WYZWANIA WOBEC STATYSTYKI JAKO NAUKI

Prof. zw. dr hab. Czesław Domański
Katedra Metod Statystycznych
Uniwersytet Łódzki

WYZWANIA WOBEC STATYSTYKI JAKO NAUKI

1. Wprowadzenie
2. Wyzwania
3. Big data
4. Zastosowanie metod statystycznych w big data – wizualizacja i analizy funkcjonujących danych ekonomicznych
5. Podsumowanie

1. Wprowadzenie

Widzialny Wszechświat ma rozmiar kilkunastu miliardów lat świetlnych. To około 10^{26} (1 z 26 zerami) metra. Z kolei najmniejsze struktury, których istnienia jesteśmy pewni, to budujące między innymi protony i neutrony, kwarki. Mają rozmiar kilku attometrów, czyli 10^{-18} metra. Najmniejsze i największe obserwowane przez człowieka obiekty dzielą od siebie aż 44 rzędy wielkości!

Kwarki są o

100 000 000 000 000 000 000 000 000 000 000 000 000 000

razy mniejsze od największego obiektu dociekań naukowców.

Nasz świat mieści się w tych 44 zerach.

Są w nim:

- cząstki elementarne,
- żywe organizmy i ich DNA,
- Ziemia i inne planety,
- są gwiazdy, galaktyki i gromady galaktyk.

A gdzieś w tym wszystkim jest człowiek. Jedyna istota, która chce wiedzieć i chce to wszystko zrozumieć.

2. Wyzwania

1) Analiza danych wielkich rozmiarów i wyciąganie z nich wniosków – wiąże się z rozwojem biznesu informatycznego.

Jeśli w jakimś banku odbywają się miliardy transakcji dziennie, to na podstawie ich analizy chcielibyśmy coś wiedzieć:

- czy system jest szczelny,
- czy są ataki hakerskie,
- jacy klienci korzystają z banku najczęściej,
- które miasta przodują itd.

2) Archiwizacja olbrzymiej ilości tych danych.

3. Big data

Wybrane jednostki informacji

Nazwa	Skrót	Ilość bajtów
Kilobajt	KB	1000
Petabajt	PB	1000000000000000
Eksabajt	EB	1000000000000000000
Zettabajt	ZB	1000000000000000000000

Źródła danych:

- Klasyczne bazy danych
- Informatyka w zarządzaniu
- Urządzenia mobilne
- Portale internetowe
- Sieci sensorów
- Big Science
- Planowanie przestrzenne

