Metody redukcji obciążenia estymatorów w badaniach wykorzystujących big data

Na przykładzie badania ofert pracy

Maciej Beręsewicz, Marcin Szymkowiak
Katedra Statystyki, Uniwersytet Ekonomiczny w Poznaniu
Ośrodek Statystyki Małych Obszarów, Urząd Statystyczny w Poznaniu

II Kongres Statystyki Polskiej, 10-12 lipca, Warszawa

Plan prezentacji

1. Wprowadzenie
   - Badanie ofert pracy w Internecie
   - Big data w statystyce publicznej
   - Mechanizmy powstawania braków danych
   - Obciążenie spowodowane błędem autoselekcji
2. W jaki sposób zredukować obciążenie?
   - Podejście quasi-randomizacyjne
   - Podejście predykcyjne/modelowe
3. Przykłady z badania ofert pracy
   - Badanie Popytu na Pracę
   - Badanie Kapitału Ludzkiego – 2010-2015
4. Literatura

Wprowadzenie – kontekst rynku pracy

- Coraz większy odsetek odmów udziału w badaniach reprezentacyjnych; próby niwelacji obciążenia respondentów.
- Wzrost zapotrzebowania na bieżącą informację na niskich poziomach agregacji.
- Zmiana paradygmatu w statystyce publicznej – wykorzystanie istniejących źródeł danych (np. rejestry administracyjne).
- Nowe, alternatywne źródła danych – Internet, big data..., które można analizować w sposób automatyczny.
- Potrzeba analizy podaży i popytu na rynku pracy (np. Barometr zawodów, zawody deficytowe, Prognozowanie zatrudnienia).
- Nowe źródła wymagają rzetelnej oceny oraz przedstawienia możliwości ich wykorzystania w statystyce, w szczególności statystyce publicznej.

Wprowadzenie – miejsca poszukiwania pracy

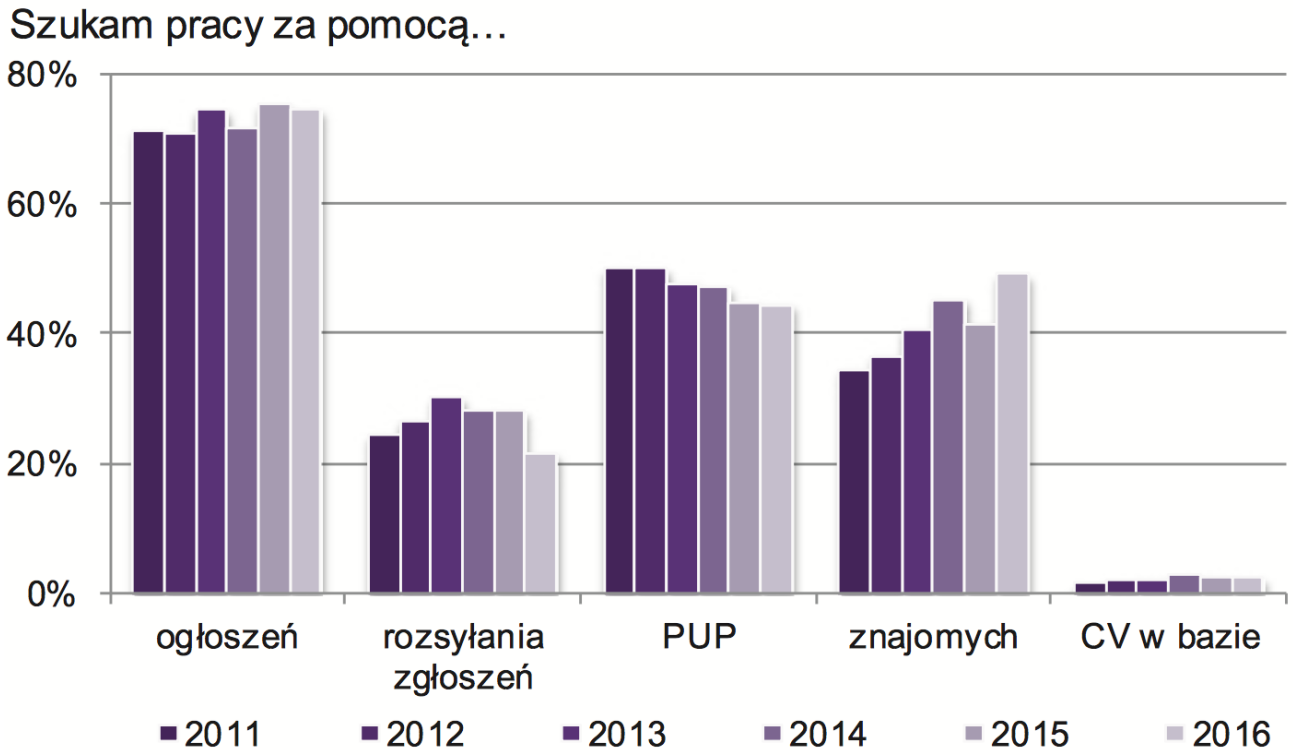
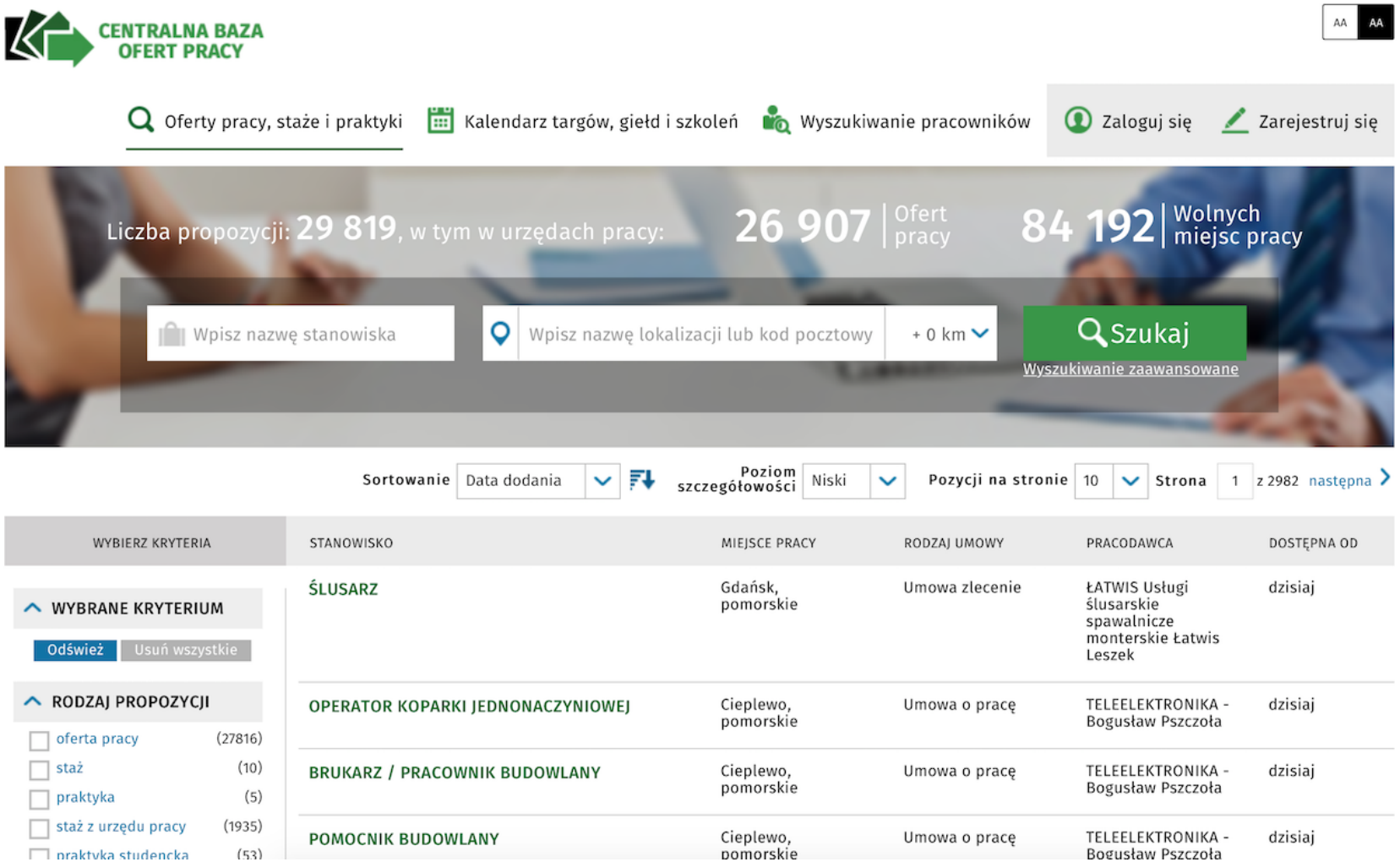


