METODOLOGIA SPISU LUDNOŚCI I MIESZKAŃ 2011

1. Źródła danych

Powszechny spis ludności i mieszkań 2011 został przeprowadzony metodą mieszaną tj. poprzez pozyskiwanie danych ze źródeł administracyjnych (rejestrów i systemów informacyjnych) oraz zbieranie danych bezpośrednio od ludności w ramach badania reprezentacyjnego oraz tzw. badania pełnego. Oprócz tego przeprowadzone zostały dwa pełne badania, obejmujące osoby przebywające w obiektach zbiorowego zakwaterowania oraz osoby bezdomne. Zastosowane rozwiązania miały przede wszystkim zmniejszyć koszty spisu oraz obciążenie osób objętych spisem, przy zachowaniu wysokiej jakości wyników spisu.

W ustawie o NSP 2011 przyjęte zostało założenie jak najszerszego wykorzystania systemów informacyjnych administracji publicznej, jako źródeł danych do spisu, co w konsekwencji oznaczało, że informacje przewidziane do zebrania w trakcie spisu pobrane zostały przede wszystkim z dostępnych źródeł administracyjnych, a następnie wykorzystane do przygotowania i aktualizacji wykazu adresowo-mieszkaniowego, następnie do utworzenia operatu adresowo-mieszkaniowego do losowania próby do badania reprezentacyjnego oraz jako bezpośrednie źródło danych spisowych. Dane niewystępujące w systemach informacyjnych administracji publicznej lub niespełniające wymogów jakości danych statystycznych zebrano od osób objętych spisem. Jednak w tym przypadku przewidziano zastosowanie nowoczesnych technik gromadzenia danych w celu wyeliminowania formularzy papierowych.

2. Prace przygotowawcze do spisu

Ważnym obszarem prac przygotowawczych do NSP 2011 były działania związane z tworzeniem elektronicznego wykazu adresowo-mieszkaniowego na potrzeby spisu. Wykaz adresowo-mieszkaniowy przygotowany został w oparciu o rejestr TERYT (Krajowy Rejestr Urzędowy Podziału Terytorialnego Kraju) z wykorzystaniem danych pochodzących z innych źródeł, m.in. z państwowego zasobu geodezyjnego i kartograficznego w zakresie lokalizacji przestrzennej budynków. W związku z połączeniem tych dwóch źródeł do identyfikatorów adresowych budynków z rejestru TERYT dołączone zostały współrzędne geodezyjne x,y (punkty adresowe). W zakresie informacji o osobach na potrzeby wykazu adresowo-mieszkaniowego wykorzystane były gminne zbiory meldunkowe, które zostały połączone z systemem NOBC, tj. systemem identyfikacji adresowej ulic, nieruchomości, budynków i mieszkań stanowiącym część rejestru TERYT. Przygotowane w tym trybie zestawienie budynków, mieszkań i osób było weryfikowane przez urzędy gmin podczas aktualizacji przedspisowej. Na tym etapie została ustalona zbiorowość budynków, mieszkań i obiektów zbiorowego zakwaterowania podlegających spisaniu, poprawność ich adresów oraz przyporządkowanie osób do poszczególnych mieszkań i obiektów.

W trakcie obchodu przedspisowego rachmistrzowie dokonali weryfikacji wszystkich punktów adresowych, które znalazły się w obszarze przydzielonych obwodów, porównując je z pozycjami ujętymi w wykazie adresowo-mieszkaniowym do obchodu. Przedmiotem obchodu było zatwierdzenie lub modyfikacja punktów adresowych wyszczególnionych w wykazie, a także usunięcie nieistniejących punktów adresowych lub ewentualne dodanie nowych nieodnotowanych w wykazie, na których znalazły się budynki mieszkalne, budynki niemieszkalne z mieszkaniami lub budynki obiektów zbiorowego zakwaterowania, np. dom studencki. W odróżnieniu do poprzednich spisów ludności i mieszkań rachmistrzowie nie odwiedzali mieszkań w celu dokonania weryfikacji informacji dotyczących nazwiska i imienia głównego lokatora i liczby osób podlegających spisowi.

Rachmistrzowie korzystali z map cyfrowych, zainstalowanych na przenośnych urządzeniach elektronicznych typu handheld. Kartograficzna dokumentacja przedspisowa przygotowana została w formie elektronicznej i wykorzystywała narzędzia GIS (*Geographic Information Systems*). Na mapach cyfrowych zostały zaznaczone obwody spisowe z budynkami i punktami adresowymi. Mapa cyfrowa była sprzęgnięta z elektronicznym, opisanym wyżej wykazem adresowo-mieszkaniowym zawierającym adresy budynków, mieszkań i obiektów zbiorowego zakwaterowania przeznaczonych do spisania. Tak przygotowany moduł elektroniczny dokumentacji wyjściowej był dostępny na wszystkich poziomach zarządzania spisem. Aplikacja sygnalizowała na mapie aktualne położenie rachmistrza, a w przypadku stwierdzenia przez niego, że w terenie znajduje się nieujęty w operacie spisowym i nienaniesiony na mapie cyfrowej punkt adresowy, pod którym mieszkają ludzie, umożliwiała wprowadzenie tego punktu na mapę cyfrową za pomocą urządzenia GPS, zainstalowanego na przenośnym urządzeniu elektronicznym, oraz dopisanie do operatu brakującego mieszkania, czy innego obiektu.

