METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XVI/4, 2015, str. 280 – 289

MODELE PROGNOZ EKONOMETRYCZNYCH

Krzysztof Zmarzłowski, Marek Karwański
Katedra Informatyki
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: Krzysztof_Zmarzlowski@sggw.pl, Marek_Karwanski@sggw.pl

Streszczenie: Krótko i średnioterminowe prognozy często oparte są na różnych modelach ekonometrycznych. Dla modeli stosowanych do pojedynczych spółek, mamy do dyspozycji szereg miar, pozwalających porównywać je od strony dokładności uzyskiwanych rezultatów. Sytuacja komplikuje się, gdy prognozy dotyczą grupy spółek bądź sektorów gospodarczych. W pracy autorzy proponują nowoczesne narzędzie graficzne oparte na krzywej REC (Regression Error Characteristic). Detaliczne wyniki stosowania tej metody oceny modeli zostaną zaprezentowane w zastosowaniu do polskich firm z sektora budowlanego, notowanych na giełdzie.

Słowa kluczowe: krzywa REC, modele prognostyczne, sektor giełdowy

WSTĘP

Wybór najlepszego z modeli powiązany jest ze spełnieniem szeregu założeń statystycznych. Z punktu widzenia analitycznego, ważniejsza jest jednak szybsza ocena danego modelu i możliwości jego wykorzystania w dalszych etapach analizy. Istotną rzeczą jest to, aby najlepszy model spełniał zarówno najistotniejsze założenia statystyczne, jak również, aby z mnogości wskaźników go opisujących mieć możliwość wyboru jednego, który w jasny i prosty sposób mówi nam o dopasowaniu danego modelu do danych. W niniejszej pracy autorzy skupili się głównie na drugiej części powyższego problemu. W tym celu postanowiono skorzystać z jednej z nowszych metod oceny dopasowania modelu do danych - krzywej REC (Regression Error Characteristic).

Głównym materiałem badawczym były dane pochodzące z dwóch niezależnych źródeł. Pierwszym zestawem zbioru badawczego były dane dotyczące wskaźników ekonomicznych spółek notowanych na Polskiej Giełdzie Papierów Wartościowych (GPW) udostępnione przez serwis internetowy Notoria.

Obejmowały one 271 firm z różnych sektorów giełdowych. Ze względu na istotę porównania modeli prognostycznych dla całego sektora, wybrano do badania sektor spółek budowlanych, obejmujący 16 firm. Dane dotyczyły okresu od X 1998 do XII 2005 i miały charakter kwartalny. Zakres danych obejmował ok. 900 różnych wskaźników ekonomicznych, z czego duża większość była składowymi pozostałych. Drugim źródłem danych były ceny zamknięcia akcji pobrane z serwisu Reuters. W tym przypadku były to notowania dzienne (obejmujące wyłącznie dni robocze).

METODOLOGIA BADAŃ

Analiza danych i wybór odpowiedniego modelu przebiegały w następującym porządku:

1. Przygotowanie i oczyszczenie danych.
2. Selekcja zmiennych (wskaźników kapitałowych) metodą ekspercką.
3. Budowa modeli ekonometrycznych.
4. Porównanie i ocena modeli.

Przygotowanie i oczyszczenie danych

Do analizy wykorzystano dwa źródła danych. Dotyczyły one wielu aspektów rynku i zbierane były według różnych metodologii, dlatego poddane zostały różnym procedurom przygotowania wstępnego. Pierwszym źródłem były szeregi czasowe notowań (dziennych) spółek, uzyskane z systemu Reuters. Do analizy wzięto tzw. „notowania czyste", czyli po usunięciu wpływu: splitów, dywidend, praw poboru, nabycia, objęcia oraz denominacji. W kilku przypadkach występowały braki danych, które zostały poddane procedurze uzupełniania. Do uzupełniania braków danych użyty został model średniej ruchomej o szerokości 4-obserwacji. Model został wybrany ze względu na to, że w badaniu skupiono się bardziej na relacji średnich w długich horyzontach czasowych i dlatego spłaszczenie zmienności nie było aż tak kluczowym problemem. Drugim źródłem danych były informacje miesięczne o spółkach z serwisu Notorii, dotyczące wskaźników finansowych spółek notowanych na warszawskiej giełdzie.

