Miary korelacji

Janusz Miśkiewicz

Instytut Fizyki Teoretycznej, Uniwersytetu Wrocławskiego,
pl. M.Borna 9, 50-204 Wrocław, Poland

5 Ogólnopolskie Sympozjum FENS
Warszwa, 2010

Zagadnienie

- Istotą układów ekonomicznych jest ich wzajemne oddziaływanie (konkurencja, bądź współdziałanie).
- W fizyce zwykle budujemy model, który jest następnie weryfikowany doswiadczalnie.
- W gospodarce przedsiębiorstwa wzajemne relacje otaczają tajemnicą.

Zagadnienie

- W ekonomii istotnymi i naturalnymi są pytania o wzajemne relacje:
  - Czy dane podmioty są od siebie zależne?
  - Jeżeli tak to który jest dominujący?
  - Czy są niezależne wzajemnie, ale współzależne od podmiotu trzeciego.
  - Jaki jest stopień zależności?

Zagadnienie

- W ekonomii istotnymi i naturalnymi są pytania o wzajemne relacje:
  - Czy dane podmioty są od siebie zależne?
  - Jeżeli tak to który jest dominujący?
  - Czy są niezależne wzajemnie, ale współzależne od podmiotu trzeciego.
  - Jaki jest stopień zależności?

Korelacje

$$A = f(B)$$

Algorytm standardowy

- Miara odległości
- Macierz odległości
- Drzewo MST
- Własności otrzymanego drzewa MST: podmioty dominujące, klasyfikacja gałęzi przemysłu, analiza hierarchii itp.

Odległość ultrametryczna (UD)

- Definicja

$$DU(A,B)_{(t,T)} = \sqrt{\frac{1}{2}(1 - corr_{(t,T)}(A,B))}, \tag{1}$$

$$corr_{(t,T)}(A,B) = \frac{\langle AB\rangle_{(t,T)} - \langle A\rangle_{(t,T)}\langle B\rangle_{(t,T)}}{\sqrt{(\langle A^2\rangle_{(t,T)} - \langle A\rangle^2_{(t,T)})(\langle B^2\rangle_{(t,T)} - \langle B\rangle^2_{(t,T)})}}, \tag{2}$$

R. Mantegna, H. E. Stanley "An Introduction to Econophysics", Cambridge University Press, 2000

Własności UD

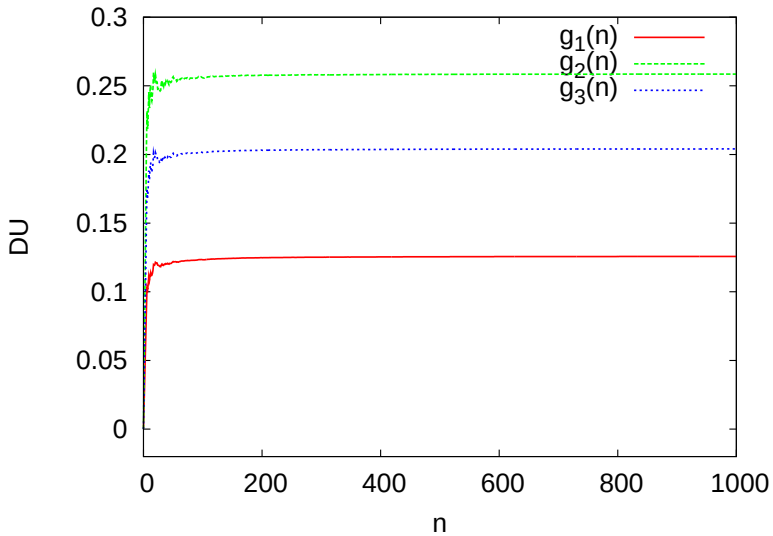
Zalety

- Właściwie klasyfikuje podmioty w kontekście optymalizacji portfela.
- Wynika z teorii portfela optymalnego.
- Powszechnie stosowana.
- Weryfikuje korelacje liniowe.