3. Badanie pełne

Badanie zostało przeprowadzone w oparciu o pozyskane do celów statystycznych rejestry administracyjne i systemy informacyjne. Do przeprowadzenia NSP zgromadzono informacje m.in. z następujących systemów centralnych i rozproszonych: Ministerstwa Finansów, Ministerstwa Spraw Wewnętrznych i Administracji (zbioru PESEL) Zakładu Ubezpieczeń Społecznych, Kasy Rolniczego Ubezpieczenia Społecznego, Narodowego Funduszu Zdrowia (Centralnego Wykazu Ubezpieczonych), urzędów gmin (gminnych zbiorów meldunkowych).

W wyniku szczegółowego przeanalizowania pozyskanych danych, połączono zebrane informacje tworząc wykaz podmiotowy do przeprowadzenia badania spisowego. Uzyskane tą metodą dane zostały również użyte do wygenerowania podpowiedzi do ankiet spisowych (na podstawie tzw. *Master Recordu*) i przyspieszenia w ten sposób zbierania informacji od respondentów. W dalszej kolejności metoda ta posłużyła do wygenerowania wyników NSP.

Podstawowym źródłem informacji na temat budynków i mieszkań w NSP 2011 była tzw. baza budynków. Została ona utworzona w wyniku połączenia danych pozyskanych w ramach NSP 2002, badań statystycznych min. dotyczących wydanych pozwoleń na budowę oraz budynkach mieszkalnych i mieszkaniach w budynkach niemieszkalnych oddanych do użytkowania oraz Elektronicznej Karty Budynku, za pomocą której każdy podmiot zarządzający lub administrujący budynkami przekazał informacje dotyczące wybranych cech i parametrów poszczególnych budynków wielomieszkaniowych.

Dodatkowym źródłem danych były informacje zebrane przez rachmistrzów w trakcie obchodu przedspisowego, które umożliwiły aktualizację bazy poprzez dopisanie budynków niefigurujących w bazie, usunięcie nieistniejących, czy zmianę ich rodzaju (np. z mieszkalnego na niemieszkalny).

Szerszy zakres danych – zarówno o osobach jak i mieszkaniach – został pozyskany w ramach badania reprezentacyjnego.

4. Badanie reprezentacyjne

Przeprowadzone w ramach NSP badanie reprezentacyjne, dostarczyło danych, które nie są gromadzone w rejestrach i systemach informacyjnych. Badanie zostało przeprowadzone na próbie losowej ok. 20% mieszkań w skali kraju. Jednostką losowania było mieszkanie, a dokładniej jego adres.

Operat losowania mieszkań, który został utworzony w oparciu o rejestry i systemy informacyjne, został odpowiednio uporządkowany i podzielony na poszczególne warstwy. Najważniejsze cechy, jakie brano pod uwagę przy alokacji próby pomiędzy poszczególne jednostki podziału administracyjnego to:

– liczba osób zameldowanych w mieszkaniu,

– występowanie dużych skupisk mieszkaniowych w dużych miastach tzw. „blokowisk”

– występowanie osoby pracującej w mieszkaniu,

– występowanie emeryta lub rencisty w przypadku braku pracującego,

– występowanie bezrobotnego w mieszkaniu w przypadku braku ww. osób,

– uwzględnienie mieszkań z użytkownikiem gospodarstwa rolnego.

Zasady tworzenia operatu losowania mieszkań

Etapy tworzenia operatu losowania mieszkań zawierały:

1. Zintegrowanego wykazu adresowo-mieszkaniowego w oparciu o zestawienie adresowo-mieszkaniowe,
2. Weryfikacja wykazu obchodem przedspisowym,
3. Uzupełnienie wykazu dodatkowymi zmiennymi i utworzenie operatu losowania do badania reprezentacyjnego

Dodatkowo do wykazu adresowo-mieszkaniowego w Operacyjnej Bazie Mikrodanych zostały wprowadzone informacje pochodzące z innych rejestrów i systemów informacyjnych lub baz statystycznych jak: zbiory Zakładu Ubezpieczeń Społecznych (CRU, EiR oraz CRPS), Kasy Rolniczego Ubezpieczenia Społecznego (KRUS), Krajowej Ewidencji Podatników (KEP), Narodowego Funduszu Zdrowia (NFZ), Elektronicznego Krajowego Systemu Monitoringu Orzekania o Niepełnosprawności (EKSMoON), PESEL, spisu ludności i mieszkań 2002, bazy budynkowej oraz Wykazu bazowego do PSR 2010. Celem włączenia do wykazu dodatkowych informacji była poprawa jakości danych zgromadzonych w tym wykazie oraz zapewnienie informacji na potrzeby utworzenia operatu losowania. W trakcie weryfikacji danych w wykazie zostały ustalone kryteria (zmienne oraz ich wartości), które zostały wykorzystane do tworzenia bardziej jednorodnych grup (warstwowania) jednostek badania.

Przy tworzeniu operatu losowania pominięto:

a) obiekty zbiorowego zamieszkania niezawierające mieszkań,
b) obiekty zamknięte oraz mieszkania na terenie obiektów zamkniętych,
c) mieszkania w obiektach zbiorowego zakwaterowania stanowiących obiekty zamknięte,
d) pomieszczenia prowizoryczne,
e) mieszkania niezamieszkane, zniszczone na skutek klęsk żywiołowych w szczególności powodzi.