Figure 1: Sposób poszukiwania pracy przez osoby bezrobotne

Źródło: Badanie bezrobotnych realizowane przez Narodowy Bank Polski w powiatowych urzędach pracy.

Centralna Baza Ofert Pracy



Centralna Baza Ofert Pracy

Zakres informacyjny portalu CBOP jest następujący:

- dane pracodawcy,
- liczba wolnych miejsc pracy,
- nazwę zawodu i zakres obowiązków,
- miejsce pracy i zakres godzinowy,
- wynagrodzenie,
- rodzaj umowy,
- wymagania dotyczące danej oferty,
- terytorium, na którym ważna jest oferta: Polska, UE itd.

Po wypełnieniu formularza zostaje on wysłany do urzędu. Po weryfikacji zgodności oferty przez pracownika urzędu może ona pojawić się w systemie PUP. Urzędnik zajmujący się daną ofertą pracy powinien dzwonić co kilka dni do pracodawcy celu potwierdzenia aktualności wakatu.

Wprowadzenie – źródła danych o popycie na pracę

- Źródła statystyczne (ze statystyki publicznej)
  - Sprawozdawczość przedsiębiorstw – Z-05 (Badanie popytu na pracę) `http://form.stat.gov.pl/formularze/2018/passive/Z-05.pdf`.
- Źródła niestatystyczne (spoza statystyki publicznej)
  - Narodowy Bank Polski – Badanie ankietowe rynku pracy: (1) badanie Badanie osób bezrobotnych (próba losowa); (2) Badanie przedsiębiorstw (próba celowa).
  - Badanie Kapitału Ludzkiego (PARP) – m.in. badanie pracodawców i ofert pracy; badanie w latach 2010-2014, kontynuacja od 2016 roku.
  - Oferty pracy w administracji publicznej i jednostkach publicznych (np. BIP, Służba Cywilna)
  - Oferty pracy w Urzędach Pracy (m.in. Centralna Baza Ofert Pracy),
  - Oferty pracy publikowane w Internecie (m.in. LinkedIn, OLX, pracuj.pl).

Wprowadzenie

Big data – próba nielosowa będąca wynikiem autoselekcji jednostek badanej populacji.

Big data dla statystyki publicznej to:

- wtórne, niezbadane źródło danych (...big noise?),
- potencjalne źródło wspierające istniejące źródła i badania,
- potencjalne źródło rozszerzające istniejące statystyki.

Wprowadzenie – wybrana (ostatnia) literatura

- Beręsewicz (2017). *A Two-Step Procedure to Measure Representativeness of Internet Data Sources*. International Statistical Review, 85(3), 473-493.
- Blumenstock (2016). *Fighting Poverty with Data*, Science, 353(6301), 753-754.
- Carroll, Murphy, Hanley, Dempsey & Dunne (2018). *Household Classification Using Smart Meter Data*. Journal of Official Statistics, 34(1), 1–25.
- Chen, Valliant, & Elliott (2018). *Model-assisted calibration of non-probability sample survey data using adaptive LASSO*. Survey Methodology, 44(1).
- Elliott & Valliant (2017). *Inference for Nonprobability Samples*. Statistical Science, 32(2), 249–264.
- Schmid, Bruckschen, Salvati, & Zbiranski (2017). *Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal*. Journal of the Royal Statistical Society: Series A, 180(4), 1163-1190.
- Sikov (2018). *A brief reivew of approaches to non-ignorable non-response*. International Statistical Review (dostępna online).