Własności UD

Wady

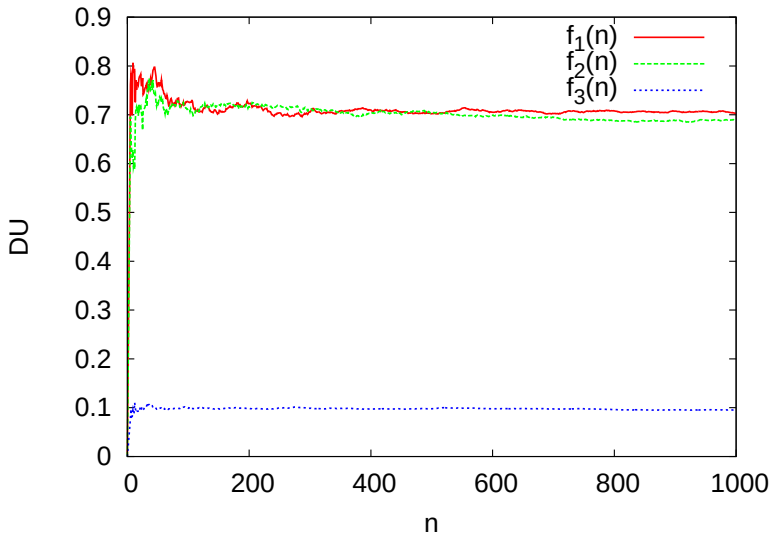
- Odległość ultrametryczna rów.(1) bada korelacje liniowe
  - $a_i = i + w(0.5)$, $b_i = a_i^2 + w(0.5)$; $g_1(n) = DU(A, B)$
  - $a_i = i + w(0.5)$, $b_i = a_i^3 + w(0.5)$; $g_2(n) = DU(A, B)$
  - $a_i = i + w(0.5)$, $b_i = a_i^4 + w(0.5)$; $g_3(n) = DU(A, B)$



Własności UD

Wady

- Odległość ta jest bardzo wrażliwa na zakłócenia (szum).
  - $\langle A \rangle = 0.5$, $\langle B \rangle = 0.5$; $f_1(n) = DU(A, B)$
  - $\langle A \rangle = 0.5$, $\langle B \rangle = 0.0$; $f_2(n) = DU(A, B)$
  - $\langle A \rangle = 0.5$, $b_i = 5 * a_i + w(0.5)$; $f_3(n) = DU(A, B)$



Własności UD

Wady

- Niech X będzie zmienną losową o skończonej warincji i funkcją rozkładu pr. $f(x)$ symetryczną względem wartości średniej tzn. $f(x) = f(-x)$ dla $x \in (-\infty, \infty)$. Wtedy definując zmienną losową $Y = |X|$ otrzymujemy

$$corr(X, Y) = 0$$

Własności UD

Wady

- Jeżeli układ badany potraktujemy schematycznie:

$$\text{Układ 1} \Leftrightarrow \text{Układ 2}$$

Własności UD

Wady

- Jeżeli układ badany potraktujemy schematycznie:

$$\text{Układ 1} \Leftrightarrow \text{Układ 2}$$
$$\Uparrow \qquad \Uparrow$$
$$\text{szum} \qquad \text{szum}$$