W konsekwencji do operatu weszły mieszkania zamieszkane, w których osoby były zameldowane lub przebywały bez meldunku oraz te mieszkania niezamieszkane, które były w trakcie remontu, zmiany lokatora lub nowo wybudowane. Budynki, które występowały pod jednym adresem były brane do losowania, jeżeli ich liczba nie przekroczyła dwóch. Operat losowania mieszkań został zaktualizowany w trakcie trwania obchodu przedspisowego, w wyniku którego do badania dołączono budynki jednomieszkaniowe, które nie znalazły się wcześniej w wykazie.

Zmienne uwzględnione w operacie losowania

Dla potrzeb badania reprezentacyjnego w wykazie dopisano do każdego mieszkania zestaw informacji zawarty w operacie losowania, spośród których wybrano ostateczne kryteria warstwowania:

a) symbol terytorialny,

b) położenie budynku/mieszkania na terenie objętym powodzią lub inna klęską żywiołową,

c) liczba mieszkań w budynku wielomieszkaniowym,

d) położenie mieszkania w budynku wielomieszkaniowym,

e) rok zakończenia budowy budynku,

f) liczba osób w mieszkaniu (0, 1, 2, 3, 4, 5, 6, 7 i więcej) – po dodatkowych badaniach,

g) liczba osób zameldowanych na pobyt stały – po dodatkowych badaniach,

h) występowanie cudzoziemca w mieszkaniu,

i) występowanie osoby pracującej (posiadanie ubezpieczenia), emeryta lub rencisty w mieszkaniu oraz osoby bezrobotnej,

j) występowanie osoby niepełnosprawnej,

k) występowanie użytkownika gospodarstwa rolnego oraz powierzchni użytków rolnych,

l) położenie mieszkania w gminie o udziale społeczności mniejszościowej (narodowej lub etnicznej) co najmniej 10%.

Wartości zmiennych, jako kryteria warstwowania zostały zróżnicowane dla obszarów miejskich i wiejskich. Wybór zmiennych został ostatecznie ustalony w trakcie losowania próbnego.

Alokacja próby w poszczególnych powiatach[1]

Podstawowym celem badania reprezentacyjnego realizowanego w ramach spisu 2011 jest uzyskanie informacji o sytuacji społeczno-demograficznej na poziomie powiatu. Konieczne było więc dokonanie podziału założonej 20%-owej próby mieszkań dla Polski pomiędzy powiaty. Dokonano tego przy wykorzystaniu metody alokacji pierwiastkowej. Metoda ta stanowi kompromis pomiędzy alokacją proporcjonalną a alokacją zapewniającą jednakową precyzję dla subpopulacji. Przy założeniu, że zastosowano by proporcjonalne losowanie próby, w każdym powiecie próba stanowiłaby 20 % populacji. Ponieważ precyzja wyników tj. wielkość błędu losowego zależy od liczebności próby, uzyskano by w efekcie

[1] Materiał roboczy opracowany przez zespól matematyków pod kierunkiem Bronisława Lednickiego.

niedostateczną precyzję dla wielu małych powiatów. Z kolei, w metodzie alternatywnej uzyskujemy, w przybliżeniu, jednakową precyzję wyników dla wszystkich powiatów, ale za cenę istotnego „spłaszczenia" liczebności próby. W efekcie liczebność próby, a tym samym pracochłonność przy realizacji spisu byłaby mało zróżnicowana pomiędzy dużymi i małymi powiatami. Z tych powodów jako metodę rozdziału próby przyjęto alokację pierwiastkową, w której np. liczba mieszkań losowanych w p-tym powiecie jest proporcjonalna do pierwiastka kwadratowego z populacyjnej liczby mieszkań i wyrażona wzorem:

$$n_p = n^* * \frac{\sqrt{N_p}}{\sum_p \sqrt{N_p}}$$

gdzie:

n* – założona liczebność próby dla Polski,
Np – liczba mieszkań w p-tym powiecie,

Według powyższego wzoru obliczone zostały liczebności prób dla poszczególnych powiatów. Potraktowanie Warszawy, tak jak każdego innego powiatu, mogłoby w konsekwencji spowodować niemożność uzyskania w miarę precyzyjnych wyników dla poszczególnych dzielnic. Jako ostateczną próbę dla miasta Warszawy przyjęto wartość średnią z powyższych wartości, tj. 87 500 mieszkań. Wartość ta została odjęta od założonej liczebności próby dla Polski tj. od 2 631 tys., po czym dokonano alokacji pierwiastkowej dla 378 powiatów. Próby w najmniejszych powiatach tj. beskidzkim i sejneńskim wynoszą po ok. 3200 mieszkań (49% populacji), zaś największa próba (poza Warszawą) jest w m. Łodzi – 23 000 mieszkań, czyli 6,9% liczby mieszkań w tym mieście – powiecie.

W Warszawie, Łodzi, Krakowie, Wrocławiu i w Poznaniu podział próby pomiędzy poszczególne dzielnice i delegatury dokonany został również metodą pierwiastkową. Poza Warszawą, w pozostałych miastach, ze względu na małą liczbę delegatur (4 lub 5) nie zachodziła potrzeba zwiększania liczebności próby, tak jak w przypadku Warszawy. Ustalone dla poszczególnych delegatur w tych miastach liczebności prób zapewniają uzyskanie precyzji nie gorszej niż w najmniejszych powiatach.