Mechanizmy powstawania braków danych

Rubin (1976) zaproponował następującą klasyfikację mechanizmów powstawania braków danych (ang. missing data mechanisms)

- Missing Completely at Random (MCAR)

$$Pr(R_i = 1|\boldsymbol{x}, \boldsymbol{y}) = const., \tag{1}$$

- Missing at Random (MAR)

$$Pr(R_i = 1|\boldsymbol{x}_i, \boldsymbol{y}_i) = Pr(R_i = 1|\boldsymbol{x}_i), \tag{2}$$

- Not Missing at Random (NMAR)

$$Pr(R_i = 1|\boldsymbol{x}_i, \boldsymbol{y}_i) \neq Pr(R_i = 1|\boldsymbol{x}_i), \tag{3}$$

gdzie $R = \{0, 1\}$ zmienna informująca czy dana jednostka udzieliła odpowiedzi, $Pr()$ to prawdopodobieństwo odpowiedzi, $\boldsymbol{x}_i$ to wektor zmiennych pomocniczych, a $\boldsymbol{y}_i$ wektor zmiennych celu.

Obciążenie spowodowane błędem autoselekcji

Załóżmy, że jesteśmy zainteresowani wartością średnią $\bar{Y} = N^{-1} \sum_{i=1}^{N} Y_k$ pewnej cechy Y.

Jeżeli w tym celu wykorzystamy wyłącznie jednostki obserwowane w źródle *big data* to wartość przeciętna dana jest równa $\bar{Y}_I = N_I^{-1} \sum_{i=1}^{N} I_i Y_i$, gdzie I oznacza inkluzję jednostki i do źródła big data.

W takim przypadku obciążenie estymatora $\bar{y}_s$ dane jest wzorem (4)

$$Bias(\bar{y}_s) = E(\bar{y}_s) - \bar{Y} \approx \bar{Y}_I^* - \bar{Y}_I = \frac{corr(\rho, Y) s(\rho) s(Y)}{\bar{\rho}}, \qquad (4)$$

gdzie $E(\bar{y}_s) \approx \bar{Y}_I^* = \frac{1}{N_i \bar{\rho}} \sum_{i=1}^{N} \rho_i I_i Y_i$, $\rho_i = Pr(R_i = 1 | \boldsymbol{x}_i)$ to skłonność do odpowiedzi, $\bar{\rho}$ to przeciętna skłonność do odpowiedzi, $corr(.)$ to współczynnik korelacji liniowej Pearsona, a $s(.)$ to odchylenie standardowe (por. Bethlehem, 2010).

Obciążenie, a zależności między x, Y i R

Table 1: Redukcja obciążenia, a zależności między x, Y i R

	Słaba zależność (x, Y)	Silna zależność (x, Y)
Słaba zależność (x, R)	Mały wpływ na redukcję obciążenia	Mały wpływ na redukcję obciążenia
Silna zależność (x, R)	Mały wpływ na redukcję obciążenia	**Redukcja obciążenia**

Źródło: Opracowanie własne na podstawie Zhang, Thomsen i Kleven (2013).

Wnioski:

- Potrzebne są bardzo dobre zmienne pomocnicze – silne zależności z R i y.
- Potrzebne są rozkłady na poziomie populacji (np. średnie, wartości globalne).

W takim razie, w jaki sposób można zredukować obciążenie związane z wykorzystaniem big data?

Wprowadzenie – publikacja w Eurostat Working Papers

W jaki sposób możemy zredukować obciążenie?

W literaturze przedmiotu możemy znaleźć dwa główne podejścia do redukcji obciążenia w próbach nielosowych:

- **Podejście quasi-randomizacyjne** (ang. quasi-randomization approach) – w którym dążymy do uzyskania wag mających na celu zredukować różnice między próbą, a populacją. Następnie te wagi wykorzystujemy do estymacji.
- **Podejście predykcyjne / modelowe** (ang. prediction / superpopulation model approach) – w którym zakładamy, że estymacja odbywa się na podstawie założonego modelu, który budujemy na próbie i aplikujemy dla jednostek spoza próby.

... a także ich kombinację.

Jednakże, podejścia te nie gwarantują 100% redukcji obciążenia.

Podejście quasi-randomizacyjne

W podejściu quasi-randomizacyjnym chcielibyśmy otrzymać wagi, które będą uwzględniały następujące elementy korygujące:

- błąd pokrycia (oznaczony przez c_i),
- błąd autoselekcji (oznaczony przez ρ_i),
- błąd braku reprezentatywności ze względu na znane rozkłady zmiennych pomocniczych (oznaczone przez ww_i).

Załóżmy, że jesteśmy zainteresowani wartością globalną zmiennej Y, którą w przypadku podejścia pseudo-randomizacyjnego możemy określić następująco:

$$\hat{Y} = \sum_i w_i \rho_i^{-1} c_i^{-1} Y_i. \tag{5}$$

W jaki sposób otrzymać te wagi?

W jaki sposób otrzymać wagi w podejściu quasi-randomizacyjnym?

W literaturze można znaleźć dwa podejścia (Elliott i Valliant, 2017), które wykorzystują:

- **badanie referencyjne**, które realizowane jest jednocześnie z wykorzystaniem próby nielosowej/big data. Idea polega na:
  - złączeniu dwóch prób (losowej i nielosowej) oraz oszacowaniu modelu, który przewidywać będzie czy dana jednostka znajduje się w źródle big data;
- **sample matching**, realizowana na poziomie jednostkowym lub domeny
  - *łączenie na poziomie jednostki* – wykorzystanie propensity scores (Rosenbaum & Rubin 1983) do połączenia jednostek,
  - *łączenie na poziomie domeny* – sprowadzenie rozkładów brzegowych zmiennych pomocniczych do wartości znanych z populacji – np. przez kalibrację (Devill & Särndal, 1992).

