WIADOMOŚCI STATYSTYCZNE

PL ISSN 0043-518X
e-ISSN 2543-8476

Rok LXIII
12 (691) 2018, 5—22

STUDIA METODOLOGICZNE

Janusz DYGASZEWICZ
Bolesław SZAFRAŃSKI

Badania statystyczne — ujęcie modelowe

Streszczenie. *Doświadczenia zarówno z obszaru zadań badawczo-rozwojowych, jak i z przedsięwzięć projektowo-wdrożeniowych dotyczących wsparcia informatycznego produkcji statystycznej wskazują na zbyt małe, w stosunku do potencjalnych możliwości, wykorzystanie rozwiniętych metod modelowania matematycznego. Celem opracowania jest wykazanie, że efekty modelowania matematycznego w dziedzinie badań statystycznych mogą nie tylko przyczynić się do podniesienia efektywności przetwarzania danych w statystyce publicznej, lecz także wpływać na jakość wymagań funkcjonalnych odnośnie do wsparcia informatycznego badań statystycznych. Cel ten zrealizowano poprzez omówienie ogólnego modelu matematycznego badania statystycznego, ze szczególnym uwzględnieniem podstawowych faz produkcji statystycznej (zbierania, przetwarzania, analizy i udostępniania danych statystycznych), a także poprzez wskazanie zadań optymalizacyjnych i korzyści z nich wynikających w przypadku problemów, które mogą występować w procesie projektowania wsparcia informatycznego. Dla potwierdzenia użyteczności przedstawionego podejścia zaprezentowano — w postaci diagramu UML — koncepcję integracji efektów modelowania matematycznego i tradycyjnego projektowania wsparcia informatycznego.*

Słowa kluczowe: badanie statystyczne, modelowanie matematyczne, wsparcie informatyczne.

JEL: C02, C18

Współczesna statystyka publiczna to złożony układ organizacyjno-techniczny produkcji statystycznej, służącej realizacji założonego celu badań statystycznych. Produkcję tę wspierają metody i narzędzia informatyczne. Analiza wyma-

gań integracyjnych statystyki publicznej oraz rozpoznanie zaleceń metodycznych dla projektowania systemów informatycznych (Kisielnicki, 2017) uzasadniają potrzebę prowadzenia prac mających na celu zwiększenie roli modelowania matematycznego w kształtowaniu architektury oraz zasad działania mechanizmów, które wpływają na efektywność funkcjonowania środowiska produkcji statystycznej. Należy podkreślić, że zarówno w przypadku zadań badawczo-rozwojowych, jak i przedsięwzięć projektowo-wdrożeniowych dotyczących wsparcia informatycznego obserwuje się zbyt małe w stosunku do możliwości wykorzystywanie rozwiniętych metod modelowania matematycznego oraz dorobku inżynierii systemowej w metodach zarządzania projektami wytwarzania wsparcia informatycznego[1]. Przyczyną takiego stanu rzeczy jest przede wszystkim to, że stosowane metody projektowania systemów informatycznych nie zawierają skutecznych mechanizmów absorpcji efektów uzyskanych dzięki budowie i badaniu modeli matematycznych. Nie ma bowiem uznanego języka komunikacji między obszarami modelowania matematycznego a obszarem metodyk zarządzania projektami informatycznymi[2].

W ogólnym przypadku można zauważyć, że efekty modelowania matematycznego procesów zachodzących w systemie statystyki publicznej mogą być użyteczne zarówno na poziomie ogólnym (np. w trakcie tworzenia lub weryfikacji założeń, ograniczeń i wymagań formułowanych w stosunku do organizacji badań statystycznych), jak i na poziomie szczegółowych decyzji projektowych (np. w trakcie podejmowania decyzji dotyczących identyfikacji doboru kanałów gromadzenia danych). Na poziomie ogólnym efekty modelowania matematycznego wpływają na kształt wymagań funkcjonalnych, podczas gdy na poziomie szczegółowym mogą stanowić wsparcie dla zespołu projektowego w podejmowaniu decyzji projektowych, mających znaczenie dla spełnienia kryteriów czasowych (np. czasu udostępnienia danych), jakościowych, niezawodnościowych lub kosztowych procesu produkcji statystycznej. W artykule skupiono się na kryteriach kosztowych i wskazano potrzebę uwzględnienia w procesach projektowania badań statystycznych wyników uzyskanych np. na drodze rozwiązywania praktycznych zagadnień optymalizacyjnych (Chudy, 2014). Dzięki takiemu podejściu w tworzeniu koncepcji realizacji badań statystycznych możliwe jest spójne wykorzystanie wiedzy pochodzącej z dwóch źródeł — od praktyków o wieloletnim doświadczeniu w statystyce publicznej oraz od analityków, którzy formułują i rozwiązują problemy z wykorzystaniem modelowania matematycznego.

Biorąc powyższe pod uwagę, głównym celem artykułu jest wykazanie, że efekty modelowania matematycznego w dziedzinie badań statystycznych mogą

[1] Odnosi się to nawet do najnowszych i zaawansowanych metod. Można także zauważyć, że programy szkoleniowe ich dotyczące nie obejmują zagadnień wykorzystania modelowania matematycznego w procesach projektowania systemów informatycznych.

[2] Dodatkową barierą jest niewątpliwie to, że obok specjalistów z dziedziny statystyki matematycznej w planowaniu masowych badań statystycznych nie uczestniczą specjaliści z doświadczeniem w wykorzystywaniu modelowania matematycznego do optymalizacji infrastruktury realizacji tych badań, czyli architektury platform informatycznego wsparcia badań statystycznych.

nie tylko przyczynić się do podniesienia efektywności przetwarzania danych w statystyce publicznej, lecz także wpływać na jakość wymagań funkcjonalnych odnośnie do wsparcia informatycznego tych badań. Cel zrealizowano poprzez omówienie ogólnego modelu matematycznego badania statystycznego, z uwzględnieniem podstawowych faz produkcji statystycznej[3], a także poprzez wskazanie zadań optymalizacyjnych i korzyści z nich wynikających w przypadku problemów, jakie mogą występować w procesie projektowania wsparcia informatycznego.

