Zielona Góra, 14 stycznia 2019 r.

prof. dr hab. inż. Dariusz Uciński
Instytut Sterowania i Systemów Informatycznych
Uniwersytet Zielonogórski

RECENZJA
rozprawy doktorskiej Pana mgr. Piotra Kopki
pt. *Zastosowanie metodologii bayesowskiej w stochastycznej rekonstrukcji zdarzeń na przykładzie uwolnień niebezpiecznych gazów do atmosfery*
opracowana na wniosek Rady Naukowej Instytutu Badań Systemowych Polskiej Akademii Nauk

I. Obszar problemowy rozprawy

Problem identyfikacji źródeł emisji niebezpiecznych gazów do atmosfery jest od wielu lat przedmiotem intensywnych badań ze względu na jego zastosowania zarówno w monitorowaniu i sterowaniu jakością atmosfery, jak również w zagadnieniach związanych z zapewnieniem bezpieczeństwa. Prognozuje się, że do końca pierwszej połowy XXI wieku połowa populacji Ziemi będzie żyła w obszarach miejskich. Podstawową kwestią jest towarzyszący temu wzrost zanieczyszczenia powietrza w wyniku zwiększonej emisji pyłów, dwutlenku siarki, tlenku azotu, tlenku węgla i metali ciężkich, co stanowi bezpośredni skutek stale rosnących emisji wskutek intensywnego ruchu drogowego, zwiększonego zapotrzebowania na zużycie energii oraz ciągłego zmniejszania otwartych przestrzeni zieleni. Integracja systemów monitorowania z nowoczesnymi technikami modelowania i optymalizacji ma na celu zmniejszenie narażenia na zanieczyszczenie powietrza i zmniejszenie poziomu substancji toksycznych poprzez m.in. projektowanie budynków nowej generacji, a ogólniej, nowe rozwiązania urbanizacyjne oraz inteligentne zarządzanie informacją w miastach.

Problem identyfikacji źródeł niebezpiecznych substancji stał się również istotny z uwagi na potencjalne zagrożenia atakami biochemicznymi powodującymi rozprzestrzenianie niebezpiecznych substancji w postaci aerozolu lub wycieki niebezpiecznych materiałów biochemicznych przewożonych przez pojazdy naziemne lub latające. W typowym scenariuszu, po wystąpieniu pewnego rodzaju skażenia chemicznego lub biologicznego, pojawia się chmura toksycznego materiału, której ewolucję określają procesy fizyczne i chemiczne, warunki pogodowe oraz topografia terenu. Dynamikę zmian opisują wyrafinowane czasoprzestrzenne modele matematyczne, ujmujące kluczowe aspekty zachodzących zjawisk dyfuzji i transportu oraz charakter wymuszeń, takie jak panujące warunki pogodowe czy też warunki brzegowe związane ze skomplikowaną geografią terenu. Najczęściej bazują one na formalizmie równaniach różniczkowych cząstkowych określonych w odpowiednio zdefiniowanych obszarach przestrzennych i przy różnego typu zadanych warunkach brzegowych.

Pomiarów stężeń substancji toksycznych dokonuje się za pomocą sieci sensorów

ulokowanych w rozważanym obszarze przestrzennym. Warunki meteorologiczne określa się na podstawie pomiarów wykonywanych przez stacje meteorologiczne. Celem jest użycie tych obserwacji w połączeniu z modelem matematycznym ewolucji chmury gazu w celu detekcji i lokalizacji źródła zanieczyszczenia, a w konsekwencji, np. prognozowania przyszłej ewolucji tej chmury.

Problem identyfikacji źródeł jest jednym z klasy problemów odwrotnych, redukującym się do pewnego zadania identyfikacji (parametrycznej lub nieparametrycznej). Należy więc zagwarantować warunki identyfikowalności parametrów źródła, co nie jest takie oczywiste, jeśli weźmie się pod uwagę skończony wymiar przestrzeni wyjść i nieskończenie wymiarowy charakter przestrzeni stanów i/lub parametrów modeli z czasoprzestrzenną dynamiką. Innym problemem teoretycznym jest zagwarantowanie istnienia estymatora parametrów źródła i jego ciągłej zależności od danych, co przekłada się na wymóg zwartości przestrzeni parametrów. Jeśli tych aspektów nie potraktuje się we właściwy sposób, proces estymacji może okazać się tzw. problemem źle postawionym (*ang.* ill-posed problem), co oznacza, że nawet niewielkie zaszumienie danych może prowadzić do bardzo dużych błędów estymacji, a ponadtowiele różnych zestawów wartości parametrów współgra z zaobserwowanymi danymi pomiarowymi. Spowodowało to rozwój technik regularyzacji (spośród których najbardziej znana jest regularyzacja Tichonowa), które kalibrują parametry modelu nie tylko poprzez minimalizację miary niedopasowania pomiarów i wyjścia modelu, ale również nakładając karę na niepożądane cechy estymat. Rzadko kiedy jednak uwzględniają one statystyczne aspekty problemu estymacji.

