

**NeuroMet2010**

# Wpływ przygotowania danych – czyszczenia danych na jakość modelowania

dr inż. Andrzej Kochański  
dr inż. Jacek Kozłowski – prelegent

POLITECHNIKA WARSZAWSKA

# Uzupełnianie brakujących danych

---

Uzupełnianie proste

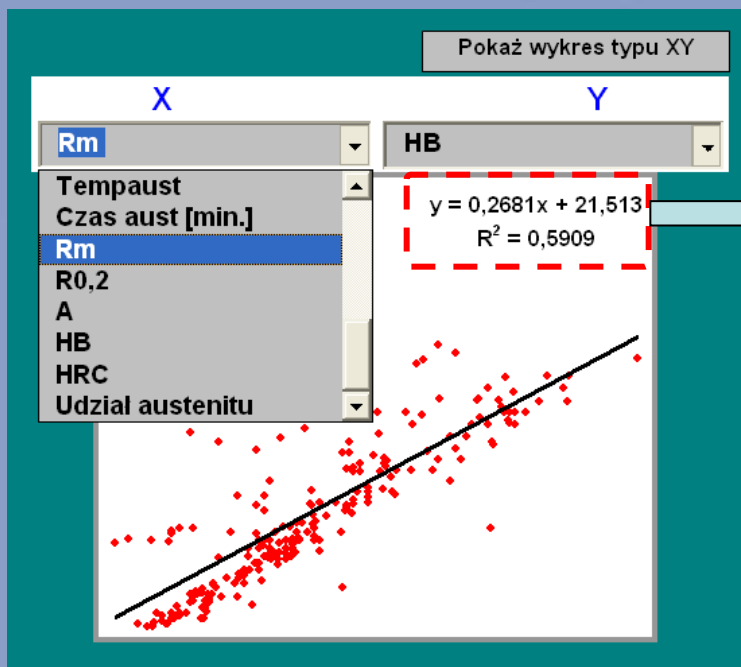
- ✓ Metody statystyczne (**średnia**, mediana...)

Uzupełnianie złożone

- ✓ Jednowymiarowa regresja liniowa



# Jednowymiarowa regresja liniowa



✓ Wyznaczenie zależności funkcyjnej oraz współczynnika DETERMINACJI dla wszystkich par zmiennych

✓  $Y = a \cdot X + b$   
Uszeregowanie według malejącej wartości współczynnika

✓ Ręczne lub automatyczne uzupełnienie brakujących danych

Wykonaj obliczenia

| X       | Y       | R2    | ERRśr | % danych |
|---------|---------|-------|-------|----------|
| HB      | HRC     | 0,932 | 4%    | 9%       |
| HRC     | HB      | 0,932 | 4%    | 9%       |
| Rm      | R0,2    | 0,779 | 6%    | 57%      |
| R0,2    | Rm      | 0,779 | 6%    | 57%      |
| Temp pi | HB      | 0,646 | 9%    | 31%      |
| HB      | Temp pi | 0,646 | 8%    | 31%      |
| R0,2    | HB      | 0,591 | 11%   | 24%      |
| HB      | R0,2    | 0,591 | 8%    | 24%      |
| Rm      | HB      | 0,590 | 9%    | 26%      |
| HB      | Rm      | 0,590 | 8%    | 26%      |
| R0,2    | HRC     | 0,564 | 9%    | 15%      |