MODEL ZAKRESU BADANIA STATYSTYCZNEGO

Pierwszym elementem modelu badania statystycznego jest zbiór obiektów[4], które podlegają badaniu:

$$\mathbf{O} = \{o_1, o_2, \dots, o_n, \dots, o_N\}$$

Elementami zbioru $\mathbf{O}$ w badaniach statystycznych są zwykle konkretne osoby fizyczne bądź prawne, np. instytucje, pojedyncze podmioty gospodarcze i inne rzeczywiste byty podlegające badaniom statystycznym. Modelowo przyjmuje się, że skład zbioru $\mathbf{O}$ jest określony. W rzeczywistości w przypadku niektórych rodzajów badań statystycznych (np. spisów powszechnych czy badań reprezentacyjnych) liczność obiektów może być nieznana, a nawet ustalana dopiero na podstawie prowadzonego badania. W niniejszym artykule nie rozważa się zagadnień doboru próby podlegającej badaniu, np. jej losowania.

Obiekty poddawane badaniu statystycznemu są różnych typów i tworzą klasy obiektów. Zbiór klas obiektów danego badania statystycznego definiuje się jako:

$$\mathbf{K} = \{k_1, k_2, \dots, k_m, \dots, k_M\}$$

Elementami zbioru $\mathbf{K}$ są takie kategorie pojęć (metadanych), jak przykładowo: osoba fizyczna, osoba fizyczna prowadząca działalność gospodarczą, emeryt, placówka oświatowa, bank itp.

Pierwsza decyzja definiująca zakres badania statystycznego dotyczy wskazania klas obiektów, które mają być objęte badaniem. Dla uproszczenia, z uwagi na koncepcyjny charakter modelu, przyjęto, że zbiory obiektów różnych klas są

[3] Pod pojęciem procesu produkcji statystycznej należy rozumieć całokształt działalności polegającej na projektowaniu badania statystycznego, zbieraniu danych i ich przetwarzaniu oraz uzyskiwaniu wyników. Pod pojęciem wsparcia informatycznego badania statystycznego rozumiemy natomiast metody i narzędzia informatyki wykorzystywane w poszczególnych fazach procesu produkcji statystycznej, zaprojektowane i wykonane zgodnie ze specyfikacją wymagań poszczególnych faz procesu produkcji statystycznej.

[4] Dla podniesienia czytelności formalne zapisy nazw zbiorów zostały pogrubione.

rozłączne. Nie ograniczy to ogólności rozważań, pod warunkiem że obiekty modelowe ze zbioru $\mathbf{O}$ dotyczące tych samych bytów rzeczywistych będą wyspecyfikowane odrębnie dla poszczególnych klas. Praktycznie wywoła to tylko taki skutek, że dany byt rzeczywisty objęty badaniem będzie mógł wystąpić w zbiorze $\mathbf{O}$ nawet wielokrotnie, np. pierwszy raz jako konkretna osoba fizyczna i drugi raz jako konkretna osoba fizyczna prowadząca działalność gospodarczą.

Konsekwencją decyzji o zaliczeniu danych klas do badania będzie opisanie w jego wynikach odpowiednich obiektów wyselekcjonowanych na podstawie funkcji α:

$$\alpha: \mathbf{O} \rightarrow \mathbf{K}$$

Druga decyzja (po uprzednim wskazaniu klas obiektów), precyzująca zakres badania statystycznego, polega na zdefiniowaniu zbioru wynikowych cech informacyjnych badania dla poszczególnych klas (inaczej — atrybutów klas obiektów).

Zakłada się, że wszystkie rozważane cechy informacyjne (w ujęciu metadanych, czyli atrybutów badanych klas obiektów, a nie ich wartości) tworzą zbiór $\mathbf{Y}$**:**

$$\mathbf{Y} = \{y_1, y_2, \dots, y_i, \dots, y_I\}$$

W konkretnym badaniu decyduje się, które cechy będą podlegały badaniu dla poszczególnych klas[5], czyli definiowana jest funkcja β:

$$\beta: \mathbf{K} \rightarrow 2^{\mathbf{Y}}$$

co oznacza, że każdej klasie ze zbioru $\mathbf{K}$ przyporządkowany jest podzbiór zbioru $\mathbf{Y}$**:**

$$\forall_{k_m \in \mathbf{K}} \beta(k_m) \subset \mathbf{Y}$$

Nie zakłada się rozłączności cech dla różnych klas. Przykładowo cecha *Nr PESEL* może być cechą zarówno dla klasy *osoba fizyczna*, jak i dla klasy *osoba fizyczna prowadząca działalność gospodarczą*.

Na podstawie przyjętych definicji α i β można zauważyć, że:

$$\forall_{o_n \in \mathbf{O}} \beta\big(\alpha(o_n)\big) \subset \mathbf{Y}$$

przy czym:

$$\forall_{o_{n_1}, o_{n_2} \in \mathbf{O}} \left(\alpha(o_{n_1}) = \alpha(o_{n_2})\right) \Rightarrow \left(\beta\left(\alpha(o_{n_1})\right) = \beta\left(\alpha(o_{n_2})\right)\right)$$

[5] Zbiór potęgowy zbioru $\mathbf{Y}$**,** umownie zapisywany w postaci $2^{\mathbf{Y}}$, to zbiór wszystkich jego podzbiorów.

czyli w przypadku różnych obiektów należących do tej samej klasy badane cechy są identyczne, ponieważ decyzja o zakresie badania statystycznego odnosi się nie do pojedynczych obiektów, lecz do ich klas.

Wyniki badania statystycznego tworzą zbiór **W:**

$$\mathbf{W} = \{w_1, w_2, \dots, w_j, \dots, w_J\}$$

Zbiór wyników jest podzbiorem iloczynu kartezjańskiego:

$$\mathbf{W} \subset \mathbf{O} \times \mathbf{Y} \times \mathbf{M} \times \mathbf{R}$$

gdzie:

$\mathbf{O} = \{o_1, o_2, \dots, o_n, \dots, o_N\}$ — zbiór obiektów podlegających badaniu statystycznemu,
$\mathbf{Y} = \{y_1, y_2, \dots, y_i, \dots, y_I\}$ — zbiór cech informacyjnych badania statystycznego,
$\mathbf{M} = \{\mu_1, \mu_2, \dots, \mu_s, \dots, \mu_S\}$ — zbiór miar wartości cech,
$\mathbf{R} = \{r_1, r_2, \dots, r_t, \dots, r_T\}$ — zbiór cech wyrażonych w odpowiednich jednostkach miar.

