dr Jörg Verstraete
Zakład Modelowania Komputerowego
Instytut Badań Systemowych
Polskiej Akademii Nauk
ul. Newelska 6
01-447 Warszawa

Warszawa, czerwiec 2018
(Załącznik 2a)

AUTOREFERAT

przedstawiający opis podstawowego osiągnięcia i pozostałego dorobku naukowego w związku z ubieganiem się o nadanie stopnia doktora habilitowanego

1. Dane osobowe

Imię:	Jörg Lodewijk Verstraete
Miejsce/data urodzenia:	Oostende, Belgia / June 1, 1976
ResearcherID:	F-9175-2013
Orcid/ScopusID:	0000-0002-7772-984X
Alternatywne klucze wyszukiwania w bazach publikacji:	Verstraete Jörg, Verstraete Joerg, Verstraete Jorg, Verstraete J.

2. Posiadane dyplomy i stopnie naukowe

11.2009	Instytut Badań Systemowych Polskiej Akademii Nauk *Nostryfikacja stopnia doktora* – rozprawa: *Fuzzy Modelling of Spatial Information*
17.04.2007	Ghent University, Belgia Stopień doktora w dziedzinie: nauki techniczne, dyscyplina: informatyka (**Engineering, Computer sciences)** Tytuł rozprawy: *Fuzzy Modelling of Spatial Information* (rozprawa w języku angielskim) Promotorzy: prof. em. dr. R. De Caluwe, prof. dr. G. De Tré
07.07.1999	Ghent University, Belgia Tytuł magistra w dziedzinie **Informatyka** (dyplom z wyróżnieniem)
02.07.1997	Ghent University, Belgia Ukończenie studiów pierwszego stopnia (bachelor) w dziedzinie **Informatyka**, *(dyplom z wyróżnieniem)*

3. Zatrudnienie w jednostkach naukowych

01.08.1999 - Ghent University, Dept. of Telecommunications and Information processing (TELIN)
- 30.04.2000 stanowisko naukowe (*Researcher)*

01.05.2000 - Ghent University, Dept. of Telecommunications and Information processing (TELIN)
- 30.04.2007 asystent (*Assistant)*

01.07.2007 - IncGEO vzw
- 30.06.2008 deweloper oprogramowania (*Developer)*

01.07.2008 - Ghent University, Dept. of Telecommunications and Information processing (TELIN)
- 30.11.2008 Deweloper oprogramowania (*Project developer)*

01.12.2008 - Ghent University, Dept. of Telecommunications and Information processing (TELIN)
- 19.04.2009 Deweloper oprogramowania (*Project developer via interim contract)*

20.04.2009 - Ghent University, Dept. of Telecommunications and Information processing (TELIN)
- 31.05.2009 Deweloper oprogramowania (*Software developer via interim contract)*

20.04.2009 - Onze Lieve Vrouw Presentatie Lokeren (higher grade secondary education)
- 30.06.2009 Nauczyciel informatyki (60% etatu)

30.04.2009 - Ghent University, Instituut Voor Permanente Vorming
- 25.06.2009 Wykładowca *„ICT Software and Data processing", część III (studia wieczorowe)*

01.02.2011 Wydział Matematyki i Nauk Informacyjnych Politechniki Warszawskiej
- 30.06.2016 Wykładowca/ asystent kursu Bazy danych (język wykładowy angielski)

21.02.2017 - University of Santiago De Compostela, CiTIUS
- 21.02.2019 *Stanowisko PostDoc*
(data planowana)

04.11.2009 - Instytut Badań Systemowych Polskiej Akademii Nauk
- do dzisiaj *Adiunkt*

4. Podstawowe osiągnięcie naukowe

Jako osiągnięcie naukowe, o którym mowa w art. 16 ust. 2 ustawy z dn. 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz stopniach i tytule w zakresie sztuki (Dz.U. 2003 nr 65 poz. 595 z późń. zm.) niniejszym wskazuję

monografię, której jestem jedynym autorem, zatytułowaną

Artificial intelligent methods for handling spatial data
Fuzzy rulebase systems and gridded data problems

Książka została przesłana do wydawnictwa Springer, do serii „Studies in Fuzziness and Soft Computing". Springer poinformował 26 czerwca 2018 r, że książka jest przyjęte do druku.

