jejím odstranění vzrostly hodnoty RMSE jen nepatrně. Odhady regresních parametrů u proměnných $\widehat{Nl}_{ST}$ a $\widehat{Nc}_{ST}$ jsou záporné. Podíváme se, co by s kvalitou modelu udělalo odstranění proměnné $\widehat{Nl}_{ST}$. V tomto případě hodnoty RMSE významněji vzrostly, tudíž proměnnou $\widehat{Nl}_{ST}$ vrátíme zpět do modelu. Zkusíme ještě odstranit proměnnou $\widehat{Nc}_{ST}$, v tomto případě je nárůst hodnot RMSE minimální, tedy ji z modelu odstraníme. Dostaneme model

```
 > M1_cut <- betareg(tau ~ G + Nl, subset(data_ST, CS>g_c), link="logit")
> coeftest(M1_cut)

z test of coefficients:

               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  8.4057e-02  1.7502e-02   4.8028 1.565e-06 ***
G            5.0486e-03  5.7923e-05  87.1609 < 2.2e-16 ***
Nl          -2.1816e+00  2.6403e-02 -82.6250 < 2.2e-16 ***
(phi)        4.9362e+00  5.3575e-02  92.1350 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M1_cut)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.0952813
> rmse(predict(M1_cut, subset(data_ST_test, data_ST_test$CS>g_c))
+   *subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.09603047
> rmse(predict(M1_cut, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.0584482
```

Je potřeba si uvědomit, že první dvě hodnoty RMSE jsou počítány pouze na datech použitých při vytváření cut modelu. Vzhledem k tomu, že došlo k odstranění zejména těch hodin, kdy je predikce shodná se skutečností, případně hodin, kdy je odchylka predikce od skutečnosti minimální, není možné tyto hodnoty RMSE z cut modelu srovnávat s hodnotami RMSE z modelu, ve kterém jsme použili všechna data. Pro možné porovnání byla spočtena i hodnota RMSE na

všech testovacích datech. Tu již můžeme porovnat s hodnotou RMSE z předchozího modelu. Z porovnání vyplývá, že cut model přinesl významné zlepšení.

Zkusme se ještě podívat, zdali by nebylo vhodné nějak transformovat dominantní vysvětlující proměnnou $\widehat{G}_{ST}$. Jednou z možností by bylo místo ní uvažovat proměnnou, kterou označíme $\widehat{G}_{cs,ST}$, již z původní proměnné dostaneme vztahem

$$\widehat{G}_{cs,ST} = \frac{\widehat{G}_{ST}}{\widehat{cs}_{sm,ST}}.$$

Tato nová proměnná by mohla přispět ke zlepšení modelu, neboť hodnoty původní proměnné $\widehat{G}_{ST}$ jsou závislé na dni a hodině. U proměnné $\widehat{G}_{cs,ST}$ jsme tuto závislost odstranili. Podíváme se, zdali tato transformace vede k těsnější závislosti vysvětlované proměnné τ_{ST} na vysvětlující proměnné $\widehat{G}_{cs,ST}$. Podíváme se na Pearsonovy korelační koeficienty

```
 > cor(data_ST$tau, data_ST$G)
[1] 0.7411362
> cor(data_ST$tau, data_ST$G_CS)
[1] 0.870778
```

z nichž vidíme, že úvaha o zlepšení závislosti by nemusela být špatná. Zkusme tedy provést náhradu nejprve pro úplný model na celých datech. Zde se ale i při zjednodušení modelu dostaneme vždy do situace, kdy máme kladný odhad regresního parametru u proměnné týkající se oblačnosti. Ponecháme-li v modelu pouze jedinou vysvětlující proměnnou $\widehat{G}_{cs,ST}$, je hodnota RMSE na testovacích datech vyšší než v modelu s proměnnou $\widehat{G}_{ST}$. Zkusme přejít ke cut modelu. Z úplného modelu ze stejných důvodu jako u předešlého cut modelu opět postupně odstraníme proměnné $\widehat{Nh}_{ST}$, $\widehat{Nm}_{ST}$, $\widehat{t}_{ST}$ a $\widehat{Nc}_{ST}$ a dostaneme model

```
 > M2_cut <- betareg(tau ~ G_CS + Nl, subset(data_ST, CS>g_c), link="logit")
> coeftest(M2_cut)

z test of coefficients:

               Estimate  Std. Error z value  Pr(>|z|)
(Intercept) -0.58023611  0.02030168 -28.581 < 2.2e-16 ***
G_CS         0.00447679  0.00004306 103.967 < 2.2e-16 ***
Nl          -1.65022684  0.02653714 -62.186 < 2.2e-16 ***
(phi)        5.76085087  0.06314298  91.235 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M2_cut)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09236592
> rmse(predict(M2_cut, subset(data_ST_test, data_ST_test$CS>g_c))
+   *subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.1038185
> rmse(predict(M2_cut, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.06284514
```

Porovnáním s původním cut modelem zjistíme, že na trénovací množině dat nám hodnota RMSE klesla, ale na testovací množině dat hodnota RMSE významně vzrostla.

Podívejme se ještě na hodnoty MAE a PBIAS u cut modelu `M1_cut`.

```
 > mae(predict(M1_cut, data_ST)*data_ST$CS, data_ST$p)
[1] 0.02876144
> mae(predict(M1_cut, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.0280738
> pbias(predict(M1_cut)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 1.7
```

Hodnoty $MAE_{M1\_cut,tr}$ oproti $MAE_{ref,tr}$ a $MAE_{M1\_cut,test}$ oproti $MAE_{ref,test}$ klesly, což jsme chtěli. Hodnota u PBIAS sice není rovna 0, ale vychýlení modelu `M1_cut` je akceptovatelné.

Zkusme ještě zjistit, zdali by se nevyplatilo místo proměnné $\widehat{G}_{ST}$ použít nějakou její transformaci. Zkusíme ji nahradit buď odmocninou, tj. proměnnou $\sqrt{\widehat{G}_{ST}}$, případně logaritmem, tj. proměnnou $\log(\widehat{G}_{ST} + c)$. Zde ale musíme díky možným nulovým hodnotám přičíst nějakou konstantu. Použijeme $c = 0{,}001$.

Pracujme rovnou s cut modelem, jelikož výše jsme zjistili, že toto má pozitivní přínos. V případě odmocninové transformace proměnné $\widehat{G}_{ST}$ dostaneme cut model `M3_sqrt`

```
 > M3_sqrt <- betareg(tau ~ sqrt(G) + Nl, subset(data_ST, CS>g_c),
+ link="logit")
> coeftest(M3_cut_sqrt)

z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.855639   0.025351 -33.751 < 2.2e-16 ***
sqrt(G)      0.145756   0.001585  91.961 < 2.2e-16 ***
Nl          -2.120679   0.026128 -81.166 < 2.2e-16 ***
(phi)        5.178443   0.056352  91.894 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M3_sqrt)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09459988
> rmse(predict(M3_sqrt, subset(data_ST_test, data_ST_test$CS>g_c))
+   *subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.09242507
> rmse(predict(M3_sqrt, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.05633668
> mae(predict(M3_sqrt, data_ST)*data_ST$CS, data_ST$p)
[1] 0.02812879
> mae(predict(M3_sqrt, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.02713243
> pbias(predict(M3_sqrt)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 2.8
```

Model `M3_sqrt` oproti modelu `M1_cut` vykazuje zlepšení v RMSE a MAE, nicméně za cenu zvětšení vychýlení.

Při použití logaritmické transformace dostaneme model `M3_log`

```
 > M3_log <- betareg(tau ~ log(G+0.001) + Nl, subset(data_ST, CS>g_c),
+ link="logit")
> coeftest(M3_log)

z test of coefficients:

                Estimate Std. Error z value  Pr(>|z|)
(Intercept)    -3.1522572  0.0472895 -66.659 < 2.2e-16 ***
log(G + 0.001)  0.8428703  0.0088205  95.559 < 2.2e-16 ***
Nl             -2.1530065  0.0260962 -82.503 < 2.2e-16 ***
(phi)           5.1099562  0.0555404  92.004 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M3_log)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09669092
> rmse(predict(M3_log, subset(data_ST_test, data_ST_test$CS>g_c))
+   *subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.0921785
> rmse(predict(M3_log, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.05629787
> mae(predict(M3_log, data_ST)*data_ST$CS, data_ST$p)
[1] 0.02879189
> mae(predict(M3_log, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.02800297
> pbias(predict(M3_log)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 3.4
```

Ten je podle hodnot RMSE lepší než model `M1_cut`, nicméně oproti modelu `M3_sqrt` jsou hodnoty RMSE a MAE nepatrně vyšší, stejně tak i vychýlení.

Zkusme ještě použít stejné transformace i na proměnné $\widehat{Nl}_{ST}$, případně i proměnnou $\widehat{Nc}_{ST}$. V případě odmocninové transformace dostaneme model `M3_sqrt2`

```
 > M3_sqrt2 <- betareg(tau ~ sqrt(G) + sqrt(Nl), subset(data_ST, CS>g_c),
+ link="logit")
> coeftest(M3_sqrt2)

z test of coefficients:

              Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.5942130  0.0271587 -21.879 < 2.2e-16 ***
sqrt(G)      0.1492836  0.0015594  95.734 < 2.2e-16 ***
sqrt(Nl)    -2.0594527  0.0246381 -83.588 < 2.2e-16 ***
(phi)        5.3141861  0.0579438  91.713 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M3_sqrt2)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09438436
> rmse(predict(M3_sqrt2, subset(data_ST_test, data_ST_test$CS>g_c))
+   *subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.09214902
```

```
> rmse(predict(M3_sqrt2, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.05622508
> mae(predict(M3_sqrt2, data_ST)*data_ST$CS, data_ST$p)
[1] 0.02828475
> mae(predict(M3_sqrt2, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.02695761
> pbias(predict(M3_sqrt2)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 2.4
```

Ten oproti modelu `M3_sqrt` vykazuje drobné zlepšení v RMSE a MAE, navíc se snížilo i vychýlení, takže dále budeme uvažovat model `M3_sqrt2`.

U logaritmické transformace, kdy k hodnotám $\widehat{Nl}_{ST}$ bude lepší díky maximálním hodnotám rovným 1 přičíst 0,5, nicméně tento model je podle hodnot RMSE i MAE horší než model `M3_log`, takže jej zamítneme.

Na závěr této části si ještě ukážeme hodnoty RMSE, MAE a PBIAS pro modely `M1_cut_Nc`, `M3_sqrt2_Nc` a `M3_log_Nc`, které se od původních liší tím, že jsme mezi vysvětlující proměnné přidali i proměnnou $\widehat{Nc}_{ST}$. U všech tří modelů její přidání dává smysl, neboť odhad regresního parametru je záporný, tudíž vyřazení této proměnné mohlo být unáhlené.

```
 > rmse(predict(M1_cut_Nc)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09493904
> rmse(predict(M1_cut_Nc, subset(data_ST_test, data_ST_test$CS>g_c))
+   *subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.09606677
> rmse(predict(M1_cut_Nc, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.05839921
> mae(predict(M1_cut_Nc, data_ST)*data_ST$CS, data_ST$p)
[1] 0.02842424
> mae(predict(M1_cut_Nc, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.0275336
> pbias(predict(M1_cut_Nc)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.6
> rmse(predict(M3_sqrt2_Nc)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09360341
> rmse(predict(M3_sqrt2_Nc, subset(data_ST_test, data_ST_test$CS>g_c))
+   *subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.09250796
> rmse(predict(M3_sqrt2_Nc, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.05635919
> mae(predict(M3_sqrt2_Nc, data_ST)*data_ST$CS, data_ST$p)
[1] 0.02793754
> mae(predict(M3_sqrt2_Nc, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.02647189
> pbias(predict(M3_sqrt2_Nc)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 1.5
> rmse(predict(M3_log_Nc)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09494245
```

```
> rmse(predict(M3_log_Nc, subset(data_ST_test, data_ST_test$CS>g_c))
+   *subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.09257292
> rmse(predict(M3_log_Nc, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.05639063
> mae(predict(M3_log_Nc, data_ST)*data_ST$CS, data_ST$p)
[1] 0.02798592
> mae(predict(M3_log_Nc, data_ST_test)*data_ST_test$CS, data_ST_test$p)
[1] 0.02699443
> pbias(predict(M3_log_Nc)*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 2
```

Přidáním proměnné $\widehat{Nc}_{ST}$ do původních modelů nepatrně poklesla většina hodnot RMSE a MAE. Tam kde vzrostly, se jednalo o drobný nárůst. Zároveň se nám povedlo snížit vychýlení modelů, kdy zejména u modelu `M1_cut_Nc` se hodnota PBIAS již blíží k 0. Zejména z tohoto důvodu můžeme usoudit, že vynechání proměnné $\widehat{Nc}_{ST}$ nebylo vhodné, proto ji v modelech ponecháme a v dalších částech budeme pracovat s modely `M1_cut_Nc`, `M3_sqrt2_Nc` a `M3_log_Nc`.