Proces pozyskania danych wejściowych i ich różnorakich transformacji ma prowadzić do zdefiniowania funkcji częściowej γ, która określa poszukiwane wyniki badania statystycznego zgodnie z jego celem:

$$\gamma\colon \mathbf{O} \times \mathbf{Y} \overset{/}{\rightarrow} \mathbf{W} \subset \mathbf{O} \times \mathbf{Y} \times \mathbf{M} \times \mathbf{R}$$

$$\gamma\left(o_{n_j}, y_{i_j}\right) = w_j = \left(o_{n_j}, y_{i_j}, \mu_{s_j}, r_{t_j}\right)$$

Stąd pojedyncze wyniki badania tworzą czwórki uporządkowane:

$$w_j = \left(o_{n_j}, y_{i_j}, \mu_{s_j}, r_{t_j}\right)$$

przy czym spełniony jest warunek, że dla danej pary $\left(o_{n_j}, y_{i_j}\right)$ stosuje się tylko jedną miarę wartości oraz że:

$$\forall_{w_j \in \mathbf{W}} \left(r_{t_j} \text{ wyrażone jest w jednostkach miary } \mu_{s_j}\right)$$

Funkcja częściowa γ jest określona dla pary (o_n, y_i) jedynie wtedy, gdy:

$$y_i \in \beta\left(\alpha(o_n)\right)$$

czyli funkcja częściowa γ jest określona jedynie dla niektórych par (obiekt, cecha); wyjątkiem byłoby, gdyby badanie obejmowało dla wszystkich klas $\mathbf{K}$ obiektów $\mathbf{O}$ wszystkie cechy zbioru $\mathbf{Y}$ lub gdyby opisywano jednorodne badanie statystyczne obejmujące jedną klasę.

Podsumowując, pojedynczy wynik badania statystycznego zawiera identyfikatory obiektu i cechy, które podlegają badaniu, jednostkę miary oraz wyrażoną w tej jednostce miary zidentyfikowaną wartość cechy. W takim rozumieniu, jak omówiono to powyżej, liczność zbioru wyników badania statystycznego $\mathbf{W}$ jest równa:

$$|\mathbf{W}| = \sum_{n=1}^{N} \left|\beta\left(\alpha(o_n)\right)\right|$$

KANAŁY UZYSKIWANIA, TRANSFORMACJI I UDOSTĘPNIANIA DANYCH

Model uzyskiwania danych wejściowych

W artykule używa się określenia *kanały uzyskiwania danych wejściowych*, które oznacza zarówno źródła istniejących danych wejściowych, jak i metody ich uzyskania na rzecz danego badania statystycznego. Badanie ma na celu ustalenie — przy wykorzystaniu wszystkich możliwych kanałów — wartości wszystkich cech dla wszystkich obiektów, które zaliczono do zakresu tego badania. Warto zauważyć, że osiągnięcie tak postawionego celu jest w praktyce trudne (Stefanowicz, 2004). Często zachodzi ponadto konieczność posłużenia się niestandardowymi metodami uzupełniania wyników. Zastosowanie znajdują tu metody interpolacji matematycznej, imputacji, szacowania, prognozowania statystycznego czy ostatecznie metody eksperckie.

Potencjalne kanały uzyskiwania danych wejściowych badania statystycznego tworzą zbiór:

$$\mathbf{U} = \{u_1, u_2, \dots, u_l, \dots, u_L\}$$

Tymi kanałami można zdobyć dane wejściowe o obiektach tworzących zbiór $\mathbf{O}$, które należą do klas wyodrębnionych w zbiorze $\mathbf{K}$.

Uzyskiwane cechy informacyjne (analogicznie jak w odniesieniu do zbioru $\mathbf{Y}$ — w ujęciu metadanych, czyli atrybutów badanych klas obiektów, a nie ich wartości) tworzą zbiór $\mathbf{X}$:

$$\mathbf{X} = \{x_1, x_2, \dots, x_l, \dots, x_L\}$$

Cechy danych wejściowych tworzących zbiór $\mathbf{X}$ nie są w ogólnym przypadku tożsame z cechami informacyjnymi wynikowymi, tworzącymi zbiór $\mathbf{Y}$. Cechy

wynikowe są produktem transformacji cech wejściowych. W części modelu obejmującej jedynie fazę uzyskiwania danych wejściowych można przyjąć, że zbiór $\mathbf{X}$ jest tożsamy ze zbiorem $\mathbf{Y}$. Rozważa się tu bowiem jedynie zagadnienie wyboru kanałów uzyskiwania danych. Omawiany w tej części model matematyczny badania statystycznego nie obejmuje faz procesu badawczego, w których następuje przetwarzanie cech zbioru danych wejściowych $\mathbf{X}$ w cechy zbioru danych wynikowych $\mathbf{Y}$. W rezultacie pojęcie zbioru $\mathbf{X}$ nie będzie tu używane, a cechy wejściowe i wynikowe będą reprezentowane przez zbiór $\mathbf{Y}$. Konsekwentnie, wartości pozyskiwanych cech informacyjnych tworzą zbiór $\mathbf{W}$.

Możliwości informacyjne poszczególnych kanałów określa funkcja λ:

$$\lambda: \mathbf{O} \times \mathbf{Y} \to 2^{\mathbf{U}}$$

Funkcja λ może być częściowa, co znaczy, że nie dla każdej pary (o_n, y_i) jest określona, ale jeżeli parę (o_n, y_i) zakwalifikowano do zakresu danego badania statystycznego, to:

$$y_i \in \beta\big(\alpha(o_n)\big), \text{ to } \lambda(o_n, y_i) \neq \Phi$$

Powyższy warunek oznacza, że dla każdej pary $\big((\text{obiekt}, \text{cecha}) \equiv (o_n, y_i)\big)$ zakwalifikowanej do zakresu badania statystycznego istnieje co najmniej jeden kanał, z którego można pozyskać wartość cechy y_i obiektu o_n.

Warto nadmienić, że w niektórych sytuacjach rzeczywistych, np. gdy żaden z kanałów nie ma zdolności udostępnienia wartości cechy informacyjnej dla obiektu, można rozważać zastosowanie odmiany modelu, w której dane zdobywa się metodami niestandardowymi. Zróżnicowanie kanałów badania statystycznego może prowadzić do wyróżnienia takich odmian modelu matematycznego, jak:

- z kanałem dominującym (model może dostarczyć wszystkie dane);
- z kanałami równorzędnymi (model ze zróżnicowanymi kosztami);
- z uwzględnieniem niepewności danych;
- z weryfikacją porównawczą danych.

