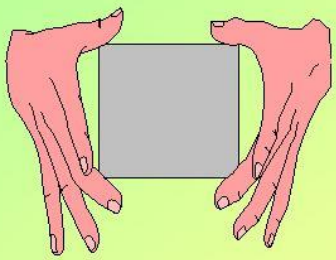
Klasyczna Metoda Najmniejszych Kwadratów



METODY APROKSYMACJI

MATEUSZ WAGA

Gimnazjum im. Jana Matejki w Zabierzowie

Mateusz Waga
Gimnazjum im. Jana Matejki, Zabierzów

SPIS TREŚCI

Mateusz Waga
Gimnazjum im. Jana Matejki, Zabierzów

1 WSTĘP

W naukach statystycznych na podstawie pomiarów otrzymujemy pary liczb, które jak przypuszczamy, są ze sobą powiązane jakąś zależnością funkcyjną $y = f(x)$ lub korelacją, np. średnia ocen w klasie w zależności od długości nauczania, wielkości produkcji w zależności od miesięcy itp. Sensownym posunięciem jest znalezienie takiej krzywej, która w możliwie najlepszy sposób przybliża punktu uzyskane w wyniku pomiarów. Znajdowanie takich krzywych jest celem teorii aproksymacji. Bazując na wyznaczonej krzywej, która określa trend rozwojowy danego zjawiska możemy prognozować przyszłe wartości (np. wielkość produkcji czy średnią ocen w klasie).

Podobne problemy spotykamy w wielu innych dziedzinach nauki. Ekonometria, która zajmuje się mierzeniem i przewidywaniem zjawisk zachodzących w gospodarce, w dużej mierze korzysta z teorii aproksymacji. W politologii wyznaczenie trendów poparcia dla poszczególnych partii politycznych ma kluczowe znaczenie przy ustalaniu strategii wyborczych. Także w naukach przyrodniczych wykonujemy często eksperymenty polegające na pomiarach par wielkości, które, jak przypuszczamy, są ze sobą powiązane jakąś zależnością funkcyjną , np. wydłużenie sprężyny w zależności od wiszącego na niej ciężaru. Metody aproksymacyjne sprawdzają się także w analizowaniu procesów demograficznych, w których prognozujemy ilość mieszkańców w danych krajach.

Przedmiotem bieżącego opracowania będzie metoda najmniejszych kwadratów, która stanowi szczególny przypadek teorii aproksymacji. Wyniki uzyskane z danych statystycznych będziemy aproksymować linią prostą. Jest to najprostsza metoda, która dobrze się sprawdza kiedy na podstawie danych można wyznaczyć linie trendu. W przypadku kiedy dane nie układają się wzdłuż linii prostej należy zastosować inne bardziej skomplikowane metody np. aproksymacje wielomianową.