Jest to praca multidyscyplinarna, łącząca techniki sztucznej inteligencji z przetwarzaniem danych geograficznych. Praca ta bazuje na wiedzy zdobytej podczas pracy nad doktoratem oraz pracy w Instytucie Badań Naukowych PAN, znacznie ją rozszerzając. Wspomniana wyżej monografia prezentuje nowe algorytmy przetwarzania danych geograficznych oraz zastosowania zbiorów rozmytych w zagadnieniach przetwarzania danych przestrzennych.

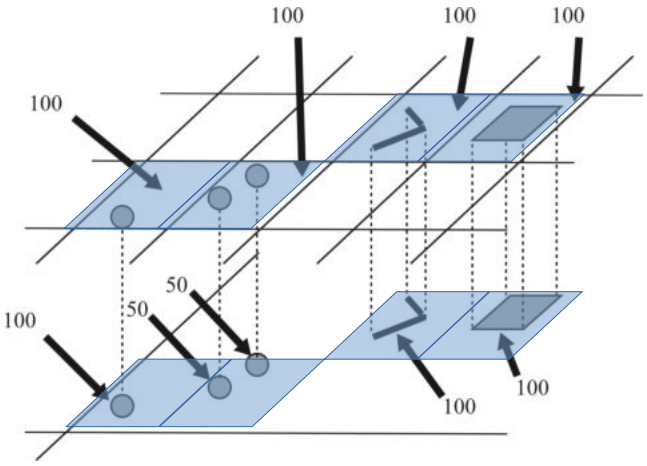
4.1 Cel naukowy

Głównym celem przeprowadzonych badań jest opracowanie metodologii, która pozwoli na dokonywanie przekształceń danych przestrzennych, opisanych na rastrach, bez wprowadzania dodatkowych niedokładności i niepewności danych oraz realizuje dezagregację przestrzenną, w szczególności problem zmiany definicji siatki, w stopniu znacznie lepszym niż dotychczas stosowane techniki. Obecnie istnieje bardzo ograniczony zestaw narzędzi do przekształcania siatek danych rastrowych, bazujący głównie na założeniu o statystycznym rozkładzie danych. Zwykle zakłada się jednolite i równomierne rozłożenie danych, i to niejednokrotnie pomimo dostępności dodatkowej wiedzy na temat kształtu tego rozkładu. Wspomniane tu dodatkowe informacje o rozkładzie są często zawarte w innych źródłach danych, niejednokrotnie reprezentowanych w innym formacie (na przykład zdefiniowanych na innej siatce itp.), stąd połączenie tych źródeł może nastręczać dodatkowych trudności. Metody sztucznej inteligencji oraz metody przetwarzania danych niepewnych lub niepełnych okazują się pomocne w przypadku, gdy nie znamy bazowego rozkładu danych w rozważanej przestrzeni. Metodologia opracowana w trakcie przeprowadzonych badań tworzy system, stanowiący sekwencję czynności (*workflow*), która symuluje inteligentne rozumowanie eksperta, nie wymagając przy tym interakcji z człowiekiem.

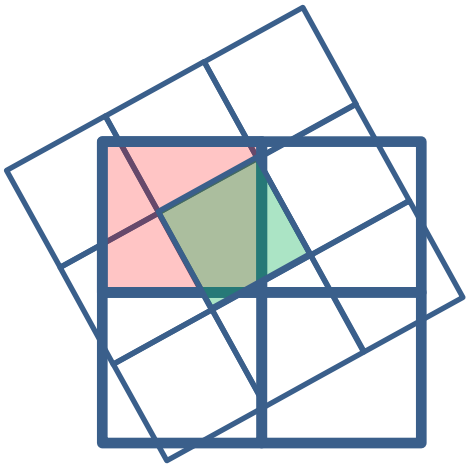
Coraz częściej zauważa się korzyści płynące z powiązania danych z informacją o ich lokalizacji, co prowadzi do rozwoju badań nad przestrzennymi aspektami danych. Sprzyja temu rozwój technologii pozwalający na zwiększenie dokładności pomiarów, miniaturyzacja oraz spadek cen urządzeń służących do lokalizowania obiektów. Obecnie prawie każdy telefon komórkowy jest