Model przydziału kanału uzyskiwania danych wejściowych

Decyzja o przydziale kanału uzyskiwania danych wejściowych w badaniu statystycznym ma postać:

$$\tau: \mathbf{O} \times \mathbf{Y} \to \mathbf{U}$$

przy czym:

- funkcja częściowa τ jest określona dla wszystkich par (o_n, y_i) zakwalifikowanych do zakresu danego badania statystycznego, czyli gdy $y_i \in \beta\big(\alpha(o_n)\big)$.

Innymi słowy, istnieje możliwość uzyskania wszystkich wartości wynikających z ustalonego zakresu badania statystycznego;

- przydział kanału do uzyskania wartości dla pary (o_n, y_i) może nastąpić jedynie wtedy, gdy jest to możliwe do zrealizowania w tym kanale, czyli:

$$\forall_{o_n \in \mathbf{O}^\tau} \left(o_n, \beta\left(\alpha(o_n)\right)\right) \in \lambda\left(o_n, \beta\left(\alpha(o_n)\right)\right)$$

Model transformacji danych

Badane cechy informacyjne tworzą zbiór $\mathbf{Y}$**:**

$$\mathbf{Y} = \{y_1, y_2, \ldots, y_i, \ldots, y_I\}$$

Cechy danych wejściowych tworzących zbiór $\mathbf{X}$ nie są w ogólnym przypadku tożsame z cechami informacyjnymi wynikowymi tworzącymi zbiór $\mathbf{Y}$**.** Cechy wynikowe stanowią produkt transformacji cech wejściowych.

Po uzyskaniu wartości cech informacji wejściowych $\mathbf{X}$ badania statystycznego uruchomiony zostaje proces transformacji, w wyniku którego otrzymywany jest zbiór cech informacji wynikowych $\mathbf{Y}$ tego badania. W procesie transformacji można potencjalnie wykorzystać niektóre z metod tworzących zbiór $\boldsymbol{\Omega}$**:**

$$\boldsymbol{\Omega} = \{\omega_1, \omega_2, \ldots, \omega_f, \ldots, \omega_F\}$$

Funkcjonalność metod transformacji określona jest funkcją σ:

$$\sigma: \mathbf{Y} \to 2^{\boldsymbol{\Omega}}$$

czyli dla wynikowej cechy informacyjnej y_i ze zbioru $\mathbf{Y}$ można zastosować jedną z $\sigma(y_i)$ metod transformacji $\left(\forall_{y_i \in \mathbf{Y}} \sigma(y_i) \subset \boldsymbol{\Omega}\right)$.

W ogólności sformułowane jest zapotrzebowanie na określone cechy informacji wejściowych, które trzeba uzyskać, aby w wyniku transformacji otrzymać daną cechę wynikową:

$$\upsilon: \mathbf{Y} \times \boldsymbol{\Omega} \to 2^{\mathbf{X}}$$

Funkcja υ jest funkcją częściową, określoną dla par $\left(y_i, \omega_f\right)$, gdy $\omega_f \in \sigma(y_i)$.

W tym miejscu — dla poprawienia czytelności formalnych zapisów — zastosowano pewne uproszczenie. Założono mianowicie, że dana informacyjna cecha wynikowa ze zbioru $\mathbf{Y}$ generuje zapotrzebowanie na związane z nią cechy informacyjne wejściowe ze zbioru $\mathbf{X}$**,** niezależnie od tego, jakiego obiektu ze zbioru $\mathbf{O}$ dotyczą. W rzeczywistości relacje te mogą być inne, ale można wówczas opisane metody traktować jako odmienne i uproszczenie nie zmieni szczegółowości rozważań.

W fazie transformacji danych wejściowych w wynikowe podejmuje się decyzję o formalnej postaci:

$$\pi: \mathbf{Y} \rightarrow \mathbf{\Omega}$$

czyli dla każdej cechy wynikowej wybiera się metodę transformacji.

Transformacja danych wejściowych wybraną metodą (z cechami informacyjnymi ze zbioru $\mathbf{X}$) w dane wynikowe (z cechami informacyjnymi ze zbioru $\mathbf{Y}$) generuje koszty także w fazie transformacji.

W zależności od koncepcji ich szacowania można zastosować bardziej lub mniej złożony moduł obliczeniowy, ale ważniejsze jest to, że poza kosztami ponoszonymi w fazie transformacji (które zależą bezpośrednio od jej metody), wybór metody przetwarzania ma pośredni wpływ na koszty ponoszone w fazie pozyskiwania danych wejściowych badania. Wynika to z wpływu wyboru metody na zakres koniecznych do uzyskania danych wejściowych. Określa to funkcja $\upsilon(\upsilon: \mathbf{Y} \times \mathbf{\Omega} \rightarrow 2^{\mathbf{X}})$. Dopiero po ustaleniu zakresu uzyskiwanych danych wejściowych możliwy jest wybór kanałów uzyskiwania danych i szacowanie ponoszonych kosztów.

Model udostępniania wyników

Udostępniane cechy informacyjne tworzą zbiór $\mathbf{Y}$:

$$\mathbf{Y} = \{y_1, y_2, \ldots, y_i, \ldots, y_I\}$$

W procesie udostępniania danych wynikowych badania statystycznego można wykorzystywać różne metody (kanały) udostępniania. Potencjalne kanały udostępniania i upowszechniania danych wynikowych danego badania statystycznego tworzą zbiór $\mathbf{Z}$:

$$\mathbf{Z} = \{z_1, z_2, \ldots, z_h, \ldots, z_H\}$$

Możliwości funkcjonalne poszczególnych kanałów definiuje funkcja η, określająca, które kanały udostępniania mogą służyć do upowszechnienia cechy informacyjnej y_i o obiekcie o_n:

$$\eta: \mathbf{O} \times \mathbf{Y} \rightarrow 2^{\mathbf{Z}}$$

Decyzja o przydziale kanału udostępniania danych wynikowych w badaniu statystycznym ma postać:

$$\varepsilon: \mathbf{O} \times \mathbf{Y} \rightarrow \mathbf{Z}$$

OSZACOWANIA KOSZTOWE I OPTYMALIZACJA BADAŃ STATYSTYCZNYCH

Wyróżniono trzy kategorie kosztów uzyskiwania danych wejściowych w badaniu statystycznym:

- koszt ogólny wykorzystania kanału — w przypadku użycia kanału do uzyskania jakiejś liczby wartości cech informacyjnych. W ujęciu formalnym koszt ogólny wykorzystania kanału u_l powstaje, gdy:

$$\exists_{(o_n,y_i)}\tau(o_n,y_i)=u_l$$

- koszt pośredni wykorzystania kanału do uzyskania jakiejś liczby wartości cech informacyjnych określonej klasy obiektów. W ujęciu formalnym koszt pośredni wykorzystania kanału u_l dla potrzeb klasy obiektów k_m powstaje, gdy:

$$\exists_{(o_n,y_i)}(\tau(o_n,y_i)=u_l)\wedge(\alpha(o_n)=k_m)$$

- koszt jednostkowy wykorzystania kanału do uzyskania wartości konkretnej cechy informacyjnej konkretnego obiektu. W ujęciu formalnym koszt jednostkowy wykorzystania kanału u_l w celu pozyskania wartości konkretnej cechy konkretnego obiektu (o_n,y_i) powstaje, gdy:

$$\tau(o_n,y_i)=u_l$$

Koszt całkowity w fazie uzyskiwania danych wejściowych to suma wszystkich wyróżnionych rodzajów kosztów poniesionych we wszystkich przydzielonych kanałach uzyskiwania danych wejściowych.