5.3.3 Beta regrese na p

Druhým modelem, který vyzkoušíme, bude beta regresní model s vysvětlovanou proměnnou p_{ST}. Postupovat budeme obdobně jako v předchozím případě. Nejprve zkusíme beta regresní model na celých datech a netransformovaných veličinách. I po zjednodušení modelu na model `M4 <- betareg(p ~ G + Nl, data_ST, link="logit")` dostaneme hodnoty RMSE a MAE významně vyšší než u cut modelů s vysvětlovanou proměnnou τ, takže opět přejdeme ke cut modelu. Nejdříve jej vyzkoušíme na netransformovaných veličinách. Bohužel i v tomto případě nedošlo k žádnému zlepšení hodnot RMSE a MAE, modely jsou výrazně horší než modely s vysvětlovanou proměnnou τ.

Zkusme přejít k modelům, kde zkusíme vysvětlující proměnné transformovat. Proměnná $\widehat{G}_{cs,ST}$ je v případě vysvětlované proměnné p zjevně méně vhodná než proměnná $\widehat{G}_{ST}$, což potvrdily i Pearsonovy korelační koeficienty

```
 > cor(data_ST$p, data_ST$G)
[1] 0.9284008
> cor(data_ST$p, data_ST$G_CS)
[1] 0.7543527
```

Z tohoto důvodu přejdeme rovnou k logaritmické transformaci proměnné $\widehat{G}_{ST}$. Zde dostaneme model `M5_log`, do kterého jsme proměnnou $\widehat{Nc}_{ST}$ nezařadili díky kladnému odhadu regresního parametru

```
 > M5_log <- betareg(p ~ log(G+0.001) + Nl, subset(data_ST, CS>g_c),
+ link="logit")
> coeftest(M5_log)

z test of coefficients:
```

```
                 Estimate Std. Error  z value  Pr(>|z|)
(Intercept)    -5.9782811  0.0349840 -170.886 < 2.2e-16 ***
log(G + 0.001)  1.0681179  0.0062775  170.151 < 2.2e-16 ***
Nl             -0.6870247  0.0167830  -40.936 < 2.2e-16 ***
(phi)          16.9475585  0.1892370   89.557 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M5_log), subset(data_ST$p, data_ST$CS>g_c))
[1] 0.1012145
> rmse(predict(M5_log, subset(data_ST_test, data_ST_test$CS>g_c)),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.1179605
> rmse(predict(M5_log, data_ST_test), data_ST_test$p)
[1] 0.07301418
> mae(predict(M5_log, data_ST), data_ST$p)
[1] 0.03210671
> mae(predict(M5_log, data_ST_test), data_ST_test$p)
[1] 0.03574132
> pbias(predict(M5_log), subset(data_ST$p, data_ST$CS>g_c))
[1] -0.2
```

Tento model má velmi malé vychýlení a hodnoty RMSE a MAE se oproti modelům bez transformované vysvětlující proměnné $\widehat{G}_{ST}$ zmenšily, nicméně od jeho použití nás odrazuje skutečnost, že tyto hodnoty jsou stále vyšší než u modelů s vysvětlovanou proměnnou τ. Transformujeme-li v modelu i proměnnou $\widehat{Nl}_{ST}$, hodnoty RMSE a MAE jsou velmi podobné hodnotám z původního modelu.

Zkusme ještě odmocninovou transformaci $\widehat{G}_{ST}$. Tento model vychází v hodnotách RMSE a MAE hůře než model `M5_log`, obdnobně i model s $\sqrt{\widehat{Nl}_{ST}}$, takže odmocninovou transformaci zamítneme.

Vzhledem k tomu, že jsme při těchto predikcích nijak nevyužili informace z Clear sky modelu, může se v tomto případě stát, že nějaké predikce jsou vyšší než je v daném čase horní hranice odhadnutá Clear sky modelem. Zkusme predikované hodnoty podmínit tím, že se v situaci, kdy predikovaná hodnota bude větší než hranice daná Clear sky modelem, vezme za predikci hranice Clear sky modelu. Uvažujme prozatím nejlepší model z této části `M5_log`, pro který dostaneme

```
 > j <- length(data_ST$D)
> pr_tr_M5_log <- rep(0,j)
> for (i in 1:j)
+ pr_tr_M5_log[i] <- min(predict(M5_log, data_ST[i,]), data_ST$CS[i])
> j <- length(data_ST_test$D)
> pr_te_M5_log <- rep(0,j)
> for (i in 1:j)
+ pr_te_M5_log[i] <- min(predict(M5_log, data_ST_test[i,]),data_ST_test$CS[i])
> XXX <- data.frame(pr_tr_M5_log, p = data_ST$p, CS = data_ST$CS)
> rmse(subset(XXX$pr_tr_M5_log, XXX$CS>g_c), subset(XXX$p, XXX$CS>g_c))
[1] 0.09723319
> rmse(pr_te_M5_log, data_ST_test$p)
[1] 0.06997445
> mae(pr_tr_M5_log,data_ST$p)
[1] 0.02836319
> mae(pr_te_M5_log,data_ST_test$p)
[1] 0.03298164
```

I při tomto ošetření predikce jsou hodnoty RMSE a MAE vyšší než u modelů s vysvětlovanou proměnnou τ.

Závěrem této části můžeme říci, že pro beta regresní model přináší Clear sky model $\widehat{cs}_{sm,ST}$ spolu se změnou vysvětlované proměnné p na τ kladný přínos v podobě lepších modelů. Beta regresní modely s vysvětlovanou proměnnou p již nebudeme dále uvažovat.

5.3.4 Lineární regrese na transformovanou vysvětlovanou proměnnou τ

Třetím modelem, který vyzkoušíme, bude lineární regresní model s vhodně transformovanou vysvětlovanou proměnnou τ_{ST}. Pro transformaci vysvětlující proměnné budeme muset najít nějakou vhodnou funkci, která transformuje interval $(0, 1)$ na celou reálnou osu, tj. funkci $f(\cdot) : (0, 1) \mapsto \mathbb{R}$, která navíc musí být striktně rostoucí a dvakrát diferencovatelná. Na takové funkce jsme již narazili v kapitole 3.2, když jsme uváděli příklad vhodných linkových funkcí pro beta regresi.

Začněme nejdřív s logitovou funkcí, což je funkce $f(x) = \log(x/(1-x))$. Postupovat budeme podobně jako v předchozích případech. Nejprve zkusíme model na celých datech a netransformovaných veličinách. Tento model `M6_full <- lm (log(tau/(1-tau)) ~ G + T + Nl + Nc + Nm + Nh, data_ST)` ovšem rychle zamítneme, neboť zejména v zimním období predikuje sotva zaznamenatelnou výrobu, což je dobře vidět z obrázku 5.1. Zejména „sloupec“ predikovaných hodnot okolo 0, přičemž se skutečné hodnoty procentuálního měření výkonu pohybují až okolo hodnot blízkých 0,5, nesvědčí o vhodnosti modelu.

V tomto případě bude nejspíš nutný přechod ke cut modelu. Po vynechání nevýznamné proměnné $\widehat{Nh}_{ST}$, proměnné $\widehat{Nm}_{ST}$ s kladným odhadem regresního parametru a teploty $\widehat{t}_{ST}$ dostaneme model

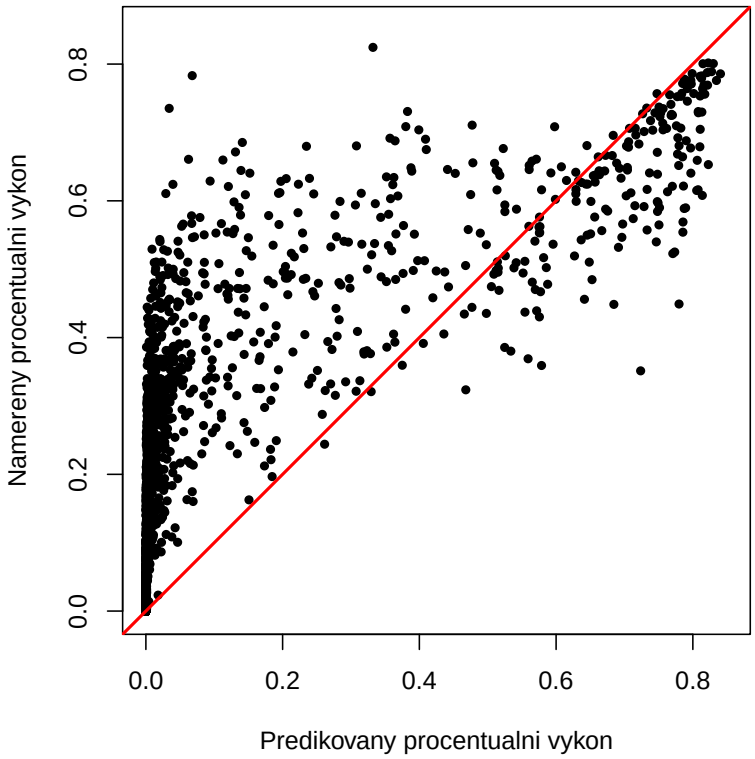
```
 > summary(M6_cut)

Call:
lm(formula = logit(tau) ~ G + Nl + Nc, data = subset(data_ST,
    CS > g_c))

Residuals:
    Min      1Q  Median      3Q     Max
-5.1256 -0.7953 -0.0461  0.6074 11.0976

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.0836570  0.0497434   21.79   <2e-16 ***
G            0.0057604  0.0001039   55.46   <2e-16 ***
Nl          -2.4450927  0.0492522  -49.64   <2e-16 ***
Nc          -1.3530530  0.0559636  -24.18   <2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.463 on 15711 degrees of freedom
Multiple R-squared:  0.5486,    Adjusted R-squared:  0.5485
F-statistic:  6364 on 3 and 15711 DF,  p-value: < 2.2e-16
```