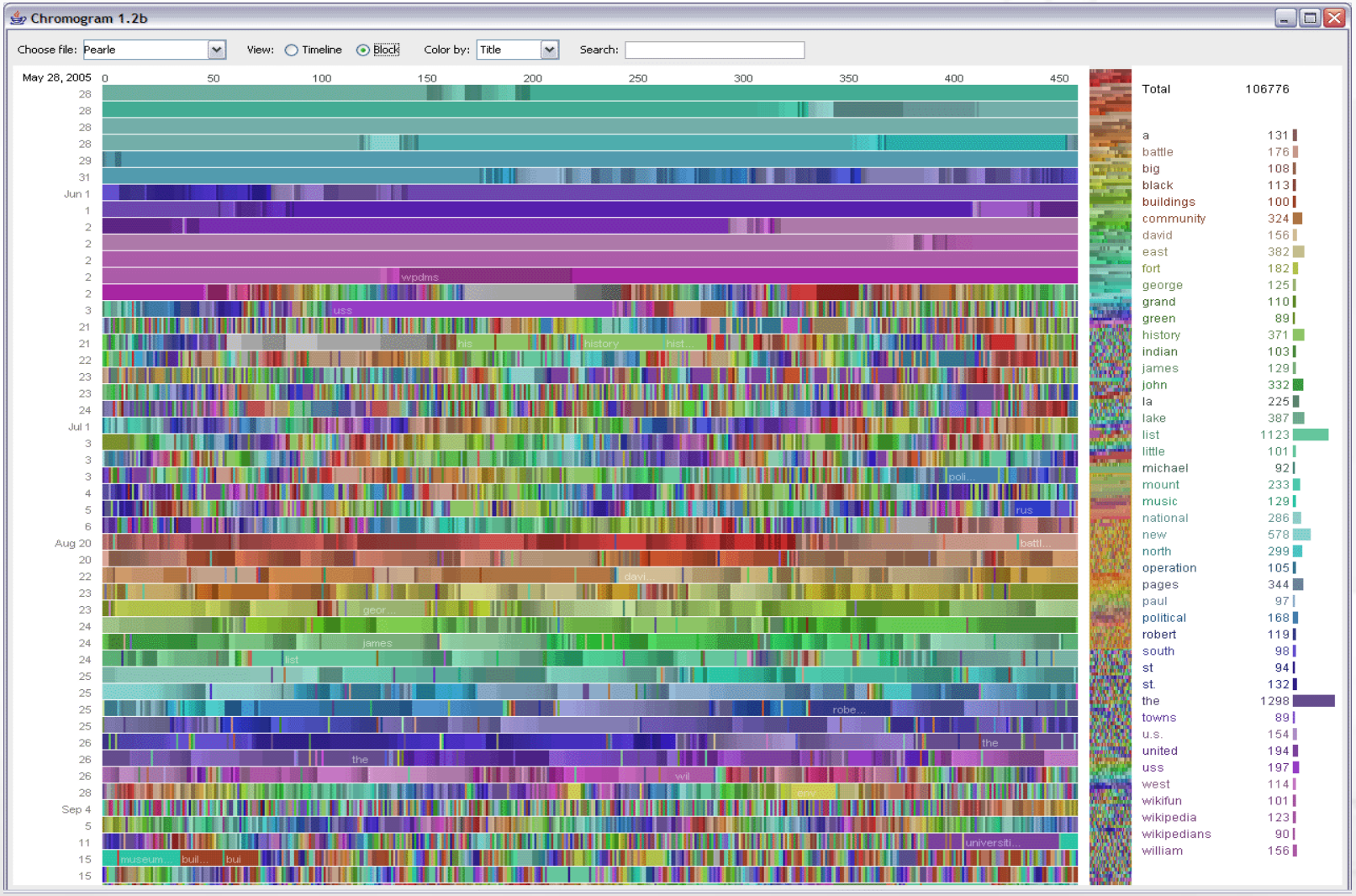
Czym jest Big Data?

- Cecha kluczowa – zbiory zbyt duże do analizy klasycznymi metodami;
- Klasyczna analiza może być wykonalna, ale niewydajna;
- Podział tych zbiorów na mniejsze jest niepożądany.

Big Data – 4V

- Volume - duża ilość danych;
- Variety - duża różnorodność danych;
- Velocity - duża szybkość pojawiania się nowych danych i ich analizy w czasie rzeczywistym;
- Value - znacząca wartość danych dla biznesu.

