PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Marcin Błażejowski
Wyższa Szkoła Bankowa w Toruniu

Paweł Kufel, Tadeusz Kufel
Uniwersytet Mikołaja Kopernika w Toruniu

INTEGRACJA ŚRODOWISK OBLICZENIOWYCH OPROGRAMOWANIA GRETL I *R*

Streszczenie: Celem artykułu jest przedstawienie baz danych dla oprogramowania GRETL (*GNU Regression, Econometric and Time-series Library*) dla danych zaimportowanych z Banku Danych Regionalnych GUS. Utworzone banki danych dla oprogramowania GRETL w podziale terytorialnym powiatowym i wojewódzkim dotyczą ponad 1,5 tys. szeregów dla okresu od 1999 do 2007 r. 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 integrowanego z oprogramowania GRETL pakietu ***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 GUS. Utworzone banki danych[2] dla oprogramowania GRETL w podziale terytorialnym powiatowym i wojewódzkim zawierają 1,5 tys. szeregów dla lat od 1999 do 2007. 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***. Wybór oprogramowania GRETL i ***R*** podyktowany był bezpłatnym dostępem dla każdego użytkownika, ponieważ są to pakiety typu *open source*.

[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).

2. 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*** bardzo mocno rozszerza spektrum możliwych do wykonania analiz[3].

Istnieją trzy sposoby integracji GRETL-a 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 poprzez 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 na dalszą pracę w programie ***R*** w oknie konsoli, wykorzystując podłączoną bazę danych z programu GRETL.

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 przez menu ***Plik / Pliki poleceń skryptowych / nowy plik skryptowy / skrypty gretla***. Należy jednak pamiętać, że w tym przypadku należy ustawić zmienną środowiskową „R_HOME", która wskazuje na umiejscowienie „etc/Renviron" w drzewie katalogowym (np. w środowisku GNU/Linux dla większości przypadków będzie to „/usr/lib/R").

3. Przykłady dla metod klasyfikacji – prezentacje kartograficzne

Ważnym typem integracji środowisk obliczeniowych GRETL oraz ***R*** jest ich łączne wykorzystanie w ten sposób, że każde z nich wykorzystywane jest do takich analiz, które w danym środowisku są możliwe lub łatwiejsze do wykonania. Przykładem tego typu połączenia jest kreślenie map w środowisku ***R*** na podstawie wy-

[3] Szerszy opis podstaw pracy z pakietem ***R*** przedstawia praca: [Biecek 2008], a wielowymiarową analizę danych w pakiecie ***R*** praca: [Walesiek, Gatnar 2009].

ników modeli ekonometrycznych oszacowanych w GRETL-u. Nie jest to więc konkurencyjne wykorzystanie GRETL-a w stosunku do ***R***a (w środowisku ***R*** także można szacować modele ekonometryczne), ale wykorzystanie komplementarne: w sposób optymalny użyte są możliwości obu z omawianych środowisk.

W pierwszym przykładzie zostaną wyznaczone odległości Mahalanobisa od środków ciężkości dla 376 powiatów w Polsce. Wykorzystane zostały następujące cechy badanych obiektów w 2008 r.[4]: stopa bezrobocia, wydane pozwolenia na budowę oraz liczba nowych mieszkań oddanych do użytku. Odległości zostały wyznaczone w programie GRETL według formuły [Kufel 2007, s. 64-66]:

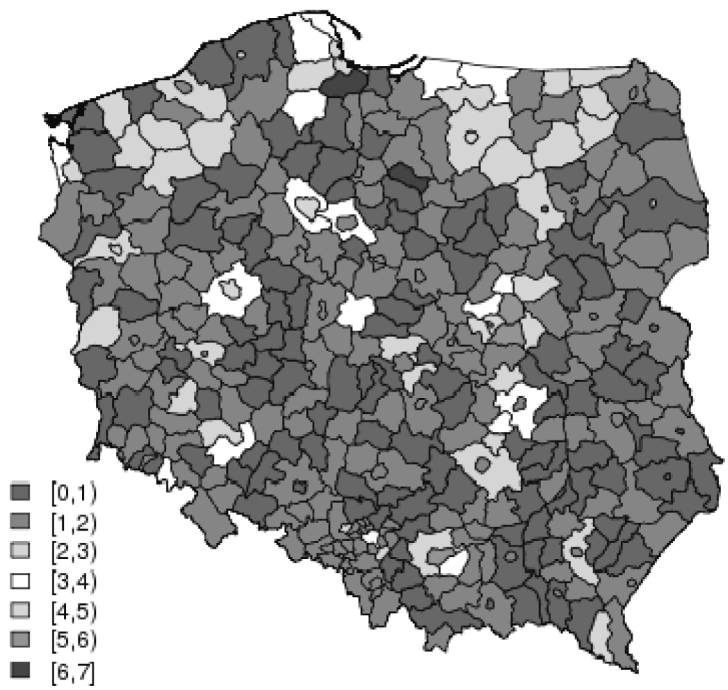
$$MD_i = \left(x_i - \bar{x}\right) s^{-1} \left(x_i - \bar{x}\right)^T ,$$