Własności UD

Wady

Zakładając, że szum pojawia się w obu szeregach czasowych oraz że jest to szum biały,

$$A = \widehat{A} + W_A,\ B = \widehat{B} + W_B$$

bezpośrednim rachunkiem można pokazać, że

$$DU(A,B)_{(t,T)} =$$

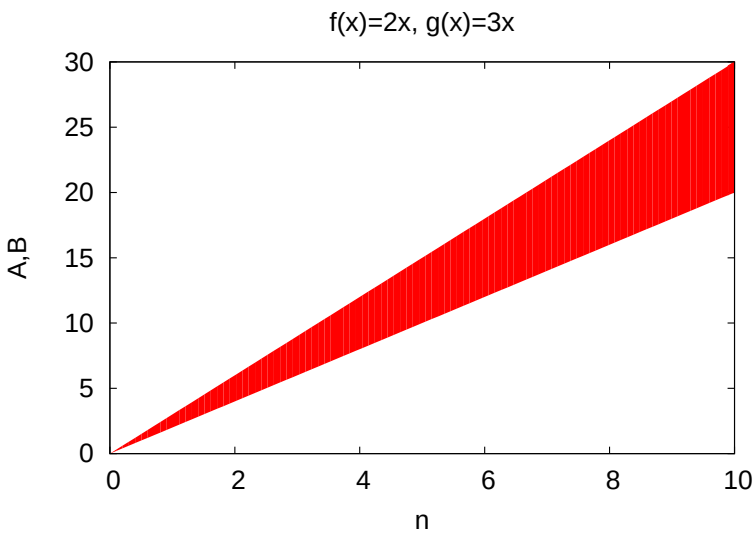
$$\sqrt{\frac{1}{2}(1 - \frac{\langle AB\rangle - \langle A\rangle\langle B\rangle}{\sqrt{(\langle A^2\rangle + \langle W_A^2\rangle - \langle A\rangle^2)(\langle B^2\rangle + \langle W_B^2\rangle - \langle B\rangle^2)}})}.$$

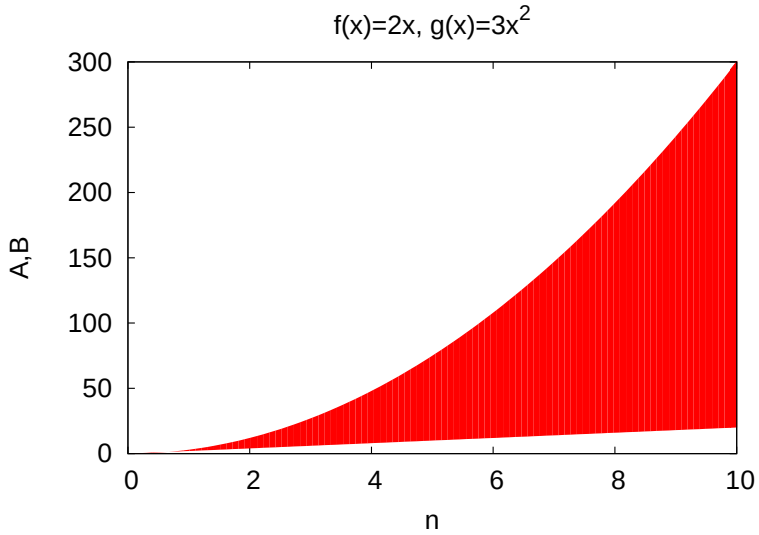
Odległości alternatywne

- Odległość Manhattan i odległości pochodne.
  - W podstawowej formie: $D_M(A,B) = \sum_{i=1}^{n} |a_i - b_i|$
  - oraz uśredniona po długości szeregu: $DM(A,B) = \frac{1}{n} \sum_{i=1}^{n} |a_i - b_i|$
- Zalety
  - Większa odporność na szum: np. dla $a_i > b_i > 0$ zakłócenie w postaci $A + W, B + W$, gdzie W jest białym szumem, ulegnie zredukowaniu.
  - Ponadto dla miary D_M należy zauważyć, że jest ona funkcją długości szeregu czasowego.