wyposażony w GPS, ale dostępne są również inne metody lokalizacji (np. badając zasięg sieci WI-FI). Jednocześnie rozwój technologii, zarówno w obszarze sprzętu, jak i oprogramowania, pozwala na wydajne gromadzenie większych ilości danych. Wzrost ilości gromadzonych danych oraz zwiększenie ich dokładności zwiększa także potencjał badawczy i stawia nowe wymagania odnośnie analizy danych. Pojawianją się nowe pytania: o to jak dane są połączone, jak na siebie wpływają, itp. Oczekuje się, że podając nowoczesnym systemom komputerowym więcej danych, otrzymamy nowe, nieodkryte wcześniej powiązania i rezultaty, które dla człowieka nie są widoczne. Przykład takiego systemu jest opisany w [1], gdzie badacze starają się połączyć dane o zanieczyszczeniu powietrza, ze statystycznymi informacjami o zaludnieniu; w [2] wspomniane dane są dalej łączone ze statystycznym stanem zdrowia i efektami wypadków drogowych, w celu określenia zalet i wad poruszania się rowerem w miastach, gdzie występuje znaczne zanieczyszczeniu powietrza. Inny przykład jest opisany w [3] gdzie badacze starają się przewidzieć obszary zagrożone epidemią cholery, bazując na rozkładzie temperatur, wilgotności, zasolenia wody oraz danych demograficznych. Problemem w tym przykładzie jest to, że niektóre czynniki, jak na przykład zasolenie, nie mogą być mierzone, a są estymowane z innych wartości, co wprowadza niepewności do rozważanego modelu.

W przestrzennych bazach danych i specjalizowanych systemach informacji geograficznej, używane są różne reprezentacje modelu danych [4,5]: wektorowy (*vector* lub *feature-based*), rastrowy (raster, siatki) oraz numeryczny model terenu. Wybór modelu zależy od typu i charakteru danych. Modele wektorowe wykorzystują podstawowe struktury geometryczne do reprezentacji rzeczywistych obiektów (np. fizycznych obiektów, takich jak drogi, ale także koncepcyjnych obiektów jak np. granice administracyjne). To sprawia, że modele te są bardzo użyteczne do modelowania środowiska i pozwalają na bardzo dokładne modelowanie świata rzeczywistego, wymaga to jednak wystarczająco dobrej jakości danych. Siatki lub modele rastrowe pozwalają na modelowanie numerycznych właściwości, których wartości zmieniają się dla każdej lokalizacji. Dane określone na siatkach modelują pewną cechę poprzez podzielenie regionu zainteresowania na pewną liczbę komórek i przypisanie każdej z tych komórek jednej wartości liczbowej, reprezentującej wartość w tej komórce. Model rastrowy zapewnia zdyskretyzowane oszacowanie ciągłej numerycznej cechy. Przykładem takich danych jest zanieczyszczenie środowiska (opisane w np. [6]), gdzie dla każdej lokalizacji możliwe jest określenie ilości poszczególnych zanieczyszczeń, ale nie jest możliwe zmierzenie natężenia danego zanieczyszczenia w każdym miejscu. Model rastrowy pozwala na określenie dyskretnej siatki wartości zanieczyszczeń. Ostatnimi typami modeli w przestrzennych bazach danych są numeryczne modele terenu, które mają bardzo specyficzne zastosowanie, dotyczące ukształtowania terenu i używają połączenia punktów próbkowanych (*sample*) z linearną interpolacją, poprzez konstruowanie nieregularnej siatki trójkątów.



Rysunek 1: Modele rastrowe i możliwy rozkład danych w poszczególnych komórkach.



Rysunek 2: Przykład niedopasowanych siatek

We wspomnianej pracy rozważany jest głównie rastrowy model danych. Skupiając się na modelach rastrowych, ważne jest, aby zdać sobie sprawę, że definicja siatki (czyli jej rozmiar i kształt komórek) determinuje przestrzenną dokładność reprezentacji, w której mniejsze komórki oznaczają większą przestrzenną dokładność. Definicja siatki w dużym stopniu zależy od zastosowania oraz od przestrzennej rozdzielczości, na której modelowane dane mogą być zbierane lub generowane. Drugim ważnym aspektem modeli rastrowych jest to, że wewnątrz pojedynczej komórki nie ma żadnej informacji na temat przestrzennego rozkładu danych. Wartość numeryczna, związana z komórką, jest wartością zagregowaną, która jest uznawana za reprezentanta całej przestrzeni należącej do komórki. Nie oznacza to jednak, że na całej powierzchni komórki dane są rozłożone równomiernie. Problem ten jest zilustrowany na rysunku 1., gdzie zaprezentowano różne rozkłady danych, agregowane do tej samej wartości w komórce. Z punktu widzenia siatki rastrowej, wszystkie te przypadki są w praktyce tym samym. Jeśli jednak zajdzie potrzeba przekształcenia siatki - na przykład zmiany rozmiaru siatki na gęstszą - nastąpi widoczna utrata precyzji i zwiększenie niepewności wartości. Przy dokonywaniu sekwencji wielokrotnych przekształceń, może nastąpić znaczące pogorszenie jakości danych wynikowych, szczególnie jeżeli geometrie siatek są różne i nie ma przekształcenia jeden-do-jeden dla poszczególnych siatek (Rys. 2.).

