dr Marcin CZUPRYNA
Szkoła Główna Handlowa w Warszawie
e-mail: mczupr@sgh.waw.pl

dr Przemysław SZUFEL
Szkoła Główna Handlowa w Warszawie
e-mail: pszufe@sgh.waw.pl

dr hab. Bogumił KAMIŃSKI
Szkoła Główna Handlowa w Warszawie
e-mail: bkamins@sgh.waw.pl

mgr Anna WIERTLEWSKA
Szkoła Główna Handlowa w Warszawie
e-mail: awiert@sgh.waw.pl

DOI: **10.15290/ose.2017.03.87.03**

O ESTYMACJI PREFERENCJI W SZTUCZNYCH SIECIACH SPOŁECZNYCH[1]

Streszczenie

W artykule rozważano scenariusz, w którym administracja publiczna wykorzystuje internetową platformę społecznościową do komunikacji z obywatelami i uzyskiwania informacji o ich preferencjach. Z platformy tej korzysta tylko część całej populacji (subpopulacja), co powoduje, że preferencje obserwowane na platformie mogą być niereprezentatywne dla całego społeczeństwa. W niniejszym opracowaniu uwzględniono dwa problemy związane z brakiem reprezentatywności preferencji, tj.: (1) odmienną strukturę demograficzną populacji i subpopulacji oraz (2) różnice w procesie dynamiki preferencji w całej populacji i subpopulacji wyrażającej swoje opinie na platformie społecznościowej.

Dane wykorzystane w analizie obejmują informacje o aktywności użytkowników na platformie społecznościowej, ich dane socjodemograficzne oraz dane o populacji pochodzące ze spisu powszechnego. W celu badania dynamiki preferencji skonstruowano wieloagentowy model symulacyjny, w którym sieć społeczną przedstawiono za pomocą nieskierowanego grafu, gdzie węzły reprezentują obywateli, a łuki ich relacje społeczne.

[1] Niniejsze prace badawcze zostały zrealizowane w ramach projektu ROUTE-TO-PA (*Raising Open and User-friendly Transparency-Enabling Technologies for Public Administrations*) [http://routetopa.eu/], który jest finansowany ze środków Europejskiego Programu w Zakresie Badań Naukowych i Innowacji „Horizon 2020" na podstawie umowy o dotację nr 645860. Autorzy wyrażają również podziękowanie anonimowym recenzentom za ich uwagi dotyczące treści artykułu.

W procesie analizy najpierw jest generowana sztuczna populacja i na niej jest symulowana dynamika preferencji. Następnie losowo, metodą kuli śnieżnej (ang. *snowballsampling*) są wybierane różne niereprezentatywne subpopulacje, na których są testowane algorytmy uogólniania preferencji przez odtwarzanie dynamiki całej populacji. Miarą jakości modelu jest zgodność preferencji między subpopulacją a całą populacją. Rezultaty przeprowadzonych symulacji wskazały na skuteczność zastosowanej metody: wraz z kolejnymi krokami symulacji wzrasta zgodność między populacją rzeczywistą a syntetyczną. Okazało się również, że najistotniejszymi determinantami błędów uogólniania preferencji są model dyfuzji preferencji oraz waga opinii własnej agenta.

Słowa kluczowe: dynamika preferencji, modelowanie sieci społecznych, symulacje wieloagentowe

PREFERENCE ELICITATION IN SYNTHETIC SOCIAL NETWORKS

Summary

The paper considers a scenario in which public administration (PA) uses an online social platform to collect information on citizens' preferences. However, the opinions of the sub-population that uses the online platform might be not representative. The author develops a method for generalization of the dynamics of the preferences observed on the social platform onto the entire population. The available data include information collected by the PA from the online platform (assuming that it is run and administered by the PA) and census data regarding the population. Hence, the PA has access to basic personal data of platform users (e.g. gender and age), position in the online social network, and opinions revealed on the platform. The online users' data can be analyzed along with the aggregated census data on the entire population. The author has implemented a multi-agent simulation model that takes into account the distribution of personal attributes, social network data, and opinion diffusion dynamics. The analysis involves showing how different algorithms enable generalization of preferences collected by the online platform to the entire population. The results of the analysis prove that the proposed method is efficient in the preference elicitation process – with each simulation step, the preference congruence level between real and synthetic populations increases. The main determinants of preference elicitation errors include the preference diffusion model and the weight of the agents' own opinions.

Key words: preference dynamics, social network modelling, agent-based simulation

JEL: C6, C8, C9, R5

1. Wstęp

Celem pracy jest konstrukcja metod uogólniania preferencji z niereprezentatywnej podpopulacji na całą populację z uwzględnieniem procesu dyfuzji preferencji w sieciach społecznościowych. Potrzeba stworzenia takiego podejścia powstała w trakcie realizacji projektu ROUTE-TO-PA finansowanego w ramach programu Unii Europejskiej „Horizon 2020" (numer grantu 645860). Głównym produktem projektu ROUTE-TO-PA jest platforma społecznościowa Social Platform for Open Data (SPOD, http://spod.routetopa.eu/). Platforma SPOD umożliwia interakcję pomiędzy obywatelami oraz interakcję obywateli z administracją publiczną. Interakcje te mają opierać się na otwartych danych, a mianowicie: administracja publiczna udostępnia obywatelom na platformie SPOD informacje dotyczące alokacji oraz wydatków środków publicznych. Na ich podstawie obywatele mogą monitorować i kontrolować podejmowane przez nią działania administracyjne i tym samym wpływać na wzrost ich efektywności.

Jednocześnie platforma SPOD jest wykorzystywana przez administrację publiczną w celu zbierania informacji o preferencjach obywateli. Preferencje i opinie wyrażane przez użytkowników portalu, którzy stanowią pewną część całego społeczeństwa (*subpopulację*), administracja publiczna uogólnia na całą populację. Dzięki temu, bazując na pełniejszej wiedzy o potrzebach i preferencjach społeczeństwa, jest w stanie efektywniej podejmować wszelkie decyzje administracyjne.

Dane wykorzystane w analizie obejmują informacje o aktywności użytkowników na platformie społecznościowej, ich dane socjodemograficzne podane w procesie rejestracji oraz dane o całej populacji pochodzące ze spisów powszechnych. Dostępne dane o użytkownikach obejmują: płeć, wiek, status społeczny, informacje o zatrudnieniu i dynamikę wyrażanych opinii.

Założenie, że rozkłady cech społeczno-demograficznych wśród użytkowników internetowego portalu społecznego są zbliżone do rozkładu w całej populacji nie musi być prawdziwe. Rozkłady te mogą różnić się między populacją a subpopulacją, co przekłada się na niereprezentatywność. Należy w tym miejscu wyróżnić dwa błędy systematyczne, charakterystyczne dla opisywanego problemu: *selection bias* oraz *persuasiveness bias*. *Selection bias* wiąże się z niereprezentatywną dla całego społeczeństwa próbą użytkowników portalu internetowego, na podstawie której mają być dokonywane uogólnienia na całą populację. Z kolei, *persuasiveness bias* dotyczy sytuacji, w której kilku przekonujących użytkowników portalu może mieć znaczący wpływ na dynamikę preferencji i tok całej dyskusji.