Koszty ponoszone w trzech najdroższych fazach badania statystycznego (uzyskiwania danych wejściowych, transformacji i udostępniania danych wynikowych) zależą — w ujęciu formalnym — od trzech funkcji:

- wyboru kanałów uzyskiwania danych wejściowych:

$$\tau\colon \mathbf{O}\times\mathbf{Y}\to\mathbf{U}$$

- wyboru metod transformacji danych:

$$\pi\colon \mathbf{Y}\to\mathbf{\Omega}$$

- wyboru kanałów udostępniania wyników:

$$\varepsilon\colon \mathbf{O}\times\mathbf{Y}\to\mathbf{Z}$$

Określenie funkcji τ, π i ε, przy których szacowany koszt jest najmniejszy (ogólniej: przy których osiąga się optimum), optymalizuje badanie statystyczne.

Zastosowanie matematycznego modelu optymalizacyjnego do wyboru najlepszych kanałów uzyskiwania danych wejściowych i udostępniania danych wynikowych oraz wyboru najlepszych metod transformacji danych wymagałoby kosztownych prac przygotowawczych. W związku z powyższym należy dokonać oszacowania, jakie korzyści można osiągnąć w wyniku optymalizacji prowadzenia badań — co uzasadniałoby racjonalność podjęcia prac przygotowawczych[6].

INTEGRACJA MODELOWANIA ARCHITEKTONICZNEGO I MATEMATYCZNEGO W PROCESIE ZARZĄDZANIA PROJEKTOWANIEM WSPARCIA INFORMATYCZNEGO

Relacje między tradycyjnym podejściem do planowania badania statystycznego a możliwymi do wykorzystania metodami i narzędziami modelowania matematycznego przedstawiono w postaci zapisanych notacją UML[7] diagramów aktywności oraz przypadków użycia (Wrycza, 2006). Zapewnia to przejrzystość i zwięzłość wniosków z przeprowadzonych rozważań na temat integracji modelowania architektonicznego i matematycznego oraz pozwala na syntetyczne przedstawienie płynących z nich rekomendacji.

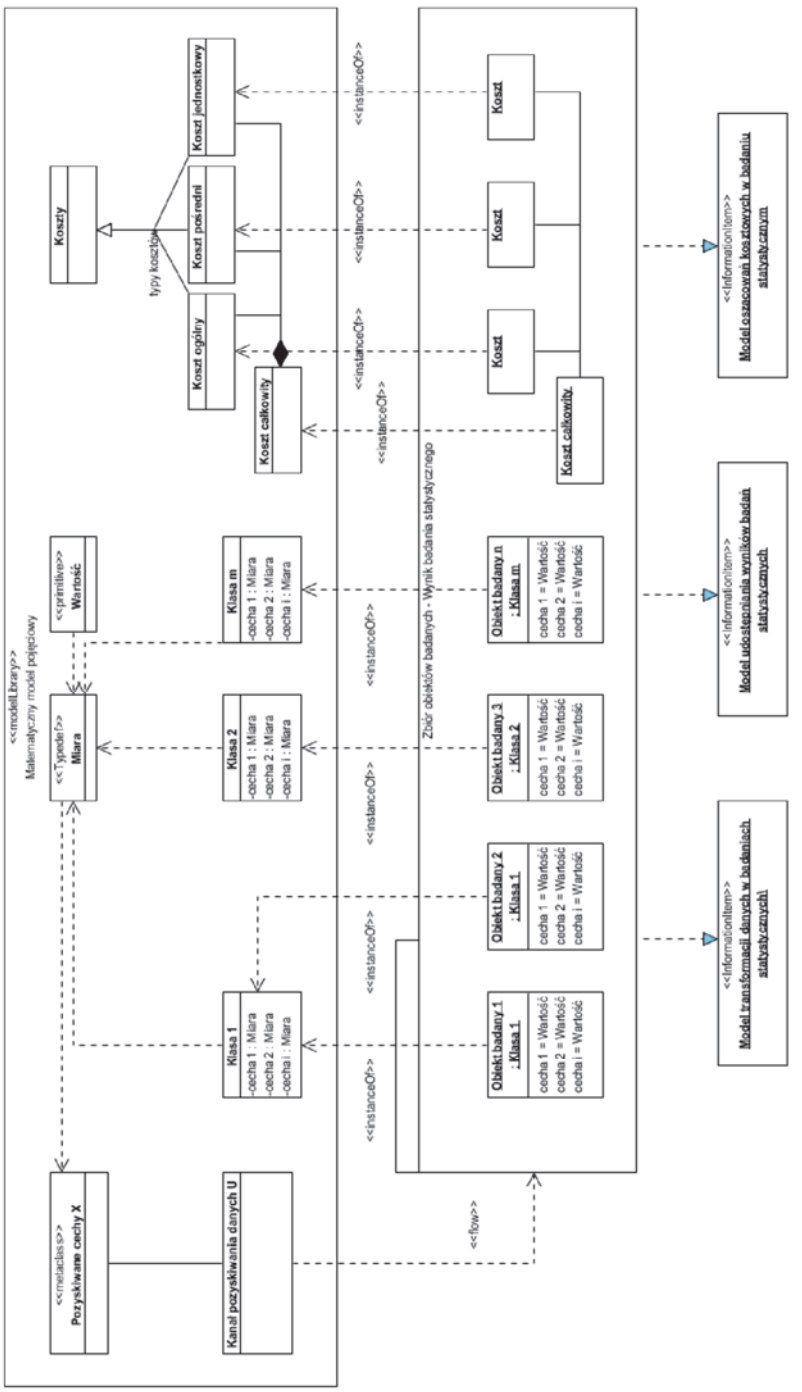
Diagramy 1 i 2 ilustrują strukturę wcześniej zaprezentowanych modeli matematycznych oraz zalecany sposób włączenia modelowania matematycznego do procesu projektowania i realizacji badań statystycznych. Oba diagramy wskazują na związki między metodami stosowanymi w modelowaniu matematycznym a niezbędnymi dla ich zastosowania składnicami danych. W praktyce oznacza to, że zespół projektowy planujący badanie statystyczne musi w warstwie koncepcyjnej przewidzieć zgromadzenie wiarygodnych reprezentatywnych danych niezbędnych do zastosowania np. określonej metody optymalizacji, a w warstwie technologicznej musi dysponować biblioteką modułów programowych implementujących przewidziane do wykorzystania metody optymalizacji. W przyszłości obok modeli wymienionych na diagramach 1 i 2 powinny pojawić się kolejne, zidentyfikowane w procesie analizy projektowania oczekiwanej funkcjonalności wsparcia informatycznego dla procesów produkcji statystycznej.