Kalibracja

1. Metoda zaproponowana przez Devilla i Särndala (1992). Polega na ustaleniu tzw. wag kalibracyjnych, które mają zminimalizować odległość między wagami wejściowymi, a nowymi wagami, które mają spełnić pewne warunki ograniczające.
2. Uzyskane wagi po zsumowaniu powinny odtwarzać znane wartości globalne lub przeciętne.
3. W przypadku źródeł big data, wagi wejściowe (klasycznie będące odwrotnością prawdopodobieństwa inkluzji do próby) mogą być wektorem składającym się z samych 1 (por. podejście do kalibracji w rejestrach administracyjnych), N/n, gdzie n to wielkość próby, a N to wielkość populacji lub odwrotnością skłonności do odpowiedzi ρ_i^{-1}.

Kalibracja – w jaki sposób określić wagi

(C1) Znaleźć minimum funkcji celu (tutaj funkcja χ^2):

$$D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i} \longrightarrow \min, \tag{6}$$

(C2) Przy spełnieniu tzw. równań kalibracyjnych

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{7}$$

(C3) oraz ograniczeń na wagi:

$$L \leqslant \frac{w_i}{d_i} \leqslant U, \quad \text{gdzie: } L < 1 \text{ i } U > 1, \ i = 1, \ldots, n. \tag{8}$$

Estymatory modelowo-kalibrowane

Wu & Sitter (2001) zaproponowali podejście modelowe do kalibracji określane pojęciem *model-calibration*, które należy do rodziny estymatorów wspieranych modelem (ang. *model-assisted estimators*). W skrócie, podejście to jest zbliżone do klasycznej kalibracji:

(C1) Znaleźć minimum funkcji celu (tutaj funkcja χ^2):

$$D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i} \longrightarrow \min, \tag{9}$$

(C2.1) Przy spełnieniu warunku

$$\sum_{i=1}^{n} w_i = N \tag{10}$$

(C2.2) oraz

$$\sum_{i=1}^{n} w_i \hat{\mu}_i = \sum_{i=1}^{N} \hat{\mu}_i \tag{11}$$

gdzie $E(y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\beta})$ oraz $V(y_i|\boldsymbol{x}_i) = v_i^2 \sigma^2$.

Podejście predykcyjne/modelowe

W podejściu predykcyjnym/modelowym:

- Zakładamy, że model zbudowany na źródle big data działa dla populacji, tj.

$$f_s(Y_i|\boldsymbol{x}_i) = \frac{Pr(R_i = 1|\boldsymbol{x}_i, Y_i) f_P(Y_i|\boldsymbol{x}_i)}{Pr(R_i = 1|\boldsymbol{x}_i)}. \tag{12}$$

  W tym przypadku nie zakładamy żadnej postaci modelu $f_s(Y_i|\boldsymbol{x}_i)$
- Następnie, w przypadku estymacji wartości globalnej Y dokonujemy sumy wartości y_i z próby oraz wartości przewidywanych dla jednostek spoza próby $\hat{y}_i$.

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{i \notin s} \hat{y}_i. \tag{13}$$

Por. Sikov (2018) w kontekście budowania modelu, gdy występuje mechanizm NMAR.

Badanie popytu na pracę

Badana zmienna

Liczba wolnych miejsc pracy w ostatnim dniu kwartału sprawozdawczego według Klasyfikacji Zawodów i Specjalności.

Wolne miejsca pracy

to miejsca pracy powstałe w wyniku ruchu zatrudnionych bądź nowo utworzone, w stosunku do których spełnione zostały jednocześnie trzy warunki:

1. miejsca pracy w dniu sprawozdawczym były faktycznie nieobsadzone,
2. pracodawca czynił starania, aby znaleźć osoby chętne do podjęcia pracy,
3. w przypadku znalezienia właściwych kandydatów, pracodawca byłby gotów do natychmiastowego przyjęcia tych osób.

Badanie popytu na pracę

Populacja

Jednostki o liczbie pracujących 1 i więcej osób z sekcji C, F, G, H, J, K, M, N, O, P, Q, R oraz S.

Dobór próby

- Wielkość próby: około 100 tys. jednostek,
- Dwa schematy losowania: (1) dla jednostek do 9 osób, (2) dla jednostek powyżej 9 osób.
- Dla pierwszej grupy:
  - Warstwy: sekcji PKD,
  - Metoda: losowanie proste warstwowe z alokacją proporcjonalną do precyzji szacunków (taka sama według sekcji PKD),
  - Wielkość próby: 50 tysięcy.
- Dla drugiej grupy:
  - Warstwy: sekcji PKD i województwo (304 warstwy)
  - Metoda: dobór losowy warstwowy, proporcjonalny do liczby pracujących z ustalonym progiem liczby zatrudnionych,
  - Wielkość próby: 50 tysięcy.

Badanie Kapitału Ludzkiego – Podstawowe informacje

Bilans Kapitału Ludzkiego (BKL) w latach 2010-2015 realizowany był przez Polską Agencję Rozwoju Przedsiębiorczości (PARP) razem z Centrum Ewaluacji i Analiz Polityk Publicznych Uniwersytetu Jagiellońskiego (CEiAPP UJ)

Celem było poznanie zasobów kompetencyjnych polskiego rynku pracy (kompetencje zgłaszane przez pracodawców i oferowane przez uczestników podaży na rynku pracy).

Stałe moduły badania:

- badanie pracodawców;
- badanie ofert pracy;
- badanie ludności;
- badanie instytucji szkoleniowych.

Badanie ofert pracy – metodyka

- **Jednostka badania**: oferta pracy na terenie Polski (z wykluczeniem staży i praktyk dla studentów i uczniów)
- **Źródła danych**:
  - Careerjet.pl (75%) – ogólnokrajowy portal internetowy pośrednictwa pracy (wyszukiwarka ofert pracy),
  - Powiatowe Urzędy Pracy (25%) wylosowane do badania osób bezrobotnych (10 placówek PUP na województwo).
- **Sposób doboru próby**: wyczerpująca, oferty aktualne w wybranym dniu (w przypadku PUP oferta miała być aktualna w dniu przeprowadzania badania, a w przypadku Careerjet.pl w pierwszej kolejności kodowane były oferty zarejestrowane w tym dniu).
- **Liczebność próby założonej**: minimum 20 000 (m.in. wszystkie oferty pracy z PUP oraz oferty z Careerjet.pl).