W ostatnich latach alternatywą stało się wnioskowanie bayesowskie, które w naturalny sposób umożliwia uwzględnienie wstępnej wiedzy o rozkładach estymowanych parametrów i/ lub stanu początkowego (w postaci ich rozkładów apriorycznych) przed zaobserwowaniem jakichkolwiek danych. W połączeniu z adekwatnym opisem niepewności obserwacji oraz odwzorowaniem z przestrzeni parametrów do przestrzeni wyjść (określonym przez równanie stanu, w tym przypadku równanie różniczkowe cząstkowe, i równanie wyjścia opisujące model obserwacji dokonywanych przez sensory), umożliwia to wyznaczenie rozkładu *a posteriori* parametrów źródła, odzwierciedlającego stopień zaufania do ich wartości. Rozkład ten nie jest jednak dany w postaci zwartej formuły do bezpośredniego wykorzystania, a można z niego jedynie próbkować, najczęściej poprzez budowę łańcucha Markowa Monte Carlo, który ma rozkład równowagowy pokrywający się ze wspomnianym rozkładem *a posteriori*. Podejście, podsumowane w klasycznej już monografii Jari Kaipio i Erkki Somersalo (*Statistical and Computational Inverse Problems*, Springer, 2005), doczekało się w minionej dekadzie uogólnienia na estymację parametrów nieskończenie wymiarowych (np. stanu początkowego), przede wszystkim dzięki pracom Andrew M. Stuarta z University of Warwick (jego artykuł pt. *Inverse problems: A Bayesian perspective*, Acta Numerica, 2010, pp. 452–559 ma ponad 750 cytowań w Google Scholar). Co więcej, pojawiło się dużo wartościowych prac poświęconych planowaniu eksperymentu dla tego typu problemów, rozumianemu przede wszystkim jako rozmieszczanie czujników pomiarowych w sposób maksymalizujący zawartość informacji o estymowanych parametrach w gromadzonych danych. W tym kontekście warto wymienić przede wszystkim prace Omara Ghattasa, Alena Alexanderiana, Noemi Petry i Goerga Stadlera z University of Texas at Austin.

Pewna dojrzałość osiągnięta przez bayesowskie metody identyfikacji źródeł skażeń nie

oznacza bynajmniej, że zmierzono się ze wszystkimi problemami, jakie na tym polu przynoszą zastosowania, tym bardziej że gwałtowny rozwój sensoryki, zwłaszcza w kontekście mobilnych sieci sensorowych, generuje coraz to nowe możliwości zastosowań wraz z towarzyszącymi im wyzwaniami. Nadal wyzwaniem jest problem zastosowania tej metodologii w kontekście rzeczywistych, a nie symulowanych danych, kiedy nie są znane dokładne rozkłady wielkości obarczonych niepewnością pomiarową, a modele matematyczne, mimo skomplikowania, są tylko do pewnego stopnia odzwierciedleniem procesów fizycznych. Wreszcie dużym wyzwaniem pozostają nietrywialne aspekty informatyczne, dotyczące redukcji dużych mocy potrzebnych do rozwiązywania intensywnych zadań obliczeniowych. Wyznaczenie wartości prawdopodobieństwa w jednym punkcie przestrzeni parametrów wymaga numerycznego rozwiązania problemu prostego (czyli numerycznego rozwiązania równania różniczkowego cząstkowego modelującego rozprzestrzenianie się zanieczyszczenia) i może pochłaniać wiele godzin na dużym superkomputerze jeśli rozważa się trzy wymiary przestrzenne i modele stosowane w obliczeniowej mechanice płynów (równania Naviera-Stokesa). Liczba tego typu ewaluacji potrzebna do adekwatnego próbkowania z rozkładu *a posteriori* bardzo wielu parametrów (jest tak w sytuacji, gdy estymuje się również stan początkowy) przez konwencjonalne łańcuchy Markowa Monte-Carlo może sięgać milion i więcej. Z tego powodu coraz popularniejsze stają się przybliżone obliczenia bayesowskie (*ang.* Approximate Bayesian Computing, ABC), unikające kosztownego wyznaczania wartości funkcji wiarygodności. Stosowano je do tej pory przede wszystkim w naukach biologicznych (genetyka populacyjna, ekologia, epidemiologia i biologia systemowa), a jej zastosowanie w identyfikacji źródeł zanieczyszczeń nie jest jeszcze rozpowszechnione.

Właśnie w tym kontekście recenzowana praca Pana mgr. Piotra Kopki, poświęcona w całości bayesowskim metodom rozwiązywania zadań rekonstrukcji parametrów źródła uwolnienia niebezpiecznych gazów do atmosfery jest pozycją bardzo ambitną i aktualną. Zasadniczo, oryginalny pomysł Autora polega na adaptacji i modyfikacji algorytmu przybliżonych obliczeń bayesowskich na potrzeby identyfikacji parametrów źródła uwolnienia niebezpiecznego gazu z wykorzystaniem czasoprzestrzennych modeli symulacyjnych jego rozprzestrzeniania się. Co więcej, obszerną część rozprawy zajmuje praktyczna weryfikacja działania zaproponowanej techniki w oparciu o dane rzeczywiste pochodzące z trzech eksperymentów, dla których do tej pory nie przeprowadzano tego typu rekonstrukcji, z analizą porównawczą z konwencjonalnymi metodami Monte Carlo wykorzystującymi łańcuchy Markowa.

Biorąc pod uwagę wszystkie wymienione czynniki, sformułowane na str. 12 cele i tezę pracy, jak również wynikające z nich zadania szczegółowe, są jasne i dobrze określone. Sprowadzają się one do wykazania, że bayesowskie algorytmy próbkowania z rozkładu *a posteriori* parametrów źródła uwolnienia niebezpiecznego gazu, skonstruowanego na podstawie asymilacji pomiarów jego stężenia z sieci sensorów i danych ze stacji meteorologicznych, w połączeniu z zaawansowanymi modelami symulacyjnymi czasoprzestrzennej dynamiki zachodzących procesów fizycznych, umożliwia dość precyzyjne oszacowanie parametrów źródła i jego niepewności dla bardzo złożonych danych rzeczywistych, w stosunkowo krótkim czasie i w oparciu o niskobudżetowe zasoby sprzętowe. Efektem rozprawy ma być zestaw efektywnych algorytmów próbkowania z możliwością identyfikacji parametrów zmiennych w czasie. Tak zarysowaną problematykę rozprawy uważam za istotną i nadzwyczaj aktualną, o rezultatach mogących otworzyć nowy nurt badań nad zagadnieniami

odwrotnymi dla układów z czasoprzestrzenną dynamiką. Fakt ten przesądza o pozytywnej ocenie wybranego tematu jako przedmiotu opiniowanej rozprawy doktorskiej.