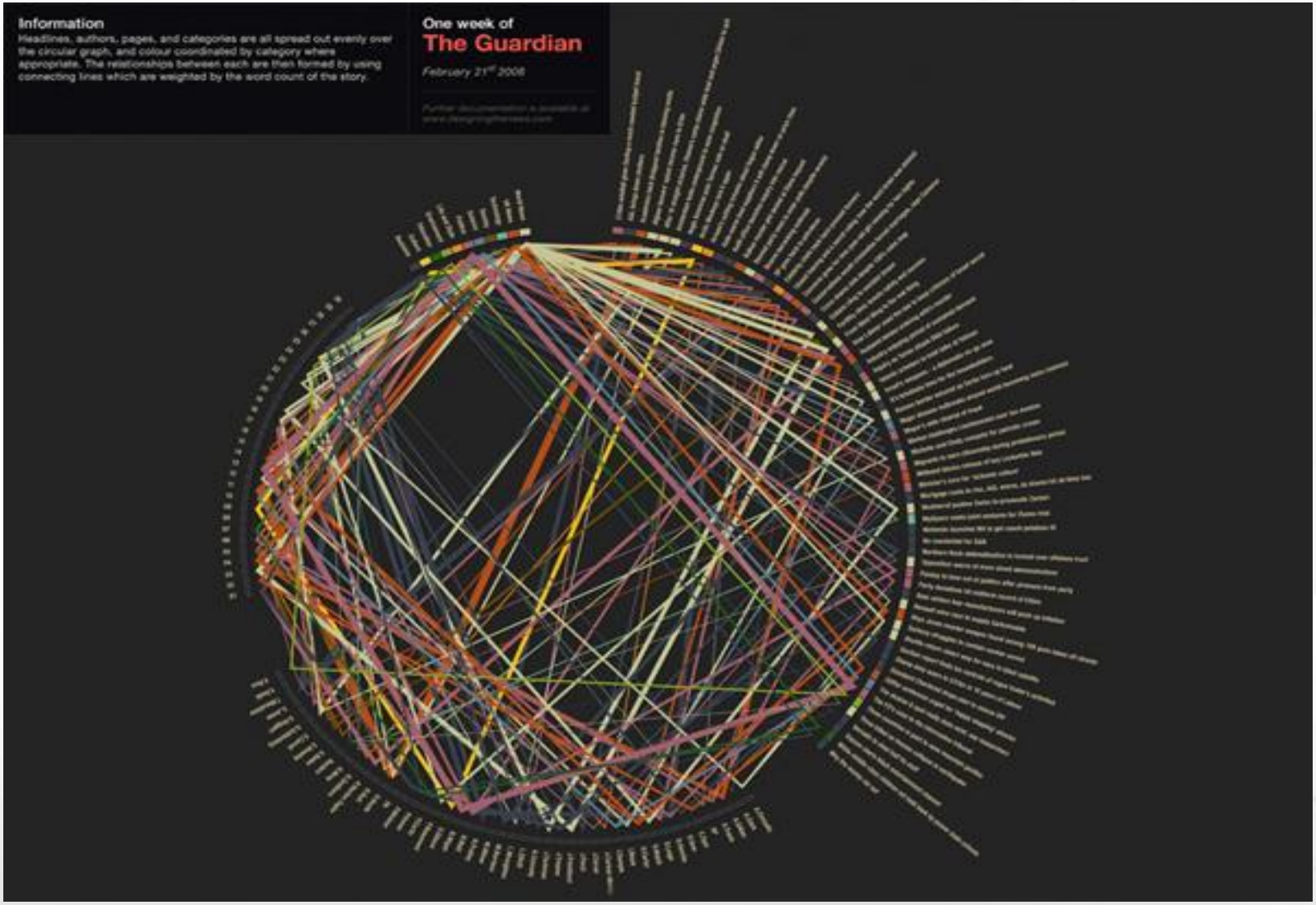
Chromogram



[IBM]