Odległości alternatywne

Poczyńmy obserwację:





Klasyfikacja korelacji

Aproksymując dyskretną zmienną długości ciągu zmienną ciągłą i zakładając, że $a_i > b_i$

$$D_M(A,B)(n) \simeq \int_0^n (a(t) - b(t))\, dt$$

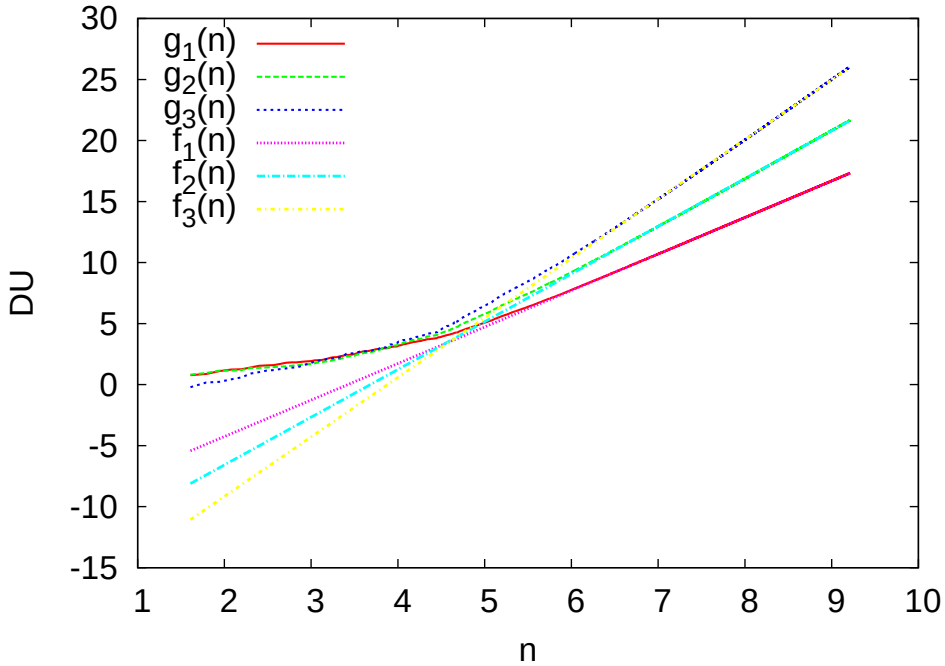
Wtedy funkcję korelacji można znaleźć jako:

$$f(n) = \frac{d(D_M(A,B)(n))}{dn},$$

Klasyfikacja korelacji

Szeregi badane:

$$x_i = t_i + w(0.5),$$
$$y_i^1 = x_i^2 + w(0.5),$$
$$y_i^2 = x_i^3 + w(0.5),$$
$$y_i^3 = x_i^4 + w(0.5).$$