Podaj próg współczynnika R2 do wypełnienia

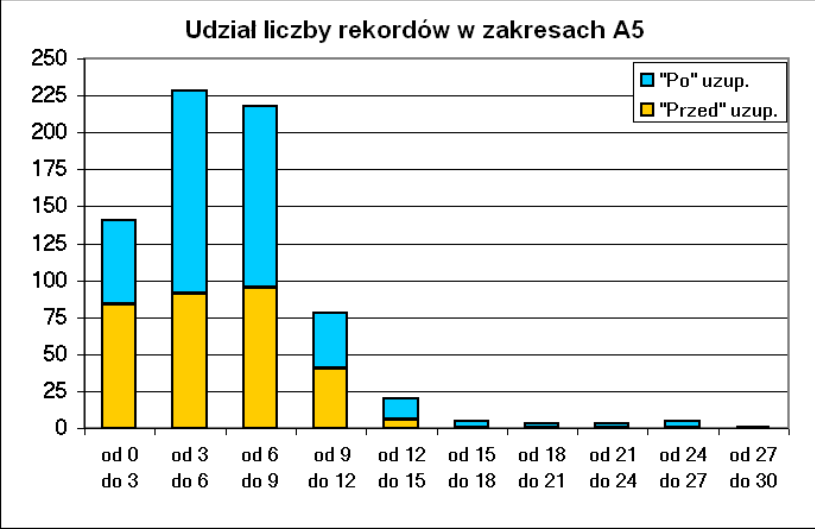
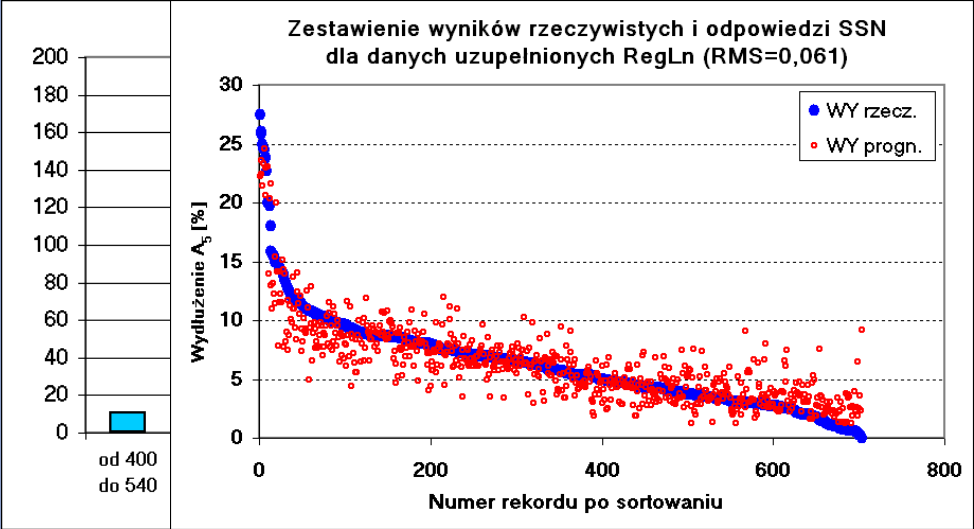
0,41

Uzupełnij braki "Y" (wybrane)

Uzupełnij braki "Y" (próg)

# Udział braków

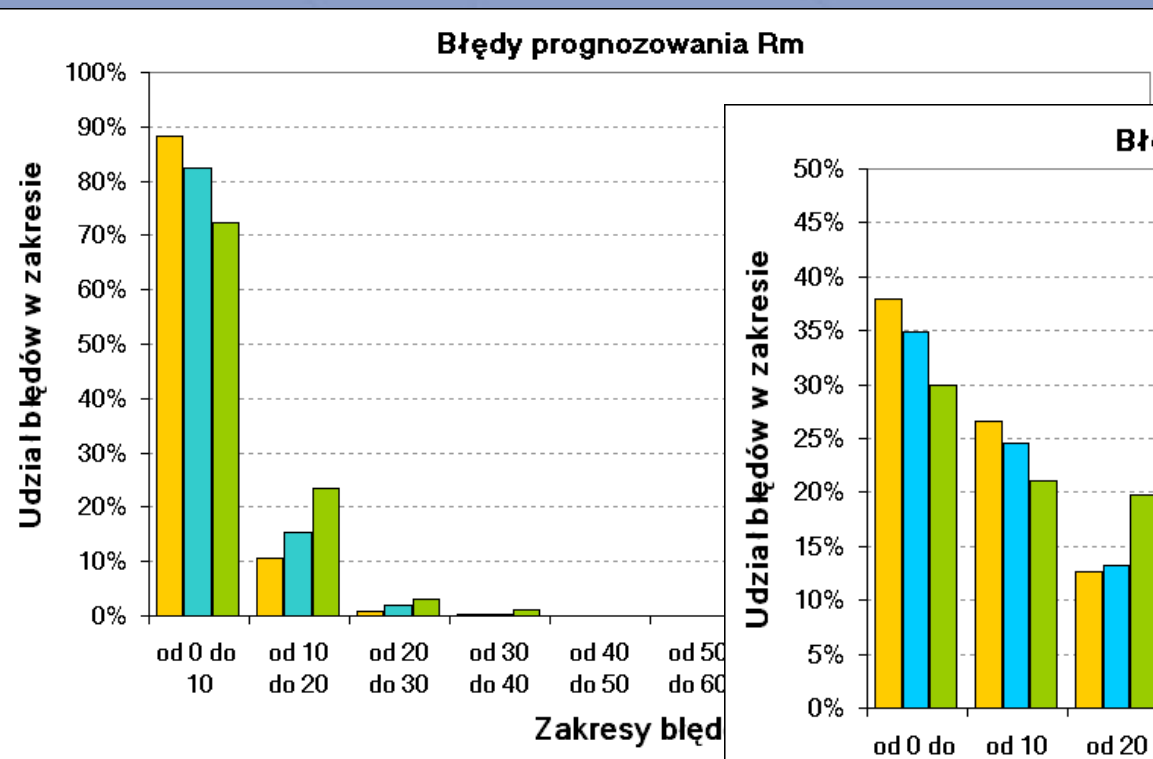
|                | C                 | Si  | S   | P   | Mg  | Mn  | Ni   | Cu   | Mo   | Temp<br>pi<br>[°C] | Czas<br>pi<br>[min.] | Temp<br>aust.<br>[°C] | Czas<br>aust.<br>[min.] | Rm<br>[MPa] | A <sub>5</sub><br>[%] | Udział<br>rekordów<br>pełnych<br>[%] |
|----------------|-------------------|-----|-----|-----|-----|-----|------|------|------|--------------------|----------------------|-----------------------|-------------------------|-------------|-----------------------|--------------------------------------|
|                | Udział braków [%] |     |     |     |     |     |      |      |      |                    |                      |                       |                         |             |                       |                                      |
| Przed uzupełn. | 0,5               | 0,5 | 4,2 | 4,2 | 4,4 | 2,2 | 12,0 | 12,8 | 28,1 | 2,4                | 2,5                  | 1,2                   | 3,5                     | 26,9        | 30,6                  | 34,8                                 |
| Po uzupełn.    | 0,5               | 0,5 | 4,2 | 4,2 | 4,4 | 2,2 | 12,0 | 12,8 | 4,3  | 1,3                | 2,5                  | 0,8                   | 3,5                     | 1,3         | 0,8                   | 76,2                                 |