Schemat losowania próby

W celu wylosowania, w każdym z powiatów, próby o ustalonej wcześniej liczebności zastosowany został schemat losowania jednostopniowego warstwowego. Jednostki losowania – mieszkania zostały przed losowaniem pogrupowane w warstwy w celu zwiększenia efektywności losowania. Zastosowano zróżnicowane podejście do warstwowania w zależności od typu powiatu i gminy.

W miastach – powiatach oraz w dzielnicach wymienionych wyżej 5 największych miast, a także w innych wyróżnionych większych miastach niebędących powiatami, w pierwszym etapie podział mieszkań na dwie kategorie:

a) mieszkania w „blokowiskach”,

b) pozostałe mieszkania.

Kryterium podziału na powyższe kategorie został, przypisany do każdego mieszkania, wskaźnik wyrażający liczbę mieszkań w budynku, w którym znajduje się dane mieszkanie. Do „blokowisk” zaliczone zostały mieszkania o wartości tego wskaźnika powyżej jego mediany. Następnie w każdej z powyższych grup mieszkania zostały powarstwowane ze względu na liczbę osób w mieszkaniu, po czym w ramach tak utworzonych kategorii nastąpił dalszy podział na 4 grupy ze względu na:

a) występowanie osoby pracującej w mieszkaniu,

b) występowanie emeryta lub rencisty w przypadku braku pracującego,

c) występowanie bezrobotnego w mieszkaniu w przypadku braku w/w osób,

d) mieszkania z innymi osobami.

Ponadto, w mieszkaniach poza „blokowiskami”, dodano kategorię: mieszkanie z użytkownikiem gospodarstwa rolnego.

Warstwowanie według liczby osób zamieszkałych w danym mieszkaniu było istotnym czynnik mającym wpływ na precyzję wyników, zniwelowało bowiem negatywne efekty zróżnicowanej liczby osób zamieszkałych w mieszkaniach. Efektywność tego etapu warstwowania zależała od korelacji liczby osób w mieszkaniu zapisanej w operacie losowania ze stanem rzeczywistym. Warstwowanie ze względu na pozostałe zmienne miało z kolei pozytywny wpływ na wyniki spisu związane z aktywnością ekonomiczną.

W pozostałych powiatach w pierwszym etapie założono warstwowanie według gmin, przy czym gminy miejsko – wiejskie traktowane były, jako dwie oddzielne gminy. W mniejszych miastach pominięty został etap podziału na mieszkania w tzw. blokowiska i pozostałe mieszkania, a warstwowanie odbywało się analogicznie jak w dużych miastach poza blokowiskami.

W gminach wiejskich w pierwszym etapie mieszkania podzielone zostały na dwie kategorie:

a) mieszkania z użytkownikiem gospodarstwa rolnego,

b) pozostałe mieszkania.

W kategorii pierwszej wykonano warstwowanie wg liczby osób zamieszkałych w mieszkaniu, a następnie, w miarę możliwości, dalsze warstwowanie wg powierzchni gospodarstwa rolnego (dwie ewentualnie trzy grupy obszarowe w zależności od liczby mieszkań z użytkownikiem gospodarstwa w gminie). W przypadku drugiej kategorii mieszkań wykonano warstwowanie analogiczne jak w małych miastach.

Przedstawiona wyżej koncepcja warstwowania mogła prowadzić do utworzenia w niektórych powiatach bardzo rozdrobnionych warstw. W związku z tym program losujący przy tworzeniu warstw sprawdzał jednocześnie czy tworzone w powyższy sposób warstwy spełniały wymóg minimalnej liczby mieszkań. Warstwy zawierające zbyt mało mieszkań zostały łączone z sąsiednimi. Założono odwrotną niż przy tworzeniu, hierarchię kryteriów tzw. sklejania warstw.

Po utworzeniu warstw ustalone zostały liczebności prób do wylosowania w poszczególnych warstwach. Przyjęto proporcjonalną alokację próby pomiędzy warstwy. Oznacza to m.in., że ustalona w wyniku alokacji pierwiastkowej frakcja losowania w danym powiecie obowiązywała we wszystkich gminach tego powiatu i w warstwach utworzonych wewnątrz gmin.

Efektem tych działań było wylosowanie, spośród prawie 13,5 mln mieszkań, próby liczącej ponad 2,7 mln mieszkań. Utworzono prawie 70,5 tys. warstw, zaś wielkość próby w poszczególnych warstwach wahała się od niemal 6% do ponad 49%.

Zakres tematyczny badania reprezentacyjnego w NSP 2011 uwzględniał sześć dużych obszarów tematycznych:

- ludność i jej charakterystyka demograficzno-społeczna,
- aktywność ekonomiczna,
- migracje wewnętrzne i zagraniczne ludności,
- narodowość i wyznanie,
- gospodarstwa domowe i rodziny
- oraz budynki i mieszkania.

W ramach tych obszarów można wyróżnić 15 tematów badawczych (p. punkt 3 w rozdziale I). Badanie reprezentacyjne w NSP 2011 stanowi komplementarną całość do badania pełnego, przeprowadzanego w oparciu o rejestry i systemy informacyjne.

