

NEJPOROVNÁNÍ MODELOVÁNÍ VYBRANÝCH ČASOVÝCH ŘAD V SYSTÉMECH SAS A STATISTICA

COMPARISON OF SELECTED TIME SERIES MODELLING IN SAS AND STATISTICA SYSTEMS

Zdeněk Louda

Anotace:

Systémy SAS a STATISTICA jsou jedny z nejrozšířenějších systémů pro statistickou analýzu dat na světě, oblast analýzy časových řad je jedním z nejužívanějších statistických nástrojů. Příspěvek by měl zmapovat a porovnat možnosti pro práci s časovými řadami v systémech SAS a STATISTICA a dále by měl seznámit s výsledky analýz, modelování a prognózování vybraných časových řad na agrárním trhu za použití obou systémů.

Klíčová slova:

Časová řada, analýza, model, prognóza, systémy SAS a STATISTICA.

Summary:

Systems SAS and STATISTICA are one of the most popular systems in the world for statistical data analysis, time series analysis is the most widely used among the statistical tools. Article should map and compare possibilities in the area of time series in SAS and STATISTICA systems and next should acquaint with results of analysis, modelling and forecasting of selected time series on agrarian market in both systems.

Key words:

Time series, analysis, model, forecast, SAS and STATISTICA systems.

ÚVOD A CÍL

Snaha zkoumat minulý vývoj nejrůznorodějších společenských jevů, poučit se z historie a z daných zkušeností a poznatků těžit v budoucnosti, jsou známy již od pradávna. V současné společnosti je tento letitý princip často reprezentován analýzou časových řad a jejich prognózováním do budoucna a tato oblast statistických analýz je jedním z nejužívanějších statistických nástrojů v nejrůznějších sférách lidského života. Během několika posledních desetiletí se chopily statistických analýz softwarové firmy a neustálým iterativním procesem zlepšování a inovací vytvořily produkty, z nichž jedněmi z nerozšířenějších jsou dnes systémy SAS a STATISTICA.

Cílem příspěvku je zmapovat a porovnat možnosti systémů SAS a STATISTICA v oblasti analýz a prognózování časových řad, neboť analýza časových řad bude jednou ze stěžejních oblastí statistických analýz v disertační práci s názvem „Návrh a ověření metod statistické analýzy pro podnikatelské a marketingové rozhodování na trhu s agrárními produkty“. Moduly pro práci s časovými řadami v obou systémech budou nejprve obecně charakterizovány a popsány a posléze v následující praktické části budou shrnuty výsledky analýz, modelování a prognózování vybraných časových řad na agrárním trhu za použití obou systémů.

Analyzovaná data vybrané časové řady pocházejí ze Situační a výhledové zprávy - Zelenina vydané MZe ČR v prosinci roku 2002. Bylo vybráno 7 časových řad spotřebitelských cen

(SC) na trhu se zeleninou (v Kč/kg), a to konkrétně spotřebitelské ceny v období leden 1996 - říjen 2002 bílého hlávkového zelí, cibule, květáku, mrkve, salátových okurek, papriky a rajčat. Příspěvek předpokládá teoretické znalosti pokročilých modelů časových řad.

METODY

Prostředí systému SAS pro analýzu časových řad

Velkou uživatelskou výhodou statistického prostředí modulu ETS programového balíku SAS je, že uživatel prakticky stačí pouze zadat vstupní data řádně připravená pro počítačové procedury. Uživatel musí ještě subjektivně rozdělit řadu na testovací a kontrolní část, ale jinak veškerá výpočtová složitost statistických postupů odpadá a uživatel se musí probrat jen „kupou výstupů“, kterou si však může podle svého uvážení „korigovat“. Modul ETS nabízí také možnost diagnostikovat časovou řadu z hlediska přítomnosti trendu, sezónnosti a nutnosti logaritmické transformace dat a poté může vybírat jen z těch modelů nabízených programem, které vyhovují zjištěné diagnóze. Ovšem jsou již zaznamenány i pochybnosti o naprosté spolehlivosti tohoto diagnostování (modely vytvořené na základě takového doporučení nemusí vždy vést k nalezení nejlepšího modelu pro daný případ).