W celu modelowania dynamiki preferencji takich problemów, została skonstruowana innowacyjna metoda wykorzystująca podejście symulacyjne: wieloagentowy model dynamiki preferencji (ang. *open data governance model* – ODGM). Jego zadaniem jest z jednej strony dostarczenie administracji publicznej informacji na temat aktywności użytkowników platformy, w tym wyrażane przez nich opinie i preferencje oraz ich pozycja w sieci społecznej, a z drugiej strony ma on w efektywny sposób uogólniać te preferencje na całą populację. Model dynamiki preferencji opiera się na symulacjach wieloagentowych, w wyniku których jest możliwe przeprowadzenie zaawansowanej analizy statystycznej i wizualizacji wykonanych eksperymentów symulacyjnych. Został on zbudowany w środowisku MASON i napisany w języku Java. Wykonaniu analiz statystycznych i wizualizacji posłużyły pakiety GNU R i Python.

Na potrzeby niniejszego artykułu następujące słowa: *subpopulacja, próba* i *podpopulacja* oraz *populacja sztuczna* i *populacja syntetyczna* będą używane zamiennie. Analogicznie, wyrażenia *administracja publiczna* i *samorząd* będą traktowane jako synonimy. *Uogólnianie* preferencji jest opisywane także jako ich *generalizowanie* czy *odtwarzanie*, a pojęcia: *platforma społecznościowa, platforma społeczna* i *platforma internetowa* odnoszą się do strony internetowej umożliwiającej interakcję pomiędzy obywatelami oraz interakcję obywateli z administracją publiczną.

Struktura artykułu jest następująca: w rozdziale 2. przedstawiono problem estymowania dynamiki preferencji w sztucznych sieciach społecznych; w rozdziale 3. opisano wykorzystaną metodologię, w rozdziale 4. zaprezentowano zastosowane narzędzia oraz wyliczono kolejne etapy modelowania, a w rozdziale 5. ukazano wyniki analizy symula-

cyjnego modelu wieloagentowego przeprowadzonej na danych dotyczących populacji włoskiego miasta Prato, znajdującego się w regionie Toskanii.

2. Problem rekonstrukcji dynamiki preferencji w sztucznych sieciach społecznych

Dobra komunikacja na linii *samorząd – mieszkańcy* jest niezbędna w celu lepszego zrozumienia potrzeb i preferencji mieszkańców przez administrację publiczną. Z kolei, lepsze zrozumienie obywateli przyczynia się do podejmowania bardziej efektywnych decyzji administracyjnych i do prowadzenia świadomej polityki samorządowej. Odpowiedzią na potrzebę poprawnej komunikacji są platformy społecznościowe, które stanowią miejsce dyskusji na tematy decyzji administracyjnych pomiędzy mieszkańcami (C2C – *citizen-to-citizen*) oraz mieszkańców z administracją (C2G – *citizen-to-government*). W literaturze pokazano, że platformy społeczne promują kulturę przejrzystości, otwartości informacji i w efekcie sprzyjają zmniejszaniu korupcji [Bertot, 2010]. Co więcej, platforma społecznościowa to nie tylko miejsce komunikacji samorządów z mieszkańcami, ale również miejsce dyskusji samych mieszkańców na tematy administracyjne, co umożliwia samorządom bezpośrednią obserwację wymiany zdań pomiędzy obywatelami i śledzenie ich opinii. Dzięki platformom administracja publiczna ma wgląd do preferencji mieszkańców w wielu obszarach, więc może zwracać uwagę na priorytetowe kwestie, to jest takie, które w danym momencie są najważniejsze dla obywateli i o których gorąco dyskutują. Koncepcja platformy społecznościowej jest zgodna z ideami *open government* i *open data*.

Administracja publiczna na podstawie danych z platformy powinna móc wyciągnąć wnioski o rozkładzie preferencji dla całej populacji. Subpopulacja aktywna na platformie nie stanowi jednak całej populacji. Użytkownicy platform internetowych nie są dobrą reprezentacją całej populacji i jej charakterystyk, takich jak: płeć, wiek czy poziom wynagrodzenia. Wnioski wyprowadzone wyłącznie na podstawie preferencji subpopulacji, czyli użytkowników platformy, obarczone są ryzykiem stronniczości. Dotyczy to takich kwestii, jak opinia większości społeczeństwa bądź rozkład zróżnicowania opinii mieszkańców. Ryzyko zmniejsza się wraz ze wzrostem liczby użytkowników platformy, jednakże pozostaje na wysokim poziomie, zwłaszcza w początkowych fazach życia portalu. W związku z powyższym, aby uogólnić jakiekolwiek informacje z platformy, administracja publiczna powinna uwzględnić jakościowy i ilościowy charakter niejednorodności respondentów w odniesieniu do takich aspektów, jak społeczna czy demograficzna struktura subpopulacji i populacji.

Zwykle w klasycznych metodach statystycznych, w celu poznania opinii populacji na dany temat, jest przeprowadzana ankieta, dzięki której respondentów wybiera się tak, aby próba była reprezentatywna dla danej populacji, tj. jej struktura powinna być zbieżna ze strukturą populacji w maksymalnie wielu wymiarach. Próba jest konstruowana zgodnie z zamysłem i potrzebami badaczy. W przypadku portalu społecznościowego sytuacja jest całkowicie odmienna: nie ma kontroli nad tym, kto się na niej rejestruje i kto z niej korzysta. Można powiedzieć, że próba jest samoistnie tworzona. Struktura użyt-

kowników platform nie jest tożsama ze strukturą ze spisów ludności, ale można założyć, że koreluje z tymi danymi przynajmniej wzdłuż niektórych wymiarów, takich jak: wiek, przychód bądź skłonność do wyrażania opinii (otwartość, radykalizm). Mimo korelacji, nie jest możliwe uogólnienie opinii wyrażanych na portalach społecznościowych na całą populację przy wykorzystaniu klasycznych metod statystycznych.

Poza cechami charakterystycznymi dla spisów ludności, portale społecznościowe dostarczają nowego wymiaru informacji, która jest zawarta w połączeniach i sieciach pomiędzy użytkownikami. Informacje te są odzwierciedlane poprzez wyrażane przez mieszkańców poglądy na dany temat oraz dyskusje pomiędzy użytkownikami, często nieznającymi się nawzajem. Oddziaływania on-line mogą mieć charakter *bezpośredni*, na przykład, gdy dwóch obywateli oddziałuje na siebie podczas dyskusji na dany temat albo *pośredni*, na przykład, gdy określeni użytkownicy omawiają ten sam temat na innym forum z innymi użytkownikami albo publikują swoje opinie publicznie i każdy użytkownik może je zobaczyć. Na potrzeby artykułu należy przyjąć założenie, że połączenia między obywatelami, ustalone za pośrednictwem platformy internetowej, nawiązują do sytuacji, w której dwoje obywateli bierze udział w dyskusji nad danym tematem, niezależnie od tego, czy oddziałują między sobą bezpośrednio czy pośrednio.