Uzupełnianie braków danych

Analiza kompletności danych pochodzących z Notorii wykazała występowanie braków danych w około 7% przypadków. Zdecydowano się na uzupełnienie ich przy użyciu podejścia tzw. „model based" [Raghunathan T. i in. 2001]. Bazował on na założeniu, że łączny rozkład prawdopodobieństwa można przedstawić w postaci czynnikowej:

$$f(Y_1, Y_2, \dots, Y_k | X, \theta_1, \dots \theta_k) = f_1(Y_1 | X, \theta_1) \dots f_k(Y_k | X, Y_1, \dots Y_{k-1}, \theta_k) \quad (1)$$

gdzie f_j $j=1,2,...,k$ są ciągłymi rozkładami prawdopodobieństw warunkowych, a θ_j parametrami tych rozkładów. Prawdopodobieństwa warunkowe przybliżano za pomocą równań regresyjnych z nieznanymi parametrami $\theta_{,j}$ dodatkowo wykorzystano założenie, że prawdopodobieństwa a'priori $\pi(\theta) \propto 1$. Modele regresyjne należały do jednej z poniższych klas:

1. Normalnej regresji liniowej (po odpowiedniej transformacji np.: Boxa-Coxa) dla skal ilorazowych,
2. Regresji logistycznej dla skali binarnej,
3. Uogólnionego modelu logitowego dla skali porządkowej,
4. Dwustopniowego modelu: regresja logistyczna(I-poziom), regresja liniowa (II-poziom).

Każdy etap uzupełniania braków danych składał się z c-faz. W 1-fazie wybierana była zmienna Y_1 z najmniejszą liczbą braków, uzupełnienie braku następowało na podstawie odpowiedniego równania regresyjnego, następnie korzystając z faktu, że rozkład parametrów θ jest płaski, odpowiednia wartość losowana była z odpowiedniego predykcyjnego rozkładu a'posteriori. Następnie wybierana była kolejna kolumna Y_2 itd.

Usunięcie składowej sezonowej

Wskaźniki zbierane były w cyklu miesięcznym i z tego powodu charakteryzowały się czynnikiem sezonowym. Dane historyczne tzw. „okno obserwacji" mogło zawierać istotną składową sezonową. Z tego względu obserwacje zostały poddane procedurze tzw. od-sezonowania. Do od-sezonowania wykorzystana została procedura X-12-ARIMA [US Census Bureau's 2013]. Ogólna postać modelu zastosowanego do od-sezonowania szeregów czasowych:

$$Y_t = Tr_t \cdot Sn_t \cdot Cl_t \cdot \varepsilon_t \tag{2}$$

gdzie t- czas, Y_t- zmienna prognozowana opisująca zjawisko w chwili czasu t, Tr_t- tendencja rozwojowa (trend) lub funkcja, Sn_t- wahania sezonowe, Cl_t- wahania cykliczne, ε_t- zmienna losowa w chwili czasu t (w analizie klasycznych szeregów nie wnika się w strukturę składowej losowej).

Proces dekompozycji składa się z szeregu etapów. Identyfikacja składowych nie jest zazwyczaj jednoznaczna, stąd poszczególne etapy stanowią kolejne przybliżenia i pozwalają ocenić poprawność całego procesu.

Dane miesięczne zostały przeliczone na dane dzienne z użyciem metody trendu liniowego. Pozwoliło to na połączenie dziennych notowań z miesięcznymi wskaźnikami.

Selekcja zmiennych

Z uwagi na dużą liczbę czynników, konieczne było dokonanie selekcji zmiennych opisujących spółki [Cheng i in. 2010]. Wstępna selekcja oparta została na opiniach eksperckich. Wstępna liczba potencjalnych zmiennych wynosiła ponad 900 wskaźników ekonomicznych.