Badanie ofert pracy – zbierano następujące informacje

- Numer porządkowy.
- Podregion, Województwo (w tym cała Polska).
- Numer referencyjny ogłoszenia, Data publikacji ogłoszenia.
- Powiatowy Urząd Pracy (PUP).
- Portal careerjet.pl (dokładne źródło publikacji ogłoszenia np. pracuj.pl, gazetapraca.pl, karierawfinansach.pl, itd.)
- Dane teleadresowe instytucji pośrednictwa pracy, Dane teleadresowe pracodawcy bądź konkretnego działu w firmie.
- Nazwa pracodawcy (jeżeli jest dostępna).
- Profil działalności/ PKD 2007; czy pośrednik.
- Klasyfikacja stanowiska pracy (według Klasyfikacja Zawodów i Specjalności opracowanej na podstawie ISCO-08).
- Oryginalna nazwa zamieszczona w ofercie.
- Oferowane wynagrodzenie, Rodzaj umowy o pracę.
- Wykształcenie według oczekiwanego poziomu i kierunku.
- Wymagane doświadczenie/staż pracy.
- Wymagane certyfikaty, uprawnienia do wykonywania zawodu.
- Nazwa wymaganego dodatkowego zasobu np. samochód, komputer.
- Wymagane kompetencje.

Badanie ofert pracy – jakość

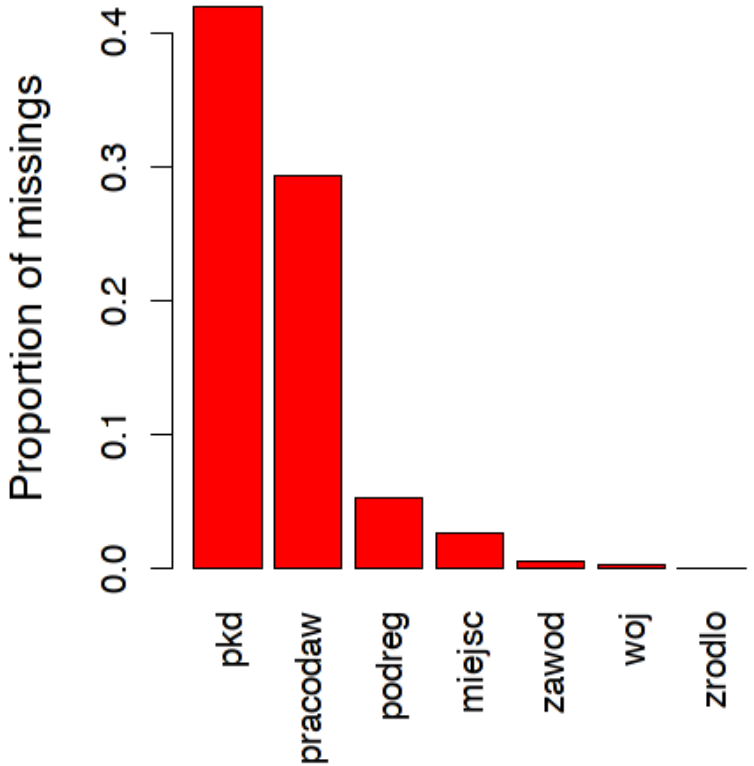
- **Kryteria selekcji ofert obejmowały**:
  - odpowiednią jakość danych (usuwane były: oferty, dla których niemożliwe było ustalenie terytorium, na którym prowadzona była rekrutacja i/lub ustalenie znajomości miejsca wykonywania pracy; oferty zawiera- jące szczątkowe informacje);
  - semantyczny obszar analizy (usuwane były ogłoszenia dotyczące stażu i praktyk);
  - geograficzny obszar analizy (usuwane były ogłoszenia oferujące pracę poza terytorium Polski).
- **Zapewnienie unikalności ofert**:
  - ten sam dzień umieszczenia ogłoszenia;
  - to samo źródło ogłoszenia;
  - tę samą miejscowość;
  - to samo województwo;
  - ten sam numer referencyjny;
  - tę samą nazwę firmy;
  - ten sam zawód.

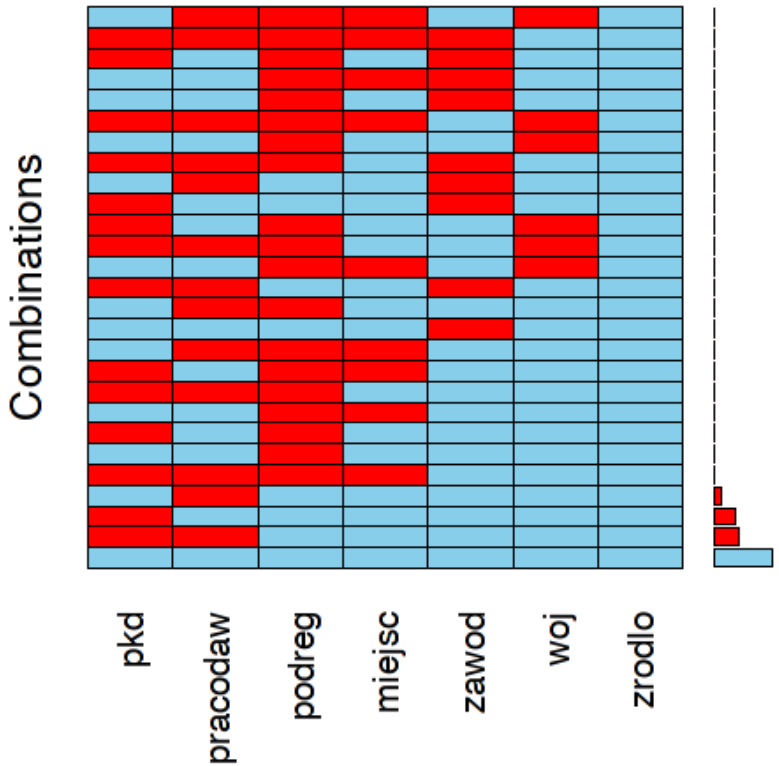
Badanie ofert pracy – wielkość próby

Table 2: Wielkość próby w poszczególnych badaniach i jakość kodowania

Rok	2010	2011	2012	2013	2014
Dzień	10 września	28 marca	26 marca	25 marca	28 marca
PUP	?	2 012	2 812	382	696
CBOP	8 198	5 004	4 440	5 232	7 846
Careerjet.pl	11 811	13 618	14 342	14 467	12 914
Ogółem	20 009	20 634	21 594	20 081	21 456
Kodowanie	0.72	0.89	0.96	0.96	0.963