Uogólnianie preferencji jest klasycznym problemem w statystycznej i ekonomicznej literaturze o następującej postaci: badacz chce poznać preferencje całej populacji, ale posiada jedynie preferencje małej próby. W idealnym przypadku, gdy próba jest losowa, można uogólnić wyniki bezpośrednio na całą populację (średnia w próbie jest nieobciążonym estymatorem średniej w populacji) i następnie można obliczyć błąd estymacji. Jeśli próba nie jest reprezentatywna, to wnioskowanie o populacji nie jest możliwe lub wnioskowanie takie wymaga zastosowania innych narzędzi statystycznych. W szczególności wyniki mogą zostać odpowiednio przeskalowane, wykorzystując statystyczne metody. W kontekście sieci społecznościowych sytuacja jest jeszcze bardziej skomplikowana, ponieważ stronniczość nie występuje wyłącznie w odniesieniu do różnych rozkładów subpopulacji i populacji (*selectionbias*), lecz także ze względu na procesy społeczne, czyli zmiany opinii w konsekwencji interakcji społecznych wśród obywateli, które mogą mieć odmienną formę dla subpopulacji i populacji (*persuasiveness bias*). Formalnie oznacza to, że dyfuzja (rozprzestrzenianie się) preferencji w subpopulacji jest odmienna od dyfuzji preferencji w całej populacji. W najgorszym scenariuszu ta odmienna forma procesów społecznych może prowadzić do jeszcze większego błędu niż ten wynikający z różnych rozkładów cech socjodemograficznych w subpopulacji i populacji. W takich sytuacjach tradycyjne miary statystyczne wraz z poprawką na korekcję błędu zwracałyby wyraźnie gorsze wyniki od zaproponowanej metody prezentowanej w niniejszym artykule.

W pracy uwzględniono dwa problemy związane z brakiem reprezentatywności preferencji: (1) odmienną strukturę demograficzną populacji i subpopulacji oraz (2) różnice w procesie dyfuzji preferencji w całej populacji i subpopulacji wyrażającej swoje opinie na platformie społecznościowej.

Narzędziami wykorzystywanymi do modelowania systemów społeczno-gospodarczych są systemy wieloagentowe, opisywane w literaturze. Systemy społeczno-gospodarcze są klasyfikowane jako systemy złożone, co oznacza, że system jako całość wykazuje od-

mienne, zagregowane cechy w skali makro, od tych, które można wywnioskować z prostego sumowania cech na poziomie mikro (poszczególnych działań indywidualnych jednostek, gospodarstw domowych, przedsiębiorstw i instytucji, które stanowią części całej populacji – makrosystemu).

Różnice w postępowaniu na poziomie makro i mikro wynikają z wzajemnych interakcji pomiędzy jednostkami. Dlatego też, w celu efektywnego modelowania złożonych systemów społeczno-gospodarczych, nie wystarczy uchwycić zachowania poszczególnych elementów na poziomie mikro i następnie je zagregować, lecz ważne jest zrozumienie i odzwierciedlenie ogólnej dynamiki systemu [Axtell, 2007; Tesfatsion, 2002]. Zasada ta stanowi bazę modelowania wieloagentowego, którego metodologia umożliwia badaczom ilościowe wyjaśnianie złożoności zjawisk społecznych i gospodarczych.

Za pomocą modeli wieloagentowych można objaśnić zachowania obserwowane w skali makro, które występują w wyniku oddziaływania działań w skali mikro (np. efekty sieciowe). Takie podejście konstruowania modeli jest określane jako metoda oddolna (ang. *bottom-up*) [Oeffner, 2009], co oznacza, że model jest projektowany na poziomie mikro, na którym interakcje i zachowania poszczególnych agentów zostały określone, a następnie na poziomie makro jest obserwowana dynamika jako wyłaniający się rezultat modelu symulacji [Tesfatsion, 2002; Pyka, Fagiolo, 2005]. Same interakcje w modelu wiążą się z tym, że agenci bezpośrednio na siebie oddziałują, a ich decyzje zależą od przeszłych i obecnych wyborów dokonywanych przez siebie i innych agentów [Fagiolo, 1998]. Interakcje te są ściśle nieliniowe, a kryteria wyboru w modelach wieloagentowych są złożone i obejmują wiele wymiarów. Ponadto, mogą pojawić się pętle sprzężenia zwrotnego pomiędzy poziomami mikro i makro. Wszystkie te cechy modeli wieloagentowych wpływają na endogeniczność i niestacjonarność systemów, które prowadzą do wyłaniania się nowych wzorców zachowań. Nowe wzorce z kolei wymuszają adaptację agentów do nowego środowiska i napędzają ich uczenie się, które jest zaimplementowane w wielu modelach wieloagentowych [Windrum, Moneta, 2007].

Modelowanie wieloagentowe uchyla, występujące w standardowym modelowaniu ekonomicznym, założenie o jednorodności populacji agentów, w którym każda osoba, gospodarstwo domowe, firma itd. są identyczne i w pełni racjonalne. Przez pełną racjonalność należy rozumieć to, że posiada pełną wiedzę, na jej podstawie podejmuje optymalne decyzje i nie ponosi kosztów w procesie podejmowania decyzji. Takie podejście z pewnością nie jest empirycznie uzasadnione, chociaż w niektórych przypadkach jest wystarczające i zapewnia zadowalającą moc predykcyjną. Jednak, gdy celem analizy jest wyjaśnienie efektów interakcji pomiędzy agentami, to kluczowe jest uznanie, że agenci są różni i nie do końca racjonalni, czyli założenie o ich jednorodności należy uchylić.

Zasadniczą cechą modelu wieloagentowego jest to, że zawiera on wiele heterogenicznych elementów, tj.: indywidualnych jednostek, gospodarstw domowych, rodzin, firm itp., które dostosowują swoje działania do dynamicznie zmieniającego się środowiska. Zazwyczaj agenci tworzą hierarchie, np. grupa osób stanowi gospodarstwo domowe oraz połączenia, np. sieci społeczne. Te trzy elementy, tj. heterogeniczność, zachowanie adaptacyjne i skomplikowane relacje między jednostkami, sugerują, że choć teoretycznie

jest możliwe zapisanie pełnej, matematycznej specyfikacji tego modelu, w praktyce nie jest to możliwe. Co więcej, właśnie w praktyce to kod komputerowy jest powszechnie stosowaną i akceptowaną metodą szczegółowej specyfikacji takich modeli. Ponadto, nie tylko określenie specyfikacji modelu jest skomplikowane. Przy rozwiązywaniu takich modeli wręcz niemożliwe jest korzystanie ze standardowych narzędzi matematycznych. Alternatywnie wykorzystuje się symulację komputerową. Podsumowując, specyfikacja modelu wieloagentowego ze względu na jego złożoność nie jest jednoznaczna (ang. *explicit*), czyli nie jest to model matematyczny, ale domyślna (ang. *implicit*) – jest to kod komputerowy. Analogicznie, metoda analizy nie jest dedukcyjna (dowodzenie twierdzeń), ale indukcyjna (analiza statystyczna wychodząca z symulacji komputerowej) [Kamiński, 2012].