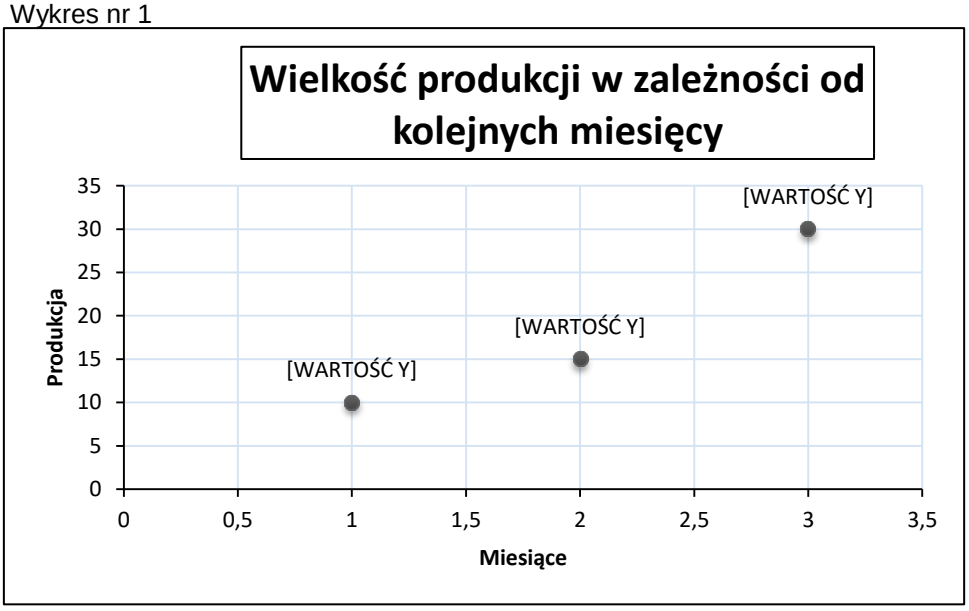
2 MODEL MATEMATYCZNY

Załóżmy, że mamy trzy pary danych pochodzących z pomiarów statystycznych przedstawionych poniżej w tabeli nr 1 (np. pomiary mogą reprezentować wielkość produkcji w kolejnych miesiącach). Chcielibyśmy ustalić prognozowaną wartość produkcji w czwartym miesiącu:

Tabela nr 1

Numer miesiąca	Wielkość produkcji
1	10
2	15
3	30

Wykres nr 1



Widać, że chociaż punkty są nieco porozrzucane na skutek, różnych czynników które miały wpływ na wielkość produkcji w kolejnych miesiącach, to jednak można wyróżnić trend wzrostowy.

Równanie prostej, która będzie wyznaczać trend ma następującą postać:
$y = ax + b$,

gdzie
a, b – to współczynniki, których chwilowo nie znamy,
y – wielkość produkcji,
x – numer miesiąca.

Poszukiwanie parametrów takiej prostej, która by przechodziła możliwie najbliżej wszystkich punktów (x_i, y_i) (gdzie i oznacza numer kolejnych pomiarów) polega na minimalizacji kwadratu sumy:

Wzór nr 1

$$S(a,b) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

gdzie
y_i – dane statystyczne, w naszym przypadku rzeczywista wielkości produkcji ,
$f(x_i)$ – wartości funkcji szukanej $y = ax + b$, w naszym przypadku prognozowana wartość produkcji,
i – kolejne numery danych statystycznych,
x_i- dane statystyczne, w naszym przypadku numery miesięcy,
n – ilość pomiarów statystycznych, w naszym przypadku ilość miesięcy.

ponieważ $f(x_i) = ax_i + b$, wzór nr 1 przyjmuje następującą postać:

Wzór nr 2

$$S(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Różnice między dokładnymi wartościami y_i oraz wartościami obliczonymi z równania prostej są podnoszone do kwadratu, aby uniknąć sytuacji, że będą się nawzajem znosiły na skutek różnicy znaków. Z tego też względu przedstawiona metoda postępowania nosi nazwę metody najmniejszych kwadratów.

Dla naszego przypadku wzór nr. 2 przyjmuje postać ($n = 3$).

$$S(a,b) = \sum_{i=1}^{3} (y_i - (ax_i + b))^2$$

Dla danych z tabeli 1 wielkość funkcji $S(a,b)$ będzie równa:

$$S(a,b) = [10 - (1 * a + b)\ ]^2 + [15 - (2a + b)]^2 + [30 - (3a + b)]^2$$

Formalnie rzecz biorąc jest to funkcja dwóch zmiennych $a\ i\ b$. Interesują nas takie wartości tych zmiennych, dla których $S(a,b)$ jest minimalna. Wiadomo, że funkcja wielu zmiennych ma minimum w punkcie, dla którego pochodne cząstkowe tej funkcji po wszystkich zmiennych są równe zeru. Zatem w tym przypadku muszą być spełnione następujące warunki:

Wzór nr 3

$$\frac{\partial S(a,b)}{\partial a} = 0$$

$$\frac{\partial S(a,b)}{\partial b} = 0$$

Czyli w naszym przypadku wzór nr 3 przyjmuje postać:

Mateusz Waga
Gimnazjum im. Jana Matejki, Zabierzów

$$\frac{\partial S(a,b)}{\partial a} = 2 * [10 - (1 * a + b)\ ] * (-1)^{\square} + 2 * [15 - (2a + b)] * (-2)^{\square} + 2 * [30 - (3a + b)] * (-3)^{\square}$$

$$\frac{\partial S(a,b)}{\partial b} = 2 * [10 - (1 * a + b)\ ] * (-1)^{\square} + 2 * [15 - (2a + b)] * (-1)^{\square} + 2 * [30 - (3a + b)] * (-1)^{\square}$$