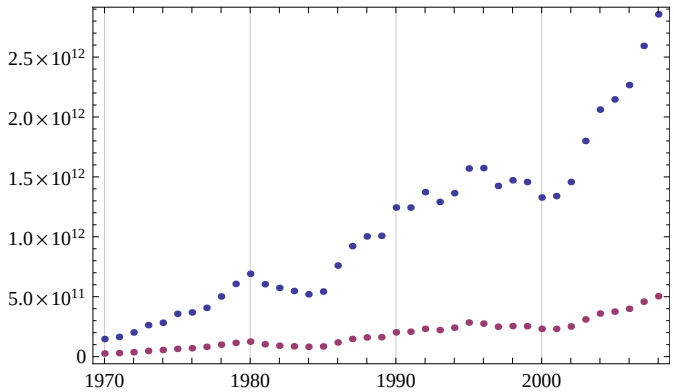


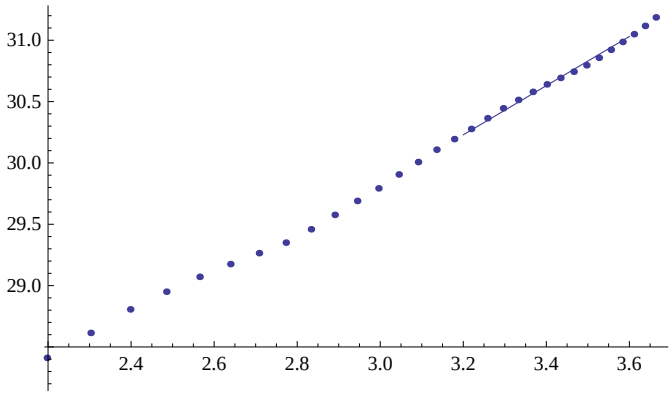
Współczynniki kierunkowe dofitowanych prostych są:
$a_1 \approx 2.91,\ a_2 \approx 3.91,\ a_3 \approx 4.87$, co obliczeniu pochodnej daje funkcje wyjściowe.

- Miara Manhattan daje możliwość oszacowania charakteru korelacji pomiędzy podmiotami.

Przykład

Porównanie PKB Francji i Belgii – odległość D_M





Parametry dofitowanej prostej: y= 2.00x+ 23.82

Odległości alternatywne

- Odległości oparte na entropii.
  - Shanonna

$$S = -\sum_i p_i \ln p_i$$

  - Indeksie Theila

$$Th_A(t,T) = \sum_{i=t-T}^{t} \left(\frac{A_i}{\sum_{j=t-T}^{t} A_j} \ln \frac{A_i}{\langle A \rangle_{(t,T)}}\right)$$

  - Kullbacka-Leiblera

$$d(p|q) = \sum p_i \ln \frac{p_i}{q_i}$$

,

Odległości alternatywne

- Entropia Shannona i index Theila transformują szeregi czasowe do szeregów entropii (zależne od długości okna czasowego) nastepnie do obliczenia odległości można zastosować zarówno odległość DU jak i DM.
- Miary oparte na entropii porównują złożoność informacyjną szeregów czasowych.
- Entropia Shannona i Kullbacka-Leiblera wymaga poznania funkcji rozkładu prawdopodobieństwa.

Struktury sieciowe

- Minimalne drzewo rozpinające (MST)
- Łańcuch dwukierunkowy (BMLP)
  Konstrukcja rozpoczyna się od znalezienia najbliższych sąsiadów. Następnie poszukuje się najbliższego sąsiada do każdego z końców i przyłączany jest bliższy z nich.
- Łańcuch jednokierunkowy (UMLP)
  Pierwszy element sieci jest narzucony, następnie do niego jest przyłączany najbliższy sąsiad, który staje się końcem sieci. Węzły są przyłączane do końca sieci.