gdzie: x_i jest wektorem obserwacji, $\bar{x}$ jest wektorem średnich dla zmiennych, s jest macierzą kowariancji zmiennych. Wyznaczenie odległości w GRETL-u realizuje się według ścieżki ***Widok / Odległość Mahalanobisa***. Następnie, wykorzystując możliwości pakietu ***R***, wyznaczone odległości dla poszczególnych powiatów naniesiono na mapę administracyjną powiatów w Polsce. W tym celu została wykorzystana darmowa mapa powiatów w Polsce, którą można pobrać z serwisu internetowego: http://www.gadm.org. Ważną kwestią jest tutaj to, aby kolejność obiektów w bazie GRETL-a była identyczna z kolejnością obiektów w bazie danych geograficznych. W celu wykreślenia mapy należy wykonać następujący kod w programie GRETL:

```
foreign language=R --send-data

# Załadowanie 4 niezbędnych bibliotek środowiska R
library(maptools)
library(sp)
library(classInt)
library(RColorBrewer)

# Wczytanie pliku shapefile z danymi geograficznymi dla powiatów
pow_shp <- readShapePoly("POL_adm2.shp",proj4string=CRS("+proj=longlat +ellps=clrk80"))

# Przesłanie bazy danych z programu GRETL do środowiska R
pow <- as.data.frame(gretldata)

# Ustawienie parametrów mapy
zmienna <- pow$mdist
kolory <- brewer.pal(przedzialy, "RdGy")
klasy <- classIntervals(zmienna, przedzialy, style = "pretty")
tabela.kolorow <- findColours(klasy, kolory)
plot(pow_shp, col=tabela.kolorow)
legend("bottomleft", legend = names(attr(tabela.kolorow, "table")),
fill = attr(tabela.kolorow, "palette"), cex = 1, bty = "n")

end foreign.
```

[4] Dane dotyczące powiatów pochodziły z Banku Danych Regionalnych GUS.

Pierwsza część zaprezentowanego skryptu zawiera informacje o wymaganych bibliotekach, które będą wykorzystywane przy kreśleniu map. Następnie wczytywany jest plik z danymi geograficznymi, w tym przypadku „POL_adm2.shp". W następnym kroku następuje przesłanie całej dostępnej w GRETL-u bazy danych do środowiska ***R***. W ostatnim etapie następuje wykreślenie mapy geograficznej na podstawie wartości zmiennej „mdist", która przedstawiona jest na rys. 1.



Rys. 1. Odległości powiatów od środków ciężkości dla 3 zmiennych w 2008 r.

Źródło: opracowanie własne.

Drugi przykład przedstawia sposób wykreślenia mapy województw w Polsce ze względu na wartości reszt modelu ekonometrycznego stopy bezrobocia w roku 2008 o następującej specyfikacji:

$$\begin{aligned} \text{st\_bezrob}_i = {} & \alpha_0 + \alpha_1 \text{sr\_mes\_wynag}_i + \alpha_2 \text{doch\_budzetu}_i + \alpha_3 \text{pow\_gos\_rol}_i \\ & + \alpha_4 \text{drogi\_publ}_i + \alpha_5 \text{naklady\_BR}_i + \alpha_6 \text{prod\_bud\_mont}_i \alpha_7 \text{turys\_zagr}_i \\ & + \alpha_8 \text{regon}_i + \alpha_9 \text{wyd\_budzet}_i + \alpha_{10} \text{naklady\_inw}_i + \xi_i \end{aligned} \quad , (1)$$

gdzie: *sr_mes_wynag* oznacza średnie miesięczne wynagrodzenie, *doch_budzetu* – dochody budżetowe, *pow_gos_rol* – powierzchnię gospodarstw rolnych, *drogi_publ* – długość dróg publicznych, *naklady_BR* – wielkość nakładów na badania

