PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Marcin Błażejowski
Wyższa Szkoła Bankowa w Toruniu
Paweł Kufel, Tadeusz Kufel
Uniwersytet Mikołaja Kopernika w Toruniu

BANK DANYCH REGIONALNYCH GUS JAKO PODSTAWA ANALIZ ILOŚCIOWYCH W OPROGRAMOWANIU GRETL I R

1. Wstęp

Celem artykułu jest zaprezentowanie baz danych dla oprogramowania GRETL[1] (*GNU Regression, Econometric and Time-series Library*) dla danych importowanych z Banku Danych Regionalnych (w skrócie: BDR) GUS. Utworzone banki danych[2] dotyczące oprogramowania GRETL w podziale terytorialnym powiatowym i wojewódzkim zawierają 1,5 tys. szeregów w odniesieniu do lat 1999-2006. Dla danych statystycznych przedstawionych w bankach zaprezentowano przykłady analiz ilościowych z zakresu ekonometrii dla danych przekrojowych w oprogramowaniu GRETL oraz klasyfikacji obiektów za pomocą funkcji integracji oprogramowania GRETL z pakietem **R**.

2. Baza danych regionalnych GUS w oprogramowaniu GRETL

Tworzenie własnych baz danych wymaga utworzenia pliku z danymi w oprogramowaniu GRETL, który zawiera pełne opisy szeregów danych[3], a następnie zapisania tego pliku jako bazy danych (*.bin) za pomocą funkcji ***Plik / Zapisz dane jako / Baza danych...(*.bin)***. Zapisanie danych do bazy wymaga wskazania wy-

[1] Oprogramowanie dostępne na stronie: gretl.sorceforge.net oraz www.kufel.torun.pl. Tłumaczenie na język polski wykonują: Tadeusz Kufel i Paweł Kufel (UMK Toruń).

[2] Banki danych dostępne są na stronie www.kufel.torun.pl. Autorem instalatora banków jest Marcin Błażejowski (WSB Toruń).

[3] Pełne opisy danych można wykonać za pomocą okna dialogowego *Edycja atrybutów* (klucz F2).

branych lub wszystkich szeregów oraz wprowadzenia tekstu opisu bazy. Postępując w przedstawiony sposób, stworzono takie bazy, jak: baza danych regionalnych GUS dla powiatów oraz województw.

Wykorzystując internetowy interfejs komunikacyjny z Bankiem Danych Regionalnych GUS[4], można zaimportować do Excela, a potem do oprogramowania GRETL obszerny zbiór informacji o jednostkach administracyjnych Polski w podziale na: regiony, województwa, podregiony, powiaty i gminy. BDR zawiera ponad 10 tys. informacji statystycznych dotyczących lat 1999-2006 w układzie danych rocznych. Bank Danych Regionalnych GUS przeniesiony do baz GRETL zawiera ponad 1,5 tys. cech z zakresu:

– stanu ludności i ruchu naturalnego,
– rynku pracy, wynagrodzeń i świadczeń społecznych,
– rolnictwa i leśnictwa, transportu i łączności,
– ochrony środowiska, gospodarki mieszkaniowej i komunalnej,
– przemysłu i budownictwa, handlu i turystyki,
– szkolnictwa podstawowego, ponadpodstawowego i wyższego,
– ochrony zdrowia, kultury i sztuki,
– dochodów i wydatków budżetowych,
– inwestycji i środków trwałych,

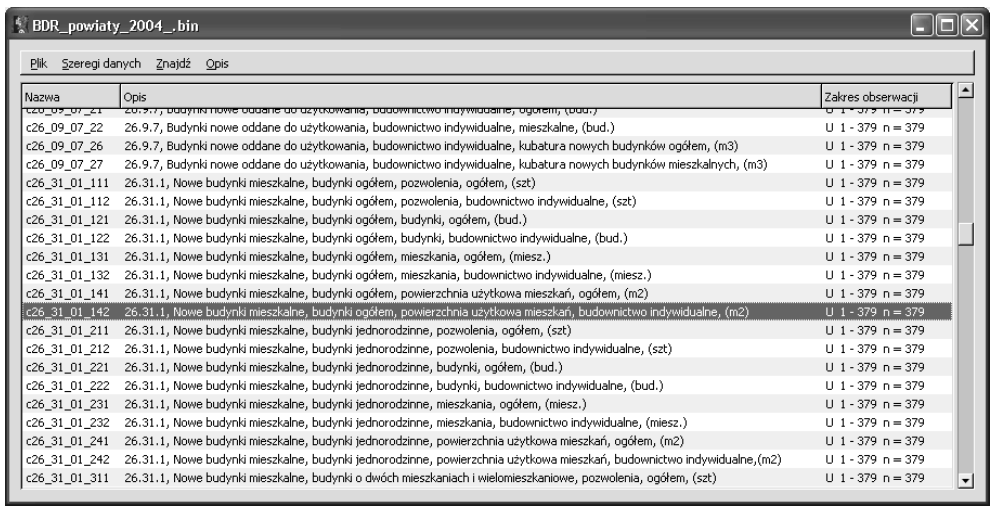
w przekrojach jednostek powiatowych (NTS4) i wojewódzkich (NTS2) dla 1999-2006.