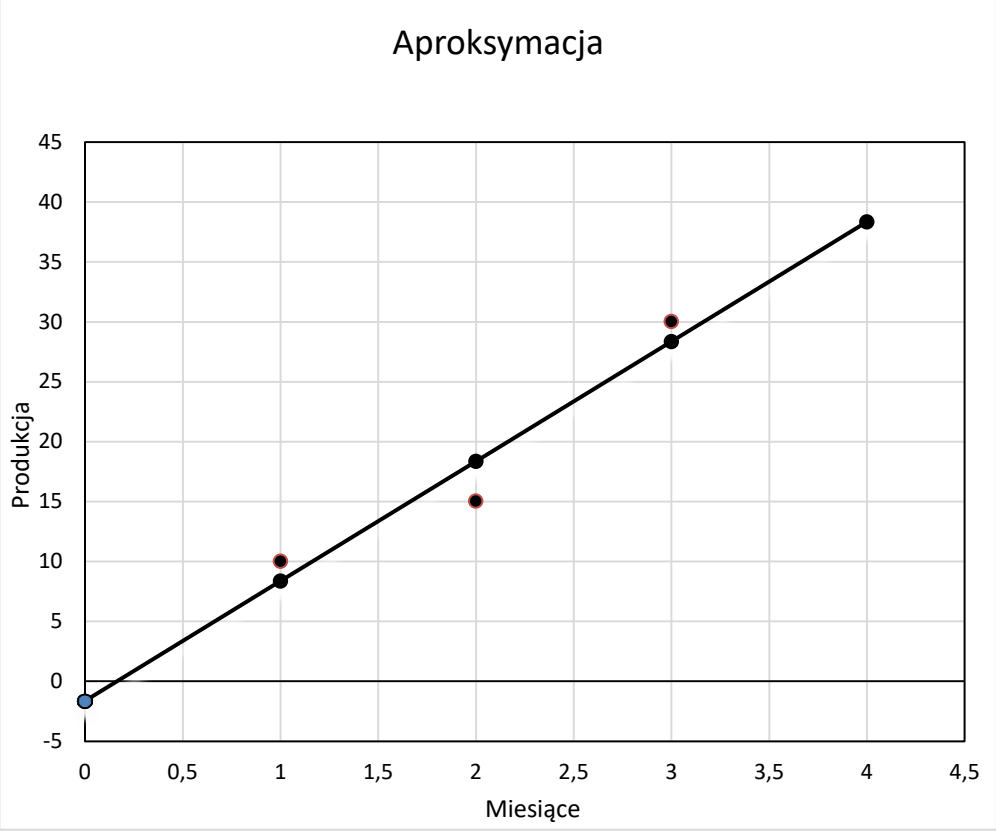
Po uproszczeniach dostajemy następujący układ równań:

$$\begin{cases} 28a + 12b = 260 \\ 12a + 6b = 110 \end{cases}$$

Którego rozwiązaniem jest $a = 10, b = -\frac{5}{3}$;
Zatem otrzymujemy wzór naszej funkcji:
$f(xi) = 10x_i - 5/3$

Na wykresie nr 2 , zostały pokazany wykres funkcji $f(x_i)$ i punkty otrzymane z pomiarów statystycznych.

Wykres nr 2



W powyższej analizy wynika, że prognozowany wzrost produkcji w czwartym miesiącu będzie wynosił około 38 produktów.

3 UOGÓLNIENIE MODELU MATEMATYCZNEG

Nauczyliśmy się już na podstawie prostej aproksymacji (opartej na trzech punktach) wyznaczać funkcję trendu na podstawie której obliczyliśmy prognozowaną wartość produkcji w czwartym miesiącu.

Powyższe zagadnienie możemy uogólnić dla większej liczby pomiarów. Można założyć, że dane statystyczne lub dane pomiarowe, które są przedmiotem aproksymacji są z reguły obarczone błędami losowymi (np. załamanie pogody mogło mieć decydujący wpływ na obniżenie produkcji w miesiącu nr 2). Przy pomocy metody najmniejszych kwadratów szeregi statystyczne oczyszcza się z błędów losowych.