II. Koncepcja oraz realizacja rozprawy

Obszerna rozprawa, napisana w języku polskim i licząca 158 stron numerowanych, składa się ze wstępu, trzech zasadniczych rozdziałów przedstawiających koncepcję proponowanych metod, trzech rozdziałów raportujących wyniki eksperymentów weryfikujących działanie tych metod w praktyce w oparciu o dane rzeczywiste, rozdziału podsumowującego, oraz dodatku opisującego zastosowane oprogramowanie symulacyjne rozprzestrzenianie się gazów w atmosferze. Załączony wykaz 76 pozycji cytowanej literatury w zasadzie odzwierciedla stan badań w zakresie tematycznym rozprawy (o pewnych brakach w tym wykazie traktuję w części poświęconej uwagom krytycznym).

Pracę rozpoczyna *Wstęp*, na który składa się przedstawienie motywacji zagadnień rozprawy, cele i teza pracy (w tym momencie już intuicyjnie jasną), oraz ogólny opis proponowanego podejścia. Rozdział kończy charakterystyka struktury pracy.

Rozdział 2 stanowi zwięzłe wprowadzenie do problematyki identyfikacji parametrów źródła uwolnienia gazu i, ogólniej, do rozwiązywania zagadnień odwrotnych. Wzmiankuje się konieczność regularyzacji zadania odwrotnego poprzez włączenie dodatkowej wstępnej informacji o estymowanych parametrach oraz bardziej szczegółowo omawia się alternatywne stochastyczne podejście do identyfikacji źródeł emisji gazu, wywodzące się z technik bayesowskich. Dokonuje się klasyfikacji problemów rekonstrukcji ze względu na skalę obserwowanego obszaru, a następnie omawia się sposób wykonywania pomiarów przez sieć sensorową wraz z indukowanymi przez nią błędami pomiarowymi, oraz bardzo ogólnie matematyczne modele rozprzestrzeniania się niebezpiecznego gazu i wynikające z ich użycia błędy modelowania. Wprowadza się graficzną reprezentację rozwiązywania zadania rekonstrukcji z przepływem informacji między jego komponentami.

Rozdział 3 stanowi krótkie wprowadzenie do wnioskowania bayesowskiego i metodologii dynamicznych modeli Bayesa stosowanych w rozprawie. Omawia się rolę rozkładu *a priori* oraz sposoby wyznaczania funkcji wiarygodności. Pokrótce charakteryzuje się rozkład *a posteriori* i wspomina pojęcie sprzężonych rozkładów apriorycznych (*ang.* conjugate priors). Stosowanie tych ostatnich implikuje, że rozkłady *a posteriori* należą do tej samej rodziny funkcji rozkładu co rozkład aprioryczny. Wartości parametrów tych rozkładów (*a priori* i *a posteriori*) są różne, a rolą bayesowskiej funkcji wiarygodności jest tu jedynie uaktualnienie apriorycznych parametrów modelu, przy zachowaniu jego funkcjonalnej postaci. Rodziny rozkładów sprzężonych były szczególnie ważnymi klasami funkcji w dobie małych mocy obliczeniowej. Niestety, takie podejście nie jest możliwe w przypadku identyfikacji parametrów źródeł uwolnień gazu z uwagi na dużą nieliniowość odwzorowania z przestrzeni parametrów do przestrzeni wyjść, realizowanego na dodatek poprzez bardzo złożone techniki symulacyjne. Wyznaczenie wartości tego odwzorowania jest konieczne do wyznaczenia wartości funkcji wiarygodności. Z konieczności należy się więc odwołać do symulacji Monte Carlo z wykorzystaniem łańcuchów Markowa, które próbkują z rozkładu *a posteriori*. Rozdział kończy graficzny model sekwencyjnego wnioskowania bayesowskiego zastosowanego w rozprawie do stochastycznej rekonstrukcji parametrów źródła uwolnienia gazu.