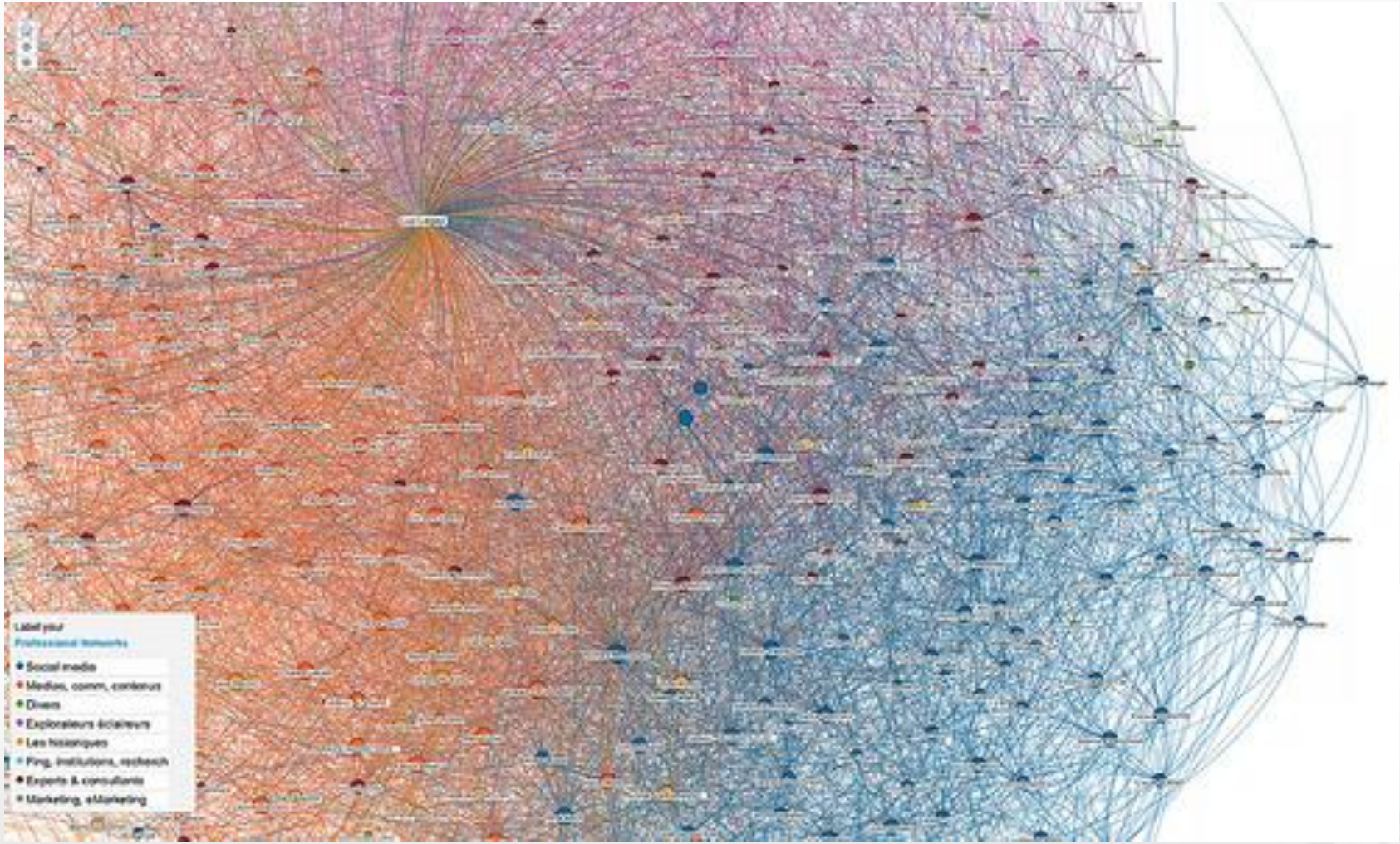
Uniwersytet ŁÓDZKI

Mapy powiązań



Uniwersytet ŁÓDZKI

Mapy powiązań



Metodologia

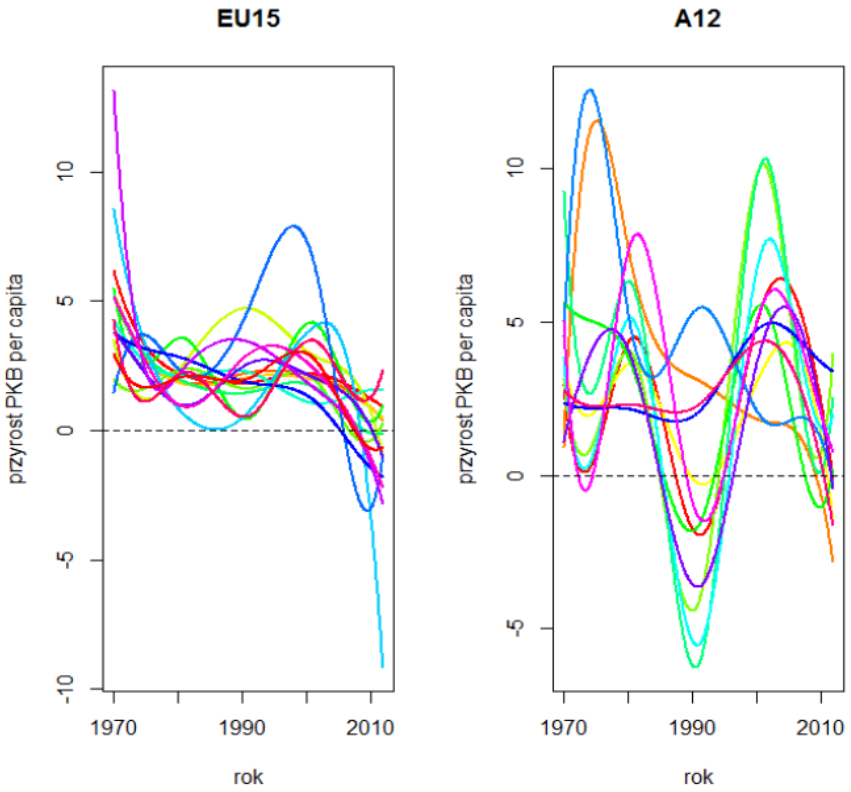
- Podział zadań (distributed computing)
- Rozłożenie pamięci
- Parallel programming
- Komputery pod konkretne zastosowania
- Uczenie maszynowe
- Rozpoznawanie wzorów
- Klasyfikatory
  - Naiwny klasyfikator bayesowski
- MapReduce:
  - Podział zadań
  - Wykonanie pracy w częściach
  - Przetworzenie wyników
  - Spójna odpowiedź
- Apache Hadoop