Połączenie danych z jednej siatki, z wartościami w drugiej siatce jest operacją, która wprowadza dużo niepewności i brak precyzji, szczególnie biorąc pod uwagę fakt, że każda komórka reprezentuje pewien przestrzenny rozkład danych, który nie może być uwzględniony. Celem niniejszych badań jest analiza możliwości stworzenia bardziej zaawansowanych technik rozwiązania tego problemu, uwzględniając fakt, że w wielu przypadkach dodatkowe zbiory danych mogą dostarczyć informacji na temat przestrzennego rozkładu danych, modelowanych na siatce. Przykładem takiego uzupełniającego zbioru danych zanieczyszczenia powietrza, jest mapa sieci dróg, jeżeli wiemy, że zanieczyszczenie jest związane z emisją przez samochody spalinowe. Dodatkowe dane, nazwane na potrzeby tej pracy danymi powiązanymi lub uzupełniającymi (*proxy*), w wielu przypadkach istnieją, jak na przykład kategoria drogi lub natężenie ruchu ulicznego w wyżej wymienionym przykładzie. Ekspert używając tej wiedzy, może potencjalnie zanalizować

dane i zdefiniować przestrzenny rozkład danych bardziej odpowiadający rzeczywistości. Jednakże, zbiory te są tak duże, że potrzebna jest technika automatyzacji tego procesu. Badania przedstawione w tej pracy, umożliwiły stworzenie sekwencji czynności – *workflow* lub przepływu pracy, używającego metod sztucznej inteligencji – a dokładnie systemów rozmytych, aby z minimalną ingerencją eksperta, wykorzystać istniejące dane powiązane. Wspomniana sekwencja czynności składa się z wielu nowatorskich metod, które rozwiązują poszczególne problemy, pojawiające się przy operacjach na danych przestrzennych i wprowadzają wiele nowych idei do obecnego stanu wiedzy. Warto nadmienić, że w trakcie pracy nad systemem, rozwinięto i zaimplementowano szereg algorytmów, których zastosowanie znacznie wykracza poza opisaną tutaj tematykę.

4.2 Osiągnięte wyniki

4.2.1 Opracowanie sekwencji czynności (*workflow*)

Operacje na danych rastrowych są określone przez algebrę Tomlina [7] i każdorazowo wymagają przyporządkowania jeden-do-jeden komórek siatek biorących udział w przekształceniu. Problem nakładania map/siatek (*map overlay problem*) jest zwykle rozwiązywany przez odwzorowanie jednej siatki (nazywanej siatką wejściową) na drugą (nazywaną siatką wyjściową). Po tej operacji istnieje przyporządkowanie jeden-do-jeden pomiędzy komórkami obydwu siatek, dzięki czemu, możliwe jest ich porównywanie, jak i wykonanie innych operacji. Oznacza to, że dla każdej komórki siatki wejściowej, wielkość wartości reprezentatywnej dla każdej z ich podziałów musi być określona. W badaniu tym założono, że modelowana wartość jest podzielna na rozważanym obszarze, na przykład gęstość zanieczyszczenia albo liczba ludzi na obszarze. W prosty sposób można przetłumaczyć dane z wielkości relatywnej (np. gęstość populacji lub koncentracja zanieczyszczenia), poprzez rozważenie wielkości obszaru, z którym dana komórka jest związana. Metodologia ta może być także użyta przy danych niepodzielnych (jak na przykład temperatury), ale przy uwzględnieniu pewnych warunków.