Rozdział 4, poświęcony metodom Monte Carlo wykorzystującym łańcuchy Markowa (*ang.* Markov Chain Monte Carlo, MCMC), przedstawia algorytmy próbkowania z rozkładu *a posteriori* zastosowane w rozprawie. Ślady początków tej klasy metod symulacyjnych sięgają lat 1944–1945 i projektu Manhattan w Los Alamos. Metoda MCMC polega na konstrukcji nieredukowalnego nieokresowego łańcucha Markowa, opisywanego pewnym jądrem przejścia, który po wystarczająco długim czasie podąży z obranego arbitralnie rozkładu początkowego do pewnego docelowego rozkładu stacjonarnego. W algorytmach tego typu możliwa jest znajomość rozkładów tylko z dokładnością do proporcjonalności, bez stałej normującej. To przesądza o ich kluczowej roli we wnioskowaniu bayesowskim, gdzie rozkład *a posteriori* jest odwrotnie proporcjonalny do rozkładu obserwacji, który w teorii można otrzymać jako rozkład brzegowy z łącznego rozkładu obserwacji i parametrów, jednak w praktyce jest to niewykonalne z uwagi na konieczność obliczania całek wielokrotnych po parametrach. W rozdziale przedstawia się klasyczny algorytm Metropolisa-Hastingsa, sekwencyjny algorytm Monte Carlo dla dynamicznych modeli Bayesa oraz bootstrapowy filtr Gordona. Sposób prezentacji tych metod jest dość bliski raportowi G. Johannesson, B. Hanley, J. Nitao pt. *Dynamic Bayesian Models via Monte Carlo – An Introduction with Examples* (Lawrence Livermore National Laboratory, 2004). Metody te służą w rozprawie porównaniu z metodami przybliżonych obliczeń bayesowskich (ABC), które Autor preferuje, wykorzystując jako podstawowe narzędzie w kolejnych rozdziałach. Rozdział uzupełnia charakterystyka idei i własności metod ABC. O sukcesie tego podejścia w praktyce przesądza właściwy dobór trzech kluczowych elementów: zestawu wartości progowych określających margines bliskości w przestrzeni wyjść, metryki definiującej odległość w przestrzeni wyjść, oraz jądra przejścia pozwalającego na generowanie nowych realizacji w przestrzeni parametrów. Szczegółowy opis doboru tych elementów dla problemu identyfikacji parametrów źródeł uwolnienia gazu znajduje się w kolejnych rozdziałach, poświęconych studiom przypadku związanym z danymi rzeczywistymi. Rozważania kończy sekwencyjna wersja podstawowego algorytmu ABC z adaptacją wag w zależności od danych, podwyższająca współczynnik akceptacji generowanych próbek, zaproponowana w rozprawie doktorskiej Fernando Bonassiego z Duke University, obecnie statystyka Google/Youtube.

Po przeczytaniu pierwszych trzech rozdziałów czytelnik posiada dobrą ogólną orientację w zakresie omawianych zagadnień, dotychczas stosowanych podejściach, oraz oryginalnych rozwiązaniach zaproponowanych przez Autora. Być może nieco obszerniejszego umotywowania i opisu wymaga jedynie część dotycząca dynamicznych modeli Bayesa, które rzadziej pojawia się w literaturze nt. MCMC i ABC.

W kolejnych trzech rozdziałach, stanowiących najbardziej wartościową część rozprawy, Autor przedstawia wyniki eksperymentów potwierdzających efektywność proponowanych przez Niego metod. Ta część stanowi wprawdzie aż połowę objętości pracy, jednak jest jeszcze bardziej interesująca od części pierwszej. Duże uznanie budzi przede wszystkim ogromny nakład pracy włożony w przeprowadzenie wszystkich badań, tak bardzo dalekich od trywialności. Ich rezultaty potwierdziły w praktyce zasadność zaproponowanej metodologii rekonstrukcji parametrów źródła dla trzech studiów przypadku dotyczących eksperymentów pomiarowych dla rzeczywistych uwolnień gazu. Parametry źródła są w nich znane, co umożliwia ocenę jakości ich estymat, jednak zasadnicza trudność polega na tym, że pomiary pochodzą z rzeczywistych procesów fizycznych, dla których modele matematyczne są tylko pewnym przybliżeniem, a wszelkie niepewności występujących w nich parametrów i danych pomiarowych nie mają podanych rozkładów. Z jednej strony stanowi to ogromne wyzwanie,

jednak z drugiej strony stanowi najlepszą weryfikację praktyczną proponowanego podejścia. Rezultaty badań przedstawiono w bardzo czytelny sposób poprzez wizualizację rozkładów *a posteriori* dla pojedynczych parametrów źródła i ich par, bayesowskie przedziały ufności dla pojedynczych parametrów i bayesowskie obszary ufności dla brzegowych rozkładów a posteriori par parametrów, estymaty maksymalnego prawdopodobieństwa *a posteriori* (*ang.* maximum a posteriori probability, MAP) oraz estymaty warunkowej wartości oczekiwanej.

W *rozdz. 5* dokonano porównania efektywności czterech algorytmów omówionych w rozdz. 4 (klasyczna i sekwencyjna wersja metody Metropolisa-Hastingsa, sekwencyjny algorytm Monte Carlo, algorytm przybliżonych obliczeń bayesowskich z adaptacją wag) dla danych z jednego dnia cyklu eksperymentów kopenhaskich (Copenhagen Tracer Experiments, 19/10/1978). Modelem symulacyjnym rozprzestrzeniania się gazu był SCIPUFF, a dane odpowiadają pomiarom z 40 sensorów i jednej stacji meteorologicznej dla trzech kroków czasowych. Wyniki porównania wskazują na przewagę zaproponowanej wersji algorytmu ABC, który dokładniej oszacowuje poziom emisji substancji (wszystkie algorytmy w podobny, dość dobry sposób oceniają dwie współrzędne położenia źródła – oceny wysokości emisji są jednak dla wszystkich metod raczej kiepskie).

Rozdział 6 podsumowuje wyniki estymacji parametrów ruchomego źródła zastosowanego w eksperymencie OLAD (*ang.* Over-Land Alongwind Dispersion) przeprowadzonego na jednym z poligonów armii USA dn. 9/09/1997. Ruch emitera realizowano przy pomocy samochodu ciężarowego poruszającego się ruchem jednostajnym wzdłuż linii prostej. Modelem symulacyjnym był ponownie SCIPUFF, a dane odpowiadały pomiarom z pięciu sensorów i sześciu stacji meteorologicznych dla 45 kroków czasowych. Ruchome źródło charakteryzowało siedem parametrów, które stosunkowo dobrze oszacowano z zastosowaniem autorskiej implementacji algorytmu ABC (tym razem nie dokonywano porównań z innymi metodami używanymi w poprzednim rozdziale). Zamieszczone tu rezultaty raportowano wcześniej w dwóch publikacjach, w których doktorant był współautorem (pierwsza z nich ukazała się w bardzo dobrym czasopiśmie *Atmospheric Environment* wydawanym przez Elsevier, a druga – w materiałach konferencyjnych wydawanych przez IOP Publishing Ltd.).