Oprogramowanie GRETL pozwala nazwać zmienną (szereg) identyfikatorem składającym się z maksymalnie 15 znaków, wśród których pierwszy znak musi być literą. Nazwę cechy – zmiennej − skonstruowano na podstawie następujących uporządkowanych numeracji: **cxx_yy_zz_w1w2w3w4w5**, gdzie: xx – numer kategorii, yy – numer grupy, zz – numer podgrupy, od w1 do w5 – numery sortowania kolejnych wymiarów. Na przykład cecha o numerze '**c26_31_01_142**' oznacza:

26. Przemysł i budownictwo (kategoria).
31. Pozwolenia na budowę (grupa).
01. Nowe budynki mieszkalne (podgrupa).
1. Budynki ogółem (wymiar 1).
4. Powierzchnia użytkowa mieszkań (wymiar 2).
2. Budownictwo indywidualne (wymiar 3), jednostka (m2).

Okno zaprezentowane na rys. 1 przedstawia przykładowy wykaz cech dostępnych w bazie danych powiatów dla 2004 r. Funkcja menu ***Szeregi danych / Import*** pozwala przenieść wybrane szeregi do obszaru roboczego oprogramowania GRETL.

[4] Bank Danych Regionalnych GUS (http://www.stat.gov.pl) jest największym w Polsce uporządkowanym zbiorem informacji o sytuacji społeczno-gospodarczej, demograficznej, społecznej oraz stanie środowiska, opisującym województwa, powiaty oraz gminy jako podmioty systemu organizacji społecznej i administracyjnej państwa, a także regiony i podregiony stanowiące elementy nomenklatury jednostek terytorialnych do celów statystycznych. Bank Danych Regionalnych gromadzi, systematycznie uzupełnia i aktualizuje informacje statystyczne o poszczególnych jednostkach podziału terytorialnego.



Rys. 1. Okna programu GRETL: otwarta baza danych na lokalnym dysku
Źródło: opracowanie własne z wykorzystaniem programu GRETL.

Bazy danych dotyczące powiatów są tworzone dla każdego roku osobno, ponieważ prawie co rok liczba powiatów, a także obszar niektórych z nich ulegały zmianie. Dlatego nie jest właściwe tworzenie wspólnej bazy w ujęciu przekrojowo-czasowym.

W celu jednoznacznej identyfikacji cech stworzono dodatkowy opis dostępny poprzez menu ***Opis*** (por. rys. 1). Plik z opisem zawiera numery i nazwy kategorii, grup i podgrup, bez wskazań znaczeniowych wymiarów.

Zbudowane bazy danych statystycznych dla oprogramowania GRETL na podstawie Banku Danych Regionalnych GUS są zebrane w specjalnym pliku instalacyjnym pod nazwą *BDR_gretl.exe* i udostępnione na stronie internetowej http://www.kufel.torun.pl. Wykonanie – zainstalowanie tego pliku na dysku lokalnym − umożliwia otwarcie okna baz na serwerze lokalnym za pomocą menu ***Plik / Pliki baz danych / Zainstalowane bazy danych...***