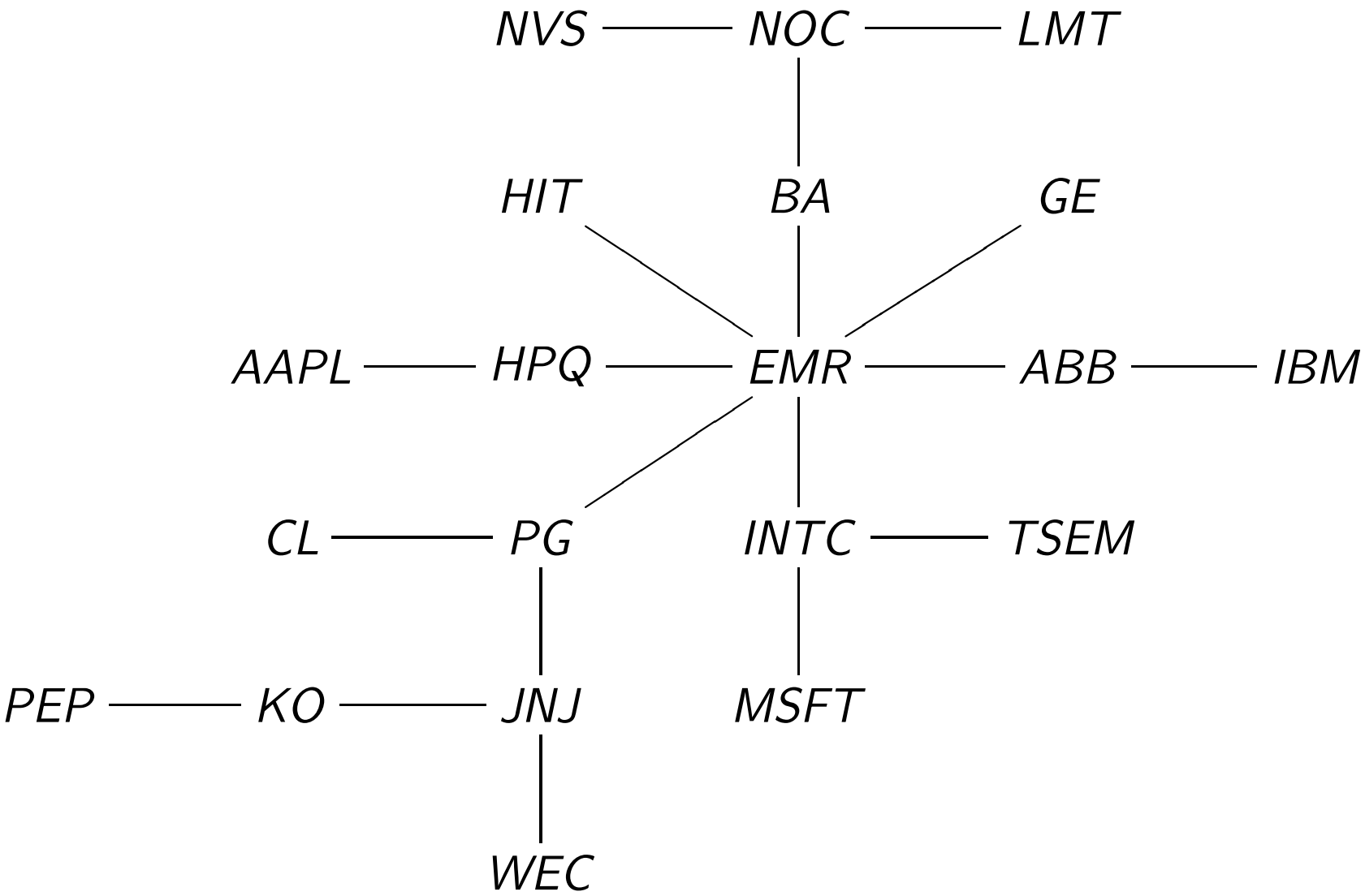
Ewolucja giełdy

Jako ilustrację własności odległości UD i ThD przedstawione zostaną własności sieci ewoluujących MST, BMLP i UMLP dla następujących grupy podmiotów giełdowych:

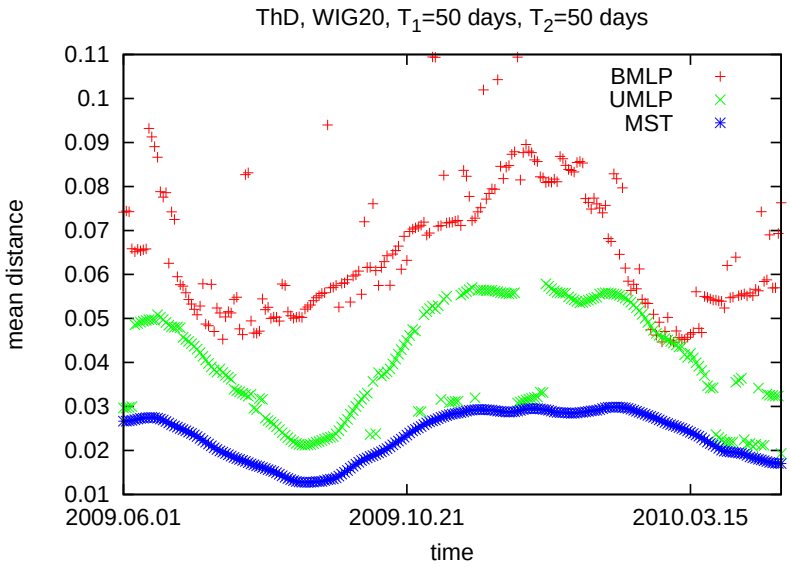
- WIG20: PEKAO, PKO BP, KGHM, PKN ORLEN, TPSA, BZ WBK, ASSECO POLAND, CEZ, GETIN HOLDING, GTC, TVN, PBG, POLIMEXMS, BRE, LOTOS, CYFROWY POLSAT, BIOTON.
- Wartości odpowiadają notowaniom zamkniecia w czasie od 05.01.2009 do 30.04.2010.