Braki danych w kluczowych zmiennych





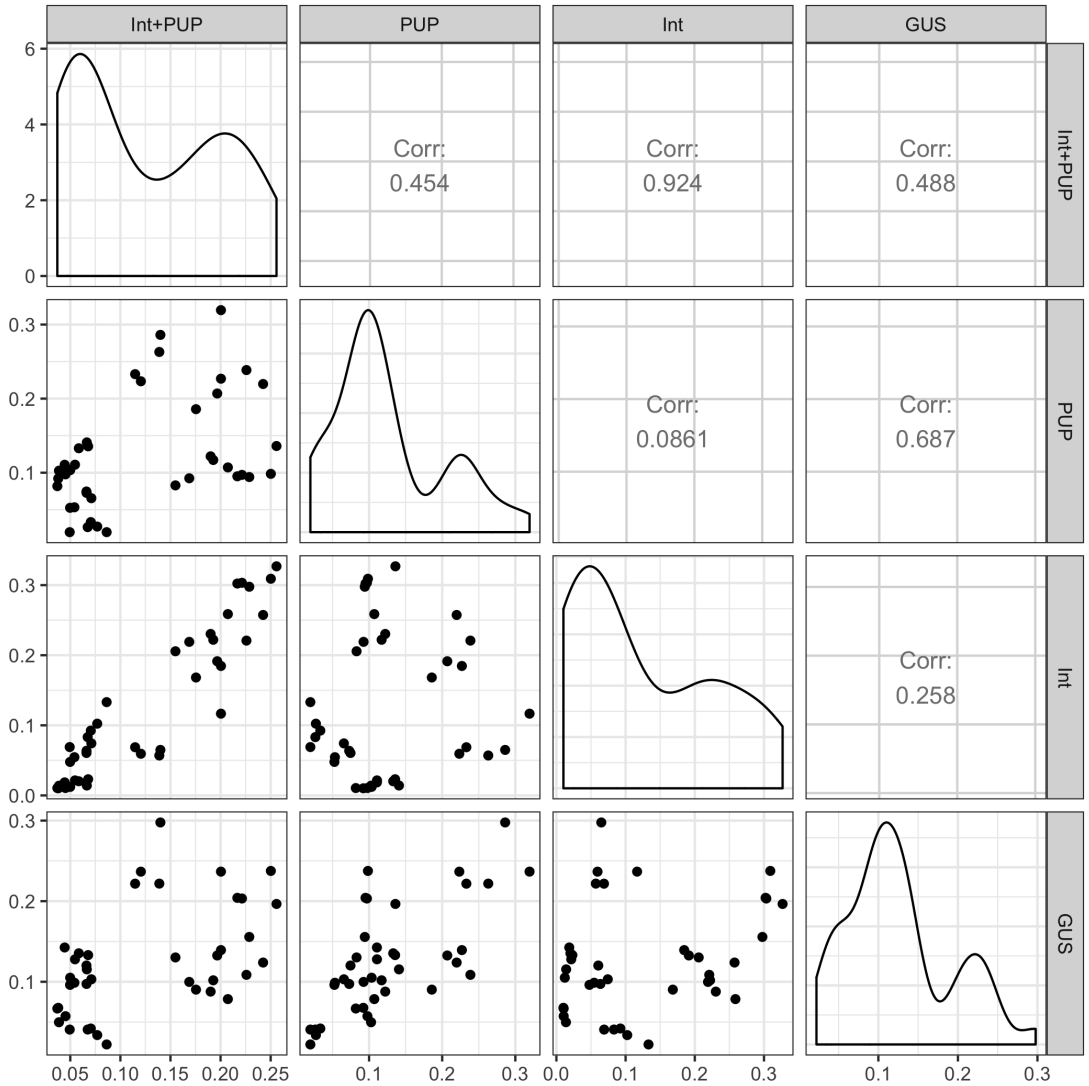
Porównanie z badaniem popytu na pracę

W jaki sposób dokonano porównania:

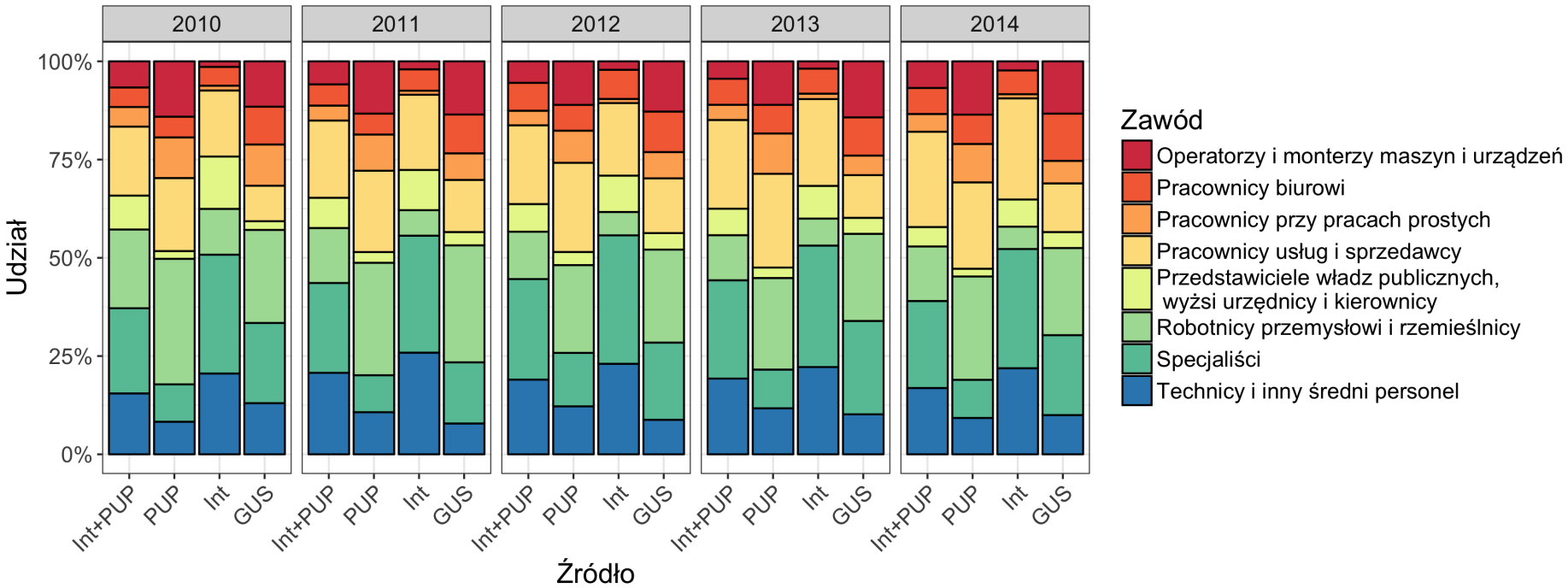
- Porównywano dane z odpowiednich kwartałów uwzględniając okresy badania,
- skupiono się jedynie na zawodach.