5. Spis osób w obiektach zbiorowego zakwaterowania oraz osób bezdomnych

W trakcie trwania spisu zebrano informacje o osobach przebywających powyżej 3 miesięcy w obiektach zbiorowego zakwaterowania, czyli w budynkach zajętych przez jeden odrębny zakład, świadczący usługi: opiekuńczo-wychowawcze, opiekuńczo-lecznicze, obiekty związane z pracą lub nauką (domy studenckie, internaty, hotele pracownicze), bądź inne, w którym to obiekcie zamieszkuje/przebywa zwykle większa liczba osób. Dane zostały pozyskane od właścicieli, administratorów lub zarządców obiektów przy wykorzystaniu specjalnej aplikacji internetowej. W szczególnych przypadkach informacje były pozyskiwane przy wsparciu pracowników wojewódzkich lub gminnych biur spisowych oraz rachmistrzów.

W dniach 15-16 kwietnia 2011 r. – przy współpracy z Pomorskim Forum na Rzecz Wychodzenia z Bezdomności – przeprowadzono również badanie osób bezdomnych.

Osoby bezdomne spisywane były przez rachmistrza na aplikacji mobilnej, w miejscu ich przebywania wskazanym przez pracowników gminnych biur spisowych w porozumieniu z placówkami udzielającymi pomocy bezdomnym. Zostały spisane osoby, które wieczór i noc w momencie spisu spędzały poza jakąkolwiek instytucją funkcjonującą całodobowo, w miejscach ich przebywania takich jak: dworce kolejowe i autobusowe oraz ich okolice, kanały i węzły ciepłownicze, ogródki działkowe, ulice, bunkry, lasy i parki, centra handlowe, parkingi, opuszczone samochody, przyczepy kempingowe, klatki schodowe, zsypy, piwnice, śmietniki, wagony i bocznice kolejowe, ogrzewalnie itp. Osoby bezdomne przebywające w obiektach zbiorowego zakwaterowania (schroniska, noclegownie i instytucje dla bezdomnych), zostały spisane przez administratorów tych obiektów.

6. Formularze wykorzystywane w spisie 2011

W spisie 2011 wykorzystywano dwa rodzaje formularzy, dostępnych wyłącznie w formie elektronicznej. Formularz długi, o szerokim zakresie tematycznym z dużą liczbą pytań (ponad 120 pytań), stosowany był w badaniu reprezentacyjnym, natomiast formularz krótki (16 pytań) znalazł zastosowanie w badaniu pełnym, przede wszystkim w celu aktualizacji danych pozyskiwanych do spisu z rejestrów i systemów informacyjnych. Formularze elektroniczne były dostępne w trybie on-line, a dodatkowo formularz krótki – także w trybie off-line. Formularze zostały przygotowane w wersji aplikacji na urządzenia przenośne typu handheld oraz w wersji aplikacji internetowej, która była wykorzystywana podczas samospisu. Elementem aplikacji elektronicznej były słowniki funkcjonujące przy pytaniach w poszczególnych obszarach tematycznych: edukacji, aktywności ekonomicznej, kraju obywatelstwa i urodzenia, migracji, przynależności narodowo-etnicznej, wyznania. Przy niektórych słownikach istniała możliwość swobodnego zapisu słownego. Dodatkowo w części adresowej formularzy elektronicznych podłączony był słownik TERYT.

Odrębne, uproszczone formularze elektroniczne były opracowane dla badania osób bezdomnych oraz osób w obiektach zbiorowego zakwaterowania.

W ramach spisu 2011 przeprowadzono także pełne badanie ludności w 86 gminach, wstępnie wytypowanych na podstawie wyników spisu 2002. Kryterium do wyróżnienia tych gmin był co najmniej 10% udział osób należących do mniejszości narodowej lub etnicznej w liczbie mieszkańców gminy w 2002 roku. Dla osób zamieszkałych lub przebywających w tych gminach w wylosowanych mieszkaniach wypełniany był formularz długi, zaś w pozostałych mieszkaniach – formularz krótki. Pytania o przynależność narodowo-etniczną oraz o język używany w kontaktach domowych zostały włączone do obu formularzy, m.in. także z tego powodu, że w żadnym z systemów administracyjnych nie występowała informacja o narodowości, możliwa do wykorzystania w spisie ludności. Dane z tego badania mają duże znaczenie dla opracowania wyników spisu w obszarze narodowości i języka, zwłaszcza w zakresie ustalania tzw. gmin mniejszościowych.

7. Spis kontrolny

W dniach od 1 do 11 lipca 2011 r. przeprowadzony został spis kontrolny do NSP 2011. Celem spisu kontrolnego 2011 było sprawdzenie kompletności przeprowadzonego spisu, poprawności danych uzyskanych w spisie oraz zgodności tych danych ze stanem faktycznym.

Spośród 2 744 tys. mieszkań, które wcześniej zostały wylosowane do badania reprezentacyjnego wylosowano 80 tys. mieszkań. Spisem kontrolnym objęto mieszkania, w których respondenci dokonali samospisu przez internet, mieszkania zostały spisane bezpośrednio przez rachmistrzów spisowych lub spisane telefonicznie – przez ankieterów, jak i takie, w których spis nie został przeprowadzony (z różnych powodów).