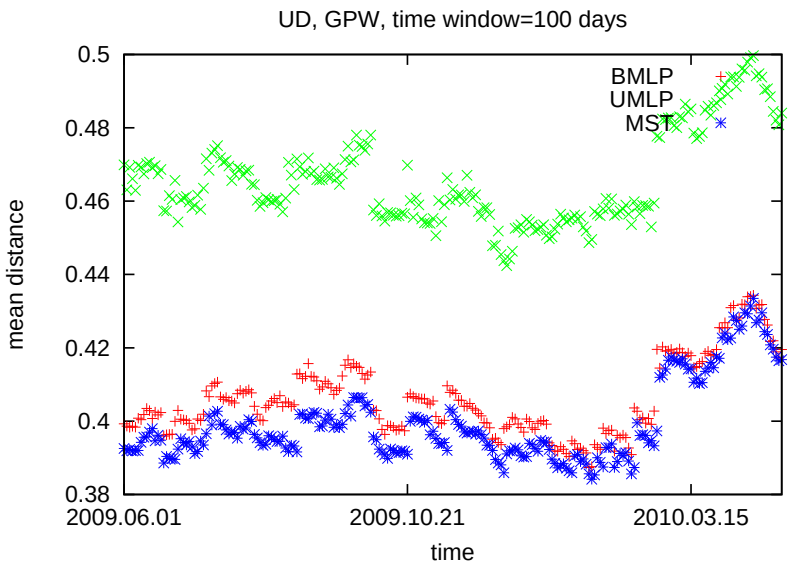
Ewolucja giełdy

- S&P 500: ABB Ltd.(ABB), Apple Inc. (AAPL), Boeing Co. (BA), the Coca-Cola Company (KO), Emerson Electric Co. (EMR), General Electric Co. (GE), Hewlett-Packard Company (HPQ), Hitachi Ltd. (HIT), IBM (IBM), Intel Corporation (INTC), Johnson & Johnson (JNJ), Lockheed Martin Corporation (LMT), Microsoft Co. (MSFT), Northrop Grumman Corporation (NOC), Novartis AG (NVS), Colgate-Palmolive Co. (CL), Pepsico Inc. (PEP), Procter & Gamble Co. (PG), Tower Semiconductor LTD. (TSEM), Wisconsin Energy Corporation Co. (WEC).
- Wartości odpowiadają notowaniom zamkniecia w czasie od 02.01.2009 to 30.04.2010.