4. Zastosowanie metod statystycznych w big data – wizualizacja i analizy funkcjonujących danych ekonomicznych

Często dane rozpatrywane w ekonomii mają bezpośrednio bądź pośrednio postać funkcji. Weźmy dla przykładu:

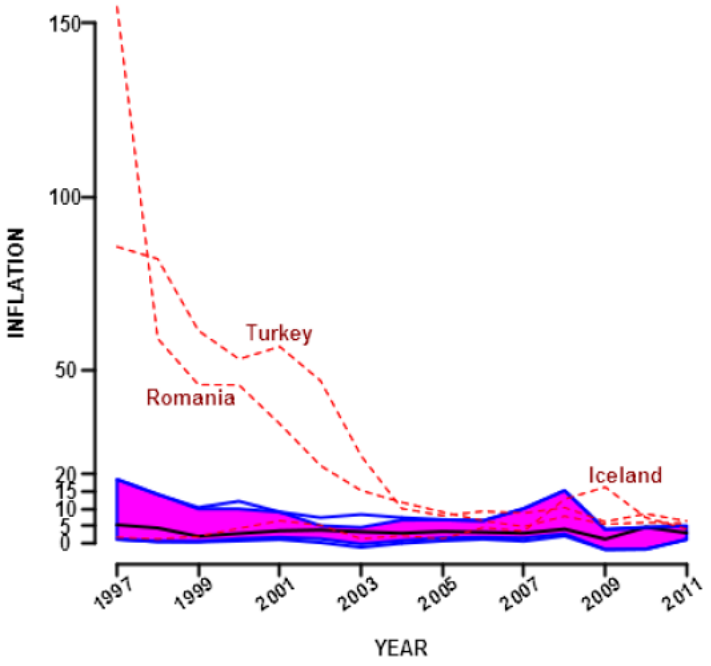
- badania ścieżek rozwoju przedsiębiorstw, trajektorii rozwoju ekonomicznego państw bądź regionów (makroekonomiczne modele wzrostu, badanie faz rozwoju przedsiębiorstwa, cyklu życia produktu – funkcjonalne PCA).
- analizy związków pomiędzy oczekiwaną stopą zwrotu z inwestycji finansowej a „wahaniem" przebiegu tej stopy zwrotu w przeszłości bądź „burzliwością" trajektorii dzisiaj" a taką charakterystyką w przeszłości – funkcjonalna regresja).
- analiza związków pomiędzy ścieżkami rozwoju (kształtem całej trajektorii) dla różnych państw, przedsiębiorstw (funkcjonalne korelacje kanoniczne).
- szacowanie funkcji gęstości, regresji dla danych panelowych (danych tworzących skupiska), grafologia, diagnostyka medyczna, statystyczna teoria kształtu (rozpoznawanie przedmiotów i zachowań niebezpiecznych na podstawie transmisji z kamer miejskiego monitoringu).

PRZYKŁADY

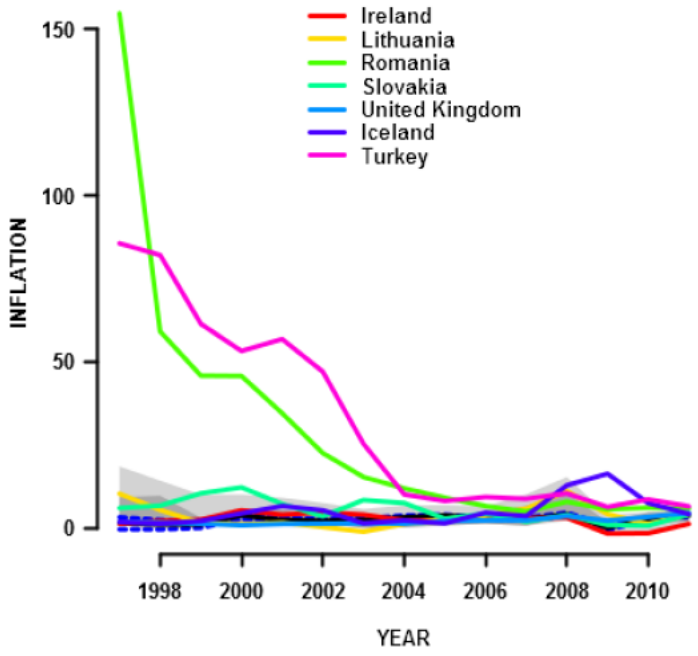


Trajektorie przyrostu PKB per capita w krajach EU15 oraz A12 w latach 1970 – 2011.