Spis kontrolny został przeprowadzany przez ankieterów poprzez telefon (metodą CATI), co spowodowało, iż obejmował on wyłącznie te mieszkania, w których przynajmniej dla jednej osoby udało się ustalić numer telefonu, przy czym nie miało znaczenia czy był to telefon stacjonarny, czy komórkowy. Formularz do spisu kontrolnego zawierał 14 pytań. Zebrane na jego podstawie informacje powinny pozwolić na dokonanie oceny błędów nielosowych, a przede wszystkim oceny błędów pokrycia (in plus oraz in minus), popełnianych z powodu podwójnego spisywania osób, ewentualnego dopisywania fikcyjnych osób oraz opuszczania osób, a także do określenia błędów treści (wynikających z braków odpowiedzi lub błędów w odpowiedziach). Przyjęte założenia oraz dane pozyskane w spisie kontrolnym, powinny pozwolić na ocenę wpływu czynników zewnętrznych, tj. rachmistrzów, ankieterów czy samych respondentów, którzy dokonali samospisu – na jakość wyników spisu ludności i mieszkań 2011. Efekty takiej analizy będą prezentowane w kolejnych ogólnopolskich i regionalnych publikacjach tematycznych, natomiast szczegółowe zasady metodologiczne spisu kontrolnego zostaną przedstawione w publikacji poświęconej metodologii spisu ludności i mieszkań 2011.

8. Techniki pozyskiwania danych (metody CAxI) oraz zbiory danych wynikowych

Podstawowym sposobem gromadzenia danych w spisie 2011 była technika CAII – Computer Assisted Internet Interview. Od 1 kwietnia do 16 czerwca 2011 r. ok. 12% populacji kraju spisało się bez udziału rachmistrzów w drodze tzw. samospisu internetowego. Ta forma zbierania danych polegała na udostępnieniu on-line formularzy elektronicznych, zawierających pytania spisowe. Formularze zawierały dane pozyskane z rejestrów i respondenci weryfikowali oraz uzupełniali je zgodnie ze stanem faktycznym w dniu 31 marca 2011 r. W szczególnych przypadkach np. braku możliwości zalogowania się do systemu, respondenci mogli skorzystać z formularzy bez naniesionych danych i wypełnić je samodzielnie w całości w trybie off-line.

CAPI – Computer Assisted Personal Interview. Od 8 kwietnia do 30 czerwca 2011 r. rachmistrzowie spisowi przeprowadzili wywiady w terenie na podstawie przydzielonej puli adresów mieszkań, wylosowanych do badania reprezentacyjnego. Odpowiedzi respondentów były w trakcie wywiadu rejestrowane na przenośnych urządzeniach elektronicznych typu handheld.

CATI – Computer Assisted Telephone Interview. Ankieterzy statystyczni przeprowadzali spis telefonicznie, uzupełniając zebrane dane za pośrednictwem dedykowanej aplikacji elektronicznej.

Zbiory danych wynikowych

W wyniku zebrania informacji objętych spisem powstały zbiory danych wynikowych z wyodrębnionymi obiektami, powiązanych ze sobą wspólnym identyfikatorem:

- Budynki
- Mieszkania
- Osoby
- Gospodarstwa domowe
- Rodziny
- Obiekty zbiorowego zakwaterowania
- Bezdomni

Zakres informacji dla powyższych obiektów został ujęty w tzw. złotym rekordzie (ZR), zawierającym zmienne informacyjne obejmujące zarówno dane ze spisu pełnego (czyli pozyskiwane z rejestrów i systemów administracyjnych), jak też z badania reprezentacyjnego. Zakres informacyjny złotego rekordu zawiera zmienne bezpośrednio pozyskane oraz wyprowadzone wtórnie.

Wyliczenie tych zmiennych odbywa się w środowisku Operacyjnej Bazy Mikrodanych (OBM) – systemie wspierającym proces spisowy. Jego głównym zadaniem jest integracja danych pochodzących z różnych źródeł. System ten wykorzystywany jest głównie na etapie: czyszczenia, walidacji, deduplikacji oraz korekty danych. W OBM następuje przekształcenie danych źródłowych na dane spisowe[2].

W kolejnych etapach przetwarzania danych złoty rekord przekazywany jest do środowiska Analitycznej Bazy Danych – bazy, w której są przechowywane odpersonalizowane dane spisowe w ostatecznej wersji. W ABM są wykonywane dalsze przekształcenia (głównie wyprowadzanie kolejnych zmiennych i obiektów pochodnych). Dodatkowo środowisko to posiada narzędzia dla prowadzenia analiz statystycznych oraz udostępniania wyników spisu. Algorytmy wyprowadzania zmiennych zawartych w złotym rekordzie zawierają wskazanie najbardziej odpowiedniego źródła danej zmiennej oraz sposobu jej wyliczenia (tj. uwzględnienia zależności merytorycznych pomiędzy zmiennymi i kontroli logicznej).

9. Łączenie danych pozyskanych z różnych źródeł

Metody przetwarzania zbiorów danych administracyjnych w zbiory statystyczne

Dla potrzeb spisu informacje pozyskiwane z rejestrów zostały poddane odpowiednim mechanizmom i procedurom zarówno informatycznym, jak też merytorycznym. Głównym celem tych działań było przekształcenie zmiennych zawartych rejestrach do postaci, która umożliwiałaby statystyczne przetwarzanie wartości zmiennych.

W ramach podejmowanych działań merytorycznych nad przygotowaniem rejestrów do wykorzystania w spisie zostały przygotowane min. następujące procedury:

- normalizacji zmiennych, której celem było doprowadzenie do jednolitego zapisu tych samych zmiennych pochodzących z różnych rejestrów;
- reguły kontroli i korekty, której celem było sprawdzenie poprawności oraz powiązań logicznych między zmiennymi w ramach jednego rejestru lub systemu informacyjnego;
- synchronizacji – niektóre rejestry zawierały informacje o faktach/zdarzeniach dotyczących osób, które należało przetworzyć i przekształcić w informacje o osobie;
- konwersji, która miała na celu doprowadzenie do ujednolicenia klasyfikacji danej zmiennej pochodzącej z różnych rejestrów niezależnie od oznaczenia jakie było stosowane w rejestrze;
- procedury deduplikacji zmiennych z rejestrów administracyjnych.