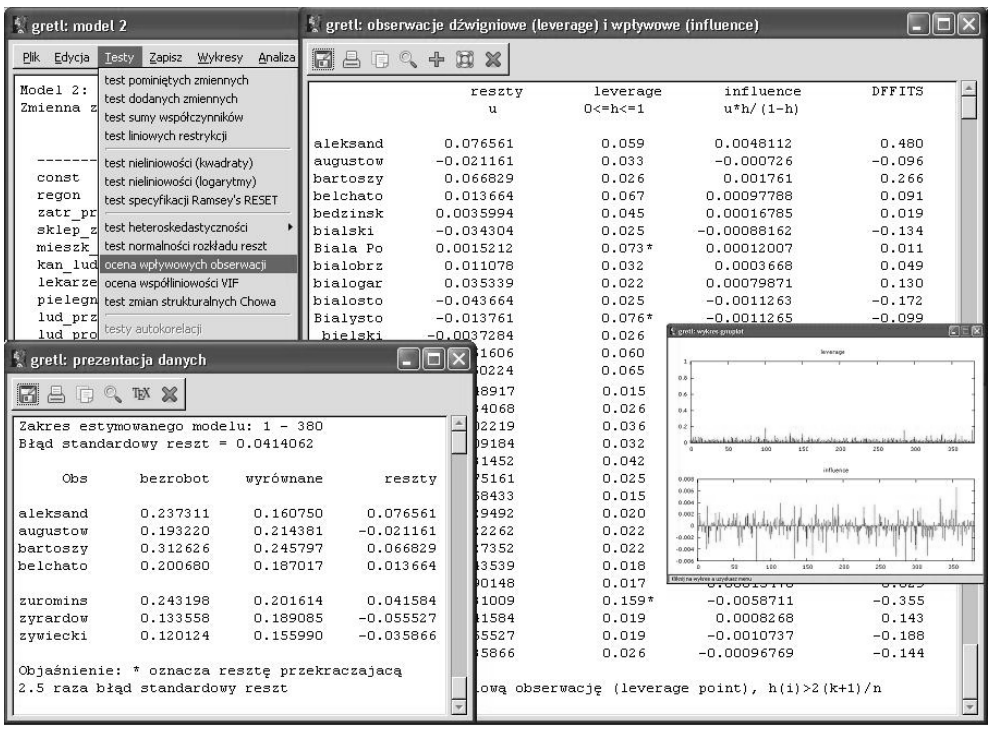
Utworzony Bank Danych Regionalnych w przekroju powiatów zawiera około 1,5 tys. szeregów z informacjami statystycznymi.

3. Ekonometria dla danych przekrojowych

Głównym celem budowy modeli ekonometrycznych dla danych przekrojowych jest analiza współzależności zjawisk. Dodatkowym celem jest dokonanie klasyfikacji obiektów na:

– dźwigniowe (*leverage point*),
– wpływowe (*influential observations*),
– nietypowe (*outliers observations*).

Cel główny − analiza współzależności − można osiągnąć dzięki zbudowaniu dowolnego modelu ekonometrycznego dla danych przekrojowych, natomiast dodatkowa analiza weryfikacyjna modelu pozwala dokonać klasyfikacji obiektów.



Rys. 2. Przykładowe okna wyników analizy obiektów dźwigniowych, wpływowych i nietypowych
Źródło: opracowanie własne z wykorzystaniem programu GRETL.

Na podstawie wartości elementów diagonalnych h_i macierzy rzutowania (*hut matrix*)[5] $H = X\left(X^T X\right)^{-1} X^T$ można dokonać oceny, które elementy są dźwigniowe (*leverage point*). Jeżeli $h_i > h = 2\left(k+1\right)/n$, to *i*-ty obiekt można traktować jako dźwigniowy, czyli usunięcie tego obiektu z szeregu obserwacji spowoduje istotną zmianę oszacowanych parametrów modelu.

Miernik $DFFITS_i$ (*different of fits*) jest kryterium wykrywania obiektów wpływowych[6], które zarazem mogą być obiektami nietypowymi. $DFFITS_i$ jest bowiem wystandaryzowaną miarą przyrostu teoretycznej wartości y_i wynikającą z pominięcia konkretnego obiektu. Wartość miernika $DFFITS_i$ jest wyznaczana ze wzoru:

[5] Por. [Davidson, MacKinnon 2004, s. 79-80; Maddala 2006, s. 529-535; *Ekonometria*… 2002, s. 166-170; Kufel 2007, s. 64-66].

[6] Por. [Davidson, MacKinnon 2004, s. 76-78; Maddala 2006, s. 537-540; Kufel 2007, s. 64-66; Kufel 2008].

$DFFITS_i = \tilde{u}_i \sqrt{\left(\frac{h_i}{1-h_i}\right)}$, gdzie: $\tilde{u}_i$ jest *i*-tą resztą studentyzowaną. Jeżeli $\left|DFFITS_i\right| > 2\sqrt{(k+1)/n}$, to *i*-ty obiekt jest wpływowy i dodatkowo może być nietypowy.