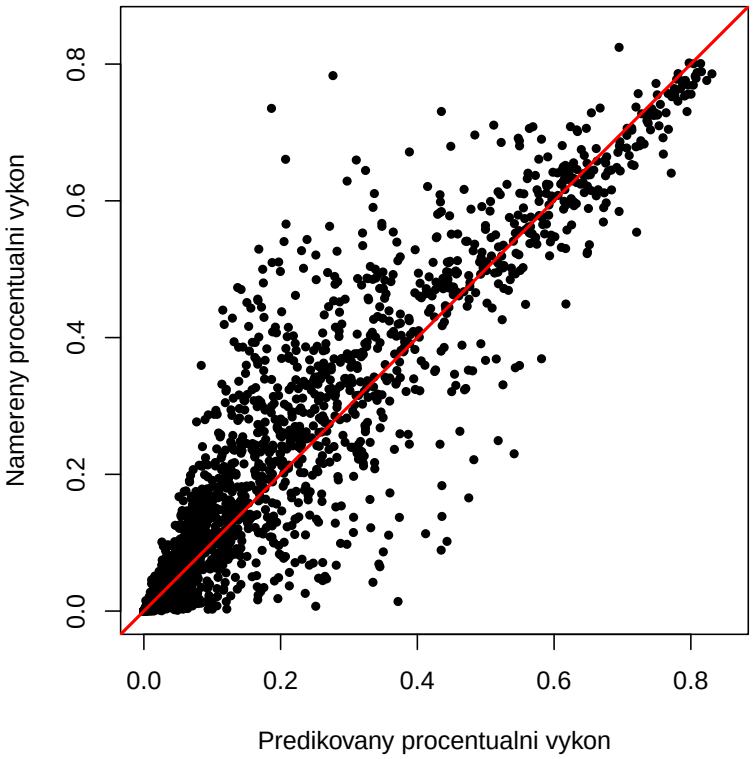


Obrázek 5.1: Predikované hodnoty $\widehat{p}_{ST}$ modelem `M6_full` proti naměřeným hodnotám p_{ST} na datech z roku 2017

```
> rmse(logit(predict(M6_cut), inverse = TRUE)*subset(data_ST$CS,
+   data_ST$CS>g_c), subset(data_ST$p, data_ST$CS>g_c))
[1] 0.100087
> rmse(logit(predict(M6_cut, subset(data_ST_test, data_ST_test$CS>g_c)),
+   inverse = TRUE)*subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.1038169
> rmse(logit(predict(M6_cut, data_ST_test), inverse = TRUE)*data_ST_test$CS,
+   data_ST_test$p)
[1] 0.06301736
> mae(logit(predict(M6_cut, data_ST), inverse = TRUE)*data_ST$CS, data_ST$p)
[1] 0.02822325
> mae(logit(predict(M6_cut, data_ST_test), inverse = TRUE)*data_ST_test$CS,
+   data_ST_test$p)
[1] 0.0284786
> pbias(logit(predict(M6_cut), inverse = TRUE)*subset(data_ST$CS,
+   data_ST$CS>g_c), subset(data_ST$p, data_ST$CS>g_c))
[1] 2.5
```

Z obrázku 5.2 vidíme, že se nám přechodem ke cut modelu podařilo odstranit „sloupec“ predikovaných hodnot okolo 0 a model již působí věrohodněji. Z tohoto důvodu budeme dále v této sekci pracovat již jen s cut modely, protože modely konstruované na všech datech nejsou vhodné.

Nyní přejdeme k modelům, kde zkusíme vhodně transformovat i vysvětlující proměnnou $\widehat{G}_{ST}$. Vzhledem k tomu, že vysvětlovaná proměnná je τ_{ST} můžeme zkusit proměnnou $\widehat{G}_{cs,ST}$. Model `M7 <- lm (logit(tau) ~ G_CS + Nl + Nc, subset(data_ST, CS>g_c))` dopadl stejně jako model `M2_cut`, kdy na trénovací



Obrázek 5.2: Predikované hodnoty $\widehat{p}_{ST}$ modelem `M6_cut` proti naměřeným hodnotám p_{ST} na datech z roku 2017

množině dat nám hodnoty RMSE a MAE klesly, zatímco na testovací množině dat tyto hodnoty vzrostly.

Jako další transformaci proměnné $\widehat{G}_{ST}$ vyzkoušíme logaritmus, který by mohl dávat smysl i díky tomu, že částečně za pomocí logaritmů transformujeme i vysvětlovanou proměnnou. Zde dostaneme model `M8_log`, o kterém by se dalo uvažovat jako o vhodném modelu.

```
 > summary(M8_log)

Call:
lm(formula = logit(tau) ~ log(G + 0.001) + Nl + Nc, data = subset(data_ST,
    CS > g_c))

Residuals:
    Min      1Q  Median      3Q     Max
-5.0214 -0.7495 -0.0679  0.5631 11.0126

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.63282    0.09666  -16.89   <2e-16 ***
log(G + 0.001)  0.82182    0.01532   53.63   <2e-16 ***
Nl             -2.33174    0.04985  -46.77   <2e-16 ***
Nc             -1.75688    0.05338  -32.91   <2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.47 on 15711 degrees of freedom
Multiple R-squared:  0.5437,    Adjusted R-squared:  0.5436
F-statistic:  6241 on 3 and 15711 DF,  p-value: < 2.2e-16
```

```
> rmse(logit(predict(M8_log), inverse = TRUE)*subset(data_ST$CS,
+   data_ST$CS>g_c), subset(data_ST$p, data_ST$CS>g_c))
[1] 0.1009114
> rmse(logit(predict(M8_log, subset(data_ST_test, data_ST_test$CS>g_c)),
+   inverse = TRUE)*subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.1013221
> rmse(logit(predict(M8_log, data_ST_test), inverse = TRUE)*data_ST_test$CS,
+   data_ST_test$p)
[1] 0.06156559
> mae(logit(predict(M8_log, data_ST), inverse = TRUE)*data_ST$CS, data_ST$p)
[1] 0.02787362
> mae(logit(predict(M8_log, data_ST_test), inverse = TRUE)*data_ST_test$CS,
+   data_ST_test$p)
[1] 0.02796807
> pbias(logit(predict(M8_log), inverse = TRUE)*subset(data_ST$CS,
+   data_ST$CS>g_c), subset(data_ST$p, data_ST$CS>g_c))
[1] 4.9
```

Zkusme ještě logaritmickou transformaci zbylých dvou proměnných $\widehat{Nl}_{ST}$ a $\widehat{Nc}_{ST}$. Dostaneme model. `M8_log2`

```
 > summary(M8_log2)

Call:
lm(formula = logit(tau) ~ log(G + 0.001) + log(Nl + 0.5) + log(Nc + 0.5),
    data = subset(data_ST, CS > g_c))

Residuals:
    Min      1Q  Median      3Q     Max
-4.8478 -0.7533 -0.0647  0.5628 11.1736

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.01526    0.07612  -52.75   <2e-16 ***
log(G + 0.001)  0.84778    0.01492   56.81   <2e-16 ***
log(Nl + 0.5)  -2.03932    0.04142  -49.23   <2e-16 ***
log(Nc + 0.5)  -1.67312    0.05137  -32.57   <2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.459 on 15711 degrees of freedom
Multiple R-squared:  0.5508,    Adjusted R-squared:  0.5507
F-statistic:  6421 on 3 and 15711 DF,  p-value: < 2.2e-16

> rmse(logit(predict(M8_log2), inverse = TRUE)*subset(data_ST$CS,
+   data_ST$CS>g_c), subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09945217
> rmse(logit(predict(M8_log2, subset(data_ST_test, data_ST_test$CS>g_c)),
+   inverse = TRUE)*subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.1009067
> rmse(logit(predict(M8_log2, data_ST_test), inverse = TRUE)*data_ST_test$CS,
+   data_ST_test$p)
[1] 0.06137223
> mae(logit(predict(M8_log2, data_ST), inverse = TRUE)*data_ST$CS, data_ST$p)
[1] 0.02778968
```

```
> mae(logit(predict(M8_log2, data_ST_test), inverse = TRUE)*data_ST_test$CS,
+   data_ST_test$p)
[1] 0.02813353
> pbias(logit(predict(M8_log2), inverse = TRUE)*subset(data_ST$CS,
+   data_ST$CS>g_c), subset(data_ST$p, data_ST$CS>g_c))
[1] 4.1
```

V porovnání s modelem `M8_log` nám klesly hodnoty u RMSE a MAE, navíc nám mírně klesla hodnota u PBIAS, přesto je vychýlení modelu stále významné.

Jako poslední u logitové funkce zkusíme odmocninou transformaci proměnné $\widehat{G}_{ST}$. Oba dva modely (jeden z transformováním pouze proměnné $\widehat{G}_{ST}$, druhý včetně transformovaných proměnných $\widehat{Nl}_{ST}$ a $\widehat{Nc}_{ST}$) vykazují oproti logaritmickým transformacím nepatrně nižší hodnoty RMSE a MAE a jsou i méně vychýlené. Nejlépe vychází model `M8_sqrt2`, nicméně i ten je stále vychýlený

```
 > summary(M8_sqrt2)

Call:
lm(formula = logit(tau) ~ sqrt(G) + sqrt(Nl) + sqrt(Nc),
    data = subset(data_ST, CS > g_c))

Residuals:
    Min      1Q  Median      3Q     Max
-4.8799 -0.7614 -0.0730  0.5815 11.5857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.509301   0.072457   7.029 2.17e-12 ***
sqrt(G)      0.178226   0.002735  65.166  < 2e-16 ***
sqrt(Nl)    -2.304419   0.045824 -50.289  < 2e-16 ***
sqrt(Nc)    -1.558015   0.065734 -23.702  < 2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.432 on 15711 degrees of freedom
Multiple R-squared:  0.5672,    Adjusted R-squared:  0.5672
F-statistic:  6865 on 3 and 15711 DF,  p-value: < 2.2e-16

> rmse(logit(predict(M8_sqrt2), inverse = TRUE)*subset(data_ST$CS,
+   data_ST$CS>g_c), subset(data_ST$p, data_ST$CS>g_c))
[1] 0.09718088
> rmse(logit(predict(M8_sqrt2, subset(data_ST_test, data_ST_test$CS>g_c)),
+   inverse = TRUE)*subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.100039
> rmse(logit(predict(M8_sqrt2, data_ST_test), inverse = TRUE)*data_ST_test$CS,
+   data_ST_test$p)
[1] 0.06090956
> mae(logit(predict(M8_sqrt2, data_ST), inverse = TRUE)*data_ST$CS, data_ST$p)
[1] 0.02751994
> mae(logit(predict(M8_sqrt2, data_ST_test), inverse = TRUE)*data_ST_test$CS,
+   data_ST_test$p)
[1] 0.02791384
> pbias(logit(predict(M8_sqrt2), inverse = TRUE)*subset(data_ST$CS,
+   data_ST$CS>g_c), subset(data_ST$p, data_ST$CS>g_c))
[1] 3.1
```

Nyní přejdeme ke druhé možné funkci $f(\cdot)$, kterou je probitová funkce, což je funkce $f(x) = \Phi^{-1}(x)$, kde $\Phi(\cdot)$ je distribuční funkce standardního normálního rozdělení. Zkusíme-li libovolný model na celých datech, budeme mít stejný problém jako u logitové funkce, takže přejdeme rovnou ke cut modelům. V případě cut modelu jsme ve všech případech v porování s modelem se stejnými vysvětlujícími proměnnými akorát s transformací τ_{ST} logitovou funkcí dostali vždy menší hodnoty RMSE, MAE i PBIAS, z toho můžeme usoudit, že probitová funkce je vhodnější na transformování vysvětlující proměnné τ_{ST}. Pro porovnání uvedeme nejlepší model `M10_sqrt2`

```
 > summary(M10_sqrt2)

Call:
lm(formula = probit(tau) ~ sqrt(G) + sqrt(Nl) + sqrt(Nc),
    data = subset(data_ST, CS > g_c))

Residuals:
    Min      1Q  Median      3Q     Max
-2.2750 -0.4028 -0.0221  0.3553  4.3960

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.090500   0.033213   2.725  0.00644 **
sqrt(G)      0.100328   0.001254  80.027  < 2e-16 ***
sqrt(Nl)    -1.282247   0.021005 -61.045  < 2e-16 ***
sqrt(Nc)    -0.680075   0.030131 -22.570  < 2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6564 on 15711 degrees of freedom
Multiple R-squared:  0.6445,    Adjusted R-squared:  0.6444
F-statistic:  9494 on 3 and 15711 DF,  p-value: < 2.2e-16

> rmse(pnorm(predict(M10_sqrt2))*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 0.0947625
> rmse(pnorm(predict(M10_sqrt2, subset(data_ST_test, data_ST_test$CS>g_c)))*
+   subset(data_ST_test$CS, data_ST_test$CS>g_c),
+   subset(data_ST_test$p, data_ST_test$CS>g_c))
[1] 0.09720882
> rmse(pnorm(predict(M10_sqrt2, data_ST_test))*data_ST_test$CS,data_ST_test$p)
[1] 0.05923055
> mae(pnorm(predict(M10_sqrt2, data_ST))*data_ST$CS, data_ST$p)
[1] 0.02741105
> mae(pnorm(predict(M10_sqrt2, data_ST_test))*data_ST_test$CS, data_ST_test$p)
[1] 0.02749322
> pbias(pnorm(predict(M10_sqrt2))*subset(data_ST$CS, data_ST$CS>g_c),
+   subset(data_ST$p, data_ST$CS>g_c))
[1] 2.1
```

Nakonec vyzkoušejme komplementární log-log funkci $f(x) = \log[-\log(1-x)]$. Libovolný model na celých datech má opět stejný problém se spoustou predikcí blízkých nule, takže opět přejdeme rovnou ke cut modelům. V případě cut modelu jsme pro všechny zmíněné transformace vysvětlujících proměnných došli k modelu, který měl nepatrně vyšší hodnoty RMSE a MAE než modely s probitovou

funkcí, proto komplementární log-log funkci dále nebudeme uvažovat.