Mając szereg punktów empirycznych $(x_1, y_1), (x_2, y_2) \ldots\ldots\ldots (x_n, y_n)$ należy a priori ustalić postać funkcji $y_t = f(x, a, b)$, a następnie na podstawie punktów empirycznych tak dobrać wartości parametrów $a\ i\ b$, aby funkcja $y = f(x, a, b)$ możliwie najlepiej "pasowała" do zaobserwowanych punktów (x_i, y_i).

Rozważania oparte na rachunku prawdopodobieństwa pozwalają uznać za najlepsze takie wartości parametrów $a\ i\ b$, dla których suma kwadratów odchyleń zaobserwowanych wartości y_i od wartości teoretycznych $y_{ti} = f(x_i, a, b)$ jest możliwie najmniejsza, tzn.:

$$\sum_{i=1}^{n} [y_i - f(x_i, a, b)]^2 = minimum$$

Podstawowym warunkiem przyjęcia przez powyższe wyrażenie wartości minimum jest to, aby pierwsze pochodne (pochodna) względem a, b, były równe zeru. Korzystając z tych warunków wyznacza się wartości a, b zależnie od punktów zaobserwowanych (x_i, y_i). W szczególnym przypadku funkcji liniowej, tzn. gdy $y = ax + b$, otrzymuje się:

Wzory 4

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

n – liczba pomiarów

4 MODEL INFORMATYCZNY

Uogólniony model matematyczny można w bardzo prosty sposób przełożyć na język informatyczny. Poniższy kod programu napisanego w języku C++ służy do obliczenia parametrów funkcji liniowej $a\ i\ b$ oraz obliczenia prognozy dla dowolnej wartości x. Program z klawiatury wczytuje ilość pomiarów oraz pary danych będących przedmiotem aproksymacji.

```
#include <iostream>
#include <math.h>

using namespace std;
float suma1(float *x,float *y,float n)
{
float suma=0;
for(int i=1;i<=n;i++)
{
suma+=*x**y;
x++;
y++;
}
return suma;
}
float suma2(float *y,float n)
{
float suma=0;
for(int i=1;i<=n;i++)
{
suma+=*y;
y++;
}
return suma;
}
float suma3(float *x,float n)
{
float suma=0;
for(int i=1;i<=n;i++)
{
suma+=*x;
x++;
}
return suma;
}
float suma4(float *x,float n)
{
float suma=0;
for(int i=1;i<=n;i++)
{
suma+=pow(*x,2);
x++;
}
return suma;
}
int main() {
int c;
cout<<"podaj ilosc danych"<<endl;
cin>>c;
float x[c],y[c],a,b,d,p;
float s1,s2,s3,s4;
for(int i=0;i<c;i++)
{
cout<<"podaj "<<i+1<<" x:";
cin>>x[i];
cout<<"podaj "<<i+1<<" y:";
cin>>y[i];
}
d=c;
float *wsk_x=&x[0];
float *wsk_y=&y[0];
s1=suma1(wsk_x,wsk_y,d);
s2=suma2(wsk_y,d);
s3=suma3(wsk_x,d);
s4=suma4(wsk_x,d);
a=(c*s1-s2*s3)/(c*s4-pow(s3,2));
b=(s4*s2-s3*s1)/(c*s4-pow(s3,2));
cout<<"a jest rowne:"<<a<<endl;
cout<<"b jest rowne:"<<b<<endl;
cout<<"Podaj wartosc x dla jakiej ma byc obliczna probnoza:"<<endl;
cin>>p;
cout<<"probnoza jest rowna:"<<a*p+b<<endl;
return 0;
}
```

Poniższe wzory zostały zapisane w kodzie programu

Wzory matematyczne:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Wzory z programu:

a=(c*s1-s2*s3)/(c*s4-pow(s3,2))

b=(s4*s2-s3*s1)/(c*s4-pow(s3,2))

gdzie:
$c - liczba\ pomiarów,$
$s1, s2, s3, s4 - kolejne\ sumy,$
$pow(s3,2) - potęga\ kwadratowa\ z\ sumy\ s3$