[2] Każde źródło informacyjne tj. rejestry, dane pozyskane z formularzy CAxI posiada swoje miejsce w Operacyjnej Bazie Mikrodanych, jako odrębna warstwa danych źródłowych. Integracja danych w bazie OBM polega na łączeniu warstw według identyfikatorów lub kluczowych atrybutów.

Działaniem wspierającym był tzw. słownik zmiennych logicznych, który zawierał wystandaryzowane nazwy zmiennych z wskazaniem źródła pochodzenia danej. W oparciu o tak przygotowane dane opracowywane były w dalszych etapach algorytmy wyliczania poszczególnych zmiennych złotego rekordu.

Procedury deduplikacji ankiet zebranych metodami CAxI

Ze względu na przyjęte rozwiązania w badaniu reprezentacyjnym i badaniu pełnym osoba mogła zostać spisana więcej niż jeden raz. Przyjęto przy tym założenie, że rozpatrywane ankiety są kompletne i zamknięte, co oznacza, że dla każdej ankiety osobowej uzyskano wszystkie informacje wynikające ze „ścieżek przejścia" w formularzu spisowym.

Procedury wyboru ankiety zostały przeprowadzone po etapie weryfikacji i ustaleniu kompletności osób w oparciu o informacje zawarte w wykazie adresowo-mieszkaniowym po spisie oraz po deduplikacji mieszkań. Proces ten polegał na wyborze zestawu „najlepszych" ankiet osobowych dla danego mieszkania tzw. „paczki osobowej" („paczka osobowa" rozumiana, jako zestaw ankiet osobowych powiązanych z danym mieszkaniem – spisanych w danym mieszkaniu). Po wybraniu odpowiednich ankiet na ich podstawie zbudowane zostały warstwy CAxI zawierające informacje pozyskane z formularzy spisowych na potrzeby wyliczania zmiennych wynikowych wyspecyfikowanych w złotym rekordzie.

Proces wyboru najlepszych ankiet dla warstw CAxI zakładał, że w zbiorze wejściowym, którym jest zbiór wszystkich ankiet w podsystemie CAxI, dostępnych było wiele ankiet mieszkaniowych dla jednego mieszkania oraz wiele ankiet osobowych dla tej samej osoby. W rezultacie wykonania procesu, na który składały się:

- procedura wyboru ankiety mieszkaniowej;
- procedura wyboru ankiety osobowej z badania reprezentacyjnego (a dokładniej wybór najlepszej „paczki osobowej" dla danego ID_MIESZKANIA);
- procedura wyboru ankiety osobowej z badania pełnego;
- procedura wyboru ankiety osobowej w przypadku występowania ankiety z badania pełnego i reprezentacyjnego;

w warstwach CAxI znalazła się jedna, najlepsza ankieta dla mieszkania i osoby.

Szczególne znaczenie miało opracowanie procedur wyznaczenia najlepszej ankiety z badaniu reprezentacyjnego. Procedury te obejmowały wykonanie algorytmów korekty oraz obliczenie wskaźników wymaganych na późniejszych etapach selekcji ankiet. Opracowane algorytmy wyliczały 13 wskaźników szczegółowych oraz 2 wskaźniki globalne, charakteryzujące cały rekord danych o osobach w warstwach CAxI ze spisu reprezentacyjnego.

Wskaźnik szczegółowy sprawdzał stan kompletności (niekompletność) danych dla każdego z 13 modułów w oparciu o zmienną lub zmienne zamykające (reprezentujące) dany moduł. Istota wskaźnika polegała na sprawdzeniu czy zmienna (zmienne) reprezentująca moduł posiadała w danym rekordzie wartość, zarówno przy założeniu, że powinna ją mieć według założonej ścieżki przejścia (czyli wartości innej zmiennej lub zmiennych implikujących posiadanie wartości przez daną zmienną), ale także wtedy gdy nie można było sprawdzić czy dana zmienna powinna mieć wartość (gdyż zmienne warunkujące ją mają braki danych).

Na podstawie wartości wskaźników szczegółowych wyliczone zostały dwa wskaźniki globalne, tj. liczba niekompletnych modułów oraz procent niekompletnych danych. Wartości wyliczonych wskaźników pozwoliły na wybór najbardziej wartościowej informacyjnie ankiety i przygotowanie zbioru unikalnych ankiet dla osoby i mieszkania.

Opracowywanie algorytmów wyprowadzania zmiennych

Dane dotyczące osób i mieszkań pozyskane w spisie 2011 na etapie opracowywania wyników tworzą dwa podstawowe zakresy informacyjne:

1) Z badania pełnego, obejmują dane odnoszące się do wszystkich mieszkańców kraju. Zakres informacji jest jednakże stosunkowo wąski pod względem tematycznym, co jest konsekwencją dostępności danych w systemach administracyjnych. Zloty rekord w części badania pełnego zawiera zmienne wyliczone na podstawie danych dostępnych w rejestrach i systemach informacyjnych oraz zweryfikowanych przez respondentów w samospisie internetowym (formularz krótki) i z badania reprezentacyjnego (formularz długi) – w części zakresowo wspólnej dla obu formularzy. Połączenie informacji z tych trzech źródeł pozwoliło przede wszystkim na ustalenie stanów i struktury ludności według płci i wieku;
2) Z badania reprezentacyjnego, pozyskane dla 20% ludności, zawierają szeroki pod względem przedmiotowym zakres danych, które po uogólnieniu na całą populację pozwalają na szeroka charakterystykę demograficzno-społeczną i ekonomiczną ludności.