Jako nejvhodnější funkce pro transformaci vysvětlované proměnné τ_{ST} se jeví probitová funkce. Z této části pro další pokračování vybereme model `M10_sqrt2`.

5.3.5 Lineární regrese na transformovanou vysvětlovanou proměnnou *p*

Posledním modelem, který vyzkoušíme, bude lineární regresní model s vhodně transformovanou vysvětlovanou proměnnou p_{ST}. Začneme s transformací logitovou funkcí. Změnou vysvětlované proměnné se pro libovolný model na celých datech problém se spoustou predikcí blízkých nule nepovedlo odstranit, takže opět přejdeme rovnou ke cut modelům. Bohužel i v tomto případě se po vyzkoušení všech zmíněných transformací, případně i změny transformační funkce vysvětlované proměnné na probitovou nebo komplementární log-log funkci, potvrdilo, že i pro lineární regresní model s transformovanou vysvětlovanou veličinou přináší Clear sky model $\widehat{cs}_{sm,ST}$ spolu se změnou vysvětlované proměnné p na τ kladný přínos v podobě lepších modelů. Toto se nezměnilo i v případě, že jsme v situaci, kdy byla predikovaná hodnota větší než hranice daná Clear sky modelem, nahradili predikci hodnotou této hranice. Lineární regresní model s transformovanou vysvětlovanou veličinou p již nebudeme dále uvažovat.

5.3.6 Rozdělení dat na homogenní podmnožiny

V předchozích čtyřech částech jsme vždy při přechodu na cut model dospěli k přesnějším odhadům regresních parametrů než v případě odhadů regresních parametrů na celých datech. To nám vnuklo myšlenku, jestli by nebylo vhodné rozdělit trénovací množinu dat na homogenní podmnožiny a na každé z nich provést odhad regresních parametrů. Jako vhodné se nabízí dvě možná dělení. První, rozdělit data na několik skupin podle hodiny H, druhé, rozdělit data na několik skupin podle hodnoty, kterou v daný den a hodinu nabývá Clear sky model $\widehat{cs}_{sm,ST}$. Předem odhadujeme, že efektivnější by mohl být druhý způsob dělení, nicméně prozkoumejme i první způsob a pak je porovnejme. V této části budeme zkoumat 4 nejlepší modely z předchozích částí, což jsou modely:

- `M1_cut_Nc <- betareg (tau ~ G + Nl + Nc)`,
- `M3_sqrt2_Nc <- betareg (tau ~ sqrt(G) + sqrt(Nl) + sqrt(Nc))`,
- `M3_log_Nc <- betareg (tau ~ log(G+0.001) + Nl + Nc)`,
- `M10_sqrt2 <- lm (probit(tau) ~ sqrt(G) + sqrt(Nl) + sqrt(Nc))`.

Pro zjednodušení a zpřehlednění je postupně pojmenujme `Ma`, `Mb`, `Mc` a `Md`.

Začněme prvním způsobem dělení trénovací množiny a to podle hodin. Jsou dva způsoby, jak lze data podle hodiny efektivně rozdělit. Buď data z každé hodiny dáme do různé skupiny, případně spolu sloučíme data z 12. a 13. hodiny, data z 11. a 14. hodiny atd., kdy dáme dohromady hodiny, ve kterých by měly být podobné podmínky pro výrobu elektřiny z FVE. Podívejme se nejprve na výsledky, kdy data rozdělíme podle hodin na 24 skupin. U všech 4 modelů máme

nejvyšší hodnoty RMSE a MAE ve 12. a 13. hodině, přičemž tyto hodnoty klesají se vzdalováním se od poledních hodin. Hodnoty PBIAS byly okolo poledne blízké nule, v ranních a večerních hodinách již indikovaly významné vychýlení. Podíváme-li se na souhrnné hodnoty RMSE a MAE a porovnáme-li je s původními hodnotami, zjistíme, že došlo k drobnému zlepšení RMSE i MAE u všech modelů. Odhadování modelu na podmnožinách podle hodin přineslo pouze mírný pozitivní efekt v modelování.

Nyní provedeme odhady regresních parametrů na stejných modelech s tím, že data nyní rozdělíme na 12 podskupin podle vzdálenosti hodiny od poledne. Porovnáme-li výsledky s předchozí studií, zjistíme, že výsledky jsou téměř totožné, takže tato změna rozdělení žádný pozitivní efekt nepřinesla. Pouze drobné zlepšení od původních modelů se dalo očekávat zejména z důvodu, že v každé hodině je stále významný rozdíl mezi zimními a letními měsíci, proto tato rozdělení trénovací množiny nebyla nijak efektivní.

Pojďme nyní provést rozdělení, které by mělo být efektivnější, a to na základě hodnot, kterou v daný den a hodinu nabývá Clear sky model $\widehat{cs}_{sm,ST}$. V tomto případě není úplně zřejmé, jak hustou síť dělení vytvořit, zejména z důvodu, abychom nedostali buď spoustu malých podmnožin, nebo naopak bychom vytvořili zbytečně malý počet skupin a data v těchto skupinách pak nebudou dostatečně homogenní. Rozdělíme data podle hodnot $\widehat{cs}_{sm,ST}$ na κ podmnožin tak, že rozdělíme interval $[0; \max \widehat{cs}_{sm,ST}(d,h)], d \in D, h \in H$ právě na κ stejně dlouhých intervalů. Vzhledem k tomu, že v případě, kdy je $\widehat{cs}_{sm,ST}(D,H) = 0$ máme $\widehat{p}_{ST} = 0$, vynecháme tato data z dělení, neboť do modelu nepřináší žádnou užitečnou informaci. Provedeme výpočet pro hodnoty $\kappa \in \{5, 10, 15, 20\}$ s tím, že podle výsledku případně provedeme výpočet i na jiném vhodnějším dělení. Pro $\kappa = 5$ jsme dostali tyto výsledky, kde `J_tr` značí počet pozorování v dané podmnožině z trénovacích dat a `J_te` značí počet pozorování v dané podmnožině z testovacích dat.

```
 > vysl_CS_Ma
       RMSE_train  RMSE_test   MAE_train    MAE_test PBIAS_train J_tr J_te
1      0.01358708 0.01238435 0.007920077 0.007168016       -18.3 8265  658
2      0.04492460 0.04183788 0.034947498 0.033031360         3.8 3070  249
3      0.07454993 0.07507906 0.058372598 0.055893749         3.4 3760  291
4      0.09875203 0.09220144 0.077237511 0.069768459         0.9 4675  366
5      0.10754750 0.11468926 0.082025375 0.084710348         0.4 5145  466
Celkem 0.05525337 0.05613099 0.026834136 0.025997992          NA    0    0
Puv.Ma 0.05757243 0.05839921 0.028424236 0.027533601         0.6    0    0

> vysl_CS_Mb
       RMSE_train  RMSE_test   MAE_train    MAE_test PBIAS_train J_tr J_te
1      0.01250765 0.01070032 0.006924153 0.006181842        -0.5 8265  658
2      0.04407509 0.04070956 0.033747986 0.031948273         4.2 3070  249
3      0.07331733 0.07466162 0.056872957 0.056099003         3.5 3760  291
4      0.09788815 0.09316333 0.075983920 0.069994166         1.0 4675  366
5      0.10797829 0.11873290 0.082302952 0.088771936         0.3 5145  466
Celkem 0.05494951 0.05719780 0.026332198 0.026306062          NA    0    0
Puv.Mb 0.05685717 0.05635919 0.027937542 0.026471891         1.5    0    0

> vysl_CS_Mc
       RMSE_train  RMSE_test   MAE_train    MAE_test PBIAS_train J_tr J_te
1      0.01350126 0.01185335 0.007552749 0.006707824         3.7 8265  658
2      0.04509150 0.04281771 0.034264872 0.033687441         5.1 3070  249
```

```
       RMSE_train  RMSE_test   MAE_train    MAE_test PBIAS_train J_tr J_te
3      0.07573464 0.07812975 0.058550947 0.058911683         4.5 3760  291
4      0.09935730 0.09436290 0.076466849 0.073272254         1.6 4675  366
5      0.10943605 0.11926608 0.082297236 0.091337513         0.6 5145  466
Celkem 0.05594432 0.05805165 0.026681963 0.027407877          NA    0    0
Puv.Mc 0.05763553 0.05639063 0.027985923 0.026994428         2.0    0    0

> vysl_CS_Md
       RMSE_train  RMSE_test   MAE_train   MAE_test PBIAS_train J_tr J_te
1      0.01412689 0.01306161 0.007811342 0.00742752        -8.4 8265  658
2      0.04412210 0.04225007 0.032475627 0.03259461         2.0 3070  249
3      0.07372369 0.07705007 0.055103963 0.05681396         2.0 3760  291
4      0.09919802 0.09759050 0.074262294 0.07210468         1.6 4675  366
5      0.10946501 0.12332677 0.081584621 0.09181149         1.9 5145  466
Celkem 0.05566449 0.05956450 0.025990432 0.02723805          NA    0    0
Puv.Md 0.05760328 0.05923055 0.027411052 0.02749322         2.1    0    0
```

Z výsledku vidíme, že na trénovací množině dat klesly hodnoty RMSE a MAE pro všechny 4 modely. Na testovací množině dat se zmenšily hodnoty RMSE a MAE pouze u modelu `Ma`, u ostatních došlo buď k velmi nepatrnému poklesu, většinou ovšem k nárůstu těchto hodnot.

Při zvětšení kroku κ až na několik výjimek hodnoty RMSE a MAE velmi mírně klesají, nicméně zjemňování dělení nepřináší žádný výrazný efekt. Velké vychýlení v první skupině je způsobené zejména díky tomu, že se v této skupině nachází okrajové hodiny dne, kdy často bývá skutečné měření rovno nule, ale díky nenulové hodnotě $\widehat{cs}_{ST}$ predikujeme mírnou výrobu. Díky tomu, že se ve jmenovateli PBIAS nulové hodnoty měření neprojeví, je výsledná procentuální odchylka velká. Nicméně skutečný dopad těchto chyb do odchylky je velmi malý.