---

[6] W celu sprawdzenia słuszności przyjętej koncepcji podjęto próbę wstępnego oszacowania korzyści, jakie mogą wynikać z zastępowania tradycyjnych sposobów uzyskiwania danych przede wszystkim kanałami wykorzystującymi zawartość rejestrów administracyjnych. Założenia, przebieg i wyniki przeprowadzonych obliczeń zostaną, z uwagi na objętość, przedstawione w kolejnym artykule.

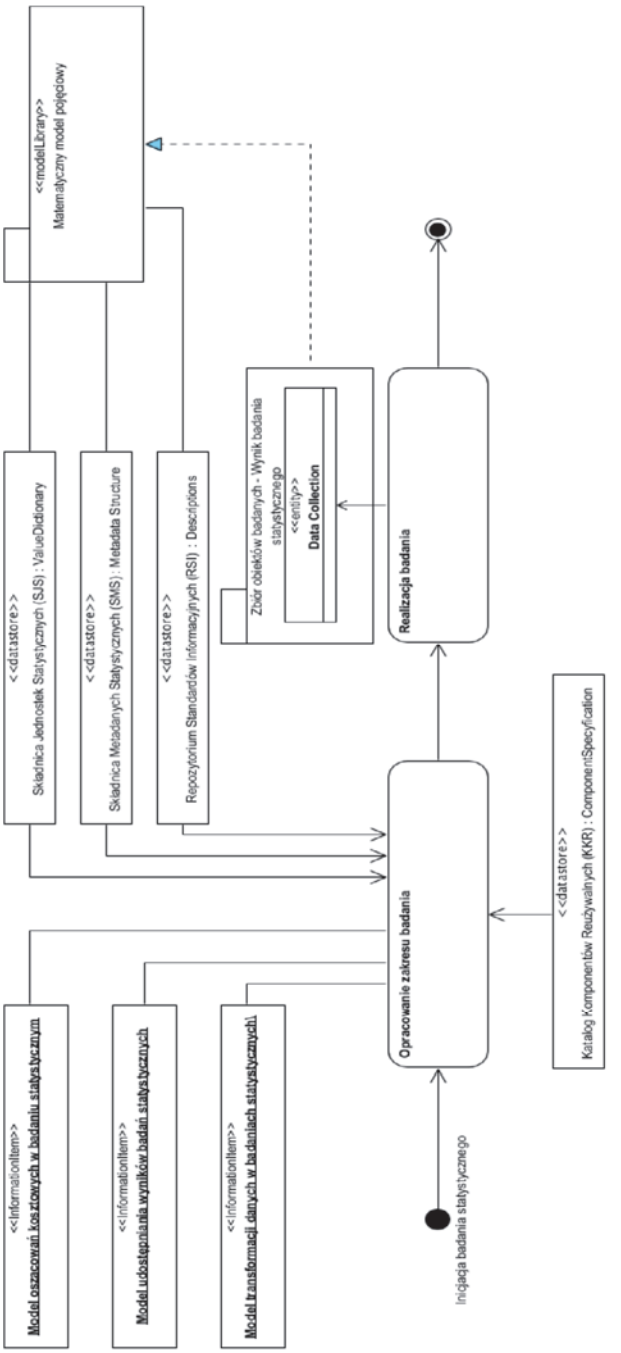
[7] UML (Unified Modeling Language) — notacja (język) służąca do modelowania wybranych fragmentów rzeczywistości, obecnie najczęściej na potrzeby tworzenia systemów informatycznych.

DIAGRAM 1. MODELOWANIE MATEMATYCZNE — STRUKTURA KLUCZOWYCH MODELI



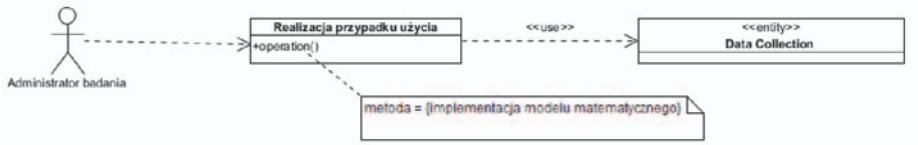
Ź r ó d ł o: opracowanie własne na podstawie: Dygaszewicz (2018).

DIAGRAM 2. INTEGRACJA MODELOWANIA ARCHITEKTONICZNEGO I MATEMATYCZNEGO W PROCESIE WSPARCIA INFORMATYCZNEGO — ZARZĄDZANIE PROJEKTEM



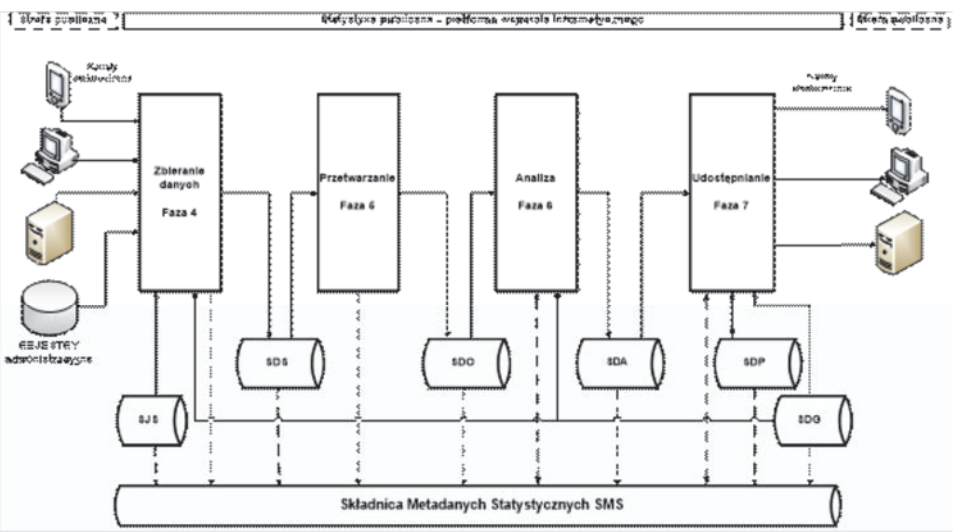
Ź r ó d ł o: jak przy diagramie 1.