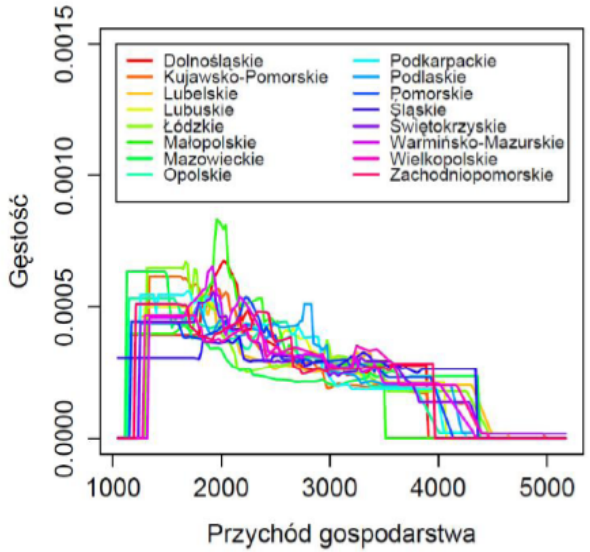
Funkcjonalny wykres pudełkowy – stopa inflacji w krajach UE w latach 1997 – 2011 (dane Eurostat).



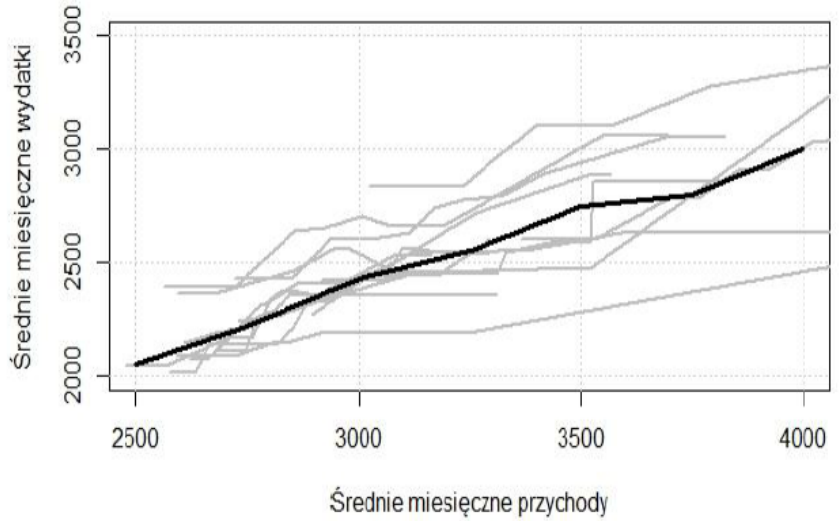
Wykres typu "tęcza" – stopa inflacji w krajach UE w latach 1997 – 2011 (dane Eurostat).