5.3.7 Závěrečný výběr modelu

Nyní přichází na řadu nejtěžší část, kterou je výběr výsledného modelu pro predikci na den $D + 1$. Všechny 4 modely jsou podle hodnot RMSE a MAE srovnatelné, což situaci neulehčuje, na druhou stranu zřejmě při jakékoliv volbě neuděláme zásadní chybu. Moji volbou je model `Ma`. Ten má nejmenší hodnoty RMSE a MAE na testovacích datech zatímco na trénovacích datech nijak nezaostává. Jeho velkou výhodnou je nejmenší vychýlení ze všech modelů. Proti modelům `Mb` a `Md` hovoří odmocninová transformace vysvětlujících proměnných, kdy se velmi těžko hledá logické vysvětlení, proč bychom měli tuto transformaci udělat. V modelu `Mc` je použita logaritická transformace proměnné $\widehat{G}_{ST}$, která ale nabývá nulových hodnot, které v tomto případě mají určitý význam. Protože zjemňování dělení nepřináší žádný velký významný efekt, zvolíme jako výsledný krok $\kappa = 10$, který je kompromisem mezi drobným zlepšením odhadu a počtem podmnožin. Výsledným modelem tedy je

```
M_ST <- betareg(tau ~ G + Nl + Nc, subset(data_ST,
       CS>deleni[i] & CS<=deleni[i+1]), link="logit")

        CS_od      CS_do   Intercept           G        Nl         Nc    (phi)
1  0.00000000 0.08507898 -1.52679740 0.067880790 -0.899326  0.2267343 1.285762
2  0.08507898 0.17015795 -0.22456052 0.022271006 -1.960592 -0.1154049 5.757104
3  0.17015795 0.25523693 -0.62902772 0.021486293 -1.736473 -0.0405865 4.779262
```

```
        CS_od      CS_do   Intercept           G        Nl         Nc    (phi)
4  0.25523693 0.34031591  0.04303311 0.011655513 -1.698942 -0.6696537 4.945573
5  0.34031591 0.42539489 -0.50621288 0.011872999 -1.493595 -0.3287910 5.504136
6  0.42539489 0.51047386 -0.08025708 0.008398100 -1.666145 -0.6909075 4.882802
7  0.51047386 0.59555284 -0.60254835 0.008641957 -1.626275 -0.1135308 5.748566
8  0.59555284 0.68063182 -0.06406100 0.005806864 -1.707667 -0.3495451 6.042936
9  0.68063182 0.76571080 -0.01095844 0.004886936 -1.838548 -0.0903325 7.573514
10 0.76571080 0.85078977  0.22905772 0.004135656 -1.923181 -0.0785607 8.465325
```

V první skupině nastala situace, že koeficient u $\widehat{Nc}_{ST}$ je kladný, proto jej v této skupině vynecháme a pro 1. skupinu dostaneme model

```
M_ST_1 <- betareg(tau ~ G + Nl, subset(data_ST,
          CS>deleni[1] & CS<=deleni[2]), link="logit")
```

```
     Intercept           G        Nl          Nc    (phi)
1  -1.52679740 0.067880790 -0.899326  0.22673426 1.285762
```

Výslednou predikci pro jednotlivé dny a hodiny je možné najít v příloze v souboru `predikce.csv`.

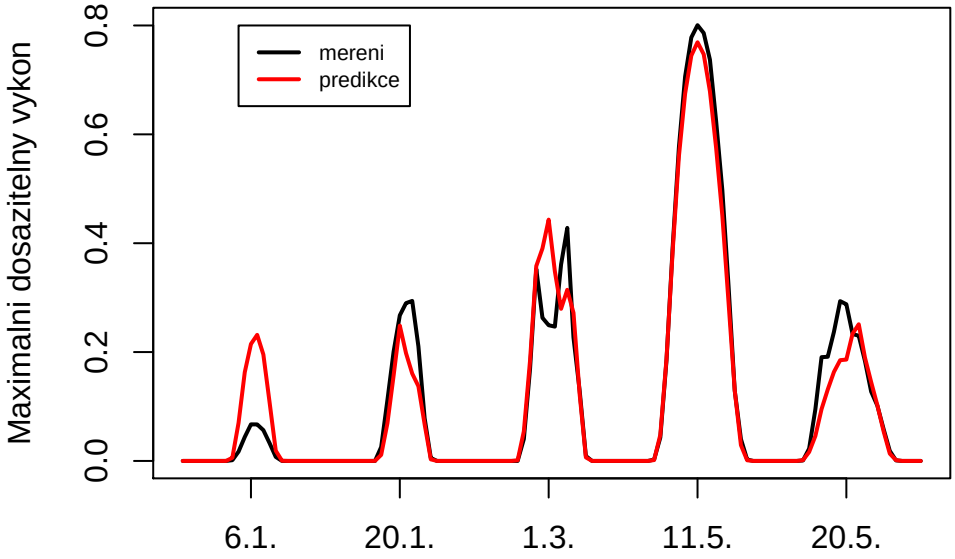
Proveďme ještě porovnání výsledného modelu s referenčním modelem. Dostaneme $IMPR_{RMSE,tr} = 54{,}43\%$, $IMPR_{RMSE,test} = 55{,}36\%$, $IMPR_{MAE,test} = 51{,}06\%$ a $IMPR_{MAE,test} = 55{,}76\%$. Z těchto čísel vidíme, že oproti referenčnímu modelu se náš model zlepši o více než 50 % ve všech kritériích.

Za výsledný model na den $D + 1$ v dalších regionech vezmeme úplně stejný model jako ve Středočeském regionu, pouze provedeme nový odhad parametrů. V tomto případě je jedinou otázkou, z jakého kraje použít data z modelu Aladin, protože pokud bychom použili data z obou krajů, dostali bychom multikolineární model. Nejlepší možností je zvolit vážený průměr těchto dat, např. pro Severomoravský region dostaneme $\widehat{G}_{SM} = \omega_1\widehat{G}_{MS} + \omega_2\widehat{G}_{OL} + \omega_3\widehat{G}_{ZL}$, kde $\omega_1 + \omega_2 + \omega_3 = 1$ a $0 \leq \omega_1, \omega_2, \omega_3 \leq 1$. Jednou z vhodnou voleb vah je zvolit jednu rovnou 1 a zbylé váhy rovno 0. Další vhodnou volbou je přiřazení každé proměnné váhu podle procentuálního zastoupení instalovaného výkonu FVE v daném kraji. Vzhledem k tomu, že procentuální zastoupení instalovaného výkonu FVE z jednotlivých krajů neznáme, ponecháme volbu vah otevřenou.

Pokud by v nějaké skupině nastala situace, kdy odhad regresního parametru $\widehat{Nc}_R$ bude kladný, přejdeme z modelu `M_R_sk <- betareg(tau ~ G + Nl + Nc)` k modelu bez této proměnné, tj. k modelu `M_R_sk <- betareg(tau ~ G + Nl)`.

Na závěr se ještě na vybraných dnech graficky podíváme, zdali predikce výsledného modelu odpovídají měření. Z roku 2017 jsme vybrali dny 6. leden - zimní den s minimální výrobou, 20. leden - zimní den s významnou výrobou, 1. březen - den, kdy máme dvě maxima výroby, jedno v dopoledních a jedno v odpoledních hodinách, 11. květen - jasný den v letním období a 20. květen - den v letním období s nízkou výrobou. Na obrázku 5.3 vidíme porovnání predikce výsledného modelu a měření. Dále ještě doplníme obrázek 5.4, kde máme predikci v porovnáním s predikcí dopadajícího slunečního záření z modelu Aladin. Z obrázků vidíme, že 6. ledna model odhadl výrobu výrazně výše, než byla skutečnost. To je ale zapříčiněno tím, že predikce dopadajícího slunečního záření je podobná predikci dne 20. ledna, takže nadhodnocení je způsobeno nepřesností predikce modelu Aladin. 20. ledna se model přiblížil měření, pokles v odpoledních hodinách byl

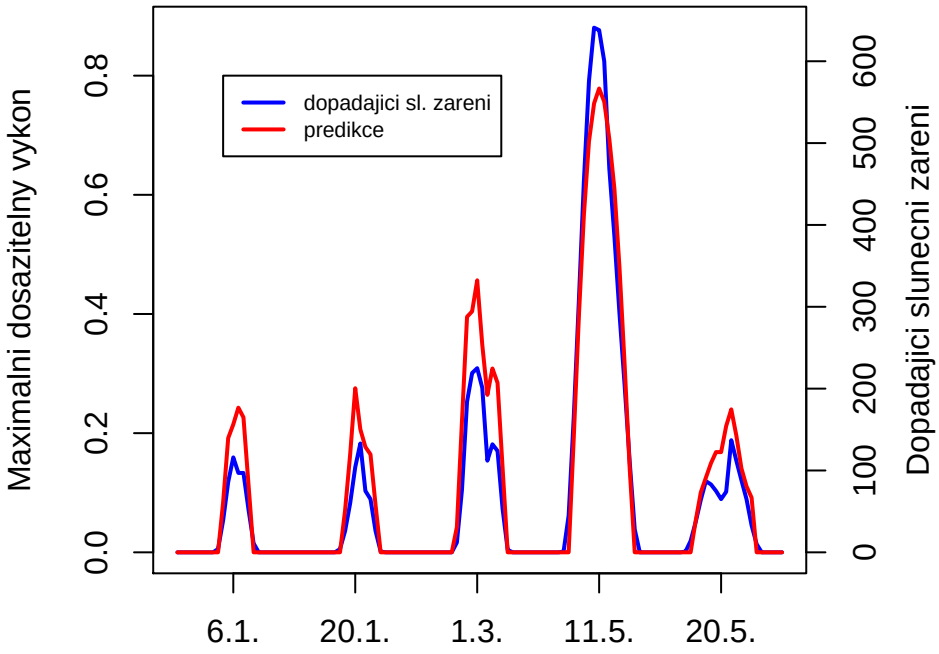
opět způsoben poklesem predikce dopadajícího slunečního záření. 1. března se odhad modelu ve výši výroby liší jen minimálně od měření, akorát v tomto případě došlo k situaci, že model predikoval vyšší výrobu v dopoledních hodinách, oproti skutečnosti, která naopak byla vyšší v odpoledních hodinách. Tato chyba je zjevně opět způsobená nepřesností predikce dopadajícího slunečního záření, kdy predikce výsledného modelu kopíruje predikci modelu Aladin. 11. května se predikce víceméně trefila do skutečnosti, 20. května vykazuje drobné odchylky, nicméně tvarově je predikce podobná měření.



Obrázek 5.3: Porovnání predikce výroby vytvořenou modelem `M_ST` a měření na vybraných dnech roku 2017 ve Středočeském regionu.

5.4 Model na den D

Nyní se podíváme na model, který bude predikovat výrobu na aktuální den. Protože již máme hotový model na den $D+1$, který zároveň funguje bez jakýchkoliv úprav i jako model na den D, můžeme modelování pojmout jako zlepšování již hotového modelu, kdy se pokusíme na blízkých hodinách v budoucnosti využít známé informace z již uplynulých hodin. Tvorbu modelu na den D si popíšeme na modelu pro Středočeský region. Nejprve si ujasníme jaká data máme v hodině H k dispozici a jaký krok k předpovědi má smysl uvažovat, abychom případně nevytvářeli modely, které nemají využití. V hodině H máme k dispozici procentuální množství naměřeného „online“ výkonu vybraných FVE z instalovaného výkonu, které jsme značili q_{ST}, z hodiny $H-1$ a všech předchozích hodin spolu s informací o procentuálním množství instalovaného výkonu vybraných FVE oproti instalovanému výkonu všech FVE, které jsme značili s_{ST}. Data s_{ST} jsou pro vytvoření modelu nepotřebná, jsou spíše informačního charakteru, kdy mohou označit ty časy, ve kterých by data z „online“ měření mohla být zkreslena malým instalovaným výkonem FVE, ze kterých máme „online“ měření k dispozici. Dále máme k dispozici měření ze stanic ČHMÚ rovněž z hodiny $H-1$ a všech předchozích, v tomto případě budeme uvažovat měření ze stanice Praha - Libuš.