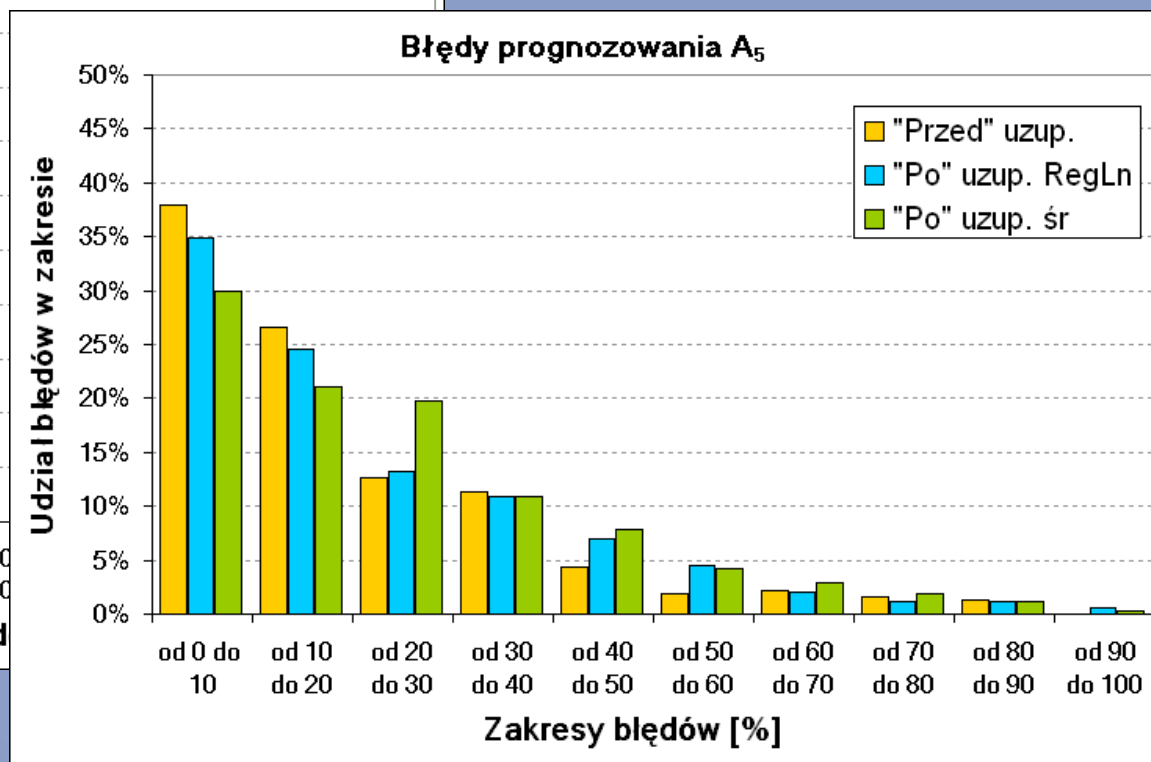
Przed uzupełn. A<sub>5</sub> ∈ (0 15%) ➡ Po uzupełn. A<sub>5</sub> ∈ (0 27%)