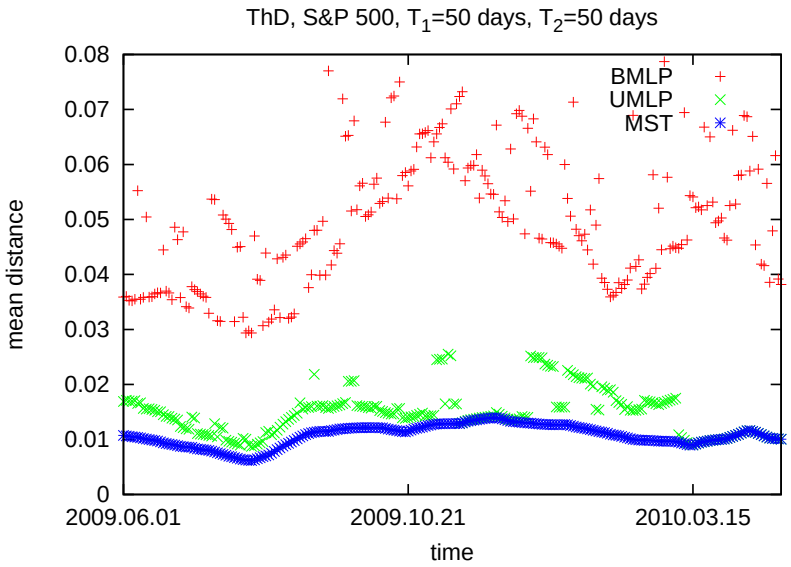

NVS
NOC
LMT
HIT
BA
GE
AAPL
HPQ
EMR
ABB
IBM
CL
PG
INTC
TSEM
PEP
KO
JNJ
MSFT
WEC

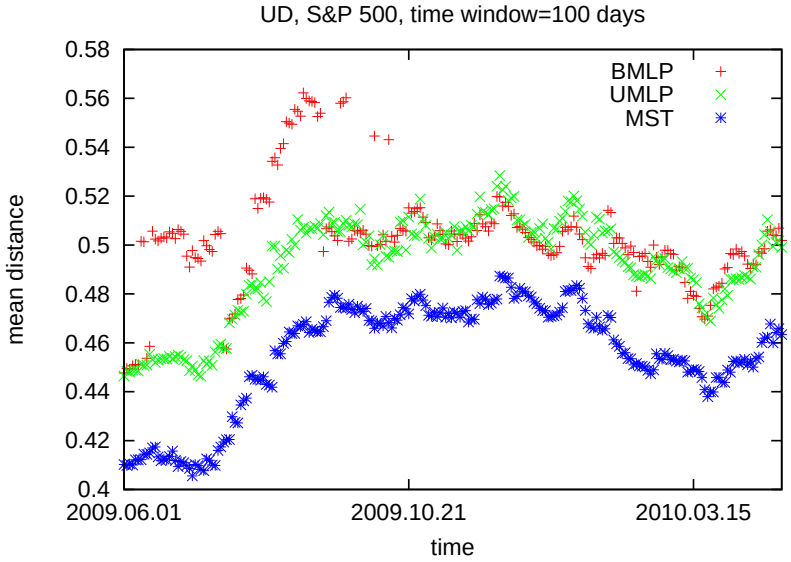
GPW


ThD, WIG20, T1=50 days, T2=50 days
BMLP
UMLP
MST
mean distance
time
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
2009.06.01
2009.10.21
2010.03.15


UD, GPW, time window=100 days
BMLP
UMLP
MST
mean distance
time
0.5
0.48
0.46
0.44
0.42
0.4
0.38
2009.06.01
2009.10.21
2010.03.15

S&P 500


ThD, S&P 500, T1=50 days, T2=50 days
BMLP
UMLP
MST
mean distance
time
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
2009.06.01
2009.10.21
2010.03.15


UD, S&P 500, time window=100 days
BMLP
UMLP
MST
mean distance
time
0.58
0.56
0.54
0.52
0.5
0.48
0.46
0.44
0.42
0.4
2009.06.01
2009.10.21
2010.03.15

Wnioski

- Odległość ultrametryczna upraszcza wybór podmiotów przy konstrukcji portfela.
- Odległość ultrametryczna bada czy istnieją korelacje liniowe, jest jednak wrażliwa na szum.
- Odległość Manchattan umożliwia kategoryzację korelacji jest też bardziej odporna na szum.
- Odległości oparte na entropii pozwalają zaobserwować obecność czynników zewnętrznych.