Budowa modeli ekonometrycznych

W praktyce, analizując dane rynkowe używa się dwóch typów modeli stochastycznych: tzw. „opartych na cenach" oraz „opartych na innowacjach" [Zumbach 2007]. W modelach „opartych na cenach" - liczy się rozkłady zwrotów. Są to np. modele analityczne, w których zakłada się, że rozkład zwrotów (r) jest rozkładem normalnym lub rozkładem Gamma. Do tej klasy zalicza się również symulacje historyczne, w których modeluje się zwroty za pomocą rozkładów historycznych. Niestety ich wadą jest stosukowo krótki horyzont czasowy.

W modelach drugiego typu podstawowym równaniem jest powiązanie zwrotów (r) z innowacjami (ϵ)

$$r(t+\delta t) = \mu(t) + \sigma(t)\epsilon(t+\delta t) \tag{3}$$

gdzie: μ(t) jest średnią zwrotów, σ(t) jest tzw. zmiennością, ϵ(t) to innowacja (losowość).

Modele oparte na innowacjach wymagają obliczenia zmienność σ. Z obserwacji rynku wynika, że modelowanie obejmuje oba wymiary jednocześnie tzn.: zarówno zwroty jak i zmienność. Warto zwrócić dodatkową uwagę na problemy rodzące się przy modelowaniu średnich i długo-terminowych horyzontów czasowych, powstające w wyniku braku dostatecznej ilości danych, pozwalających policzyć estymatory związane ze zwrotami (agregacja). Wprowadzenie procesu zmienności pozwala ominąć te trudności, jednak praktyka pokazuje, że dla długich horyzontów czasowych agregacja kolejnych zwrotów wymaga bardzo długich szeregów, umożliwiających rozsądne estymowanie modeli.

Do analizy wskaźników kapitałowych zastosowano modele regresyjne z czynnikiem losowym sterowanym innowacjami. W ramach niniejszej pracy przebadano kilka szeroko stosowanych w badaniach rynku modeli typu ARCH/GARCH [Andersen i in. 2009]. Ostatecznie wybrany został model niesymetryczny autoregresji e-GARCH. W przeciwieństwie do modeli GARCH(p,q) modele e-GARCH nie wymagają stawiania ograniczeń na parametry. Procesy e-GARCH zawsze są dodatnie. W ramach modeli e-GARCH badane były istotności poszczególnych wskaźników kapitałowych. Taka strategia nie uwzględniała wpływu kilku czynników jednocześnie, dlatego jako wartość progową przyjęto istotność = 0.50. Takie podejście umożliwiło również wstępne usunięcia czynników całkowicie niepowiązanych ze zmianami rynku. Budując modele rynku, trzeba zazwyczaj uwzględnić wiele czynników oraz relacje i zależności między nimi.

Do modelowania zależności używa się tzw. „funkcji kopuły" (ang. Copula), które pozwalają łączyć jednowymiarowe rozkłady do postaci rozkładów wielowymiarowych [McNeil i in. 2005]. Najlepiej podejść do kopuł poprzez opis przestrzeni obserwacji przy użyciu czynników. Np. obserwowane zwroty można rozpisać jako sumę czynników „ogólnych" odpowiedzialnych za korelacje i czynników specyficznych, nieskorelowanych. W przypadku rozkładów Normalnych procedura ta jest dobrze znana jako analiza czynnikowa.