Model wieloagentowy jest odzwierciedlany i analizowany za pomocą symulacji komputerowych, co z kolei wprowadza kilka ograniczeń. Najważniejszym z nich jest liczebność agentów w modelu. Modelowanie populacji składającej się z milionów jednostek na ogół jest niewykonalne (ale możliwe), gdyż wymaga ogromnych mocy obliczeniowych. Jako alternatywę konstruuje się syntetyczną populacje agentów, które z reguły zawierają mniej jednostek – np. w skalach tysięcznych. W sztucznych populacjach charakterystykę agentów wybiera się tak, aby dokładnie reprezentowali oni populację rzeczywistą. Typowym rozwiązaniem jest zebranie zagregowanych danych o rozkładzie charakterystyk jednostek w prawdziwym życiu (np.: płeć, wiek, dochód, lokalizacja) wraz z ich współzależnościami i stworzenie syntetycznej populacji, która cechuje się podobnymi rozkładami. Do metod, służących rekonstrukcji syntetycznych populacji, należą m.in.: podejście kombinatoryczne (ang. *combinatorial approach*) czy metoda Monte Carlo [zob. np. Haung, Williamson, 2001]. Ta druga została wykorzystana w analizie opisywanej w niniejszym artykule.

Ważną zaletą podejścia syntetycznej populacji w modelowaniu wieloagentowym jest to, że pozwala ono rozważać różne rzeczywiste scenariusze. Oznacza to, że można nie tylko rozważać i modelować zachowanie rzeczywistej populacji (jak na przykład w modelowaniu ekonometrycznym), ale można również rozważać scenariusze „co będzie, jeśli…", zakładając rozmaite, prawdopodobne schematy przyszłych zdarzeń. Dodatkowo, modele wieloagentowe pozwalają także analizować, w jaki sposób dany system zachowuje się w określonych okolicznościach i jakie są konsekwencje zmian w jego strukturze.

W skład modelu symulacji wieloagentowej wchodzą następujące typy elementów:

- *agenci*, przez których rozumie się obiekty o zdefiniowanym typie (np.: gospodarstwa domowe, banki, firmy czy rząd) i zaimplementowani do symulowanego środowiska gospodarczego jako podmioty autonomiczne i interaktywne. Charakteryzują się mikro-parametrami, którymi mogą różnić się (np. typ wykształcenia czy wieku). Mikro-parametry są stałe lub zmienne w stosunku do kolejnych iteracji symulacji. Każdy agent ma zbiór decyzyjnych mikro-zmiennych, które są aktualizowane zgodnie z zasadami *ex ante*, biorąc pod uwagę reguły decyzyjne w modelu;
- *struktura interakcji* definiująca, którzy agenci współdziałają ze sobą i w jaki sposób;

- *czas*, modele są symulowane w dyskretnych krokach czasowych, np.: dzień, tydzień czy miesiąc. Różne rodzaje decyzji mogą być podjęte w rozmaitych ramach czasowych,
- *makro-zmienne*, które są wynikiem określonej agregacji mikro-zmiennych. Niektóre z nich mogą być niezależnie definiowane na poziomie makro (np. stopy procentowe).

Model wieloagentowy jest zazwyczaj tak skomplikowany, że nie da się go dokładnie sparametryzować, wykorzystując dane empiryczne. Na ogół należy go kalibrować i testować jego zachowanie w stosunku do różnych wartości jego parametrów.

Ostatnim etapem modelowania wieloagentowego jest zebranie wyników wykonanych symulacji oraz meta-modelowanie. Meta-modelowanie jest kluczowym elementem analizy symulacji [Kleijnen, 2000; Santos, 2007] i polega na objaśnianiu stochastycznych relacji między parametrami wejściowymi a wyjściowymi modelowanego systemu. Meta-modele (przybliżenia) mogą być użyteczne ze względu na swoje trzy główne cechy, a mianowicie: (1) zrozumiały kształt relacji między elementami wejścia (ang. *inputs*) a wynikiem (ang. *outputs*), (2) predykcja oraz (3) optymalizacja [Barton, 1992]. Te trzy cechy meta-modeli wymagają różnego podejścia do: wyboru ich funkcjonalnej specyfikacji, konstrukcji eksperymentu symulacji i estymacji parametrów.

3. Metoda rekonstrukcji dynamiki preferencji w sieciach społecznych

W niniejszym rozdziale zaprezentowano autorską procedurę (algorytm) *rekonstrukcji dynamiki preferencji w sztucznych sieciach społecznych* wykorzystującą *wieloagentowy model symulacyjny*. Procedura ta uwzględnia metody estymacji struktury sieci połączeń społecznościowych oraz metody modelowania dynamiki opinii w sieciach opisywane w literaturze.

Proponowany algorytm wygląda następująco: w punkcie wyjścia są dostępne dane ze spisu społecznego oraz dane z internetowego portalu SPOD, w tym dane demograficzne podane przez użytkowników podczas rejestracji, a także informacje o połączeniach pomiędzy nimi (użytkownicy portalu mają możliwość wyboru kręgu swoich znajomych spośród użytkowników portalu poprzez wysyłanie lub akceptację odpowiednich zaproszeń). Ponadto, zakłada się, że każdy agent może ujawnić jedną z trzech opinii: (1) za, (2) obojętny, (3) przeciw.

W pierwszym kroku na podstawie danych ze spisu ludności jest generowana sztuczna populacja, która pod względem rozkładów cech jest zbliżona do populacji rzeczywistej. Następnie na podstawie obserwowanej próby, czyli połączeń pomiędzy użytkownikami portalu społecznościowego, jak również ich indywidualnych cech jest rekonstruowana siatka połączeń na całą syntetyczną populację. Celem określenia prawdopodobieństwa istnienia połączeń pomiędzy agentami, należącymi do syntetycznej populacji, zastosowano model regresji logistycznej. Zmiennymi objaśniającymi w modelu były pary cech socjodemograficznych, w szczególności różnice pomiędzy wartościami tych cech dla obu agentów. Wyznaczając parametry modelu, przyjęto zasadę, że im mniej agenci różnią się od siebie (np. liczbą kategorii wieku, jaka ich dzieli),

tym wyższe jest prawdopodobieństwo, że przyjaźnią się (zjawisko hemofilii). Wzór przedstawiono następująco:

$$P(y_{ij} = 1) = \frac{1}{1 + e^{\alpha_0 + \alpha_1 \times \left|x_1^i - x_1^j\right| + \cdots + \alpha_n \times \left|x_n^i - x_n^j\right|}},$$

gdzie symbolem $x^i = [x_1^i, \ldots, x_n^i]$ oznaczono wektor cech socjodemograficznych agenta i.