Obrázek 5.4: Porovnání predikce výroby vytvořené modelem `M_ST` a predikce dopadajícího slunečního záření z modelu Aladin na vybraných dnech roku 2017 ve Středočeském regionu.

Vytvářet predikci na aktuální hodinu již nemá smysl, protože v této hodině s odchylkou již stejně nic neuděláme, proto tuto situaci uvažovat nebudeme. Na hodinu $H+1$ je jedinou možností, jak ovlivnit množství nakoupené elektřiny vyrovnávací trh, nicméně zde obchody záleží výhradně na provozovateli přenosové soustavy ČEPS, a.s., tudíž možnosti vykupujícího, jak ovlivnit výši odchylky, jsou minimální. Zaměřme se proto na predikci na hodinu $H+2$ a následující hodiny, kde máme možnost nákupu příp. prodeje elektřiny na vnitrodenním trhu.

Nejprve je potřeba si uvědomit, že rozdíl mezi hodinou, na kterou chceme predikovat a hodinou, ze které máme „online“ data je minimálně tříhodinový (to je při predikci na $H+2$) Pokud je hodnota $\widehat{cs}_{sm,ST}(H-1) = 0$ nebo nule blízká, „online“ data z této doby budou nepoužitelná a model při jejich použití nejspíš jen zhoršíme. Predikce z modelu na den D, tedy budeme vytvářet pouze pro hodiny H, které současně splňují podmínku $\widehat{cs}_{sm,ST}(D,H) > 0$ a podmínku $\widehat{cs}_{sm,ST}(D,H-3) > \gamma_{cut}$, kde γ_{cut} zvolíme 0,1.

Jednou z nejčastěji užívaných metod pro predikci časových řad je ARIMAX model (autoregressive integrated moving average with exogenous input). V našem případě se tato metoda nehodí, neboť jednak chceme predikovat řadu p_{ST} pomocí řady q_{ST}, navíc i při přechodu na řadu τ_{ST} máme období, ve kterém je $\tau_{ST} = 0$. Nicméně to nám nebrání se částečně inspirovat a zařadit buď do beta regresního modelu nebo lineárního regresního modelu proměnné vyjadřující zpožděné hodnoty proměnné q_{ST}. Než začneme modelovat, podíváme se, jakou závislost má proměnná $q_{ST}(H-1)$ s proměnnými $p_{ST}(H+2)$, $p_{ST}(H+3)$ a $p_{ST}(H+4)$.

```
 > cor(subset(data_ST$q_3, data_ST$CS > 0 & data_ST$CS_3 > 0.1,
+   data_ST$s_3 > 0.1), subset(data_ST$p, data_ST$CS > 0 & data_ST$CS_3 > 0.1,
+   data_ST$s_3 > 0.1), "complete.obs")
[1] 0.4550162
```

```
> cor(subset(data_ST$q_4, data_ST$CS > 0 & data_ST$CS_4 > 0.1,
+   data_ST$s_4 > 0.1), subset(data_ST$p, data_ST$CS > 0 & data_ST$CS_4 > 0.1,
+   data_ST$s_4 > 0.1), "complete.obs")
[1] 0.2359807
> cor(subset(data_ST$q_5, data_ST$CS > 0 & data_ST$CS_5 > 0.1 ,
+   data_ST$s_5 > 0.1), subset(data_ST$p, data_ST$CS > 0 & data_ST$CS_5 > 0.1,
+   data_ST$s_5 > 0.1), "complete.obs")
[1] 0.06048124
```

Z Pearsonových korelačních koeficientů vidíme, že při tříhodinovém rozdílu bychom mohli nějakou závislost předpokládat a tím i vylepšit model, u pětihodinového rozdílu je možné, že časová vzdálenost je již příliš velká a lepší model nenajdeme. Zkusíme se ještě podívat na závislost mají proměnné $\tau_{ST}(H+2)$, $\tau_{ST}(H+3)$ a $\tau_{ST}(H+4)$ s proměnnou $q_{ST}(H-1)$. V tomto případě ale budeme muset zkoumat závislost s proměnnou $q_{ST}(H-1)/\widehat{cs}_{sm,ST}(H-1)$, neboť v předchozím případě by výsledek nedával smysl.

```
 > cor(subset(data_ST$q_3/data_ST$CS_3, data_ST$CS > 0 & data_ST$CS_3 > 0.1,
+   data_ST$s_3 > 0.1), subset(data_ST$tau, data_ST$CS > 0 &
+   data_ST$CS_3 > 0.1, data_ST$s_3 > 0.1), "complete.obs")
[1] 0.6794827
> cor(subset(data_ST$q_4/data_ST$CS_4, data_ST$CS > 0 & data_ST$CS_4 > 0.1,
+   data_ST$s_4 > 0.1), subset(data_ST$tau, data_ST$CS > 0 &
+   data_ST$CS_4 > 0.1, data_ST$s_4 > 0.1), "complete.obs")
[1] 0.6275539
> cor(subset(data_ST$q_5/data_ST$CS_5, data_ST$CS > 0 & data_ST$CS_5 > 0.1 ,
+   data_ST$s_5 > 0.1), subset(data_ST$tau, data_ST$CS > 0 &
+   data_ST$CS_5 > 0.1, data_ST$s_5 > 0.1), "complete.obs")
[1] 0.5747001
```

Zde již vidíme, že závislost můžeme předpokládat i při pětihodinovém rozdílu. Na základě Pearsonových korelačních koeficientů bude rozumné zkoumat pouze modely s vysvětlovanou proměnnou τ. Při vysvětlované proměnné p můžeme mít problém s tím, že maximální dosažitelné výkony jsou různé v hodině predikce a hodině, ze které máme data. Nejprve budeme hledat model pro predikci s krokem $k = 2$, tj. na hodinu $H + 2$, poté nejlepší model upravíme pro predikce s krokem $k = 3$ a $k = 4$

5.4.1 Beta regrese

Začněme opět beta regresním modelem. V tomto případě projdeme dvě možnosti, které můžeme použít, a to buď tvorbu modelu od začátku, nebo již zkusíme využít známé predikce z modelu `Ma_ST`, který jsme vytvořili pro predikci na den $D + 1$. V tomto případě ale již do modelu nebudeme dávat data z modelu Aladin, neboť bychom mohli mít problémy s multikolinearitou. Začneme s druhou variantou, tedy vysvětlující proměnnou bude predikce z modelu `Ma_ST`, kterou označíme $\widehat{p}_{Ma,ST}$.

Jako obvykle prozkoumáme nejprve úplný model s tím, že do něj zařadíme jen proměnné z času $H-1$, neboť informace o počasí nebo měření v čase $H-2$ je při známosti novější informace z času $H-1$ již zbytečná. V úplném modelu máme všechny proměnné významné a hodnoty RMSE a MAE oproti hodnotám z modelu

`Ma_ST` klesly, takže se zdá, že dodatečné informace přinášejí lepší model. Zkusíme z modelu vyřadit proměnnou $Nh_{LS}(H-1)$, kterou jsme v předchozí sekci vždy vyřadili jako nevýznamnou. Jejím vyřazením jsme docílili poklesu hodnot RMSE a MAE, tudíž tato proměnná bude nejspíš i zde nevýznamná. V novém modelu máme na hladině $\alpha = 0{,}01$ nevýznamnou proměnnou $G_{LS}(H-1)$, tudíž zkusíme i ji z modelu vynechat. Jejím vyřazením se hodnoty RMSE a MAE změnily jen minimálně. Zkusíme ještě z modelu vyřadit proměnnou $Nc_{LS}(H-1)$. Po jejím vyřazení ale hodnoty RMSE a MAE vzrostly tak, že je máme vyšší než byly modelu `Ma_ST`, takže vyřazení proměnné $Nc_{LS}(H-1)$ zamítneme. Při pokusu o vyřazení $q_{ST}(H-1)$ jsme dopadli stejně.

Zkusme ještě použít nějakou transformaci na vysvětlující proměnné. U toho modelu se velmi nabízí transformovat vysvětlující proměnné logitovou funkcí, neboť hodnoty $\widehat{p}_{Ma,ST}(H+2)$ a hodnoty $q_{ST}(H-1)$ se nachází na intervalu $(0,1)$. Tato transformace byla užitečná, protože hodnoty RMSE a MAE poklesly. Dostali jsme model

```
 > M1_logit <- betareg(tau ~ logit(tau_pr) + logit(q_3) + Nc_3,
+   subset(data_ST, data_ST$CS > 0 & data_ST$CS_3 > 0.1))
> coeftest(M1_logit)

z test of coefficients:

               Estimate Std. Error z value  Pr(>|z|)
(Intercept)    0.5558042  0.0208162  26.701 < 2.2e-16 ***
logit(tau_pr)  0.5771666  0.0078787  73.256 < 2.2e-16 ***
logit(q_3)     0.1912704  0.0071035  26.926 < 2.2e-16 ***
Nc_3          -0.0605714  0.0038876 -15.581 < 2.2e-16 ***
(phi)          3.4346922  0.0368450  93.220 < 2.2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> rmse(predict(M1_logit)*subset(data_ST$CS, data_ST$CS > 0 &
+   data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)), subset(data_ST$p,
+   data_ST$CS > 0 & data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)))
[1] 0.07350414
> rmse(subset(data_ST$pr, data_ST$CS > 0 & data_ST$CS_3 > 0.1 &
+   !is.na(data_ST$Nc_3)), subset(data_ST$p, data_ST$CS > 0 &
+   data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)))
[1] 0.0787369
> mae(predict(M1_logit)*subset(data_ST$CS, data_ST$CS > 0 &
+   data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)), subset(data_ST$p,
+   data_ST$CS > 0 & data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)))
[1] 0.05006921
> mae(subset(data_ST$pr, data_ST$CS > 0 & data_ST$CS_3 > 0.1 &
+   !is.na(data_ST$Nc_3)), subset(data_ST$p, data_ST$CS > 0 &
+   data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)))
[1] 0.05147474
```

Jiné transformace, např. logaritmická nebo odmocninová vedly k horším výsledkům.

Nyní zkusíme model sestavit „od začátku“, tj. za pomocí predikcí z modelu Aladin a nikoliv na $\widehat{p}_{Ma,ST}(H+2)$. V tomto případě se však s žádným modelem nedokážeme v hodnotách RMSE a MAE přiblížit modelu `M1_logit`, tudíž dále budeme uvažovat pouze tento model.

5.4.2 Lineární regrese na transformovanou vysvětlovanou proměnnou τ

Druhou možností je místo beta regrese použít lineární regresi opět na transformovanou vysvětlovanou proměnnou τ_{ST} tak, abychom dostali predikce v intervalu $(0, 1)$. Začneme jako předtím s vysvětlovanou proměnnou $\widehat{p}_{Ma,ST}(H+2)$ a opět s úplným modelem. V tomto případě máme v porovnání s modelem `Ma_ST` nepatrně nižší hodnotu RMSE a relativně větší rozdíl v hodnotách MAE. Vyřazením proměnné $Nh_{LS}(H-1)$ jsme opět docílili drobného poklesu hodnot RMSE a MAE, vyřazením na hladině $\alpha = 0{,}001$ nevýznamné proměnné $G_{LS}(H-1)$, došlo jen k nepatrnému nárůstu hodnot RMSE a MAE, tudíž ji jako v předchozím případě rovněž vynecháme. Při pokusu o vyřazení proměnné $Nc_{LS}(H-1)$ nebo $q_{ST}(H-1)$ již hodnoty RMSE a MAE vzrostly více, tudíž tyto v modelu ponecháme.