DIAGRAM 3. METAMODEL INTEGRACJI MODELI ARCHITEKTONICZNYCH I MATEMATYCZNYCH



Ź r ó d ł o: jak przy diagramie 1.

DIAGRAM 4. RELACJA MODELOWANIA MATEMATYCZNEGO DO MODELU PROCESU PRODUKCJI STATYSTYCZNEJ MPPS



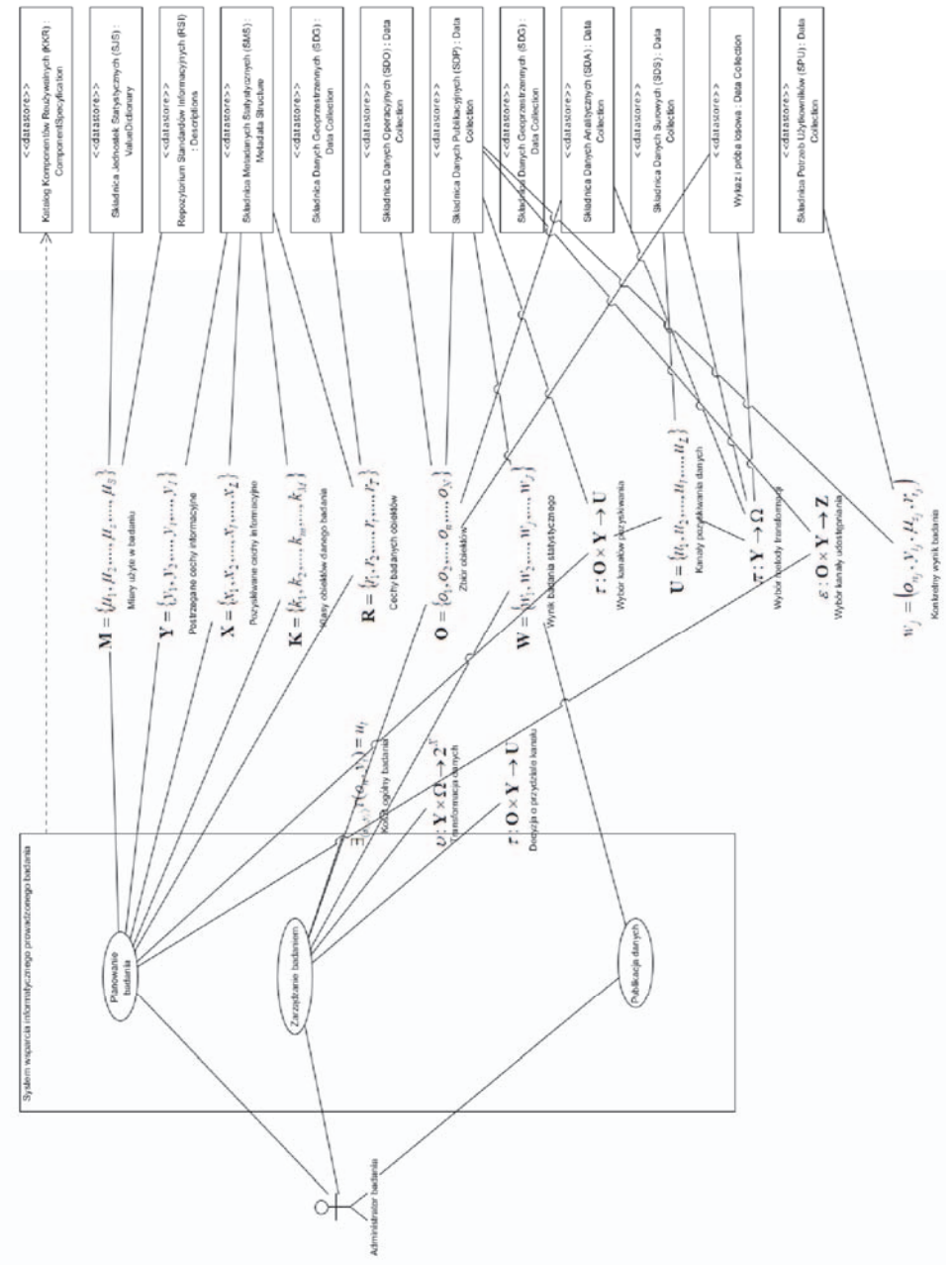
U w a g a. Składnica Metadanych Statystycznych — SMS, Składnica Danych Surowych — SDS, Składnica Danych Operacyjnych — SDO, Składnica Danych Analitycznych — SDA, Składnica Danych Publikacyjnych — SDP, Składnica Jednostek Statystycznych — SJS, Składnica Danych Geoprzestrzennych — SDG.

Ź r ó d ł o: jak przy diagramie 1.

Diagramy 3 i 4 odwołują się do ogólnego modelu matematycznego badania statystycznego. Diagram 3 w możliwie zwięzły sposób, w postaci metamodelu, ilustruje zagadnienie integracji modelowania architektonicznego i matematycznego. Diagram 4 jest jednym z możliwych diagramów przypadku użycia opisanego metamodelem przedstawionym na diagramie 3 w odniesieniu do Modelu Procesu Produkcji Statystycznej (MPPS) (Dygaszewicz, 2018), który jest polską implementacją Generycznego Modelu Procesów Statystycznych (Generic Statistical Business Proces Model — GSBPM)[8].

[8] https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0.

DIAGRAM 5. INTEGRACJA MODELI ARCHITEKTONICZNYCH I MODELI MATEMATYCZNYCH



Ź r ó d ł o: jak przy diagramie 1.