Powyższe statystyki można oszacować, wykorzystując oprogramowanie GRETL. Po oszacowaniu modelu ekonometrycznego za pomocą funkcji menu ***Modele / Klasyczna metoda najmniejszych kwadratów*** w otrzymanym oknie wyników modelu znajdują się dodatkowe funkcje pozwalające dokonać pełniejszej weryfikacji modelu. Funkcja menu ***Testy / ocena wpływowych obserwacji*** wyznacza następujące miary dla każdego *i*-tego obiektu: u_i – reszty, h_i – *leverage* (dźwigniowa wartość), $u_i{*}h_i(1 - h_i)$ – *influence* (wpływowa wartość), $DFFITS_i$ – wskaźnik wpływowej wartości. Wartości dźwigniowe i wpływowe są prezentowane na wykresie.

W oknie wyników *gretl: obserwacje dźwigniowe i wpływowe* za pomocą symbolu (*) wskazywane są dźwigniowe obiekty, a w oknie wyników *gretl: prezentacja danych* uzyskanego za pomocą menu okna modelu ***Analiza / Pokaż empiryczne, wyrównane i reszty*** wskazywane są także, za pomocą symbolu (*), nietypowe reszty przekraczające wartość $2.5{*}S_e$. Przykładowe okna wyników zaprezentowanej analizy przedstawia rys. 2.

4. Integracja oprogramowania GRETL z pakietem R

Oprogramowanie GRETL jest ukierunkowane na analizy ekonometryczne, dlatego nie zawiera zbyt wielu metod klasyfikacji. Integracja oprogramowania GRETL z pakietem **R** istotnie rozszerza liczbę i zakres analiz możliwych do wykonania[7].

Istnieją trzy sposoby integracji oprogramowania GRETL z pakietem **R**. Sposób pierwszy, najbardziej prosty, polega na podłączeniu otwartego zbioru danych, który może być podpróbą określoną przez zestaw restrykcji dla próby jako obiektu pakietu **R** o ustalonej nazwie „gretldata". Praca z podłączonym zbiorem danych wymaga tylko wywołania w menu polecenia: ***Narzędzia / Start programu R Gui***.

Sposób drugi polega na utworzeniu skryptu poleceń pakietu **R** w specjalnym oknie *gretl: edycja skryptu* ***R***, które uzyskujemy przez menu ***Plik / Pliki poleceń skryptowych / nowy plik skryptowy / skrypty R***. Utworzony skrypt **R** można wykonać w dwóch trybach:

– *nieinteraktywnym* (tylko okno wyników),
– *interaktywnym* w programie R − ostatnia linia skryptu bez polecenia *q()*.

Tryb drugi – *interaktywny* − pozwala po wykonaniu skryptu dalszą pracę w programie **R** w oknie konsoli, wykorzystując podłączoną bazę danych z programu GRETL.

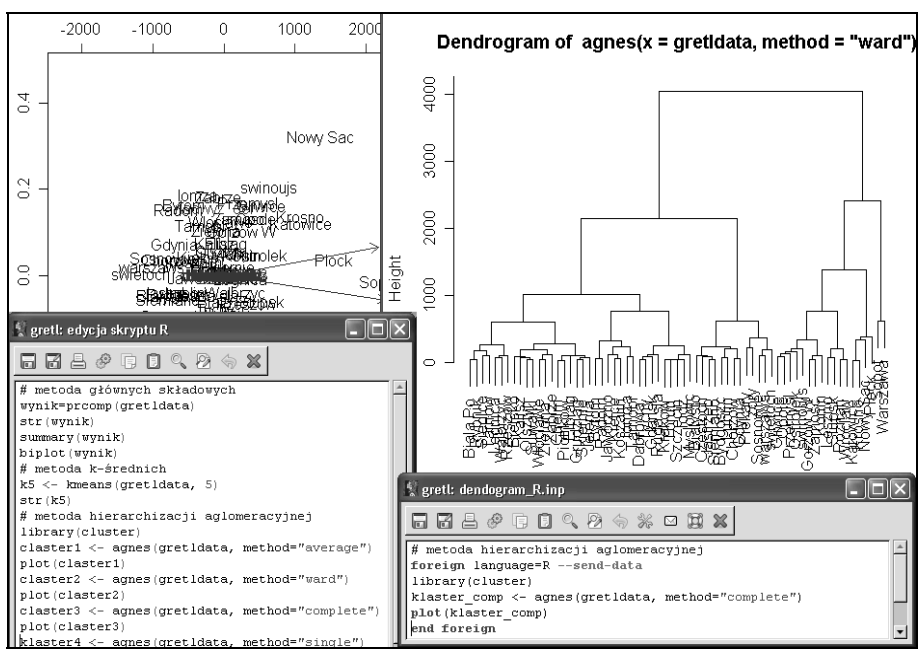
[7] Szerszy opis podstaw pracy z pakietem **R** przedstawia praca: [Biecek 2008], a wielowymiarową analizę danych w pakiecie **R** − praca: [*Statystyczna analiza*... 2008].