Zkusme opět použít nějakou transformaci i na vysvětlující proměnné. Zde se opět velmi nabízí transformace logitovou funkcí, neboť stejnou transformaci jsme použili i na vysvětlovanou proměnnou. Tato transformace byla opět užitečná, protože hodnoty RMSE a MAE poklesly. Dostali jsme model

```
 > summary(M3_logit)

Call:
lm(formula = logit(tau) ~ logit(tau_pr) + logit(q_3) + Nc_3,
    data = subset(data_ST, data_ST$CS > 0 & data_ST$CS_3 > 0.1))

Residuals:
    Min      1Q  Median      3Q     Max
-6.2098 -0.7652 -0.1240  0.4935 14.4537

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.92944    0.03889   23.90   <2e-16 ***
logit(tau_pr) 0.88603    0.01423   62.28   <2e-16 ***
logit(q_3)    0.29799    0.01315   22.65   <2e-16 ***
Nc_3         -0.10290    0.00730  -14.10   <2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.801 on 15433 degrees of freedom
  (743 observations deleted due to missingness)
Multiple R-squared:  0.462,     Adjusted R-squared:  0.4619
F-statistic:  4418 on 3 and 15433 DF,  p-value: < 2.2e-16

> rmse(logit(predict(M3_logit), inverse = TRUE)*subset(data_ST$CS, data_ST
+   $CS > 0 & data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)), subset(data_ST$p,
+   data_ST$CS > 0 & data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)))
[1] 0.07431224
> rmse(subset(data_ST$pr, data_ST$CS > 0 & data_ST$CS_3 > 0.1 &
+   !is.na(data_ST$Nc_3)), subset(data_ST$p, data_ST$CS > 0 &
+   data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)))
[1] 0.0787369
> mae(logit(predict(M3_logit), inverse = TRUE)*subset(data_ST$CS, data_ST
+   $CS > 0 & data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)), subset(data_ST$p,
+   data_ST$CS > 0 & data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)))
[1] 0.04598321
```

```
> mae(subset(data_ST$pr, data_ST$CS > 0 & data_ST$CS_3 > 0.1 &
+   !is.na(data_ST$Nc_3)), subset(data_ST$p, data_ST$CS > 0 &
+   data_ST$CS_3 > 0.1 & !is.na(data_ST$Nc_3)))
[1] 0.05147474
```

Odmocninová a logaritmická transformace v tomto případě vedla k horším hodnotám RMSE a MAE než hdonoty u modelu `M3_logit`.

Jako poslední model zkusíme sestavit lineární regresi „od začátku“, tj. nahrazením $\widehat{p}_{Ma,ST}(H+2)$ predikcemi z modelu Aladin. Opět i zde se s žádným modelem nedokážeme v hodnotách RMSE a MAE ani přiblížit modelu `M3_logit`, tudíž dále budeme uvažovat pouze tento model.

5.4.3 Rozdělení dat na homogenní podmnožiny

V minulé sekci jsme ukázali, že rozdělení trénovací množiny dat na homogenní podmnožiny a provedení odhadů regresních parametrů na každé zvlášť přineslo do modelu zlepšení. I zde rozdělíme data na několik skupin s tím, že nyní použijeme pouze dělení podle hodnoty, kterou v daný den a hodinu nabývá Clear sky model $\widehat{cs}_{sm,ST}$, přičemž použijeme stejné hranice tříd jako u modelu na den $D+1$. V této části budeme zkoumat 2 nejlepší modely z předchozích částí, což jsou modely:

- `M1_logit <- betareg (tau ~ logit(tau_pr) + logit(q_3) + Nc_3)`,
- `M3_logit <- lm (tau ~ logit(tau_pr) + logit(q_3) + Nc_3)`.

Pro zjednodušení a zpřehlednění je postupně pojmenujme `MA` a `MB`. Po provedení odhadů v jednotlivých množinách máme

```
 > vysl_MA
       RMSE_train   RMSE_test   MAE_train    MAE_test PBIAS_train
1+2    0.01138175 0.009823274 0.006543256 0.005828612        -7.7
3      0.03300547 0.033982119 0.025826988 0.025592538         1.4
4      0.04684762 0.040296872 0.036807871 0.032120397         2.8
5      0.06041154 0.062271940 0.046749866 0.046213129         0.4
6      0.07534592 0.074109784 0.058678886 0.053948377         5.6
7      0.08413794 0.082506662 0.065081106 0.061477691         0.2
8      0.09791384 0.081174995 0.075520183 0.060238699        -0.2
9      0.09769467 0.117870760 0.074688664 0.091497025        -0.3
10     0.09869583 0.097192283 0.074632337 0.071714310         0.8
Celkem 0.05055656 0.051126096 0.024178188 0.023459514         0.4
Puv.Ma 0.05466781 0.055590580 0.026326334 0.025814289         0.6

> vysl_MB
       RMSE_train  RMSE_test   MAE_train    MAE_test PBIAS_train
1+2    0.01168619 0.01013545 0.006705939 0.005927441        -8.9
3      0.03299471 0.03472665 0.025225968 0.026090484         0.8
4      0.04786550 0.04034461 0.036697652 0.031220150         3.3
5      0.06123914 0.06213088 0.046397261 0.044364378         0.8
6      0.07635697 0.07512152 0.058125459 0.054847787         5.4
7      0.08596477 0.08568459 0.064777221 0.063639305         0.5
8      0.10054846 0.08639138 0.075083600 0.061579150         1.4
9      0.10227610 0.12393441 0.075280758 0.095449837         2.1
10     0.10524972 0.10515625 0.077041374 0.076378805         3.3
Celkem 0.05239460 0.05378338 0.024266834 0.024171382         1.8
Puv.Ma 0.05466781 0.05559058 0.026326334 0.025814289         0.6
```

Z hodnot RMSE a MAE vidíme, že oba modely jsou pro predikci na $H+2$ lepší než původní model `Ma_ST`. Z modelů `MA` a `MB` je jednoznačně lepší model `MA`, který je ve všech pěti kritériích lepší než model `MB`, tudíž jej pro predikci na $H+2$ budeme považovat za finální model. Přiložíme ještě odhady regresních parametrů modelu `MA`

```
 > koef_MA1
       CS_od     CS_do Intercept        pr         q          Nc    (phi)
1+2 0.0000000 0.1701580 0.2710224 0.8424510 0.2534826  0.00000000 9.451270
3   0.1701580 0.2552369 0.5639431 0.6330659 0.4506567 -0.04198057 6.502427
4   0.2552369 0.3403159 0.6875414 0.5394426 0.4175526 -0.07182321 7.690386
5   0.3403159 0.4253949 0.7131354 0.6314603 0.3291717 -0.08378330 7.136692
6   0.4253949 0.5104739 0.9273648 0.6198448 0.2486383 -0.11933964 7.967213
7   0.5104739 0.5955528 0.8862837 0.6430679 0.2247580 -0.11977136 7.654682
8   0.5955528 0.6806318 0.9061646 0.7277394 0.1390518 -0.12694609 7.867107
9   0.6806318 0.7657108 0.7368734 0.7546474 0.1294287 -0.10040774 9.186470
10  0.7657108 0.8507898 0.6196096 0.7678702 0.2072828 -0.06442698 9.817541

> koef_MA2
       CS_od     CS_do   Intercept        pr         q    (phi)
1+2 0.0000000 0.1701580 -0.01004465 0.6689749 0.3674597 1.553016
3   0.1701580 0.2552369  0.36631135 0.6481130 0.4928501 6.431396
4   0.2552369 0.3403159  0.30899107 0.5711479 0.4811867 7.517278
5   0.3403159 0.4253949  0.26132245 0.6837862 0.3758923 6.845227
6   0.4253949 0.5104739  0.31991329 0.7241256 0.3008675 7.288156
7   0.5104739 0.5955528  0.22615165 0.7447903 0.2694532 7.094278
8   0.5955528 0.6806318  0.21965865 0.9015106 0.1447258 6.998034
9   0.6806318 0.7657108  0.18568812 0.9064701 0.1402564 8.113026
10  0.7657108 0.8507898  0.27102236 0.8424510 0.2534826 9.451270
```

V dolní části tabulky najdeme koeficienty `koef_MA2`, které jsou odhady regresních parametrů modelu `MA2 <- betareg (tau ~ logit(tau_pr)+logit(q_3)`, což je model `MA` bez vysvětlující proměnné $Nc_{LS}(H-1)$. Odhadnout tento model je nutné, neboť v některých případech nemáme měření oblačnosti $Nc_{LS}(H-1)$ k dispozici. Pokud bychom ponechali jako výsledek původní predikci modelu `Ma_ST`, potom bychom zbytečně vyloučili informaci, kterou nám dává „online“ měření $q_{ST}(H-1)$. Ve skupině 1+2 nastala již známá situace, kdy odhad regresního parametrů u proměnné $Nc_{LS}(H-1)$ byl kladný, proto jsme tuto proměnnou v této skupině z odhadu vyloučili.

Proveďme opět porovnání výsledného modelu s referenčním modelem, kterým je nyní model `Ma_ST`. Máme $IMPR_{RMSE,tr} = 12{,}15\%$, $IMPR_{RMSE,test} = 12{,}41\%$, $IMPR_{MAE,test} = 14{,}77\%$ a $IMPR_{MAE,test} = 14{,}57\%$. Z těchto čísel vidíme, že predikci na hodinu $H+2$ se nám povedlo vylepšit o více než 10 % ve všech kritériích.

5.4.4 Model na H+3 a H+4

Pro predikci na hodinu $H+3$ použijeme úplně stejný model jako pro predikci na hodinu $H+2$, akorát nahradíme $\widehat{p}_{Ma,ST}(H+2)$ za $\widehat{p}_{Ma,ST}(H+3)$. Dostaneme model, který označíme `MC`.

```
 > vysl_MC
       RMSE_train  RMSE_test   MCE_train    MCE_test PBIAS_train
1+2    0.01223241 0.01085348 0.007043722 0.006409723       -15.0
3      0.03392757 0.03448389 0.026680597 0.026388142         1.2
4      0.04808366 0.04246047 0.038207080 0.034636891         2.7
5      0.06176013 0.06367066 0.048085493 0.048127244         0.3
6      0.07769391 0.07521650 0.060539183 0.055469183         3.5
7      0.08644370 0.08608761 0.066567294 0.064075261        -0.6
8      0.11223877 0.09291621 0.086510160 0.067993218        -6.0
9      0.10475183 0.13193625 0.080801597 0.104532334        -3.6
10     0.10257590 0.10607011 0.078120907 0.078139115        -0.7
Celkem 0.05392805 0.05560038 0.025749141 0.025499344        -2.2
Puv.Ma 0.05466781 0.05559058 0.026326334 0.025814289         0.6

> koef_MC1
        CS_od     CS_do  Intercept        pr         q           Nc    (phi)
1+2 0.0000000 0.1701580 -0.1967511 0.7642103 0.2130096  0.00000000 1.488122
3   0.1701580 0.2552369  0.4343575 0.6977786 0.2996581 -0.04526332 5.860351
4   0.2552369 0.3403159  0.4897293 0.6275086 0.2890434 -0.05300954 6.738238
5   0.3403159 0.4253949  0.6424548 0.7184082 0.1913414 -0.08739982 6.395828
6   0.4253949 0.5104739  0.6769198 0.6699133 0.2076813 -0.08042308 7.237802
7   0.5104739 0.5955528  0.6528938 0.7064463 0.1774827 -0.08302783 7.182449
8   0.5955528 0.6806318  0.5707200 0.7561121 0.1603560 -0.06875596 8.142451
9   0.6806318 0.7657108  0.4771436 0.8353690 0.1160703 -0.06138472 8.545525
10  0.7657108 0.8507898  0.5650261 0.8246774 0.1546720 -0.05550981 9.971239

> koef_MC2
        CS_od     CS_do  Intercept        pr         q    (phi)
1+2 0.0000000 0.1701580 -0.1967511 0.7642103 0.2130096 1.488122
3   0.1701580 0.2552369  0.2235119 0.7092886 0.3434660 5.634972
4   0.2552369 0.3403159  0.2028010 0.6580679 0.3307547 6.734221
5   0.3403159 0.4253949  0.1656924 0.7869730 0.2211084 6.101766
6   0.4253949 0.5104739  0.2761116 0.7039854 0.2823861 6.838994
7   0.5104739 0.5955528  0.1947583 0.7771876 0.2119137 6.821763
8   0.5955528 0.6806318  0.1931193 0.8352455 0.1860675 7.834001
9   0.6806318 0.7657108  0.1621106 0.9176874 0.1251103 8.394104
10  0.7657108 0.8507898  0.2767153 0.9100431 0.1642768 9.421455
```

Ve skupině 1+2 byl opět odhad regresního parametrů u proměnné $Nc_{LS}(H-1)$ kladný, proto byl i u tohoto modelu v této skupině z odhadu vynechán. Porovnáním hodnot RMSE a MAE s původním modelem `Ma_ST` zjistíme, že hodnoty stále nepatrně klesly, takže pro predikci na hodinu $H+3$ můžeme použít model `MC`. Drobnou nevýhodou tohoto modelu je, že vykazuje mírné vychýlení.