Przegląd popularnie używanych technik rozwiązywania problemu nakładania siatek, jest zaprezentowany w [8], z czego jedyna metoda radzenia sobie z brakującymi danymi przedstawiona jest w [9]. W literaturze, najbardziej popularne podejścia zakładają jednostajny rozkład danych w każdej komórce, co jest najbardziej intuicyjnym założeniem. Algorytm zmiany siatki zgodny z tym założeniem to tzw. ważenie przestrzenne (*areal weighting*), w którym proporcjonalny udział komórki w każdej części jest liniowo związany z jego odpowiednią częścią powierzchni pola. To podejście jest bardzo intuicyjne, jednakże daje mało dokładne rezultaty, gdy rzeczywisty rozkład znacznie odbiega od założonego. Innym popularnym podejściem jest założenie ciągłego i gładkiego rozkładu na założonym obszarze [10]. Wizualnie jest to osiągane poprzez rozważenie wartości komórki w trzecim wymiarze i dopasowanie gładkiej powierzchni na otrzymanej trójwymiarowej przestrzeni. Podobnie jak przy poprzedniej metodzie, rezultaty są bardzo niedokładne, jeżeli rzeczywisty rozkład danych różni się znacznie od założonego. Zdarza się to zwłaszcza w przypadku, gdy źródłem pewnej miary jest punkt na mapie (np. źródłem zanieczyszczeń jest jedna fabryka, która zanieczyszcza pewien obszar). Inne podejścia zakładają bardziej skomplikowane rozkłady danych, poprzez użycie modeli statystycznych (np. w [11]), wymagając jednocześnie dużo wiedzy eksperckiej i dając słabe rezultaty, w przypadku słabego dopasowania modelu do

rzeczywistości. Motywacją niniejszej pracy było zauważenie, że często dostępne są dodatkowe informacje, wpływające na rozkład rozważanych danych, które mogą być użyte dla lepszego dopasowania rozkład do rzeczywistego. Możliwe jest również połączenie kilku różnych źródeł danych do jednego zbioru (np. [12]) i użycie tych dodatkowych danych do oszacowania przestrzennego rozkładu innych danych.

Poniżej zdefiniowane jest kilka pojęć, ułatwiających wyjaśnienie szczegółów opracowanej metody. Rozważmy siatkę A, składającą się z komórek a_i, które mają przypisane wartości $f(a_i)$ oraz siatkę B, składającą się z komórek b_i o wartościach $f(b_i)$. Siatki obejmują ten sam obszar mapy, ale mają inną definicję geometrii (obiekty są opisane jako inne struktury geometryczne, np. linie lub poligony). Aby porównać wartości tych siatek, należy sprowadzić jedną siatkę do formatu drugiej, np. wartości na siatce A zaprezentować na siatce B. Oznacza to znalezienie wag x_i^j takich, że dla każdej komórki b_i jej wartość jet określana w następujący sposób:

$$f(b_i)=\sum_j x_i^j f(a_j).$$

Można to uprościć, zakładając, że wartość $f(b_i)$ jest określona tylko dla tych wartości komórek z A, z którymi się pokrywa, dając

$$f(b_i)=\sum_{j|a_j\cap b_i\neq\emptyset} x_i^j f(a_j).$$

Konieczne jest zachowanie ograniczenia, że wartość całkowita w komórkach modelowanych przez siatkę B, po zmianie siatki jest taka sama, jak wartości na siatce A. Dla własności przeliczalnej oznacza to, że suma wartości w komórkach każdej z siatek musi być równa tzn. $\sum_i f(b_i)=\sum_j f(a_j)$, co z kolei można wyrazić jako założenie, że wszystkie x_i^j są większe od zera i dla wszystkich wag x_i^j dla komórki b_i muszą się sumować do wartości 1.

$$\forall i,j: x_i^j\geq 0$$
$$\forall i: \sum_j x_i^j=1$$

Zarówno dla metod ważenia przestrzennego (*areal weighting*), jak i przestrzennego wygładzania (*areal smoothing*), ograniczenia te są spełnione. Aczkolwiek, przy próbie określenia wartości pojedynczej komórki b_i, bez jawnego określania wag metody, ograniczenia te -okazują się trudne do zweryfikowania i spełnienia. Rozwiązaniem jest wprowadzenie dodatkowych kroków algorytmu – problem zmiany siatki będzie najpierw sprowadzony do problemu dezagregacji przestrzennej, która jest specjalnym przypadkiem zmiany siatki, w przypadku gdy siatka docelowa B jest tak zdefiniowana, że dzieli każdą komórkę siatki A. To eliminuje częściowe nakładanie się komórek i upraszcza ograniczenia do następującej postaci: $f(a_j)=\sum_{j|a_j\cap b_i\neq\emptyset} f(b_i)$. Takie ograniczenie jest prostsze w weryfikacji, jako ograniczenie lokalne, czyli niezależne dla każdej komórki siatki wejściowej a_j. Przekształcenie problemu zmiany siatki na problem przestrzennej dezagregacji jest pierwszym krokiem opracowanej sekwencji czynności (*workflow*). Po nałożeniu siatek można geometrycznie wyznaczyć podział każdej komórki siatki A przez linie siatki B. Otrzymana w ten sposób nieregularna siatka jest nazywana siatką segmentów i jest wstępną siatką