W reprezentacji czynnikowej mamy:

$$Y_i = c_{i1}V_1 + \cdots + c_{im}V_m + \sqrt{1 - \left(c_{i1}^2 + \cdots + c_{im}^2\right)}U_i. \quad (4)$$

gdzie V_i to czynniki ogólne, a U_i specyficzne

W modelach rynku większą rolę odgrywają tzw. kopuły Studenta. Kopuły Studenta można rozważać jako rozszerzenie kopuł Normalnych. Dla kopuł Normalnych mamy $Y_i \sim N(0, \Sigma)$, gdzie Σ oznacza stałą macierz korelacji, w przypadku kopuł Studenta elementy macierzy korelacji są zmiennymi losowymi W, które mają rozkład Chi-kwadrat z ν stopniami swobody. Czasami korzysta się z kopuł Archimedesa. Nie mają one w swoich parametrach explicite macierzy korelacji, struktura korelacyjna jest ukryta implicite w rozkładach czynników, które można „wyciągnąć" przy pomocy transformacji Laplace'a. Inny, bardziej ogólny, formalizm pozwala wprowadzić kopuły w przypadku dowolnych rozkładów jako procedurę separowania rozkładów brzegowych od korelacji opierając się na twierdzeniu Skalara [Nelsen 1998].

Drugim typem modeli wykorzystanych w badaniu były modele kointegracyjne Analiza danych rynkowych, w szczególności dla dłuższych horyzontów czasowych (powyżej 1 roku), prowadzona jest z wykorzystaniem modeli regresyjnych – kointegracyjnych (VARMAX) [Luetkepohl 2005]. Istota modelu polega na tym, że wektor kointegracyjny β jest odpowiedzialny za niestacjonarność. W przypadku modelowania danych ekonomicznych równanie kointegracyjne można utożsamić z „prawami ekonomicznymi".

Miara dopasowania modeli REC

Krzywa REC proponowana przez Bi i Bennet [Bi i in. 2003] stanowi narzędzie graficzne, które jest używane w celu porównania różnych regresyjnych modeli prognostycznych. Porównanie modeli regresyjnych opiera się na analizie reszt. Krzywa REC pozwala na graficzną prezentację błędu. Błąd jest tu zdefiniowany jako różnica pomiędzy przewidywaną warunkową wartością modelu y(x) i wartością rzeczywistej odpowiedzi y dla danego punktu (x). Różnica może być zdefiniowana zarówno w metryce kwadratowej, jak i innej.

Punktem wyjścia jest poszukiwanie modelu, który byłby „wystarczająco dobry" dla odróżnienia od "najlepszego" modelu, który z kolei należy odróżnić od "prawidłowego" modelu [Chatfield 1995]. Prawidłowy model w populacji może być nieliniowy, a proces modelowania na podstawie danych historycznych może zakładać tylko funkcje liniowe. Wystarczająco dobry model liniowy może być odpowiednim przybliżeniem właściwego modelu, mimo że prawidłowy model dla całej populacji jest bardziej skomplikowany. Z tego punktu widzenia preferowany będzie model regresyjny, dla którego funkcja błędu leży poniżej zadanego progu tolerancji.

Krzywa REC opiera się na uwzględnieniu funkcji straty dla błędów przekraczających poziom tolerancji. Zdefiniujmy funkcję precyzji acc(ε) dla poziomu tolerancji ε jako:

$$acc(\varepsilon) = \frac{\#\{(x,y):loss(f(x),y)\leq\varepsilon\}}{n} \quad (5)$$

gdzie: loss( ) jest funkcją straty (x,y) to realizacje zmiennej losowej y dla ustalonych wartości czynników x.

Krzywa REC to wykres funkcji precyzji acc(). Krzywa REC umożliwia porównanie funkcji regresji. Obszar pod krzywą może być traktowany jako miara oczekiwanej wydajności modelu regresji. Obszar nad krzywą (AOC) jest obciążonym oszacowaniem średniego oczekiwanego błędu.