Možnostmi hodnocení kvality vytvořených modelů jsou míry hodnocení kvality prognóz „ex post“, graf reziduí, grafy ACF a PACF a otestování významnosti jednotlivých parametrů modelu. Další možností při modelování a prognózování je možnost kombinovat vypočtené modely, ovšem kombinace více jak dvou modelů se nedoporučuje, neboť s rostoucím počtem parametrů těchto modelů klesá interpretovatelnost výsledného kombinovaného modelu. Dále pro prognostické využití vytvořených modelů nabízí tento modul graf vytvořeného modelu včetně prognóz na určitý počet období do budoucnosti (uživatelem předem stanovených) a tomu odpovídající tabulku prognóz, obojí včetně pásu spolehlivosti prognóz.

Prostředí systému STATISTICA pro analýzu časových řad

Analýza časových řad v systému STATISTICA je obsažena v nástrojích pro predikci a analýzu časových řad ve skupině pokročilých lineárních a nelineárních modelů. V dané nabídce jsou nejpoužívanějšími procedury tvorby ARIMA modelů a exponenciálního vyrovnávání. Uživatel si však z nabídky pro analýzu časových řad musí již na začátku vybrat, kterou oblast resp. metodu bude používat, není zde tedy přímá možnost konfrontovat modely vytvořené různými metodikami analýzy časových řad (tvorba trendů časových řad pomocí regresních funkcí v základní nabídce nástrojů pro analýzu časových řad obsažena není, lze ji však provést pomocí nástrojů regresní analýzy či pomocí grafické analýzy dané řady).

V oblasti exponenciálního vyrovnávání je nabízena možnost automatického hledání příslušných parametrů alfa, delta a gama, kdy je automaticky vybrán nejlepší model, nebo možnost síťového hledání parametrů, kdy si uživatel vybírá z několika nejlepších modelů (resp. parametrů) zjištěných danou procedurou a ohodnocenou známými charakteristikami pro hodnocení modelů. Obecně však nelze určit, která z obou možností získání parametrů exponenciálního vyrovnávání dává lepší odhady parametrů a tudíž i lepší model, a je na uživateli, který model si pro danou řadu vybere. Modul exponenciálního vyrovnávání nabízí několik obecných typů modelů, a to dle hlediska trendu modely bez trendu, s trendem lineárním, exponenciálním a tlumeným a z hlediska sezónnosti modely bez sezónnosti, s aditivní či multiplikativní sezónností – kombinací zmíněného se tedy jedná o obecných 12 typů modelů exponenciálního vyrovnávání časové řady.

V oblasti tvorby ARIMA modelů se uživatel musí „proklikat“ k nejlepšímu modelu pomocí analýzy ACF, PACF, testování statistické významnosti parametrů vypočteného modelu a případně analýzy reziduí, tudíž musí mít v oblasti tvorby ARIMA modelů patřičnou teoretickou průpravu.

Pozn. 1: Vzhledem k časové náročnosti hledání nejvhodnějšího ARIMA modelu pro zvolenou časovou řadu v systému STATISTICA bylo od této metodiky upuštěno a nejvhodnější modely byly vytvářeny pouze v modulu exponenciálního vyrovnávání.

Pozn. 2: Kvalita modelu by měla být komplexně posuzována jak všemi známými charakteristikami kvality modelu, tak i analýzou ACF, PACF a grafu reziduí, ale vzhledem k časové náročnosti takového komplexního přístupu a rozsáhlosti výstupů analýz k následnému zpracování a vyhodnocení byl postup hodnocení kvality modelu v obou porovnávaných systémech zredukován pouze na posouzení charakteristiky MAPE.