W kolejnym kroku dla syntetycznej populacji agentów wygenerowano ich opinie pierwotne, czyli opinie wyrażane przez nich po raz pierwszy, na które pozostali obywatele i wyrażone przez nich opinie nie mieli jeszcze wpływu. W tym celu wykorzystano model trinomialny, którego parametry zostały oszacowane na danych pochodzących z portalu społecznościowego: założono, że jedna z trzech możliwych opinii wyrażanych przez danego agenta *j* w rundzie 0, oznaczona symbolem $o(v_j, 0)$, zależy od jego indywidualnych cech, zgodnie ze wzorem:

$$P(o(v_j, 0) = k) = \frac{e^{\gamma_0^k + \gamma_1^k \times x_1^j + \cdots + \gamma_n^k \times x_n^j}}{\sum_{k=-1}^{1} e^{\gamma_0^k + \gamma_1^k \times x_1^j + \cdots + \gamma_n^k \times x_n^j}}.$$

Następnie dla różnych wartości parametru β, który odzwierciedla wagę przywiązania danego agenta do własnej opinii (β – waga opinii własnej agenta, 1-β –waga wpływu opinii agentów sąsiadujących), jest przeprowadzany eksperyment symulacyjny. Agenci w kolejnych krokach symulacji wchodzą w interakcje z przyjaciółmi, co przekłada się na zmianę ich opinii. Modelowanie dynamiki opinii w opisywanym algorytmie wygląda następująco: w kolejnych iteracjach preferencje agentów są aktualizowane jako liniowa średnia ważona opinii danego agenta oraz opinii agentów sąsiadujących. Przyjęte podejście należy do prostszych metod modelowania dynamiki preferencji [de Groot, 1977], w odróżnieniu od metod wykorzystujących podejście Bayesowskie [zob. np. Acemoglu, Ozdaglar, 2011]. Jednak mogłoby zostać rozszerzone np. o wagi zmienne w czasie [Krause, 2000] czy też poprzez wprowadzenie tzw. upartych agentów (ang. *stubbornagents*), którzy nie zmieniają swojej opinii pod wpływem innych agentów. Ostatecznie wynik otrzymany za pomocą eksperymentu symulacyjnego na całej populacji jest porównywany ze strukturą preferencji w obserwowanej próbie, czyli w grupie użytkowników portalu internetowego.

Przyjmijmy, że agent *j* posiada *n* sąsiadujących agentów, a rozkład ich opinii opisuje wektor π(v), zgodnie ze wzorami:

$$n = \sum_{k \in \{-1,0,1\}} n_k,$$

$$\pi(v) = \left(\frac{n_{-1}}{n}, \frac{n_0}{n}, \frac{n_1}{n}\right).$$

Dla każdego agenta i każdej kolejnej rundy rozważono trzy alternatywne sposoby aktualizacji opinii w rundzie r, takie jak:

$$o(v_j, r+1) \leftarrow k^*,$$

a) metoda średniej opinii sąsiadujących agentów:

$$s = \beta \times o(v_j, r) + (1-\beta) \times \frac{n_1 - n_{-1}}{n},$$

$$\mathrm{k}^* = \begin{cases} -1, \mathrm{s} < -0{,}33 \\ 0, -0{,}33 \leq \mathrm{s} \leq 0{,}33, \\ 1, \mathrm{s} > 0{,}33 \end{cases}$$

b) metoda dominującej opinii sąsiadujących agentów:

$$s = \beta \times o(v_j, r) + (1-\beta) \times o(max, r),$$

$$o(max, r) = \begin{cases} -1, n_{-1} = n_{max} \wedge n_{-1} \neq n_1 \\ 0, n_0 = n_{max} \vee n_{-1} = n_1 \\ 1, n_1 = n_{max} \wedge n_{-1} \neq n_1 \end{cases},$$

$$n_{max} = \max(n_{-1} n_0, n_1),$$

$$k^* = \begin{cases} -1, s < -0{,}33 \\ 0, -0{,}33 \leq s \leq 0{,}33, \\ 1, s > 0{,}33 \end{cases}$$

c) metoda polaryzującej opinii sąsiadujących agentów:

$$k^* = sign(\beta \times 10 \times o(v_j, r) + n_1 - n_{-1}).$$

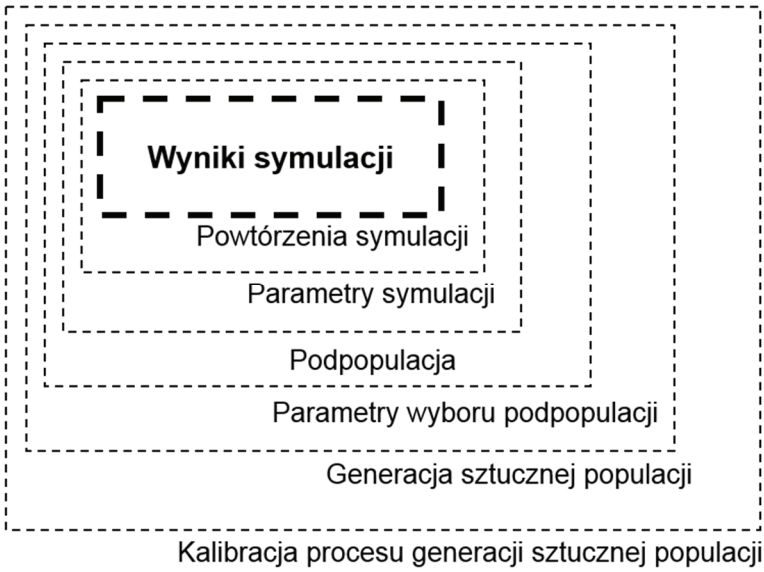
Opisany wyżej algorytm odnosi się do docelowej sytuacji, w której są znane dane z portalu SPOD. Jednakże dane te nie są jeszcze dostępne, więc analiza opisywana w niniejszym artykule musiała opierać się na sztucznie wygenerowanej próbie użytkowników: w pierwszym kroku na podstawie danych ze spisu ludności, również wygenerowano sztuczną populację (jej rozkład cech demograficznych odpowiadał populacji rzeczywistej) i na niej zasymulowano dynamikę dyfuzji (rozprzestrzeniania się) preferencji. Następnie losowo, metodą kuli śnieżnej z całej sztucznej populacji wybierano różne potencjalnie niereprezentatywne subpopulacje. Inne metody losowania próby dla sieci społecznościowych szczegółowo opisano w przytoczonej literaturze [Frank, 1974]. Dla każdej subpopulacji testowano algorytmy uogólniania preferencji przez odtwarzanie dynamiki całej populacji. Miarą jakości modelu w takim podejściu jest także zgodność preferencji wygenerowanych (symulowanych) i odtworzonych na podstawie wylosowanej subpopulacji. Zgodność taka była rozważana zarówno na wylosowanej subpopulacji, jak i wygenerowanej całej syntetycznej populacji. Ostatnim etapem było zatem zbieranie wyników wykonanych symulacji oraz meta-modelowanie. W analizie prezentowanej w niniejszym artykule meta-modelowanie ma dwa główne cele: zrozumienie i przewidywanie. W związku z tym, oczekuje się, że otrzymane meta-modele mają dwie charakterystyki: prostą in-

terpretację swojej struktury oraz moc statystyczną. W analizie opisywanej w tym artykule jako przybliżenie zastosowano lasy losowe, które należą do klasy modeli posiadających powyższe dwie pożądane właściwości.