WYNIKI BADAŃ

Pierwszym zestawem zbioru badawczego były dane udostępnione przez serwis internetowy Notoria. Aby dane nadawały się do analiz oraz modelowania konieczne było ich ujednolicenie. W tym celu skorzystano z SASowych narzędzi ETL (Extract, Transform and Load), umożliwiających wczytanie ponad 271 plików (często o różnej strukturze) oraz przeprowadzenie procesu denormalizacji danych zawartych w sprawozdaniach finansowych. Po zaczytaniu i ujednoliceniu danych źródłowych, wybrano ze wszystkich 271 spółek firmy sektora budowlanego. Poniżej została zaprezentowana lista 16 spółek, wziętych do badania:

- ATLANTIS, ELBUDOWA, ENERGOAP, ENERGOPL, ENMONTPD, INSTAL_K, KOPEX, MOST_EXP, MOST_PK, MOST_W-WA, MOST_ZAB, PEMUG, POLNORD, STALEXP, BUDIMEX, BUDOPOL.

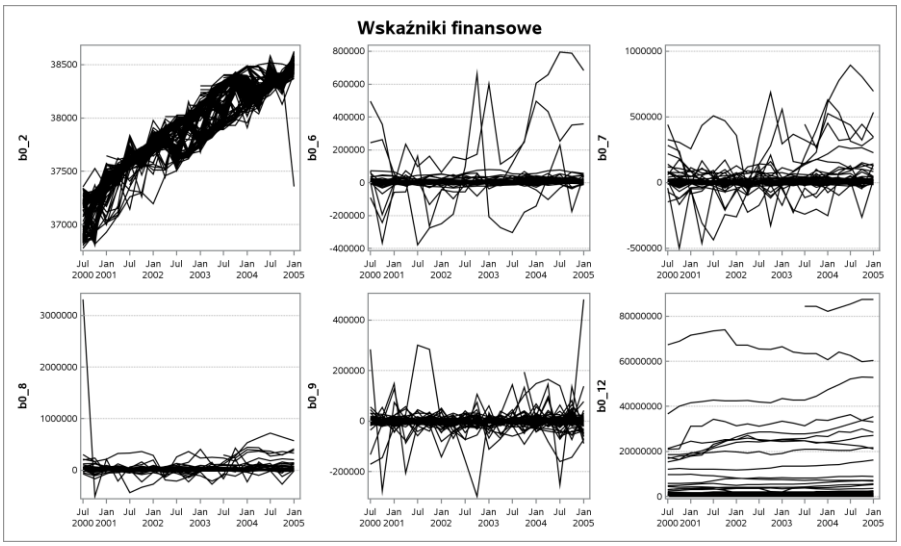
Tabela 1. Wskaźniki ekonomiczne wybrane do badania

Lp.	Nazwa zmiennej	Lp.	Nazwa zmiennej
1	Wynik z inwestycji netto	14	Koszty ogólnego zarządu
2	Zysk (strata) brutto	15	Marża zysku brutto
3	Zobowiązania i rezerwy	16	Wsk. płynności bieżącej
4	Kapitał własny (aktywa netto)	17	Wsk. płynności szybkiej
5	Liczba akcji	18	Rotacja należności
6	Wynik z dział. operacyjnej	19	Rotacja zapasów
7	Wynik z dział. inwestycyjnej	20	Wsk. pokrycia majątku
8	Wynik z dział. finansowej	21	Stopa zadłużenia
9	Kapitał własny	22	Wsk. obsługi zadłużenia
10	Zobow. - kredyty, pożyczki	23	Dług/EBITDA
11	Zobow. - emisja p. dłużnych	24	Pożyczki/Aktywa ogółem
12	Zobow. - podatki, cła, ubezp.	25	Depozyty/Aktywa ogółem
13	Koszty sprzed. prod., usług	26	Podatek dochodowy