Bardzo wartościowy jest *rozdz. 7*, dotyczący estymacji sześciu parametrów źródła w bardzo znanym eksperymencie DAPPLE (*ang.* Dispersion of Air Pollution and its Penetration into the Local Environment), przeprowadzonym w centralnej części Londynu w dn. 28/06/2007. Fundamentalną trudnością była tu konieczność uwzględnienia wielu budynków w rozważanym obszarze. Modelem symulacyjnym był QUIC, a dane odpowiadały pomiarom z 18 sensorów i jednej stacji meteorologicznych dla 10 kroków czasowych. Narzędziem oceny parametrów źródła była ponownie autorska implementacja algorytmu ABC (tu również nie dokonywano porównań z innymi metodami używanymi w rozdz. 4). Przedstawione rezultaty opublikowano ponownie w czasopiśmie *Atmospheric Environment* wydawanym przez Elseviera, a wcześniej w materiałach konferencji *Bayesian Statistics in Action* (Florencja, 2016), wydanych przez Springera.

Rozprawę kończy podsumowanie oryginalnych wyników naukowych oraz charakterystyka otwartych problemów badawczych (*rozdz. 8)*.

Główną część uzupełnia dodatek dotyczący zastosowanych modeli dyspersji stężeń gazów w atmosferze. SCIPUFF to model typu Lagrange'a stosujący zbiór obłoków gaussowskich do symulacji trójwymiarowego niestacjonarnego pola stężenia gazu. W rozprawie używano jego implementacji jako oprogramowania domeny publicznej w postaci programu PC-SCIPUFF amerykańskiej firmy Titan Corporation. Z kolei QUIC to zestaw programów opracowanych przez University of Utah oraz Los Alamos National Laboratory, również wykorzystujących modele Lagrange'a (z uwagi na ich większą szybkość względem rozwiązywania równań różniczkowych cząstkowych) oparte o wyznaczanie średniego pola wiatru, które służy następnie do obliczenia turbulentnego rozproszenia substancji niebezpiecznej w oparciu o równania spaceru losowego. Specyfiką oprogramowania jest uwzględnienie odbijania cząstek od powierzchni budynków oraz dodatkowej dyspersji wskutek poziomych niejednorodności turbulencji.

Oceniając merytorycznie całą rozprawę stwierdzam, że jest ona napisana na bardzo dobrym poziomie. Zawiera jasno sformułowany i ważny problem naukowy, oraz prezentuje poprawne rozwiązanie tego problemu, które zostało uzyskane przez Autora samodzielnie i z zastosowaniem właściwej metodologii naukowej. Na podstawie przedstawionego skrótowo omówienia treści całej rozprawy doktorskiej należy odnotować, że jej Autor wykazał się dobrymi umiejętnościami formułowania problemów naukowo-badawczych oraz ich efektywnego rozwiązywania z zastosowaniem zaawansowanych narzędzi statystyki obliczeniowej, modelowania i symulacji zanieczyszczeń atmosfery, rozwiązywania problemów odwrotnych, uczenia maszynowego i technik algorytmicznych. Już na podstawie wstępnej analizy można stwierdzić, że rozprawa stanowi dzieło wartościowe, zdecydowanie odpowiadające wymaganiom stawianym przez stosowne przepisy.

Pod względem redakcyjnym pracę napisano z dużą dbałością o szczegóły. Użyte słownictwo odpowiada powszechnie stosowanemu (o drobnych wyjątkach wzmiankuję dalej). Jak na tak dużą objętość, zawiera niewiele błędów składu, co tym bardziej koresponduje z jej dobrym poziomem merytorycznym. Na szczególne podkreślenie zasługuje to, że mimo długiej historii, publikacji polskojęzycznych poświęconych rozwiązywaniu zagadnień odwrotnych z zastosowaniem metod Monte Carlo opartych o łańcuchy Markowa jest bardzo niewiele. Rozprawa jest więc również z tego powodu warta rozpropagowania.

III. Oryginalne osiągnięcia

Chociaż podejście do identyfikacji parametrów źródła emisji niebezpiecznych gazów do atmosfery oparte o wykorzystanie łańcuchów Markowa Monte Carlo jest już znane od dawna, to jednak poszukiwanie sposobów uczynienia go jeszcze bardziej efektywnym, a w szczególności przystosowania go do rozwiązywania problemów, w których dane pochodzą z rzeczywistych pomiarów, a nie z wyidealizowanych symulacji, jest zadaniem nadzwyczaj trudnym i wciąż aktualnym. Przedstawiony w pracy matematyczny opis problemu, jego analizę oraz zaproponowane oryginale modyfikacje znanych metody i algorytmów obliczeniowych uważam za najważniejszy wkład Autora w rozważaną dziedzinę. Chociaż przedstawiony pomysł wykorzystania przybliżonych obliczeń bayesowskich (ABC) wydaje się dość naturalny biorąc pod uwagę obecne trendy statystyki obliczeniowej, jednak jego implementacja dla rozważanych scenariuszy rzeczywistych emisji gazów nie jest bynajmniej oczywista, a recenzowana praca stanowi jedną z pierwszych tak całościowych prób jego

ujęcia i formalnego uzasadnienia. Główną zaletą podejścia proponowanego w rozprawie jest duża efektywność obliczeniowa zaproponowanych wersji algorytmu ABC oraz względnie prosta możliwość jego rozwinięcia do działającego prototypu.