wynikową. Ma ona ciekawą właściwość, że nie tylko dzieli siatkę A, ale też tworzy podział siatki B. Po przestrzennej dezagregacji na siatce A, komórki siatki segmentów mogą być połączone do siatki B. Ponieważ wartości są wymierne, wystarczy zsumować je w tych segmentach i otrzymamy prawidłowe wartości w komórkach siatki B. W związku z tym, problemem, który należy rozwiązać, jest problem przestrzennej dezagregacji i w przypadku zmiany siatki można go sprowadzić do obliczenia siatki segmentów.

Celem pracy jest stworzenie systemu, który mając dostęp do dodatkowych danych i używając metod sztucznej inteligencji, automatycznie przeprowadzi operację zmiany siatki z minimalną interakcją ze strony użytkownika (eksperta). Do realizacji tego zadania, wybrano technologię rozmytego systemu regułowego ([13]). Rozmyty system regułowy to metoda z dziedziny sztucznej inteligencji i rozmytej kontroli, która łączy dane wejściowe z danymi wyjściowymi przy pomocy reguł typu:

$$\text{JEŻELI } x_1 \text{ JEST } L_1^k \text{ ORAZ } x_2 \text{ JEST } L_2^k \text{ ORAZ } \ldots x_n \text{ JEST } L_n^k \text{ TO } y \text{ JEST } L_y^k$$

W powyższym wzorze, x_i jest zmienną, zawierającą numeryczne wartości, które są częścią wyrażenia lingwistycznego (np. niski, wysoki), a wynikiem jest ocena prawdziwości wyrażenia. Wyrażenia lingwistyczne są reprezentowane przez zbiór rozmyty zdefiniowany w domenie

```
WEJŚCIE: Siatka A, Siatka B, dane pomocnicze
WYJŚCIE: Siatka B
Siatka S =siatkaSegmentów(A,B)
Generuj bazę reguł używając danych treningowych
Dla każdego segmentu s w S
        Zbiór rozmyty zr = Oceń bazę dla s
        Rezultat r = Defuzyfikuj zr
        Przypisz r do segmentu s
Dla każdej komórki c w B
        Wartość c = agregacja nakładających się segmentów z siatki S
Zwróć B
```

Algorytm 1: Idea użycia rozmytej bazy reguł dla operacji zmiany siatki.

możliwych wartości, a ocena wartości wyrażenia x_i zawiera się w przedziale [0,1].

Wyrażenie lingwistyczne ocenia stopień spełnienia warunku przez każde x_i. Oceny są agregowane i wyznaczana jest wartość wynikowa y, która odpowiada zbiorowi rozmytemu, określającemu przynależność do wyrażenia lingwistycznego L_y^k. Baza reguł zawiera pewną liczbę reguł, z których każda definiuje wynikowy zbiór rozmyty, przy czym wiele reguł może pasować do określonego zbioru danych wejściowych. Wszystkie te zbiory są następnie agregowane do pojedynczego zbioru rozmytego, który jest wynikiem wnioskowania regułowego. Następnie zbiór zostaje sprowadzony do pojedynczego wyniku — nie rozmytej (ostrej) wartości. Sposób integracji systemu regułowego z sekwencją czynności jest przedstawiony w Algorytmie 1.

Podobieństwo procesu wnioskowania rozmytego do sposobu myślenia eksperta (osoby) było powodem, dla którego została wybrana ta metodologia dla problemu nakładania siatek. Osoba taka spojrzałaby na wartości wejściowe i zaczęła rozważać: „Jeżeli wartość w tych komórkach jest

wysoka, a dodatkowe dane wskazują, że w ich części powinna być niska, to znaczy, że należy przesunąć wyższe wartości do innej lokalizacji, a więc z dużym prawdopodobieństwem rozkład nie jest równomierny". Po utworzeniu siatki segmentów, każdy segment jest częścią komórki wejściowej, więc obliczenie wartości dla danego segmentu może być przeprowadzone według następującej reguły: „Jeżeli dodatkowe dane dla tego segmentu wskazują, że wartość ta powinna być wysoka, to przypiszę temu segmentowi wyższą wartość niż innym, dzielącym tę konkretną komórkę siatki wyjściowej". Jest to dokładnie odzwierciedlenie rozmytego wnioskowania regułowego: reguły są tworzone i oceniane dla każdego segmentu, a z kompilacji wyników otrzymuje się ostateczną wartość dla każdej komórki.