5 PRZYKŁADY ZASTOSOWAŃ

5.1 Prognozowanie ilości telefonów komórkowych

Z Internetu odczytałem ilość aktywnych kart SIM w następujących latach:

n – liczba pomiarów	data	Ilość telefonów [tys.]
1	Grudzień -2015	56 253,00
2	Grudzień -2014	57 595,00
3	Grudzień - 2013	55 979,00
4	Grudzień - 2012	54 267,00
5	Grudzień - 2011	50 695,00

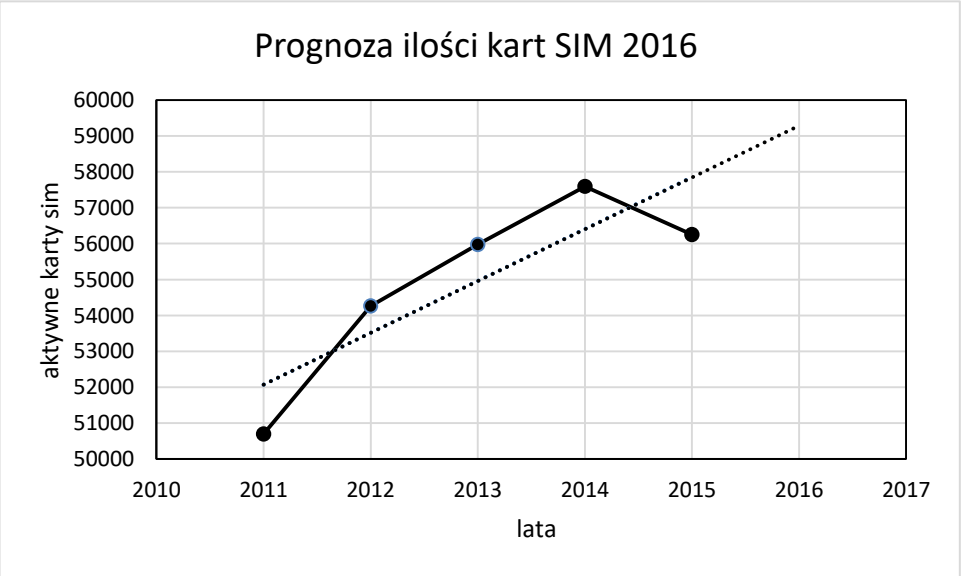
Bazując na metodzie najmniejszych kwadratów, wyznaczyłem parametry funkcji liniowej $a\ i\ b$, która podaje zależność pomiędzy ilością aktywnych kart SIM a kolejnymi latami.

Korzystając ze wzoru nr 4 obliczamy parametry funkcji liniowej dla $n = 5$.

$a = 1\,444{,}4; b = (-2\,832\,619)$ (obliczenia wykonałem z wykorzystaniem programu Excel oraz programu napisanego w języku C++ przedstawionym w rozdziale 4)

$f(x_i) = 1\,444{,}4 * x_i - 2\,832\,619$

i – kolejne pomiary.



Bazując na funkcji trendu można stwierdzić, że w grudniu 2016 prawdopodobna ilość aktywnych kart SIM będzie wynosił 59 291 (w tys.). W tym konkretnym przypadku na podstawie metody najmniejszych kwadratów można obliczyć wiarygodne wartości prognoz.

5.2 Prognozowanie ilość ludzi na świecie

Z Internetu odczytałem zmianę ilości ludzi zamieszkujących powierzchnie ziemi na przestrzeni wieków.

n – liczba pomiarów	lata	Ilość ludzi w [mln]
1	1500	450
2	1700	600
3	1800	978
4	1900	1650
5	2000	6118
6	2012	7022