Przyjmując, że głównym celem rozprawy było pokazanie, że zastosowanie zaawansowanych algorytmów próbkowania z rozkładu parametrów *a posteriori*, takich jak sekwencyjne metody Monte Carlo lub przybliżone obliczenia bayesowskie, pozwala na względnie szybką i dokładną rekonstrukcję parametrów źródła uwolnienia gazu, należy stwierdzić, że cel ten Autor osiągnął. Co więcej, weryfikacji rezultatów dokonano w oparciu o dane rzeczywiste pochodzące z uznanych benchmarków.

W szczególności, za najważniejsze rezultaty rozprawy uważam następujące:

1. zaproponowanie schematu bayesowskiej rekonstrukcji parametrów źródła wykorzystującego dynamiczne modele Lagrange'a i ukierunkowanego na pracę w trybie on-line, w miarę napływania kolejnych danych z sieci sensorowej.;

2. zastosowanie zaawansowanych algorytmów próbkowania, takich jak sekwencyjna metoda Monte Carlo oraz przybliżone obliczenia bayesowskie, dla których doświadczenia dla danych rzeczywistych potwierdzają co najmniej tak samo wysoką jakość ocen, jak dla konwencjonalnych metod próbkowania MCMC, przy jednoczesnej redukcji czasu obliczeń;

3. wprowadzenie autorskich modyfikacji algorytmu przybliżonych obliczeń bayesowskich (zmodyfikowana metryka rozbieżności z przestrzeni wyjść, oryginalna procedura adaptacji harmonogramu progów akceptacji, adaptacyjny mechanizm określania wag, model jądra przejścia dostosowany do mobilnego źródła emisji), mających na celu podwyższenie dokładności estymacji oraz skrócenie czasu obliczeń;

4. weryfikacja zaproponowanych rozwiązań w oparciu o dane rzeczywiste z trzech dużych eksperymentów obejmujących m.in. mobilne źródło emisji oraz rekonstrukcję źródła w warunkach dużej aglomeracji miejskiej z rozbudowaną infrastrukturą budynków.

Należy podkreślić, że uzyskane rezultaty są udokumentowane publikacjami zarówno w czasopismach (dwa artykuły w bardzo dobrym *Atmospheric Environment*, IF = 3.708, Elsevier, lista JCR, wg WoS kwartyl Q1 w Environmental Sciences oraz Meteorology and Atmospheric Sciences; jedna praca w Entropy, IF = 2.3, MDPI, lista JCR, wg WoS kwartyl Q2 w Multidicsiplinary Physics; publikacje w kwartalnikach *Operations Research and Decisions,* Politechnika Wrocławska, oraz *Foundations of Computing and Decision Sciences*, Politechnika Poznańska), jak i włączonymi do sześciu monografii oraz pięciu materiałów konferencyjnych wydawanych przez Springer, IOP Publishing oraz IEEE.

W podsumowaniu, należy stwierdzić, że **sformułowany cel rozprawy został osiągnięty, a jej Autor wykazał się głęboką wiedzą i umiejętnościami niezbędnymi do samodzielnego rozwiązywania problemów naukowo-technicznych w dyscyplinie *Informatyka*.**

IV. Uwagi i komentarze

Przedstawiona do recenzji praca zawiera istotną treść naukową i wiele nowych wyników. Stanowi logiczną całość począwszy od uzasadnienia praktycznego problemu, poprzez jego formalizację, aż do rozwiązania różnorakich wersji problemu z przykładami zastosowań zaproponowanej metodologii w nietrywialnych zadaniach testowych. Praca prezentuje wysoki poziom naukowy, a Autor biegle posługuje się aparatem matematycznym, m.in. w zakresie statystyki obliczeniowej, modelowania i symulacji, oraz algorytmiki.

Lektura rozprawy skłania jednak również do sformułowania następujących uwag krytycznych:

1. W pracy wykorzystuje się podejście oparte o dynamiczne modele Bayesa, które stosuje się najczęściej w sytuacji, gdy identyfikowane parametry zmieniają się w czasie. Wszystkie scenariusze rozważane w rozprawie dotyczą jednak ustalonych parametrów. Poza krótkim stwierdzeniem tego stanu rzeczy na str. 31 brak jest szerszego komentarza na ten temat. Dynamiczne modele Bayesa mogą wydawać się zbędną komplikacją, powodującą np. konieczność rozważania rozkładów łącznych określonych na przestrzeniach parametrów o coraz większym wymiarze. Co więcej, w sytuacji, w której można byłoby jednak wykorzystać moc takiego podejścia, czyli dla eksperymentu OLAD, w którym pozycja źródła rzeczywiście zmienia się w czasie, wykorzystuje się do maksimum specyfikę informację o sposobie poruszania się źródła (ruch jednostajny wzdłuż prostej o nieznanym kierunku), sprowadzając rozważania ponownie do identyfikacji zestawu parametrów stałych w czasie. Przecież w sytuacji identyfikacji mobilnego źródła tak dokładna informacja o sposobie poruszania się będzie rzadko dostępna.

2. Poza jednym przypadkiem, dotyczącym wysokości w eksperymencie DAPPLE, aprioryczne rozkłady parametrów są równomierne. Oznacza to, że znane są jedynie zakresy zmienności parametrów. W tej sytuacji na rozkład parametrów *a posteriori* decydujący wpływ będzie miała funkcja wiarygodności. Może więc dojść do sytuacji analogicznej do minimalizacji kryterium najmniejszych kwadratów bez członu regularyzującego w przypadku deterministycznych błędów pomiarowych, kiedy nawet niewielkie zaburzenia danych wyjściowych będą przekładać się na niestabilne oceny parametrów.