VÝSLEDKY

Vyhodnocení konstrukce modelů vybraných časových řad v modulu SAS/ETS

Pro tvorbu modelů časových řad v systému SAS bylo vybráno již výše zmíněných 7 časových řad spotřebitelských cen komodit agrárního trhu se zeleninou. Pro každou sledovanou časovou řadu byla provedena v modulu ETS diagnostika řady a byly vypočteny všechny modely, které odpovídají diagnostickým závěrům (v průměru bylo spočteno 7,71 modelů pro každou řadu). Ve sledovaném souboru časových řad byla v 6 případech diagnostikována možnost transformace původních dat, pouze v jediném případě byl zjištěn trend a v 6 případech byla indikována sezónnost časové řady (viz Tabulka č. 1). Dále byly pro každý model zkoumané časové řady spočteny základní charakteristiky kvality modelu (MAPE, MAE, R^2 , MSE a RMSE), z nichž však stěžejní pro výběr nejvhodnějšího modelu byla nepoužívanější - hodnota MAPE. Za nejvhodnější byl ve většině případů (u 6 komodit) vybrán některý z modelů exponenciálního vyrovnávání (nejčastěji modifikace sezónního exponenciálního vyrovnávání - viz Tabulka č. 1).

komodita	diagnostika časové řady			počet vyhov. modelů	nejvhodnější model	MAPE	vhodnost modelu vzhledem k ACF a PACF
	transf.	trend	sez.				
bíle hl. zelí	možná	ne	ano	8	Log Seasonal Exponential Smoothing	6,91884	ne
cibule	možná	ano	ne	10	Damped Trend Exponential Smoothing	7,67023	ano
květák	možná	ne	ano	8	Log Seasonal Exponential Smoothing	7,43272	ne
mrkev	možná	ne	ano	8	Log Seasonal Exponential Smoothing	6,18603	ano
okurky salátové	možná	ne	ano	8	Log Seasonal Exponential Smoothing	12,44184	ano
papriky	ne	ne	ano	4	Seasonal Exponential Smoothing	7,13163	ano
rajčata	možná	ne	ano	8	Log Seasonal Dummy	10,32547	ne

Tabulka č. 1: Vyhodnocení modelů vybraných časových řad SC na trhu se zeleninou

4 výsledné modely z celkových 7 vybraných časových řad byly shledány posouzením grafů funkcí ACF a PACF jako vhodné (viz poslední sloupec v Tabulce č. 1). V konfrontaci se standardně používaným kritériem kvality modelu a jeho vhodnosti pro případné prognózování budoucího vývoje sledované proměnné (hodnota MAPE menší nebo rovna 5% resp. 10%) nevyhovovaly pouze 2 výsledné modely (viz předposlední sloupec Tabulky č. 1). Výsledkem kombinace obou zmíněných kritérií jsou pouze 3 vhodné výsledné modely pro popis vývoje sledované proměnné a použitelné pro případnou predikci řady do budoucna (viz tučně vytisknuté komodity ve stejnojmenném sloupci Tabulky č. 1).

Vyhodnocení konstrukce modelů exponenciálního vyrovnávání vybraných časových řad v modulu exponenciálního vyrovnávání systému STATISTICA

Pro každou sledovanou časovou řadu byl v modulu exponenciálního vyrovnávání časových řad zjištěn pro každý ze zmiňovaných 12 typů modelů nejlepší model procedurou automatického odhadu parametrů a procedurou síťového hledání parametrů včetně charakteristik hodnocení kvality modelu (z nichž byla upřednostněna charakteristika MAPE).