W trakcie implementacji wyżej opisanej sekwencji czynności, pojawił się szereg problemów, wynikających ze specyficznych cech danych przestrzennych. W kolejnych sekcjach zostaną omówione poszczególne problematyczne zagadnienia przetwarzania danych przestrzennych oraz metody radzenia sobie z tymi trudnościami. Kolejne sekcje przedstawiają nowe metody, dotyczące przetwarzania danych przestrzennych, skupiając się na interpretacji danych pomocniczych oraz konstrukcji bazy reguł. W kolejnych sekcjach zakłada się, że zbiór treningowy jest dostępny w momencie uruchomienia sekwencji.

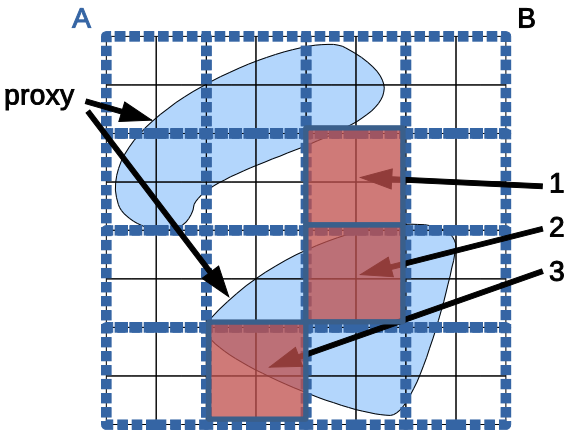
4.2.2 Parametry i zakresy danych

Pierwszym zagadnieniem przy opracowywaniu bazy reguł dla przestrzennej dezagregacji (oraz operacji zmiany siatki) jest ekstrakcja wiedzy zawartej w danych pomocniczych (*proxy*). W idealnym przypadku system powinien wykorzystać wszelkie dostępne informacje do oszacowania nieznanego rozkładu danych wejściowych. Pierwszym krokiem, aby to uzyskać, jest określenie jak dane pomocnicze są związane z danymi wejściowymi, poprzez zdefiniowanie parametrów (operacji definiujących pewne relacje pomiędzy danymi) oraz sprawdzenie czy te parametry dobrze te relacje określają. Na tej podstawie będzie możliwe wybranie najlepszych parametrów(zagadnienie to opisane jest w sekcji 4.2.3).

Przykładem parametru jest wartość określająca stopień nakładania się komórki segmentu z danymi pomocniczymi. Ta wartość jest proporcjonalnie zależna od rzeczywistego rozkładu danych wejściowych – im większy obszar segmentu pokrywa się z danymi pomocniczymi, tym wyższa wartość powinna zostać przypisana temu segmentowi. Parametr w systemie regułowym jest użyty po wyrażeniu JEŻELI i jest oznaczany jako x_1 . Przykładowa reguła może wyglądać następująco:

JEŻELI (pokrywanie się z danymi uzupełaniającymi z s) JEST L_1^k TO y JEST L_y^k

Automatyzacja procesu doboru i definiowania parametrów polega na próbie skonstruowania przez system różnych parametrów, które mogą być nieoczywiste lub mniej intuicyjne dla eksperta (użytkownika). Nawet, jeżeli korelacja pomiędzy danymi pomocniczymi i danymi wejściowymi nie wskazuje na relacje przyczynowe, nie jest to istotne dla naszego zastosowania. Jeżeli dane łączą się w jakiś identyfikowalny sposób z danymi wejściowymi, to mogą zostać użyte jako dane pomocnicze. Dzięki temu, w przeprowadzonych badaniach, możliwe było zdefiniowanie wielu różnych zmiennych, uwzględniając cechy takie jak bufory wokół geometrii oraz segmentów, wartości powiązane z danymi pomocniczymi, odległości, odwrotności odległości, i wiele innych.



Rysunek 3: Przykład ilustrujący problemy z wyznaczaniem zakresów wartości parametrów.