3. W symulacjach stosuje się najczęściej rozkłady normalne (np. w równaniu (5.5) definiującym jądro tranzycji stanów). Stwarza to jednak ryzyko wygenerowania fizycznie nierealnych wartości parametrów (np. ujemnej wysokości lub ujemnej intensywności źródła). Dużo lepsze byłoby posługiwanie się rozkładami normalnymi obciętymi do kostek zdefiniowanych przez dopuszczalne zakresy wartości parametrów.

4. Jak na rozprawę w dyscyplinie informatyka, niewiele mówi się o sposobie implementacji algorytmów, zastosowanych języków programowania i bibliotek (w przypadku algorytmów typu MCMC lub ABC dostępnych jest bezkosztowo wiele kodów, szczególnie w przypadku środowiska R lub języków Python, C++, Fortran 90 i Java). Poza dość lakoniczną informacją w rozdz. 6.4 o systemie maszyn wirtualnych

i pięciu instancjach modeli działających równolegle dla eksperymentu OLAD (Autor deklaruje, że wyniki uzyskano w czasie 2,5 h), nie charakteryzuje się ani architektury obliczeniowej, ani nie podaje szczegółowej analizy czasów dla różnych wariantów obliczeń. Porównanie tych ostatnich dla różnorakich wersji omawianych algorytmów byłoby podstawą do bardziej obiektywnej oceny efektywności rozwiązań proponowanych przez Doktoranta.

5. W rozprawie przyjęto nieco zbyt wąską perspektywę ulokowania rezultatów badań na tle innych wyników literaturowych. Brak jest odniesień do bardzo obszernych pokrewnych nurtów badań dotyczących, np. wariacyjnej asymilacji danych (np. do obszernej monografii D.G. Cacuci., I.M. Navon, M. Ionescu-Bujor, *Computational Methods for Data Evaluation and Assimilation*, CRC Press, 2013), wnioskowania bayesowskiego dla nieskończenie wymiarowych przestrzeni parametrów (chodzi chociażby o wspomniane prace Andrew M. Stuarta, w tym m.in. w R. Ghanem, D. Higdon, H. Owhadi, *Handbook of Uncertainty Quantification*, Springer, 2017), estymacji parametrów z zastosowaniem łańcuchów Markowa Monte Carlo dla modeli w postaci równań różniczkowych cząstkowych (wspomniana monografia Kaipio i Somersalo, prace Nicholasa Zabarasa z University of Warwick, a poprzednio Cornell University), prac o estymacji parametrów źródeł dla za zastosowaniem podejścia stricte deterministycznego, przy założeniu jedynie ograniczoności błędów pomiarowych (np. prace Vyacheslava Maximova z Uralskiego Oddziału Rosyjskiej Akademii Nauk w Jekaterynburgu). Brak również wzmianki o ważnym problemie optymalizacji rozmieszczania węzłów sieci sensorowej w celu maksymalnej dokładności estymat, chociaż w tym kontekście w literaturze rozważa się już zarówno podejście bayesowskie oparte o MCMC, jak i bardzo dużą liczbę parametrów (oprócz parametrów identyfikuje się również stan początkowy) – jest to obecnie bardzo aktywny nurt badań, zob. np. wspomniane prace Ghattasa, Alexanderiana, Petry i Stadlera .

6. Naturalnym pytaniem są perspektywy rozwinięcia proponowanego podejścia do sytuacji identyfikacji wielu źródeł, w skrajnym przypadku z nieznaną ich liczbą.

Ponadto pojawia się szereg nieprecyzyjności lub błędów składu:

1. Na str. 25 i wielu dalszych używa się określenia *domena* w miejsce dużo poprawniejszego słowa *obszar* (np. tu pisze się o kierunku i prędkości wiatru *w rozpatrywanej domenie*, a w pierwszej linii na str. 61 mówi się o zastosowaniu *tej samej domeny obliczeniowej*, zamiast *tego samego obszaru przestrzennego*).

2. Str. 16, wzór (2.1): wspomina się o tym, że przestrzenie parametrów oraz danych mają być przestrzeniami Hilberta bez uzasadnienia tego założenia. Kierując się taką logiką, dużo bardziej ogólne byłoby przecież założenie o przestrzeniach Banacha i takie sytuacje też często spotyka się w literaturze. Zaraz jednak potem, we wzorze (2.3) traktuje się zarówno parametry, jak i wyjście, jako elementy przestrzeni skończenie wymiarowych, bez stosownego komentarza dlaczego zrezygnowano z uprzednio wprowadzonych ogólnych ram przestrzeni Hilberta.

3. Str. 26, czwarty wiersz od dołu: zamiast pojęcia *teoretyczny szum* lepiej chyba mówić

o *błędach modelowania*?

4. Str. 31, wiersz 14 od góry: *wiarygodność globalna* $p(d_{\text{obs}}|I)$ *nazywana „ewidencją”* – angielskie *evidence* tłumaczy się w tym kontekście najczęściej jako *wiedza* lub *przesłanki*; tłumaczenie jako *ewidencja* jest bardzo dziwne.

5. Str. 38, wzór (3.19): symbol δ oznacza deltę Diraca, a nie deltę Kroneckera (chodzi o przybliżenie rozkładu ciągłego rozkładem dyskretnym, a ten wymaga użycia delt Diraca).