	Automatický odhad parametrů					
	model	alfa	delta	gama	fi	MAPE
zelí bílé hl.	TTA	1,000	0,000	-	0,290	6,1411
cibule	BTM	1,000	0,000	-	-	7,4845
květák	TTM	0,000	0,000	-	0,998	7,3184
mrkev	TTM	1,000	0,000	-	0,205	5,9685
okurky sal.	TTM	0,190	0,000	-	0,280	12,3420
paprika	LTA	0,974	0,000	0,000	-	7,2899
rajčata	BTM	0,000	0,000	-	-	9,8720

Tabulka č. 2: Vlastnosti modelů zjištěných procedurou automatického odhadu parametrů

Procedurou automatického odhadu parametrů byl ve 4 případech ze 7 vybrán jako nejvhodnější model tlumeného trendu, z nichž 3 s multiplikovanou sezónností (TTM) a jen 1 s aditivní sezónností (TTA), ve 2 případech byl jako nejvhodnější posouzen model bez trendu s multiplikovanou sezónností (BTM) a v 1 případě model lineárního trendu s aditivní sezónností (LTA - viz Tabulka č. 2). V konfrontaci s použitým kritériem kvality modelu a jeho vhodnosti pro případné prognózování budoucího vývoje sledované proměnné (hodnota MAPE menší nebo rovna 5% resp. 10%) vyhovovalo 6 ze 7 výsledných modelů (viz poslední sloupec Tabulky č. 2).

	Síťové hledání parametrů					
	model	alfa	delta	gama	fi	MAPE
zelí bílé hl.	TTM	0,900	0,100	-	0,300	6,1240
cibule	TTM	0,600	0,200	-	0,300	7,5273
květák	TTM	0,200	0,100	-	0,200	7,9335
mrkev	TTM	0,900	0,900	-	0,300	5,9067
okurky sal.	TTM	0,200	0,100	-	0,200	12,684
paprika	TTA	0,800	0,800	-	0,100	7,1748
rajčata	BTM	0,400	0,100	-	-	11,0810

Tabulka č. 3: Vlastnosti modelů zjištěných procedurou síťového hledání parametrů

Procedurou síťového hledání parametrů byl v 6 případech ze 7 vybrán jako nejvhodnější model tlumeného trendu (a pouze 1 z nich s aditivní sezónností, ostatních 5 s multiplikovanou sezónností) a pouze v jediném případě byl jako nejvhodnější posouzen model bez trendu s multiplikovanou sezónností (viz Tabulka č. 3). V konfrontaci s kritériem kvality modelu MAPE vyhovovalo pouze 5 ze 7 výsledných modelů (viz poslední sloupec Tabulky č. 3).

Z porovnání obou procedur hledání parametrů nejvhodnějších modelů exponenciálního vyrovnávání časových řad nepatrně lépe vynívá síťové hledání parametrů (pro 4 ze 7 sledovaných proměnných byl nejlepší model zjištěn právě touto procedurou – viz tučně vtištěné hodnoty MAPE a názvy proměnných v Tabulce č. 2 resp. č. 3), ale toto srovnání jen dokazuje, že lze jen těžko objektivně určit, která procedura hledání parametrů modelu dává obecně lepší výsledky.

DISKUSE

Kvalita modelů, časových řad velkou měrou závisí na vstupních datech. Velkým problémem, který znepokojuje zejména v přípravné fázi zpracovávání disertační práce, je datová základna. Data jsou často neúplná, obtížně „dohledatelná“, mnohdy nedostupná či téměř nedobytná, v oblasti analýzy časových řad se navíc negativně projevuje „mládí“ České republiky, kdy máme k dispozici jen velmi krátké časové řady, které nejsou zkresleny „centrálně plánovaným hospodářstvím“ před rokem 1989. Je však nutno brát tuto situaci jako objektivní fakt a tak k ní také přistupovat.

Určit výhody a nevýhody obou systémů není zcela jednoduché, často se jedná o subjektivní pohled uživatele (co je pro jednoho výhodou může být pro jiného nevýhodou). V dalším textu jsou shrnuty zjištěné poznatky (i když samozřejmě subjektivní) z práce v obou systémech.