Trzeci sposób, najwyższy poziom integracji z pakietem **R**, polega na włączeniu skryptu **R** do skryptu programu GRETL zawartego pomiędzy poleceniami:

foreign language=**R --send-data**
...
skrypt programu **R**
...
end foreign

Okno *gretl: polecenia skryptu* uzyskuje się poprzez menu ***Plik / Pliki poleceń skryptowych / nowy plik skryptowy / skrypty gretla***.

Rysunek 3 zawiera przykładowe okna integracji oprogramowania GRETL z pakietem **R**.



Rys. 3. Przykładowe okna skryptów i wyników analizy realizowanych w pakiecie **R** na bazie danych z oprogramowania GRETL
Źródło: opracowanie własne z wykorzystaniem programu GRETL.

Przykładowe okna skryptów i wyników dowodzą, że integracja oprogramowania GRETL z pakietem **R** jest bardzo łatwa, a trzy sposoby jej wykonania zawsze pozwalają skorzystać z banków danych przygotowanych dla oprogramowania GRETL. W ten sposób bardzo duży Bank Danych Regionalnych GUS przygotowany do pracy w oprogramowaniu GRETL jest użytecznym zbiorem danych dla metod i algorytmów funkcjonujących w pakiecie R.

5. Podsumowanie

Informacje statystyczne zawarte w Banku Danych Regionalnych GUS, a udostępnione za pomocą narzędzi oprogramowania GRETL w jego bazach danych zwiększają szanse zastosowań analiz statystycznych w odniesieniu do danych w ujęciu przekrojowym.

Funkcje integracji z pakietem R zwiększają możliwości analiz realizowanych na bazach danych oprogramowania GRETL.

Wspomaganie nauczania statystyki i ekonometrii oprogramowaniem GRETL okazuje się bardzo pomocne w nauczaniu tych przedmiotów dzięki możliwościom analizowania rzeczywistych przykładów.

Literatura

Adkins L. (2007), *Using GRETL for principles of econometrics*, 3rd edition, http://www.learneconometrics.com/gretl.html.

Biecek P. (2008), *Przewodnik po pakiecie **R***, Oficyna Wydawnicza GIS, Wrocław.

Cottrell A., Lucchetti R. 'Jack' (2008), *Gretl user's guide, GNU regression, econometrics and time series*, http://gretl.sourceforge.net.

Davidson R., MacKinnon J.G. (2004), *Econometric theory and methods*, Oxford University Press, New York, Oxford.

Ekonometria. Metody. Przykłady. Zadania (2002), red. J. Dziechciarz, AE, Wrocław.

Kufel T. (2007), *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL*, Wydawnictwo Naukowe PWN, Warszawa.

Kufel T. (2008), *Obserwacje nietypowe w procesach gospodarczych dla danych dziennych,* [w:] *Modelowanie i prognozowanie zjawisk społeczno-gospodarczych*, red. J. Pociecha, Uniwersytet Ekonomiczny, Kraków, s. 325-338.

Maddala G.S. (2006), *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.

Mixon W.J., Smith R.J. (2006), *Teaching undergraduate econometrics with GRETL*, "Journal of Applied Econometrics", vol. 21, no 7, s. 1103-1107.

*Statystyczna analiza wielowymiarowa w wykorzystaniem programu **R*** (2008), red. M. Walesiak, E. Gatnar, Wydawnictwo Naukowe PWN, Warszawa.

CENTRAL STATISTICAL OFFICE'S REGIONAL DATA BANK AS A BASIS FOR QUANTITATIVE ANALYSIS IN GRETL AND R

Summary

The purpose of this article is to present Regional Data Bank of Central Statistical Office for GRETL (GNU Regression, Econometric and Time-series Library). Built databases concern over 1500 cross-sectional series for the years 1999-2006 in district and voivodeship structure. Some quantitative analyses for this database are presented, from spatial econometrics area as well as clustering and classification, which were computed using functions of R package integrated in GRETL package.