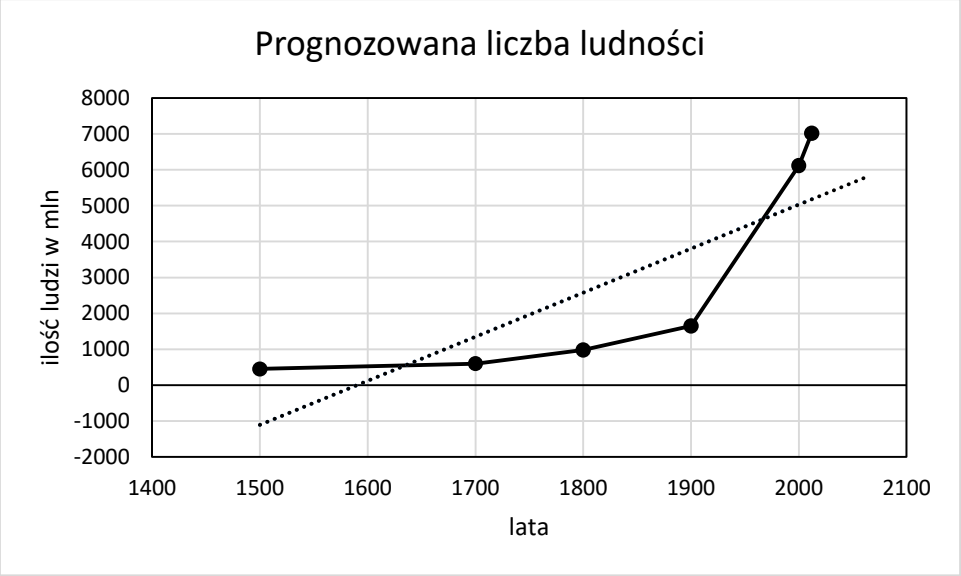
Na podstawie obliczeń (własnym programem, oraz w programie Excel) obliczyłem parametry funkcji liniowej $a\ i\ b$ (która aproksymuje wielkość ludzi na świecie) oraz obliczyłem prognozę ludności w roku 2050.

Otrzymałem następujące wyniki

$a = 12{,}27\ i\ b = (-19\ 514{,}50)$

Prognoza ilości ludzi na świecie w 2050 obliczona na podstawie powyższego modelu wynosi 5 639,92 (w milionach).

Wyniki zaprezentowano na poniższym wykresie.



Na podstawie analizy wykresu można uznać, że otrzymane wyniki prognozy opartej na metodzie najmniejszych kwadratów są błędne. W roku 2050 liczba ludności przekroczy 7 mld (aktualną liczbę ludności). Przed rokiem 1500 liczba ludności była większa od zera. Ze względu na charakter danych metoda najmniejszych kwadratów się nie sprawdza (brak zależności liniowej) i trzeba zastosować inna metodę aproksymacji co będzie przedmiotem dalszych analiz.

6 ZAKOŃCZENIE

Bardzo lubię poszerzać swoją wiedzę z matematyki. Szczególnie interesują mnie zagadnienia które można zastosować w praktyce i które nadają się do przełożenia na język programowania.

Aby zrealizować niemniejszy projekt musiałem zapoznać się z pojęciem funkcji liniowej, pochodnych cząstkowych liczonych względem dwóch zmiennych oraz poznać jedną z metod aproksymacji (metodę najmniejszych kwadratów).

Szczególną satysfakcje sprawił mi fakt że mój program (napisany w języku C++) poprawnie liczy parametry funkcji liniowej i wyniki są porównywalne z programem Excel.

Największy problem sprawiło mi zrozumienie interpretacji pochodnych (czyli wyznaczanie min lub max funkcji).

Po raz pierwszy napisałem pracę pisemną, która odnosiła się do problemów matematycznych. Mam pełną świadomość że aproksymacja w oparciu o metodę najmniejszych kwadratów nie sprawdza się w odniesieniu do wszystkich zagadnień/danych i w przyszłości będę jeszcze pracował nad innymi metodami aproksymacji (np. metodą wielomianową, wykładniczą).

7 BIGLIOGRAFIA

http://gsmonline.pl/artykuly/penetracja-telefonii-komorkowej-w-polsce-i-kw-2016 – dane o ilości telefonów komórkowych.

https://pl.wikipedia.org/wiki/Ludno%C5%9B%C4%87_%C5%9Bwiata – dane ilość ludzi na świecie

Mateusz Waga
Gimnazjum im. Jana Matejki, Zabierzów