4. Narzędzia i etapy modelowania

Wieloagentowy model symulacyjny, umożliwiający rekonstrukcję preferencji w sztucznych sieciach społecznych, został zaimplementowany dzięki wykorzystaniu następujących oprogramowań typu Free Open Source: R, Java, Python, MASON, Weka, JUNG. Na rysunku 1. zaprezentowano warstwy symulacji modelu ODGM. Zastosowane podejście pozwoliło na porównywanie wyników dla różnych parametryzacji modelu i tym samym umożliwiło testowanie wrażliwości modelu na zmiany wartości parametrów.

RYSUNEK 1.
Warstwy parametrów w wieloagentowym modelu symulacyjnym



Źródło: opracowanie własne.

Problemem w symulacyjnych modelach wieloagentowych o wysokiej liczbie parametrów może być duża liczba powtórzeń symulacji, czyli przeszukiwanie bardzo obszernej przestrzeni parametrów, co wymaga znacznej mocy obliczeniowej. Z tego względu opisywany model symulacji został przeprowadzony na klastrze obliczeniowym w chmurze Amazon Web Services. Do procesu zrównoleglania obliczeń wykorzystano narzędzie Open Grid Scheduler, a w szczególności jego implementację przeznaczoną do wykorzystania w środowisku obliczeń w chmurze – Star Cluster. Moduł symulacji został napisany w języku programowania Java, zaimplementowany w środowisku symulacyjnym

MASON i opierał się na kilku bibliotekach typu Open Source. Syntetyczna populacja, analiza i wizualizacja zostały przeprowadzone w języku GNU R, wykorzystując odpowiednie biblioteki. Poniżej opisano kolejne etapy estymacji sztucznej populacji miasta Prato i przeprowadzonych na niej symulacji.

1. Stworzenie skryptu symulacji.
2. Analiza zagregowanych danych pochodzących ze spisu ludności miasta Prato.
3. Wygenerowanie sztucznej populacji na podstawie powyższych danych.
4. Zbudowanie siatki połączeń pomiędzy agentami w sztucznej populacji.
5. Stworzenie pierwotnych preferencji w sztucznej populacji.
6. Symulacja dynamiki dyfuzji preferencji w populacji.
7. Wybór próby subpopulacji.
8. Uruchomienie właściwej symulacji, czyli przeprowadzenie określonej liczby symulacji dla danych kombinacji parametrów modelu (parametryzacji).

Kod źródłowy opracowanej implementacji modelu może być pobrany ze strony: https://bitbucket.org/pszufe/socialpreferencessimulation2/.

5. Wyniki eksperymentów symulacyjnych

Dane, na których opierała się niniejsza analiza, pochodziły ze spisu ludności we włoskim mieście Prato i z danych rocznych deklaracji podatkowych, z których pobrano informacje o dochodach obywateli. Na koniec roku 2014 Prato zamieszkiwało 191 tysięcy ludzi. W analizie wykorzystano następujące cechy społeczno-demograficzne: region zamieszkania, płeć, kategorię wiekową, zawód, stan cywilny i kategorię dochodu rocznego. Na podstawie informacji o wszystkich rozkładach brzegowych zmiennych wygenerowano reprezentatywną próbę 2 480 mieszkańców, której następnie przyporządkowano opinie pierwotne (bazując na obserwowanych społeczno-ekonomicznych cechach) oraz na której zasymulowano dyfuzję preferencji. Z wygenerowanej populacji losowano niereprezentatywne próby, na których wykonywano eksperyment symulacyjny. W celu wprowadzenia do modelu błędu reprezentatywności starsi mieszkańcy mieli tendencję do głosowania „za", a bogatsi do głosowania „przeciw".

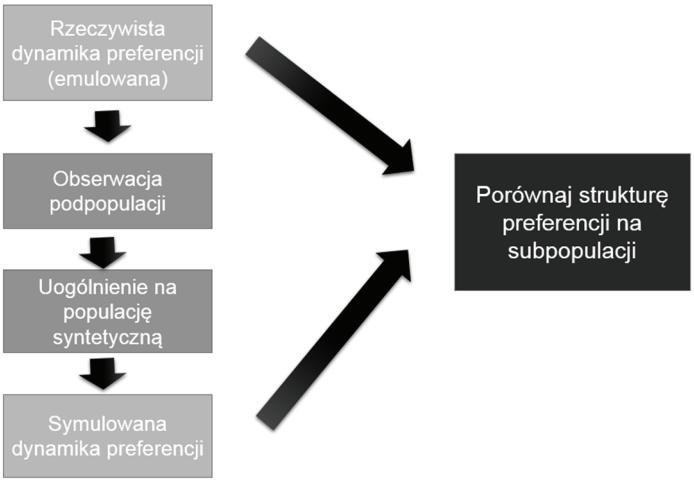
W każdym ze skończonych kroków symulacji można było obserwować dynamikę dyfuzji preferencji w populacji i w subpopulacji. W eksperymencie symulacyjnym rozważono pięciowymiarową przestrzeń parametrów, opisanych poniżej.

1. *Model dyfuzji preferencji* wskazujący na sposób, w jaki dany agent uwzględnia opinię innych agentów, z którymi jest połączony. Przyjęto trzy możliwe modele: (1) agent uwzględnia *średnią* opinię sąsiadów; (2) agent uwzględnia *dominującą* opinię sąsiadów; (3) agent *polaryzuje* opinię sąsiadów, czyli już po pierwszej rundzie symulacji musi być albo za, albo przeciw: nigdy nie może być neutralny.
2. *Średnia liczba połączeń* determinująca gęstość sieci połączeń pomiędzy agentami.
3. *Przywiązanie do opinii*, czyli parametr beta $\beta \in (0,1)$ homogeniczny dla wszystkich agentów, reprezentujący siłę przywiązania agenta do własnej opinii (β – waga opinii własnej agenta, $1-\beta$ – wpływ opinii agentów sąsiadujących). W analizie przyjęto osiem poziomów tego parametru.