# Udział błędów prognozowania

**Rm [MPa]**

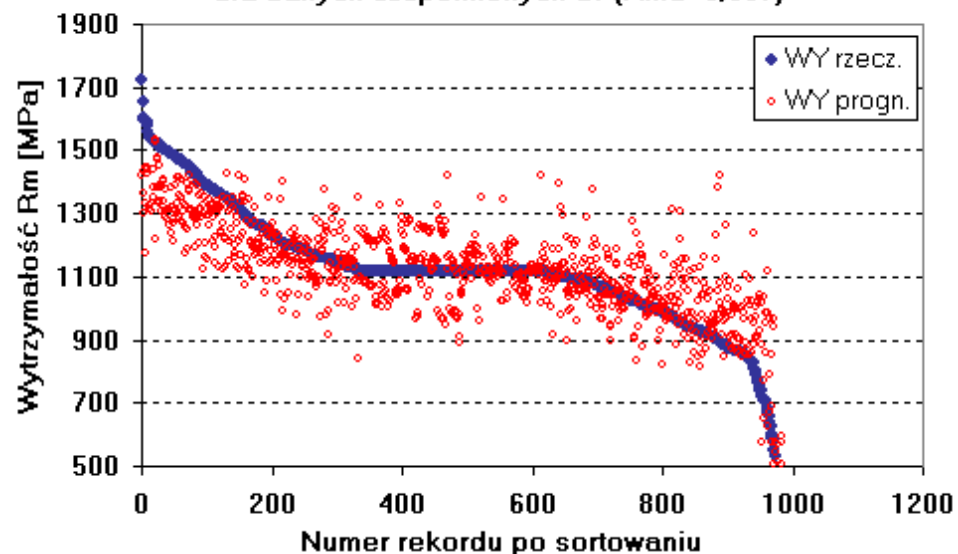


**A5 [%]**



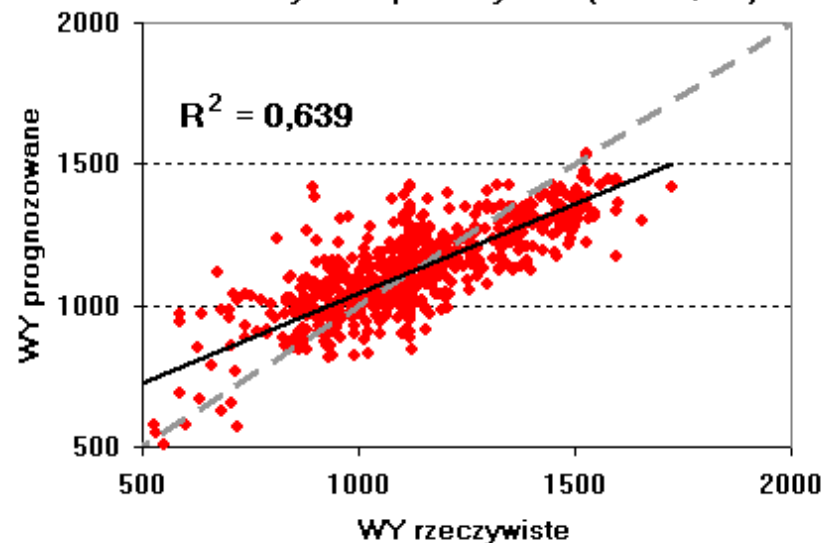
# Zestawienie wyników prognozowania

Zestawienie wyników rzeczywistych i odpowiedzi SSN dla danych uzupełnionych Sr (RMS=0,087)