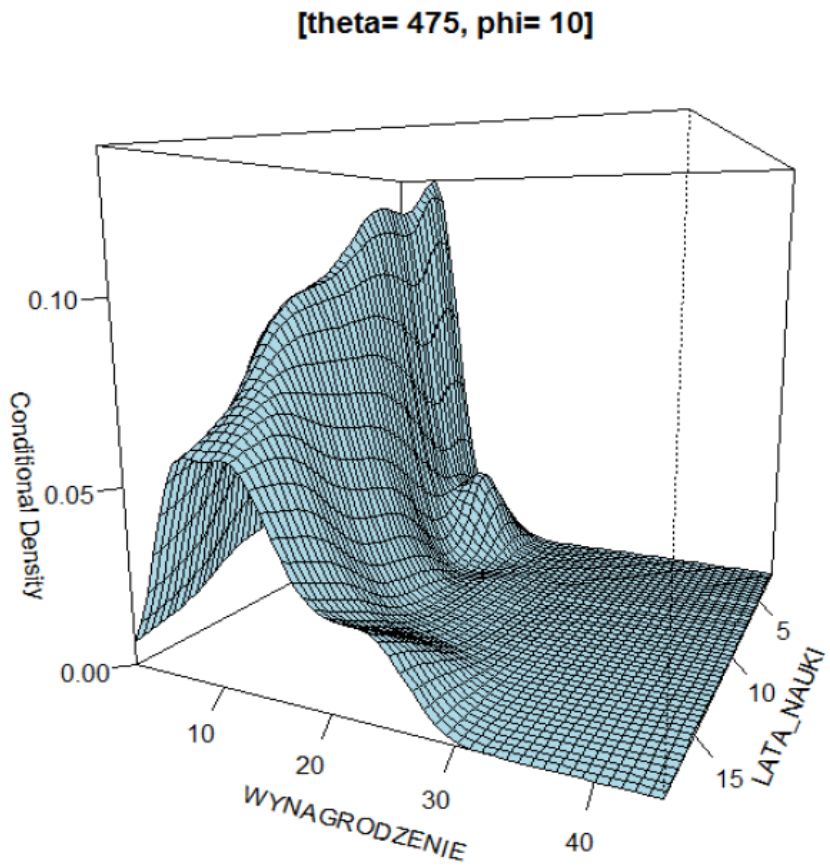


Oszacowanie gęstości prawdopod. dla przychodu centralnej części gospodarstw domowych w roku 2005 w ujęciu województw RP (dane GUS).



Wydatki vs. dochody gospodarstw domowych w ujęciu województw RP. Prosta regresja nieparametryczna dla danych panelowych (dane GUS).





Wynagrodzenie vs. lata nauki – oszacowanie jądrowe rodziny warunkowych gęstości prawdopodobieństwa.

Pionierzy funkcjonalnej analizy danych

Jim Ramsay & Bernard Silverman

POZYCJE KLASYCZNE FDA

1. Applied Functional Data Analysis, Second Edition, J. O. Ramsay and B. W. Silverman, Springer-Verlag, 2002.
2. Functional Data Analysis by J. O. Ramsay and B. W. Silverman. Book published by Springer-Verlag, 2005.
3. Functional Data Analysis with R and Matlab by J. O. Ramsay, G. Hooker and S. Graves. Book published by Springer-Verlag, 2009.

AKTUALNE KIERUNKI POSZUKIWAŃ FDA

1. Inference for Functional Data with Applications, Horvath, Lajos, Kokoszka, Piotr, Series: Springer Series in Statistics, Vol. 200, 2012, XIV
2. Nonparametric Functional Data Analysis Theory and Practice, Frédéric Ferraty, F., P. Philippe Vieu, Springer, 2006

FDA w POLSCE

1. Krzyśko, M., Górecki, T., Deręgowski, K. (2012), Jądrowa i Funkcjonalna Analiza Składowych Głównych – spotkanie PTS o. w Poznaniu.
2. Szereg zastosowań FAD w analizie sygnałów – zespoły z AGH i PW.
3. Odporna FAD w ocenie skuteczności polityk regionalnych i działań samorządów lokalnych – Kosiorowski i in. (2012), (2013).

CELE FAD z PERSPEKTYWY WYKORZYSTYWANYCH TECHNIK

- przekształcenie dyskretnych obserwacji do postaci funkcji (funkcje obserwujemy w dyskretnych chwilach) w taki sposób, aby dalsza analiza była możliwie najprostsza.

- wizualizacja danych uwypuklająca interesujące nas cechy zjawisk.

- analiza wzorców i źródeł zmienności danych.

- analiza związków zmiennymi np. za pomocą regresji skalar vs. zmienna funkcjonalna bądź zmienna funkcjonalna vs. zmienne funkcjonalne.

- porównania zjawisk, estymacja charakterystyk, wnioskowanie statystyczne.

CELE FAD z PERSPEKTYWY CELU ANALIZY

- analiza eksploracyjna (techniki odkrywania nowych cech zjawisk).

- analiza konfirmacyjna (udzielenie odpowiedzi na konkretne pytania).

- analiza predykcyjna (tworzenie schematów prognostycznych dla zjawisk).

Dziękuję za uwagę