Jako pozitivní (v modulu ETS systému **SAS**) se jeví automatická diagnostika časové řady (i když se již vyskytly pochybnosti o její spolehlivosti), fakt, že všechny metodiky tvorby modelů jsou spojené v jednom modulu a je tudíž umožněna tvorba modelů pomocí různých metodik i jejich vzájemné porovnávání (přitom však není uživateli znemožněna tvorba individuálního modelu podle jeho představ), jako negativum pak bylo shledáno značně neflexibilní, neintuitivní a „nepřátelské“ uživatelské prostředí systému lokalizovaného do anglického jazyka a nepříliš kvalitní grafické výstupy a často obtížná manipulace s nimi.

Jako výhody (v systému **STATISTICA** při analýze časových řad) byly shledány příjemné intuitivní uživatelské prostředí lokalizované do českého jazyka, grafické nástroje a výstupy (včetně jednoduché „manipulace“ s nimi) na velmi vysoké úrovni, značná automaticnost práce v modulu exponenciálního vyrovnávání časových řad, ovšem jako negativní byly shledány oddělenost metodik tvorby modelů časových řad a tudíž nemožnost přímého porovnání modelů vytvořených různými metodikami, absence regresního vyrovnávání v nabídce analýz a prognózování časových řad (tvorba trendových funkcí), absence hodnocení kvality modelů známými charakteristikami v oblasti tvorby ARIMA modelů naopak absence hodnocení kvality modelů grafy ACF a PACF v oblasti exponenciálního vyrovnávání a nepříliš zdařilý překlad některých částí výstupů a procedur.

ZÁVĚR

Značná nestabilita faktorů a vlivů na agrárních trzích velmi často způsobuje nepředvídatelné výkyvy ve vývoji sledovaných veličin. To se pak projevuje zejména v nemožnosti toto chování popsat vhodným modelem a tudíž v obtížné predikovatelnosti budoucího vývoje sledované proměnné, což bylo ověřeno i provedeným výzkumem, kdy pouze u necelé poloviny ze zkoumaných veličin byl nalezen vyhovující model jejich vývoje a případné předpovědi jejich chování v budoucnu v systému **SAS** (v systému **STATISTICA** byl uspokojivý model časové řady nalezen u 5 ze 7 proměnných, ovšem byla použita pouze metodika exponenciálního vyrovnávání a pouze míra MAPE jako kritérium kvality modelu).

Oba porovnávané systémy jsou špičkovými produkty nejen pro statistickou analýzu časových řad, každý danou oblast řeší specificky, a těžko určit, který produkt je objektivně lepší, záleží na uživateli, které prostředí a možnosti statistických analýz mu budou vyhovovat. Díky zmapování práce v modulech pro analýzu a prognózování časových řad obou systémech a výsledkům vlastního výzkumu daný příspěvek do značné míry nastínil práci s časovými řadami v systémech **SAS** a **STATISTICA** i se všemi odlišnostmi a záludnostmi.

Literatura:

1. Arlt, J: Moderní metody modelování ekonomických časových řad, Praha, Grada Publishing, 1999
2. Arlt, J; Arltová, M; Rublíková, E: Analýza ekonomických časových řad s příklady, VŠE, Praha, 2002
3. Brabenec, V; Šařecová, P; Hošková, P; Procházková, R; Louda, Z: Statistika a biometrika pro AF a ITZ, Praha, ČZU, 2004
4. Hindls, R; Hronová, S; Seger, J: Statistika pro ekonomy, Praha, Professional Publishing, 2003
5. Louda, Z: Řešené příklady v systému **STATISTICA**, Praha, ČZU, 2004
6. Prášilová, M: Předdiplomní statistický seminář – vybrané texty, Praha, ČZU, 2002
7. Svatošová, L; Hříbal, J; Volma, M: Systém **SAS** – příručka pro uživatele, ČZU, Praha, 2000
8. MZe ČR: Situační a výhledová zpráva - Zelenina, prosinec 2002

Kontaktní adresa autora:

Ing. Zdeněk Louda

Katedra statistiky, Provozně ekonomická fakulta, Česká zemědělská univerzita v Praze, Kamýcká 129, Praha 6 – Suchbátka, 16521, telefon: 2 2438 3246, email: louda@pef.czu.cz