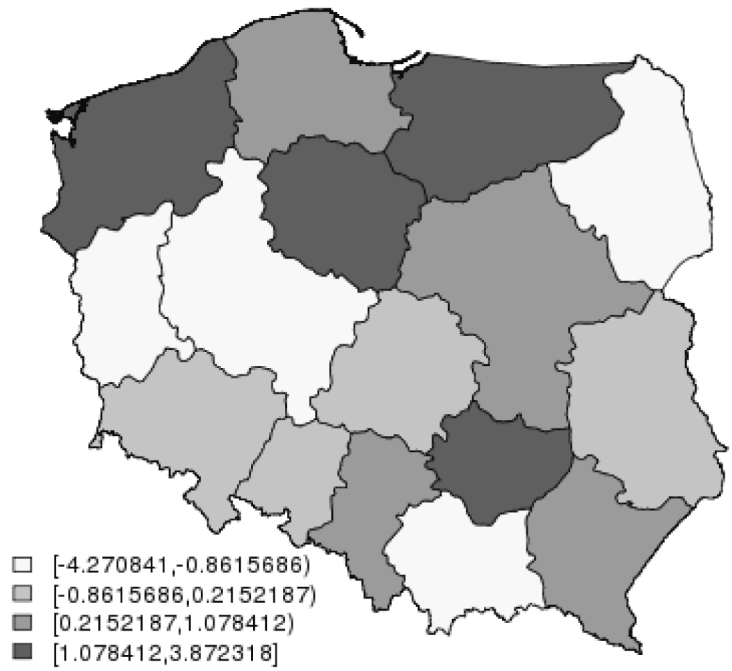
i rozwój, *prod_bud_mont* – wartość produkcji budowlano-montażowej, *turys_zagr* – liczbę turystów z zagranicy, *regon* – liczbę numerów w bazie REGON, *wyd_budzet* – wartość wydatków budżetowych, *naklady_inw* – wielkość nakładów inwestycyjnych, ξ_i zaś oznacza składnik losowy. Wszystkie uwzględnione w modelu zmienne przypadały w wyrażeniu na jednego mieszkańca. Model empiryczny, uzyskany metodą krokowej eliminacji nieistotnych zmiennych *a posteriori* na podstawie statystyk *t*-Studenta, miał postać:

$$\text{st\_bezrob}_i = 27.73 + 0.009\,\text{sr\_mes\_wynag}_i + 0.03\,\text{doch\_budzetu}_i + e_i \,.$$

Współczynnik determinacji osiągnął wartość R^2 = 0,49, wartość *p* w teście Doornika-Hansena na normalność rozkładu składnika losowego miała wartość *p* = 0,65, natomiast wartość *p* w teście White'a na heteroskedastyczność składnika losowego miała wartość *p* = 0,75. Wobec zadowalających własności modelu empirycznego postanowiono wykreślić mapę województw w Polsce ze względu na wartości reszt w modelu empirycznym. W tym celu wykonano następujący kod:

```
foreign language=R --send-data

# Załadowanie 4 niezbędnych bibliotek środowiska R
library(maptools)
library(sp)
library(classInt)
library(RColorBrewer)

# Wczytanie pliku shapefile z danymi geograficznymi dla województw
woj <- readShapePoly("POL_adm1.shp", proj4string=CRS("+proj=longlat +ellps=clrk80"))

# Przesłanie bazy danych z programu GRETL do środowiska R
woj.dane <- as.data.frame(gretldata)

# Ustawienie parametrów mapy
zmienna <- woj.dane$reszty
przedzialy <- 4
kolory <- brewer.pal(przedzialy, "BuPu") # wybór kolorów
klasy <- classIntervals(zmienna, przedzialy, style="quantile")
tabela.kolorow <- findColours(klasy, kolory)
plot(woj, col=tabela.kolorow)
legend("bottomleft", legend=names(attr(tabela.kolorow, "table")), fill=attr(tabela.kolorow, "palette"),
cex=1, bty="n")

end foreign.
```

Większość kodu pokrywa się z przykładem dla powiatów, z tą jednak różnicą, że tym razem zmienną, na podstawie której wykreślane są kolory na mapie, jest zmienna „reszty", natomiast liczba kolorów na mapie została ustalona na 4, przy czym podział na 4 kolory był podziałem kwartylowym. Rysunek 2 przedstawia województwa w Polsce ze względu na wielkość reszt z modelu dla bezrobocia w roku 2008:



Rys. 2. Wartość reszt w modelu bezrobocia dla województw w 2008 r.

Źródło: opracowanie własne.

Analiza rys. 2 wskazuje, że największe niedoszacowanie stopy bezrobocia dotyczyło województw: zachodniopomorskiego, kujawsko-pomorskiego, warmińsko-mazurskiego oraz świętokrzyskiego. Największe przeszacowanie stopy bezrobocia odnosiło się do województw: lubuskiego, wielkopolskiego, podlaskiego i małopolskiego.

Przykładowe okna skryptów i wyników dowodzą, że integracja oprogramowania GRETL z pakietem ***R*** jest bardzo prosta do realizacji, a jej trzy sposoby 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***.

4. 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 dla 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 przez możliwości analizowania rzeczywistych przykładów.

Literatura

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

Cottrell A., Lucchetti R. 'Jack', *Gretl User's Guide, GNU Regression, Econometrics and Time Series*, http://gretl.sourceforge. net, 2008.

Davidson R., MacKinnon J.G., *Econometric Theory and Methods*, Oxford University Press, New York, Oxford 2004.

Kopczewska K., Kopczewski T., Wójcik P., *Metody ilościowe w R*, CeDeWu, Warszawa 2009.

Kufel T., *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL*, PWN, Warszawa 2007.

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

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

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

Walesiak M., Gatnar E. (red.), *Statystyczna analiza wielowymiarowa z wykorzystaniem programu **R***, Wydawnictwo Naukowe PWN, Warszawa 2009.

CENTRAL STATISTICAL OFFICE'S REGIONAL DATA BANK AS THE 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.