Źródło: opracowanie własne

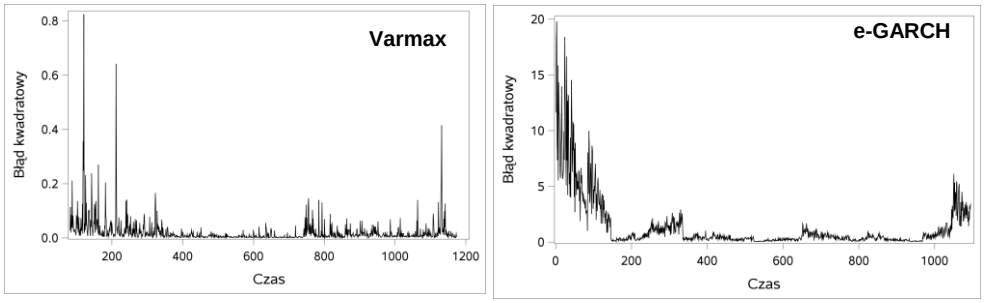
Kolejnym etapem był wybór wskaźników ekonomicznych, wchodzących do modeli, dzięki którym można było oszacować ceny akcji całego sektora budowlanego. Wybór zmiennych został dokonany na trzy sposoby: po pierwsze na podstawie dostępnych źródeł literaturowych [Gołębiewski 2009], po drugie arbitralnie i po trzecie za pomocą metod statystycznych (eliminując m.in. współliniowość zmiennych lub duże braki danych). W ten sposób otrzymano zestaw zmiennych objaśniających, wykorzystywanych w dalszym modelowaniu (Tabela 1).

Rysunek 1. Wybrane wskaźniki ekonomiczne z uzupełnionymi brakami danych



Źródło: opracowanie własne

Rysunek 2. Modele Varmax oraz e-GARCH



Źródło: opracowanie własne

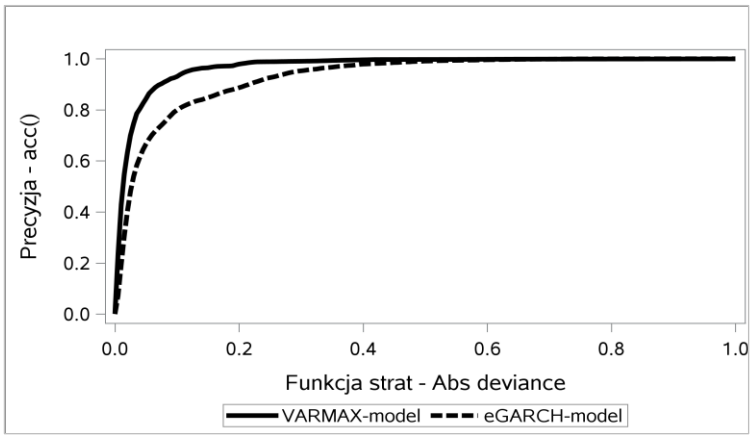
Następnym etapem analizy było uzupełnienie braków danych. W tym celu wykorzystano wspomnianą wcześniej metodologię „model based". Następnie za pomocą procedury X-12 ARIMA dane zostały zdekomponowane na trzy składowe:

sezonowość, trend oraz wahania losowe. Wybrane wskaźniki ekonomiczne, obejmujące wszystkie spółki giełdowe, zostały przedstawione na Rysunku 1. W przypadku danych dziennych, braki danych były uzupełniane średnią ruchomą ze skokiem dniowym.

Analizę porównawczą modeli przeprowadzono w oparciu o scenariusze, wygenerowane na podstawie modeli: VARMAX oraz e-GARCH. Scenariusz oznacza w tym kontekście potencjalną wartość wektora Y_t uzyskaną na podstawie modelu prognostycznego z uwzględnieniem czynnika losowego, który odzwierciedlałby naturalną zmienność rynku (Rysunek 2).

Jako model VARMAX wybrany został model z korektą błędem VECM z opóźnieniem p=1 oraz rzędem kointegracji r=1. Dobór modelu został dokonany na podstawie statystyki AIC. Z drugiej strony do porównania wybrany został model e-GARCH. Modele e-GARCH estymowane były jako modele jednowymiarowe. Z uwagi na korelacje uzyskane na podstawie modeli, reszty były korygowane przy użyciu funkcji kopuł. Do analizy użyto kopuł Studenta z 4 stopniami swobody. Uzyskane reszty zostały znormalizowane i zsumowane tak, aby otrzymać wartość błędów.