**Rm [MPa]**

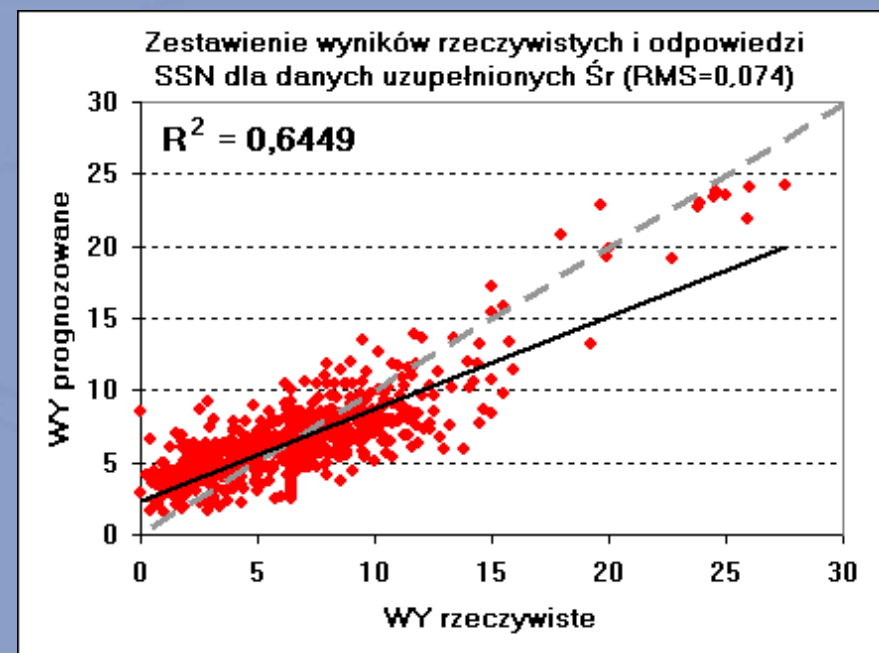
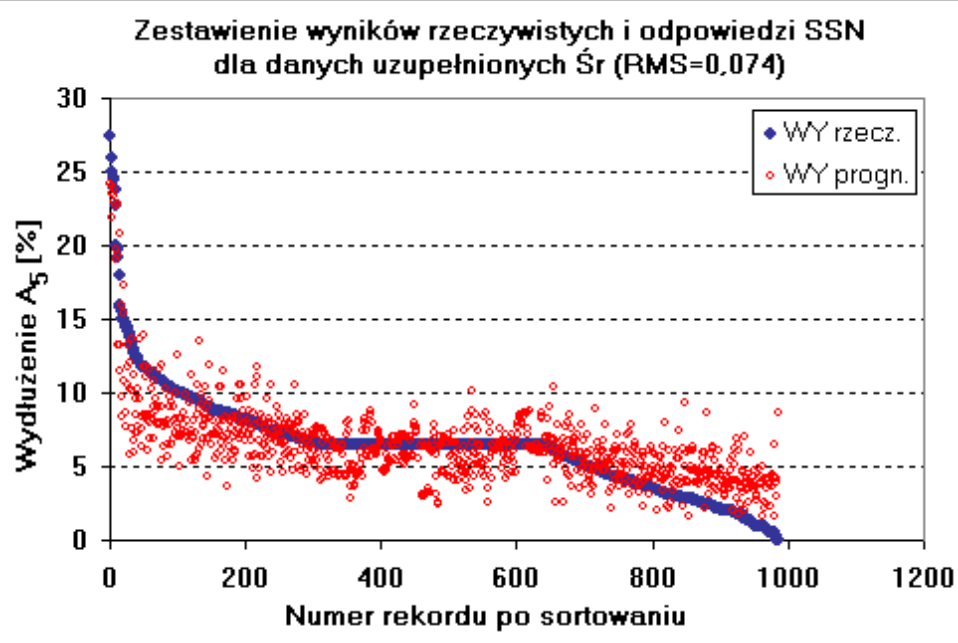
Zestawienie wyników rzeczywistych i odpowiedzi SSN dla danych uzupełnionych Sr (RMS=0,087)



| Sposób uzupełnienia | RMS  | R <sup>2</sup> |
|---------------------|------|----------------|
| Nie uzup.           | 0.06 | 0,87           |
| RegLn               | 0.07 | 0,80           |
| Średnia             | 0.09 | 0,64           |

# Zestawienie wyników prognozowania

**A5 [%]**



| Sposób uzupełnienia | RMS   | $R^2$ |
|---------------------|-------|-------|
| Nie uzup.           | 0.083 | 0,75  |
| RegLn               | 0.061 | 0,79  |
| Średnia             | 0.074 | 0,64  |

- Modele oparte na danych uzupełnionych metodami złożonymi dają lepsze wyniki prognozowania niż te oparte na danych uzupełnionych w sposób prosty.
- Prezentowane metody uzupełniania pozwalają w sposób pośredni i bezpośredni rozszerzyć zakres zmienności parametrów, jak i zwiększyć liczebność (wiarygodność) zbioru
- Miary oceny jakości modeli opartych na danych uzupełnionych są zadowalające i zbliżone do przypadków nie uzupełnionych.
- Metody złożone są użyteczne tylko dla tych przypadków, które wykazują wzajemne korelacje
- Istnieje potrzeba opracowania innych złożonych metod uzupełniania (np: opartych na wielowymiarowej regresji liniowej i nieliniowej), jak i prostych (np: k-najbliższych sąsiadów)

Dziękuję  
za uwagę