6. Str. 20, linia 11 od dołu: *modeli gaussowskich i semi-gusowskich*

7. Psedokod 4.1 na str. 43: określenie *losowanie proponowanego stanu łańcucha* jest mało klarowne. Lepsze byłoby *próbkowanie z rozkładu zastępczego* lub *próbkowanie z rozkładu proponowanego*.

8. Wzór (4.4), str. 46: jeśli ma zachodzić druga równość, zamiast $q_t(\hat{\theta}^t|\theta^{1:t-1})$ powinno być $q_t(\hat{\theta}^t|\theta^{1:t-1},d_{\text{obs}}^{1:t})$.

9. Pseudokod 4.3, str. 47: w pierwszej linii symbol *I* dla zmiennej losowej odpowiadającej indeksowi losowanej próbki koliduje z oznaczeniem *I* stosowanym wcześniej do oznaczenia wszystkich dodatkowych informacji, które mogą być użyte w modelu (zob. str. 30); w drugiej linii zamiast przecinka powinna być pionowa kreska (chodzi przecież o rozkład warunkowy); w tożsamości podanej w trzeciej linii logiczniej jest zamienić kolejnością elementy pary uporządkowanej.

10. Pseudokod 4.4, str. 48: uwagi identyczne, jak do Pseudokodu 4.3.

11. Str. 60, siódmy wiersz od góry: Stwierdzenie *wariancja* σ^2_{CTE} *została oszacowana jako podwojona wartość progu odczytu gazu* nie ma sensu z powodu tego, że wariancja i próg odczytu wyrażone są w różnych jednostkach fizycznych. Zamiast wariancji powinno się użyć odchylenia standardowego σ_{CTE}. Podobnie, w trzeciej linii od dołu na str. 61 mówi się, że *wartości* σ^2 *będą wynosiły 5% różnicy między maksymalnymi a minimalnymi wartościami poszczególnych* parametrów. Tu również σ^2 powinno się zastąpić σ.

12. Druga linia rozdz. 5.3.1 na str. 61: Algorytm Metropolisa-Hastingsa trudno nazwać, jak robi to Autor, metodą *state of the art*. Metoda jest co prawda metodą klasyczną, ale od jej powstania ponad 60 lat temu, zaproponowano rozwiązania dużo lepsze i bardziej współczesne, o czym pisze zresztą sam Autor.

13. Wzór (5.5), str. 61: Nie zgadzają się wymiary parametrów rozkładu normalnego: wartości oczekiwana jest wektorem o $4t$ elementach, więc macierz kowariancji powinna mieć rozmiar $(4t)\times(4t)$, a ma wymiar 4×4.

14. Str. 62, piąta linia od dołu. Zamiast *filtrze gordona* powinno być *filtrze Gordona*; str. 65, czwarta linia: zamiast *danymych* powinno być *danych*; str. 66, szósta linia od

dołu: zamiast *ufoności* powinno być *ufności*; str. 133, linia 13 od dołu: zamiast *Gaussowskich* powinno być *gaussowskich*; str. 134, tytuł tabeli 92: zamiast *Pasuill'a* powinno być *Pasquilla*; str. 135, trzecia linia po wzorze (9.5): zamiast *Lagrangeowska* powinno być *lagranżowska*.

15. Drugi wiersz na str. 63: Przyjmując, że obliczenia wykonuje się wg Pseudokodu 4.3, ze str. 47, zamiast $\theta_1^{1:2} \sim q_2(\theta_1^{1:2}, \theta_t^1)$ powinno być $\hat{\theta}_1^2 \sim q_2(\hat{\theta}_1^2 | \theta_t^1)$. Brak informacji o tym, jaką postać mają rozkłady q_t.

16. Wzór (5.9) na str. 64: macierz kowariancji powinna być oznaczona symbolem Σ^{k-1}. Jaką postać ma q_t ? Podobny błąd składu w zapisie macierz kowariancji występuje w równaniu (5.11) na str. 66.

17. Str. 65, dwie linie przed *Krok 1 w pseudokodzie 4.5*: Jak rozumieć stwierdzenie *jeśli przewidywania modelu są zupełnie błędne*? Na czym polega ta zupełna błędność? Podobna uwaga dotyczy trzeciej linii od góry na str. 106.

18. Str. 67, piąta linia od dołu: wyrażenie *maksymalna wartość funkcji gęstości prawdopodobieństwa* nie ma sensu z uwagi na to, że rozkład a posteriori jest rozkładem dyskretnym.

19. We wzorze (9.2) na str. 133 nie pojawia się zmienna *x*. Oznacza to chyba, że wszystko rozgrywa się w płaszczyźnie określonej równaniem $x = x_0$, o czym nie mówi się wprost. Ten wzór w trzech wymiarach nie ma sensu.

Powyższe uwagi krytyczne nie mają jednak przesadnego wpływu na ogólną opinię o recenzowanej dysertacji, którą oceniam jako bardzo wartościową. Uważam, że cele postawione przez Autora rozprawy zostały osiągnięte, poparte silnymi wynikami empirycznymi, oraz przedstawione w interesujący sposób.

V. Podsumowanie

Uwzględniając wyżej wymienione uwagi i komentarze oraz całość rozprawy doktorskiej wraz z oryginalnymi osiągnięciami naukowo-badawczymi stwierdzam, że

1. **recenzowana rozprawa doktorska Pana mgr. Piotra Kopki spełnia wszystkie wymagania ustawy o stopniach naukowych i tytule naukowym w odniesieniu do rozpraw doktorskich;**

2. **wnoszę o przyjęcie rozprawy oraz jej dopuszczenie do publicznej obrony przed Radą Naukową Instytutu Badań Systemowych Polskiej Akademii Nauk.**

Dariusz Uciński