W diagramie 5 wykorzystano zapisy matematyczne zaczerpnięte z opracowanego modelu badania statystycznego. Zapisy te wprost ukształtowały strukturę diagramu i tym samym wpływają na architekturę tworzonego wsparcia informatycznego. Należy zwrócić uwagę, że tworzą one bazę pojęciową oraz generują wymagania dotyczące składnic danych zarówno co do typów, jak i wartości[9]. Ponadto jednoznacznie wskazują na potrzebę wytworzenia i włączenia do wsparcia informatycznego oprogramowania wykorzystywanego do rozwiązywania zidentyfikowanych i sformułowanych zagadnień optymalizacyjnych w celu uzyskania parametrów ilościowych istotnych przy projektowaniu architektury wsparcia informatycznego dla określonych procesów produkcji statystycznej zgodnych z MPPS.

Podsumowanie

Analiza wymagań integracyjnych statystyki publicznej oraz wiedza na temat metod projektowania systemów informatycznych uzasadniają potrzebę prowadzenia prac mających na celu zwiększenie roli modelowania matematycznego w kształtowaniu środowiska informatycznego wykorzystywanego w statystyce publicznej. Dzięki włączeniu modelowania matematycznego do procesu projektowania wsparcia informatycznego efekty modelowania matematycznego mogą stać się czynnikami oddziaływającymi bezpośrednio na projektowanie badań statystycznych realizowanych z wykorzystaniem metod i narzędzi informatyki, np. poprzez wspomaganie efektywnego doboru kanałów uzyskiwania lub udostępniania danych. Warunkiem szerokiego wykorzystania efektów modelowania matematycznego jest stworzenie na platformie informatycznej uniwersalnego repozytorium modeli i metod rozwiązywania metodami matematycznymi problemów występujących w planowaniu i prowadzeniu badań.

Szczególna przydatność wykorzystania modelowania matematycznego wystąpi wtedy, gdy tzw. rurkowe[10] podejście do organizacji (projektowania) badań statystycznych, w którym pojedyncze badania są wspierane przez dedykowane im odseparowane programy aplikacyjne, zostanie zastąpione przez zintegrowane podejście procesowe[11], wspierane przez spójną i zintegrowaną platformę informatyczną. Warunkiem szerokiego wykorzystania efektów modelowania matema-

[9] Opracowanie i prezentacja diagramów klas wykracza poza zakres niniejszego artykułu. Tym niemniej trzeba podkreślić, że rozwinięcie zarysowanego tu podejścia skutkowałoby opracowaniem diagramów klas dla każdej składnicy wymienionej na diagramie przypadku użycia. W diagramach klas co najmniej część wyróżnionych atrybutów byłoby tożsamych ze zmiennymi modelu matematycznego.

[10] Określenie *rurkowy* jest polskim odpowiednikiem angielskiego *stovepipe* — organizacja realizuje swe procesy biznesowe w sposób odseparowany, bez współdzielenia zasobów i rozwiązań, pomimo że istnieją technologiczne możliwości zintegrowania tych procesów.

[11] Podejście procesowe jest zgodne z rekomendacjami dotyczącymi GSBPM.

tycznego jest stworzenie w ramach tej platformy uniwersalnego repozytorium modeli i metod rozwiązywania problemów (w tym omówionych wcześniej zadań optymalizacyjnych) występujących w planowaniu i prowadzeniu badań statystycznych, opracowanie i wdrożenie odpowiednich programów szkoleniowych dotyczących m.in. zasad gromadzenia danych oraz wiedzy niezbędnej do efektywnego korzystania z tych modeli i metod. Należy podkreślić, że omówione w artykule wybrane badania ilościowe dotyczące kryteriów kosztowych mogą być rozszerzone w ramach tego samego ogólnego modelu badań statystycznych o badania stymulowane innymi kryteriami, np. jakościowymi lub czasowymi.

Spośród rozważanych kierunków dalszych prac badawczych w zakresie wykorzystania modelowania matematycznego w procesach projektowania i realizacji badań statystycznych najbardziej obiecujące wydają się:

- opracowanie ram architektonicznych dla platformy informatycznej wykorzystywanej w badaniach statystycznych, które stworzą warunki do efektywnego włączenia do platformy repozytorium modeli i metod rozwiązywania problemów metodami matematycznymi;
- powiązanie dalszych prac nad wykorzystaniem modelowania matematycznego z problemami wynikającymi z rozwoju metod i technologii określanych jako Big Data. Istotą tego rodzaju prac musi być tworzenie modeli i metod, które pozwolą uzyskiwać wiarygodne statystyki w przypadku otrzymywania danych nieustrukturyzowanych, niekompletnych i do tego pochodzących ze źródeł o wiarygodności trudnej do oszacowania.

dr inż. Janusz Dygaszewicz — *GUS*
dr hab. inż. Bolesław Szafrański — *profesor WAT*

LITERATURA

Chudy, M. (2014). *Wybrane algorytmy optymalizacji.* Warszawa: Akademicka Oficyna EXIT.

Dygaszewicz, J. (2018). *Modele i metody w procesie konstruowania ram architektonicznych dla informatycznego wsparcia masowych badań statystycznych,* rozprawa doktorska (niepubl.). Biblioteka Wojskowej Akademii Technicznej.

Kisielnicki, J. (2017). *Zarządzanie projektami badawczo-rozwojowymi.* Warszawa: Wydawnictwo Nieoczywiste, GAB Media.

Stefanowicz, B. (2004). *Informacja.* Warszawa: Wydawnictwo SGH.

Wrycza, S., Marcinkowski, B., Wyrzykowski, K. (2006). *Język UML 2.0 w modelowaniu systemów informatycznych.* Gliwice: Wydawnictwo Helion.

Summary. *Experiences, both in the area of research and development tasks, as well as those from the project-implementation undertakings concerning IT support of statistical production, indicate that the use of developed mathematical modelling methods is too small in relation to potential possibilities. The aim of*

the research is to demonstrate that the effects of mathematical modelling in the field of statistical research not only can contribute to the improvement of data processing efficiency in official statistics, but also affect the quality of functional requirements for IT support for statistical surveys. This objective was achieved by discussing the general mathematical model of statistical research, with particular emphasis on the basic phases of statistical production (collection, processing, analysis and dissemination of statistical data), as well as by indicating optimization tasks and benefits resulting from problems that may occur in the process of IT support design. In order to confirm the usefulness of the presented approach, the concept of integration of the effects of mathematical modelling and the traditional design of IT support was presented in the form of a UML diagram.

Keywords: statistical survey, mathematical modelling, IT support.