Pierwszy etap wyliczenia wartości zmiennych spisowych odbywa się w środowisku operacyjnym (OBM), gdzie wyprowadzane zostają zmienne tzw. pierwotne. Algorytm każdej zmiennej wskazuje źródło pochodzenia zmiennej oraz matematyczną regułę jej wyliczania – bezpośrednie przeniesienie wartości z jednego lub z kilku źródeł lub „złożenie” wartości zmiennej z wartości kilku zmiennych.

Następny etap przetwarzania danych odbywa się w Analitycznej Bazie Mikrodanych (ABM), w której tworzone są i wyliczane kolejne zmienne, tzw. zmienne pochodne z poszczególnych obszarów tematycznych (wyliczane min. w oparciu o zmienne pierwotne) i obiekty pochodne takie jak gospodarstwa domowe i rodziny.

Wyliczanie tego typu zmiennych odbywa się w środowisku ABM ze względu na jego funkcjonalność umożliwiającą implementację bardziej skomplikowanych algorytmów. Algorytmy wyliczania zmiennych pochodnych polegają zarówno na tworzeniu agregatu zmiennej lub uzyskania ze złożenia z kilku zmiennych pierwotnych nowej zmiennej, jak również dokonania zakodowania poszczególnych typów gospodarstw domowych i rodzin, charakterystycznych ze względu na ich skład, np. gospodarstw domowych z osobami niepełnosprawnymi, emigrantami czy też osobami pracującymi, bezrobotnymi itp.

Zmienne pochodne pozwalają m.in. na łatwiejsze opracowywanie wyników, tworzenie tablic dla użytkowników oraz prowadzenie analiz.

Przyjęte założenia metodologiczne dla badania reprezentacyjnego pozwalały wybrać skrócony sposób spisania. Ta możliwość dotyczyła wyłącznie tych osób, które nie mieszkały pod danym adresem i nie zamierzały pod ten adres powrócić lub przypadków, gdy mieszkańcy nie byli w stanie odpowiedzieć na pytania z formularza spisowego odnoszące się do tych osób. W konsekwencji dla osób nieobecnych nie pozyskano odpowiedzi na większość pytań z formularza długiego. Jeżeli dana informacja była możliwa do pozyskania z rejestrów, to algorytmy wyliczania zmiennych taką informację uwzględniały. Pozostało jednak wiele braków danych, które w dalszych pracach analitycznych w miarę możliwości będą imputowane przy wykorzystaniu – mniej lub bardziej – zaawansowanych metod statystycznych.

10. Uogólnianie wyników badania reprezentacyjnego

Na etapie opracowania wyników badania reprezentacyjnego wyodrębnione zostały dwa etapy uogólniania (ważenia) wyników tego badania:

- etap I – ważenie przy zastosowaniu skorygowanych wag „z frakcji”, będących odwrotnością frakcji losowania,
- etap II – ważenie przy zastosowaniu skalibrowanych wag dla jednostek badania będących osobami.

Pierwotne wagi zostały wyznaczone jako odwrotność frakcji losowania dla 70,5 tys. warstw. Należy przypomnieć, że celem warstwowania było wyodrębnienie możliwie jednorodnych grup jednostek. W ramach każdej z warstw – wagi były identyczne. Pierwotne wagi z frakcji musiały zostać skorygowane z uwagi na 13,7% braki wypełnionych ankiet mieszkaniowych w badaniu reprezentacyjnym.

Wagi skorygowane, wyznaczone w I etapie są stosowane do uogólniania wyników spisu w zakresie mieszkań, gospodarstw domowych oraz rodzin. Natomiast dla budynków jednorodzinnych, w ramach których występują 1 lub 2 mieszkania, wyprowadzone zostały odrębne wagi.

Z uwagi na konieczność zintegrowania wyników badania reprezentacyjnego ze spisem pełnym (w zakresie podstawowych zmiennych dotyczących: płci, wieku oraz miejsca zamieszkania – poziom powiatu z wyodrębnieniem części miejskiej i wiejskiej) zaistniała potrzeba wyprowadzenia skalibrowanych wag dla poszczególnych osób.

Kalibracja jest metodą odpowiedniego doboru wag, tak aby zostały zrekompensowane straty informacji związane z występującymi brakami odpowiedzi. W NSP 2011 kalibracja miała na celu dostosowanie struktur płci i wieku pozyskanych w badaniu reprezentacyjnym do struktur ludności wg płci i wieku ustalonych w spisie pełnym, którego wyniki stanowiły populację referencyjną (odniesienia).

Kalibracja wag została wykonana przy wykorzystaniu programu CALMAR (Calibration on Margins) przez przedstawicieli środowiska naukowego Uniwersytetu Ekonomicznego w Poznaniu[3].

[3] Materiał roboczy: „Raport z opisem wyników z zakresu możliwości wykorzystania kalibracji na potrzeby korygowania wag w złotym rekordzie" pod kierunkiem prof. dr hab. Jana Paradysza.