4. *Struktura opinii początkowych:* parametr wyrażający typ opinii pierwotnej: każdy agent mógł wyrażać na początku jedną z trzech opinii: za, neutralny, przeciw.
5. *Rozmiar próby*, czyli parametr reprezentujący sposób, w jaki subpopulacja agentów jest losowana. Przyjmuje szesnaście poziomów.

Pełen iloczyn kartezjański powyższej przestrzeni parametrów zawierał 1 536 punktów (5 parametrów pomnożonych przez liczbę przyjmujących przez nie poziomów). Dla każdej parametryzacji wykonano 30 symulacji, co łącznie dało 46 080 wykonanych eksperymentów symulacyjnych. Celem tych eksperymentów był pomiar zgodności preferencji między preferencjami rzeczywistymi (w populacji rzeczywistej) a symulowanymi (w wygenerowanej, syntetycznej populacji). Zgodność jest miarą słuszności zastosowanego podejścia do uogólniania preferencji w analizowanym problemie badawczym. Rozumowanie to zostało zobrazowane na rysunku 2.

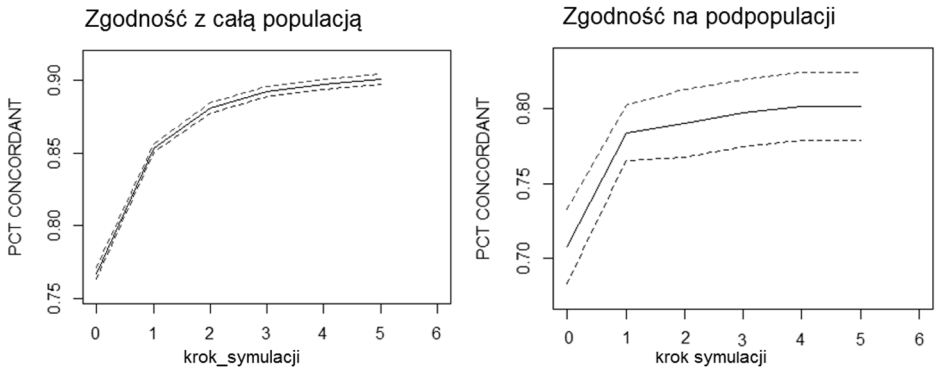
RYSUNEK 2.
Zgodność rzeczywistych i symulowanych preferencji jako miara jakości w podejściu uogólniania preferencji



Źródło: opracowanie własne.

Jak się okazało, kolejne rundy symulacji prowadzą do wzrostu zgodności preferencji między populacją rzeczywistą a odtworzoną zarówno dla całej populacji, jak i dla subpopulacji, co zostało przedstawione na rysunku 3. Wykres z lewej strony reprezentuje zgodność na całej populacji, czyli odsetek par zgodnych opinii końcowych, symulowanych w procesie tworzenia syntetycznej populacji i odtworzonych na podstawie wylosowanej próby za pomocą algorytmu opisanego w niniejszym artykule, obserwowanych na całej populacji syntetycznej. Analogiczne miary dla wylosowanej subpopulacji zobrazowano na rysunku 3. po stronie prawej. W równej mierze na jednym, jak i na drugim rysunku zgodność preferencji rośnie wraz z kolejnymi krokami symulacji, a poziom zgodności jest zadowalający (90% na poziomie populacji).

RYSUNEK 3.
Przykładowe wyniki symulacji: zgodność populacji rośnie zarówno w całej populacji, jak i w podpopulacji



Uwaga: linia ciągła przedstawia wartości średnie, natomiast linie przerywane prezentują granice przedziału, w którym mieściło się 90% uzyskanych wyników symulacji.

Źródło: opracowanie własne.

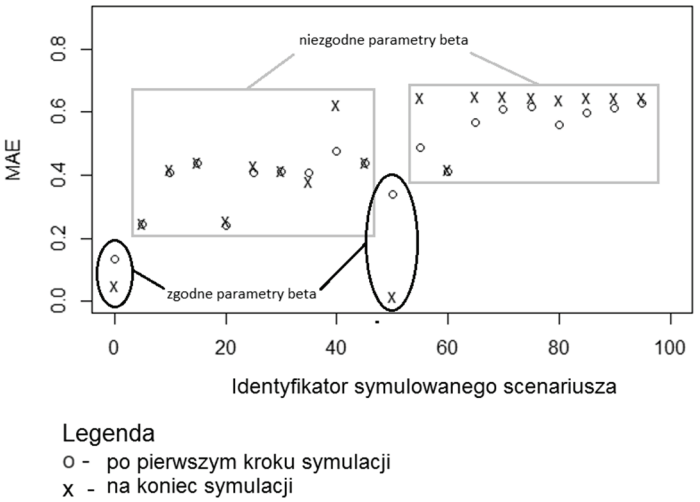
Rysunek 4. przedstawia średni błąd uogólniania preferencji. Dla niezgodnych parametrów β średni błąd absolutny uogólniania preferencji jest na wysokim poziomie zarówno na początku, jak i na końcu symulacji. Z kolei, dla zgodnych parametrów β wraz ze wzrostem liczby symulacji średni błąd odtworzenia preferencji na podpopulacji maleje.

Na ostatnim etapie za pomocą meta-modelu lasów losowych sprawdzono, który parametr jest krytyczny pod względem wpływu na błąd odtwarzania preferencji. Na rysunku 5. zostały zilustrowane wyniki porównania dwóch stanów symulacji: na początku eksperymentu symulacyjnego i na końcu, czyli po wykonaniu symulacji na całej przestrzeni parametrów. Można zauważyć, że najbardziej istotnymi determinantami błędów uogólniania preferencji są model dyfuzji preferencji oraz waga opinii własnej agenta (parametr β). Zatem typ dynamiki dyfuzji opinii i przywiązanie do opinii własnej agenta mają największy wpływ na błąd odtworzenia preferencji.

Dalsze wyniki analizy symulacji pokazały również, że błąd reprezentatywności wzrasta równolegle z tym, jak opinie agentów stają się jednorodne. Innymi słowy, spadek wariancji opinii w populacji prowadzi do wzrostu błędu reprezentatywności.

RYSUNEK 4.

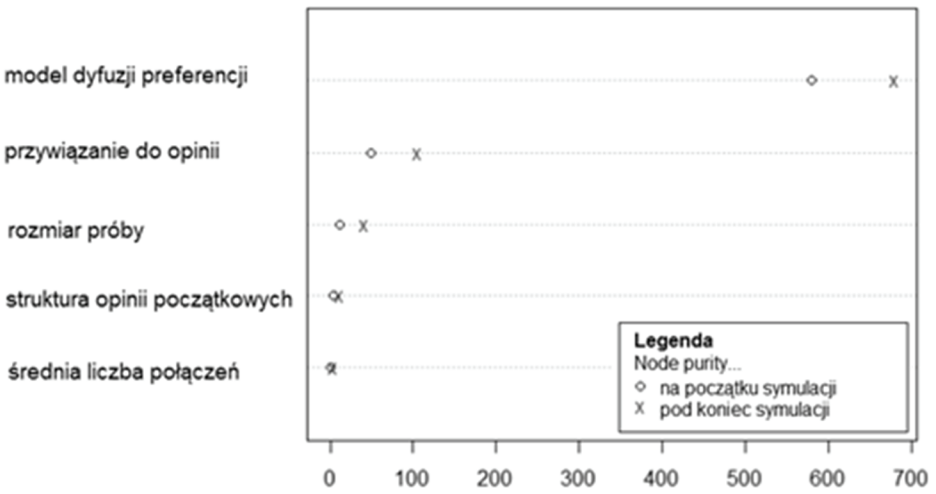
Przykładowe wyniki symulacji: średni błąd uogólniania preferencji maleje wraz ze wzrostem liczby iteracji symulacji



Źródło: opracowanie własne.