Rysunek 3. Krzywe REC dla modeli e-GARCH oraz VARMAX



Źródło: opracowanie własne

Tabela 2. Wyniki porównania numerycznego krzywych REC

Model	Area Over the REC Curve (AOC)	
	Wartość wsp.	Błąd standardowy
Varmax	0,0309	0,0010
e-GARCH	0,0710	0,0020

Źródło: opracowanie własne

Ostatnim etapem badania było porównanie modeli prognostycznych przy zastosowaniu krzywych REC. Pozwoliło ono stwierdzić, że dla różnych progów funkcji strat modele VARMAX pozwalają na uzyskanie mniejszych błędów (Rysunek 3, Tabela 2).

PODSUMOWANIE

Współczesne podejście do modelowania procesów ekonomicznych wiąże ten proces z budową modeli ilościowych, opartych na równaniach ekonometrycznych. Z tego powodu na pierwszy plan wysuwa się pomiar oceny dokładności stosowanych modeli. Niestety nie można stosować w tym celu tradycyjnej statystyki, ponieważ rozkłady odbiegają znacznie od postaci gaussowskiej.

W pracy autorzy zaprezentowali jedno z rozwiązań porównania uzyskanych modeli prognostycznych. Tym rozwiązaniem są krzywe REC, które w łatwy sposób odpowiadają na pytanie, który model daje lepsze wyniki i jest lepiej dopasowany do danych. Detaliczne wyniki stosowania tej metody zostały zaprezentowane na przykładzie sektora budowlanego polskich firm notowanych na giełdzie. Zastosowanie krzywej REC pozwoliło na oszacowanie błędów wielowymiarowych modeli. W przypadku sektora budowlanego spółek giełdowych o wiele lepszym modelem okazał się wielowymiarowy model VARMAX z korektą błędem VECM i z opóźnieniem p=1 oraz rzędem kointegracji r=1.

BIBLIOGRAFIA

Andersen T., Davis R., Kreiss J., Mikosch T. (2009) Handbook of Time Series, Springer.

Bi J., Bennet K. P. (August 2003) Regression error characteristics curves [in:] Proceedings of the AIII 20th International Conference on Machine Learning (ICML'03), pp. 43–50.

Chatfield C. (1995) Model uncertainty, data mining and statistical inference, Journal of the Royal Statistical Society, Seria A 158, pp. 419–466.

Cheng J., Lloyd J., Mildred M., Kelli A., Keith E. (February 2010) Real Longitudinal Data Analysis for Real People: Building a Good Enough Mixed Model, Stat. Med., 29(4), pp. 504–520.

Zumbach B. (2007) The RiskMetrics 2006 methodology, RiskMetrics Group.

Gołębiewski G., Tłaczała A. (2009) Analiza finansowa w teorii i w praktyce, Difin, Warszawa.

Luetkepohl H. (2005) New Introduction to Multiple Time Series Analysis, Springer.

Nelsen R. B. (1998) An Introduction to Copulas, Lectures Notes in Statistics 139, SpringerVerlag, New York.

McNeil A., Frey R., Embrechts P. (2005) Quantitative Risk Management: Concepts, Techniques and Tools, Princeton Series in Finance.

Raghunathan T. i inni (2001) A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, Survey Methodology, Vol. 27, No. 1, pp. 85-95.

US Census Bureau's (2013) The X-13ARIMA-SEATS Seasonal Adjustment Program, https://www.census.gov/srd/www/x13as/

MODEL OF ASSESMENT ECONOMETRIC FORECASTS

Abstract: The main task of the analyst is to select the optimal model. For models applied to individual companies, we have a series of measures allowing to compare them from as well as the accuracy and economic point of view. The situation becomes more sophisticated when the forecasts apply to a group of companies or economic sectors. The authors attempt to build a universal graphical tools based on the REC curve. Results of this method will be used to forecast models of selected sectors Polish companies listed on the stock exchange.

Keywords: REC curve, forecasting models, exchange sectors