Kod	Nazwa
1	Przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy
2	Specjaliści
3	Technicy i inny średni personel
4	Pracownicy biurowi
5	Pracownicy usług i sprzedawcy
7	Robotnicy przemysłowi i rzemieślnicy
8	Operatorzy i monterzy maszyn i urządzeń
9	Pracownicy przy pracach prostych

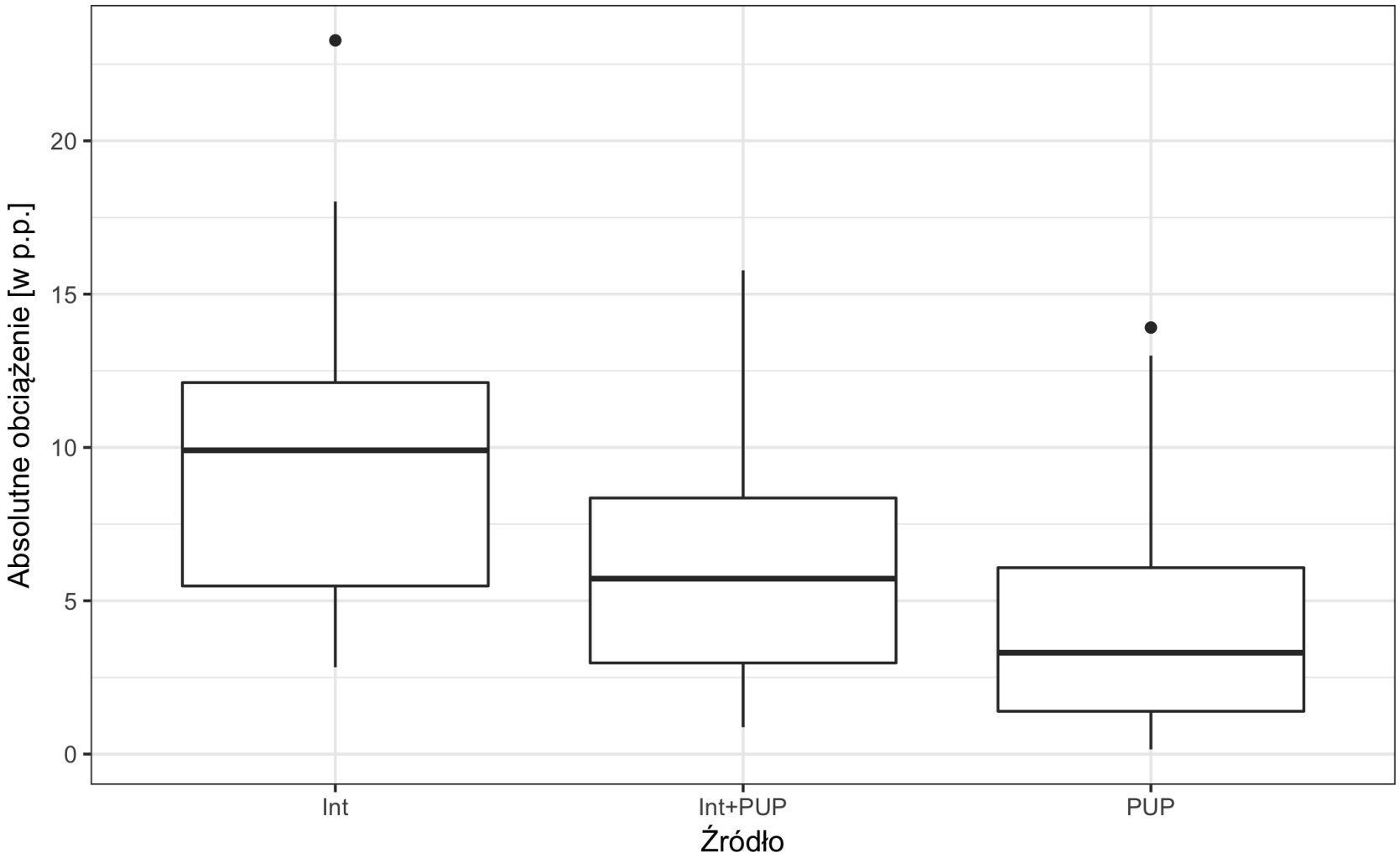
Korelacja między źródłami



Udział ofert według zawodów i źródeł



Absolutne obciążenie według źródeł

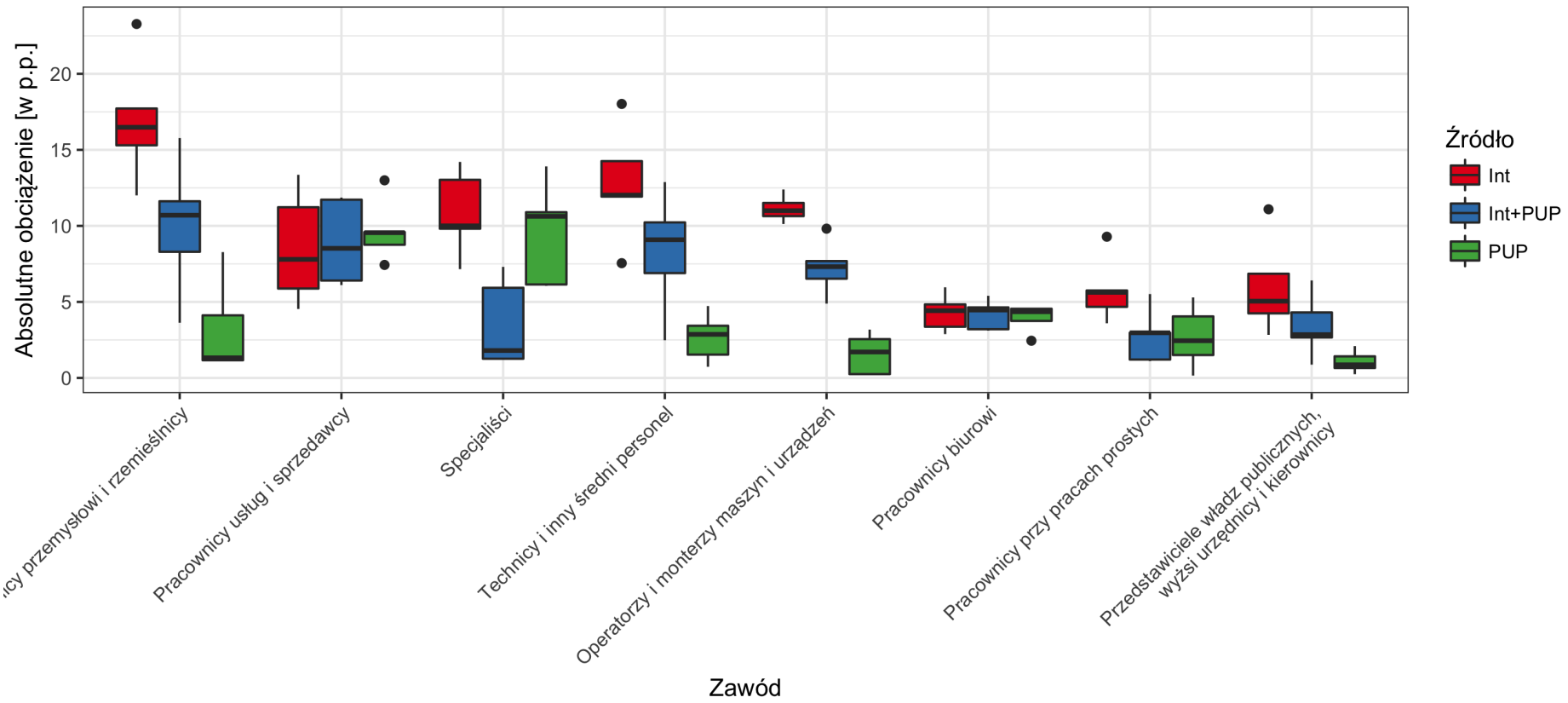


Obciążenie według źródeł

Table 3: Statystyki obciążenia według źródeł

Źródło	Min	Q1	Mediana	Średnia	Q3	Max
Względne obciążenie [w p.p]						
Int	-88.42	-75.71	1.08	23.81	94.31	499.88
PUP	-58.56	-35.23	-4.75	4.41	35.26	119.68
Int+PUP	-68.87	-45.02	-5.01	16.34	65.97	288.89
Absolutne obciążenie [w p.p.]						
Internet	2.83	5.48	9.91	9.54	12.12	23.28
PUP	0.15	1.40	3.31	4.29	6.08	13.91
Int+PUP	0.88	2.97	5.72	6.04	8.35	15.78
Względne absolutne obciążenie [w %]						
Int	27.96	50.62	82.12	93.38	94.31	499.88
PUP	1.46	17.69	35.56	37.14	51.66	119.68
Int+PUP	5.31	31.74	48.25	59.49	68.10	288.89

Absolutne obciążenie według zawodów



Podsumowanie

- Jedno z pierwszych zastosowań kalibracji w badaniu obciążenia dla prób nielosowych odnoszących się do ofert pracy w Polsce;
- Stwierdzone obciążenie może być konsekwencją nielicznego zbioru zmiennych pomocniczych (w badaniu wzięto pod uwagę tylko sekcję PKD i województwo),