RYSUNEK 5.

Czystość węzła w metodzie lasów losowych na początku i na końcu symulacji wykonanej na całej przestrzeni parametrów: 30 powtórzeń dla każdej parametryzacji. Determinanty błędów odtwarzania preferencji



Źródło: opracowanie własne.

6. Podsumowanie

W opracowaniu przedstawiono *estymowanie dynamiki preferencji w sztucznych sieciach społecznych*, którego celem było stworzenie systemu efektywnego uwzględniania preferencji obywateli przez administrację publiczną w swoich decyzjach. W skonstruowanym modelu rozważano rzeczywistą sytuację, w której za pomocą platformy społecznościowej administracja udostępnia informacje na temat podejmowanych decyzji, umożliwiając obywatelom: monitorowanie, kontrolę i wymianę zdań na temat swoich działań i wydatków publicznych. Jednocześnie na podstawie tej platformy administracja publiczna może obserwować opinie mieszkańców i analizować ich preferencje. W celu efektywnego i uwzględniającego preferencje całej populacji podejmowania decyzji administracyjnych, należy jednak uogólnić preferencje subpopulacji, czyli użytkowników portalu społecznościowego, na całą populację, a mianowicie na wszystkich obywateli. Analiza taka opiera się na danych pochodzących ze spisów ludności, a w przyszłości może także opierać się na danych z portalu społecznościowego, w szczególności na danych dotyczących: logowania przeglądanych baz danych, intensywności prowadzonych dyskusji, indywidualnych preferencji czy sieci połączeń. W zaprezentowanych w artykule wynikach wykorzystano dane o użytkownikach portalu wygenerowane w symulacjach. Zastosowanie opisanej metody do rzeczywistych danych o użytkownikach portalu będzie przedmiotem dalszych badań.

Jako metodę modelowania dyfuzji preferencji w sieciach społecznościowych wykorzystano symulacje wieloagentowe. Podejście to pozwoliło na uogólnienie informacji o preferencjach użytkowników internetowej platformy społecznościowej na całą populację, w celu umożliwienia administracji publicznej podejmowania decyzji odpowiednich dla wszystkich obywateli. W opracowaniu przedstawiono implementację praktyczną powyższego modelu do danych dotyczących prowincji Prato we Włoszech. Na podstawie danych empirycznych wygenerowano sztuczną populację liczącą 2 840 agentów, na której przeprowadzono łącznie 46 080 symulacji. Rezultaty eksperymentu symulacyjnego potwierdziły skuteczność modelu: wraz ze wzrostem liczby symulacji wzrastała zgodność preferencji między populacją rzeczywistą a syntetyczną. Zdiagnozowano także determinanty błędu uogólniania preferencji na całą populację: są to model dyfuzji preferencji oraz waga opinii własnej agenta.

Wkład autorów w powstanie artykułu

dr Marcin Czupryna – prowadzenie badań, opis wyników i przygotowanie artykułu – 25%

dr Przemysław Szufel – prowadzenie badań, opis wyników i przygotowanie artykułu – 25%

dr hab. Bogumił Kamiński – prowadzenie badań, opis wyników i przygotowanie artykułu – 25%

mgr Anna Wiertlewska – prowadzenie badań, opis wyników i przygotowanie artykułu – 25%

Literatura

Acemoglu D., Ozdaglar A., 2011, *Opinion Dynamics and Learning in Social Networks*, „Dynamic Games and Applications", vol. 1(1).

Axtell R. L., 2007, *What economic agent do: How cognition and interaction lead to emergence and complexity*, „Review Austrian Economics", no. 20, DOI 10.1007/s11138-007-0021-5.

Barton R. R., 1992, *Metamodels for simulation input-output relations*, [in:] *Proceedings of the 1992 Winter Simulation Conference*, J. Swain, D. Goldsman, R. Crain, J. Wilson (eds.), IEEE.

Bertot J. C., Jaeger P. T., Grimes J. M., 2010, *Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies*, "Government Information Quarterly", no. 27.

De Groot M. H., 1977, *Reaching a Consensus*, "Journal of the American Statistical Association", no. 69.

Fagiolo G., 1998, *Spatial interactions in dynamic decentralized economies: a review*, "The Economics of Networks", DOI 10.1007/978-3-642-72260-8_3.

Frank O., 1974, *Survey sampling in graphs*, "Journal of Statistical Planning and Inference", vol. 126.

Haung Z., Williamson P., 2001, *A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of the small-area microdata*, "Working Paper", no. 2, University of Liverpool.

Kamiński B., 2012, *Podejście wieloagentowe do modelowania rynków. Metody i zastosowania*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa.

Kamiński B., 2015, *Interval metamodels for the analysis of simulation Input – Output relations*, "Simulation Modeling Practice and Theory", no. 54.

Kleijnen J. P., Sargent R. G., 2000, *A methodology fitting and validating metamodels in simulation*, "European Journal of Operational Research", no. 120 (1).

Krause U., 2000, *A Discrete Nonlinear and Nonautonomous Model of Consensus Formation*, [in:] *Communications in Difference Equations*, J. Rakowski (ed.), Gordon and Breach, Amsterdam.

Oeffner M., 2009, *Agent – Based Keynesian Macroeconomics – An Evolutionary Model Embedded in an Agent-Based Computer Simulation*, MPRA Paper, no. 18199, The Munich University Library, Munich.

Pyka A., Fagiolo G., 2005, *Agent-based modelling: A methodolgy for Neo-Schumpeterian economics*, Discussion Paper Series, no. 272, University of Augsburg, Augsburg.

Santos I. R., Santos P. R., 2007, *Simulation metamodels for modeling output distribution parameters*, [in:] *Proceedings of the 2007 Winter Simulation Conference*, R. Barton (ed.), IEEE.

Tesfatsion L., 2002, *Agent-Based Computational Economics: Growing Economies From the Bottom Up*, "Artificial Life", vol. 8, no. 1, DOI 10.1162/106454602753694765.

Windrum P., Fagiolo G., Moneta A., 2007, *Empirical Validation of Agent-Based Models: Alternatives and Prospects*, "Journal of Artificial Societies and Social Simulation", no. 10(2).