Porovnáním modelu `MC` s referenčním modelem `Ma_ST` dostaneme vylepšení $IMPR_{RMSE,tr} = 6{,}33\%$, $IMPR_{RMSE,test} = 4{,}8\%$, $IMPR_{MAE,test} = 9{,}42\%$ a $IMPR_{MAE,test} = 7{,}4\%$. Z těchto čísel vidíme, že predikci na hodinu $H+3$ se nám povedlo vylepšit ve všech kritériích.

Pro predikci na hodinu $H+4$ opět použijeme úplně stejný model jako pro predikci na hodinu $H+2$, akorát zde nahradíme $\widehat{p}_{Ma,ST}(H+2)$ za $\widehat{p}_{Ma,ST}(H+4)$. Dostaneme model ozn. `ME`

```
 > vysl_ME
       RMSE_train  RMSE_test   MEE_train    MEE_test PBIAS_train
1+2    0.01234368 0.01099657 0.007110164 0.006484253       -14.9
3      0.03469510 0.03445142 0.027276657 0.026514267         1.3
4      0.04880497 0.04422328 0.038572917 0.036597148         2.4
5      0.06344345 0.06393041 0.048986807 0.048187334        -1.6
6      0.09090065 0.07812781 0.067703705 0.057720167        -1.9
7      0.08983373 0.09008984 0.068881676 0.068216117        -1.4
8      0.15954349 0.11670148 0.116097349 0.081657616       -17.6
9      0.14527714 0.17279731 0.109358609 0.134986494       -13.1
10     0.11252433 0.11836541 0.085250879 0.088989408        -3.0
Celkem 0.06730039 0.06526552 0.030126054 0.028980893        -7.6
Puv.Ma 0.05466781 0.05559058 0.026326334 0.025814289         0.6
```

Zde vidíme, že hodnoty RMSE a MAE jsou již významně vyšší než u původního modelu `Ma_ST`. Odsud lze usoudit, že znalost měření 5 hodin nazpět již nepřináší pozitivní efekt pro predikce, proto pro hodiny $H+4$ a následující hodiny dne D budeme používat model `Ma_ST`, který má za vysvětlující proměnné pouze predikce modelu Aladin.

Závěr

Práce je věnována tvorbě modelu pro krátkodobou predikci výroby elektrické energie z FVE. Jedná se o zdroje elektřiny, které jsou dnes ve většině zemí Evropské unie preferovány, a protože se nejedná o zdroje, u kterých můžeme ovlivnit výrobu, je (a do budoucna i bude) potřeba výrobu z těchto zdrojů predikovat.

Na začátku práce jsme představili fungování trhu s elektřinou v ČR a způsob, jakým je výroba elektřiny ve FVE podporována. Zejména informace o trhu byly vodítkem pro určení časů, v jakých je potřeba predikce provádět. V další kapitole byla představena zdrojová data, která byla používána při tvorbě modelů.

Ve třetí kapitole byly představeny kvantilová regrese a beta regrese. Tyto dvě statistické metody nejsou běžně používány, proto jsem považoval za vhodné s nimi čtenáře seznámit, neboť se nedá předpokládat, že budou každému čtenáři známy.

V další kapitole jsme navrhli postup pro odhad Clear sky modelu. Jedná se o model, který z naměřených dat odstraňuje vliv dne a hodiny, ve kterých byla konkrétní data pořízena. Před odhadem, kdy je potřeba v datech ručně vyhledat jasné dny, byl upřednostněn odhad využívající kvantilovou regresi. Získaný Clear sky model je kvalitní, před jeho využitím v praxi bude potřeba drobné manuální úpravy v místech, ve kterých se model zřejmě odchýlil od skutečnosti. Toto vychýlení může být způsobeno buď chybou v datech, což by se dalo odstranit jejich opravou a přeodhadnutím modelu, nebo tím, že v letech 2012 - 2016 nenastal v okolí takového dne žádný jasný den. Clear sky model byl poté využit i při konstruování modelů.

V závěrečné kapitole jsme se pustili do tvorby modelu. Nejdříve byl vytvořen model predikující výrobu na den $D+1$, kdy se ukázalo, že nejvhodnějším modelem byl model se třemi vysvětlujícími proměnnými - dopadajícím slunečním zářením, nízkou oblačností a celkovou oblačností. Jako vhodnější se ukázalo upřednostnění beta regrese před lineární regresí s transformovanou vysvětlovanou proměnnou. Dále se ukázalo, že použití vysvětlované proměnné τ s následnou transformací výsledné predikce na $\widehat{p}$ bylo vhodnější než použít přímo proměnnou p jako vysvětlovanou. Dále jsme zjistili, že pokud rozdělíme data do množin podle hodnot Clear sky modelu a regresní parametry budeme odhadovat pouze na datech z těchto podmnožin, dostaneme se k lepšímu výsledku, než když necháme model odhadnout na celých datech.

Ve druhé části poslední kapitoly jsme zkonstruovali model predikující výrobu na aktuální den. Zde se ukázalo, že naměřená výroba a celková oblačnost, které jsou od predikované hodiny vzdálené méně než 5 hodin, přinesou zpřesnění predikce na tyto hodiny.

Výstupem této práce určitě není nejlepší model, který lze potenciálně sestrojit. Vzhledem k tomu, že určitě existuje spousta dalších možností, jak k tvorbě modelu pro predikci výroby elektřiny z FVE přistoupit, je potřeba se tomuto problému i dále věnovat a nepovažovat současný model za konečný. Razantní zlepšení modelu se určitě nedá očekávat, neboť chyby predikce výroby elektřiny v sobě zahrnují i chyby v nepřesnosti předpovědi počasí, které se do budoucna určitě nepovede úplně odstranit.

Seznam použité literatury

[1] Energetický regulační ústav: http://www.eru.cz/cs/poze.

[2] Koenker, R., Bassett, G. Jr (1978): *Regression Quantile*. Econometrica, **46**(1), 33-50.

[3] Møller, J. K. (2006-10): *Modeling of Uncertainty in Wind Energy Forecast*. IMM, Internet:
http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4428/pdf/imm4428.pdf

[4] Koenker, R., Bassett, G. Jr (1978): *Regression Quantile*. Econometrica, **46**(1), 33-50.

[5] Bacher, P. (2008): *Short-term Solar Power Forecasting*. IMM, Internet:
http://etd.dtu.dk/thesis/211035/ep08_13_net_new_version.pdf

[6] Ferrariová, S., Cribari-Neto, F. (2004): *Beta Regression for Modelling Rates and Proportions*. Journal of Applied Statistics, **31**(7), 799-815.

[7] Simas, A., Barreto-Sousa, W., Rochová, A. (2010): *Improved Estimators for a General Class of Beta Regression Models*. Computational Statistics & Data Analysis, **54**(2), 348-366.

[8] Smithson, M., Verkuilen, J. (2006): *A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables*. Psychological Methods, **11**(1), 54-71.

[9] Koenker, R. (2017): *Quantile Regression*. Internet:
https://cran.r-project.org/web/packages/quantreg/quantreg.pdf

[10] Zeilers, A. (2016): *Beta Regression*. Internet:
https://cran.r-project.org/web/packages/betareg/betareg.pdf

[11] Zambrano-Bigiarini, M. (2014): *Goodness-of-fit functions for comparison of simulated and observed hydrological time series*. Internet:
https://cran.r-project.org/web/packages/hydroGOF/hydroGOF.pdf

Seznam obrázků

Seznam tabulek

Přílohy

Součástí práce je i příloha, ve kterých se nachází 3 soubory obsahující vstupní data, 4 soubory obsahující veškeré zdrojové kódy z programu `R` a výstupy z modelů. Soubory obsahující vstupní data jsou

- `Data_CEZ.csv`, který obsahuje data o měření výkonu z fotovoltaických elektráren. Z důvodu ochrany dat se zde nachází pouze časové řady udávající procentuální množství vyrobené elektřiny z maximálního výkonu.
- `Data_CHMU_Aladin.csv`, který obsahuje predikce z modelu Aladin. Z důvodu ochrany dat soubor obsahuje pouze 5 vybraných zimních a letních dní, aby bylo možné si poskytnout představu o těchto datech.
- `Data_CHMU_mereni.csv`, který obsahuje naměřená data z vybraných stanic ČHMÚ.. Z důvodu ochrany dat soubor opět obsahuje pouze 5 vybraných zimních a letních dní, aby bylo možné si poskytnout představu o datech.

Dále jsou v příloze i tyto soubory se zdrojovými kódy z programu `R`

- `Data_CHMU.R`, ve kterém byly provedeny výpočty Pearsonových korelačních koeficientů obou dat z ČHMÚ.
- `Clearsky.R`, ve kterém byly provedeny veškeré výpočty týkající se Clearsky modelu v kapitole 4.
- `Nasledujici_den.R`, ve kterém byly provedeny veškeré odhady modelu na den $D + 1$ v kapitole 5.3.1.
- `Aktualni_den.R`, ve kterém byly provedeny veškeré odhady modelu na den D v kapitole 5.4.

Součástí přiložených souborů jsou i soubory s výsledky

- soubor `Kriteria_6_0.9865.csv`, ve kterém jsou hodnoty kritérií pro výsledný nevyhlazený Clear sky model.
- soubory `CS_SC.csv`, `CS_ST.csv`, `CS_SM.csv`, `CS_ZC.R`, `CS_VC.csv` a `CS_OS.csv`, ve kterých je v matici výsledný nevyhlazený Clear sky model z příslušného regionu.
- soubor `CS_all.csv`, ve kterém jsou výsledky všech nevyhlazených Clear sky modelů zapsaných vektorově.
- soubor `Kriteria_sm_6_0.9865.csv`, ve kterém jsou hodnoty kritérií pro výsledný vyhlazený Clear sky model.
- soubory `CS_sm_SC.csv`, `CS_sm_ST.csv`, `CS_sm_SM.csv`, `CS_sm_ZC.csv`, `CS_sm_VC.csv` a `CS_sm_OS.csv`, ve kterých je v matici výsledný vyhlazený Clear sky model z příslušného regionu.
- soubor `CS_sm_all.csv`, ve kterém jsou výsledky všech vyhlazených Clear sky modelů zapsaných vektorově.
- soubory `predikce_Ma.csv`, `predikce_MA.csv` a `predikce_MC.csv`, které obsahují predikce pro všechny časy z příslušných modelů.