- Obciążenie może wynikać również z tego, że w badaniu GUS mogą znaleźć się oferty pracy, których nie ma ani w Internecie ani w PUP.

Literatura (wybrana) I

- Belsby, L., Bjornstad, J., & Zhang, L.-C. (2005). Modeling and Estimation Methods for Household Size in the Presence of Nonignorable Nonresponse Applied to the Norwegian Consumer Expenditure Survey. Survey Methodology, (12).
- Bethlehem, J. (2010). Selection bias in web surveys. International Statistical Review, 78(2), 161-188.
- Buelens, B., Daas, P. J. H., Burger, J., Puts, M., & van den Brakel, J. (2014). Selectivity of Big Data, (11).
- Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J. A., ... Thompson, M. (2017). Approaches to Improving Survey-Weighted Estimates. Statistical Science, 32(2), 227–248.
- Diaz, F., Gamon, M., Hofman, J. M., Kıcıman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. PloS one, 11(1), e0145406.http://doi.org/10.1214/17-STS609
- Haziza, D., & Beaumont, J.-F. (2017). Construction of Weights in Surveys: A Review. Statistical Science, 32(2), 206–226. http://doi.org/10.1214/16-STS608

Literatura (wybrana) II

- Rubin, D. B. (1976). Inference and missing data. Biometrika.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.
- Zhang, L. C., Thomsen, I., & Kleven, Ø. (2013). On the Use of Auxiliary and Paradata for Dealing With Non-sampling Errors in Household Surveys. International Statistical Review, 81(2), 270-288.

Dziękujemy za uwagę!