Bez względu na sposób zdefiniowania zmiennej x_i lub y, konieczne jest określenie zakresu możliwych wartości, na których można zdefiniować terminy lingwistyczne L_i^k. W typowych zastosowaniach baz reguł, zakresy dla x_i lub y są definiowane przez użycie ich wartości minimalnej i maksymalnej w zbiorze treningowym ([13]). Jednak w kontekście przestrzennym takie podejście daje nieprawidłowe zakresy. Powód tego najlepiej ilustruje przykład na rysunku 3, gdzie siatka A jest podzielona według komórek siatki B, przy czym dane pomocnicze są reprezentowane jako cieniowane kształty. Zakładamy, że dane pomocnicze są bardzo silnie skorelowane z wartościami na siatce wejściowej.

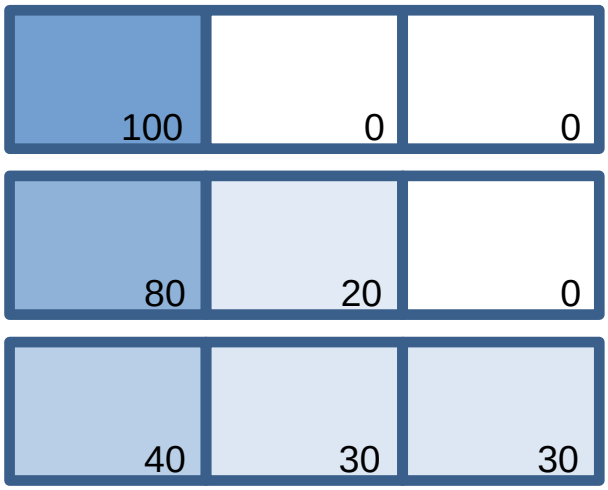
Standardowy zakres zmiennej, zdefiniowanej jako obszar pokrywania się danych pomocniczych z komórkami siatki wejściowej, na rysunku 3 byłby w przedziale $[0,1]$ i wyrażałby znaczeniowo zakres od „dane pomocnicze nie nakładają się z komórkami siatki” do „dane pomocnicze w pełni pokrywają komórkę siatki”.

Na rysunku 3 wyróżnione są trzy lokalizacje. Odnoszą się one do różnych komórek wejściowych, z których każda jest podzielona na cztery segmenty. Jak wcześniej wspomniano, celem zadania jest rozłożenie danych w komórkach wejściowych na segmenty, które ją dzielą (w tym przykładzie jest ich cztery). W lokalizacji 1 wszystkie cztery segmenty, które należą do jednej komórki wejściowej, mają bardzo niskie wartości, wyrażające stopień nakładania się z danymi pomocniczymi; najwyższa występująca wartość może wynosić zaledwie 0.05. W wyniku tego, baza reguł dla każdego z tych segmentów oceni wartość parametru x_1 prawie dokładnie tak samo- różnica w ocenie między 0 a 0.05 będzie zaniedbywalna, ponieważ są one oceniane względem pełnego zakresu $[0,1]$. Baza reguł dla każdego segmentu wyliczy takie same wartości, a system nie będzie w stanie zróżnicować tych czterech segmentów. Można by argumentować, że wyniki będą podobne, ponieważ pole obszaru nakładania się geometrii jest bardzo podobne, jednak to rozumowanie nie uwzględnia faktu, że ze względu na ogólnie niewielkie nakładanie się danych pomocniczych na daną komórkę, wartość komórki wejściowej będzie najprawdopodobniej również niska. Idealnie, tylko jeden z segmentów należących do komórki 1 powinien mieć wartość niezerową. Wynika z tego, że dane w komórce wejściowej powinny być tak odwzorowane, żeby segment z największym polem nakładania miał najwyższą wartość.

Symetryczna sytuacja występuje w pobliżu komórki 2, gdzie wszystkie oceny będą porównywalnie wysokie (na przykład w zakresie od 0.95 do 1) i ponownie system nie będzie w stanie rozróżnić tych czterech segmentów. W lokalizacji 3 problem nie występuje, ponieważ występujące wartości są rozłożone dość równomiernie w całym zakresie. Typowym rozwiązaniem tego problemu, byłoby zwiększenie czułości bazy reguł, poprzez zwiększenie liczby wyrażeń lingwistycznych do tego stopnia, żeby najlepiej dopasowane wyrażenia lingwistyczne były różne dla każdego przypadku. Wymagałoby to jednak dużej liczby wyrażeń, a co za tym idzie i reguł, co negatywnie wpłynęłoby na wydajność obliczeń. Inną opcją może być zdefiniowanie wyrażeń lingwistycznych w bardziej optymalny sposób w domenie ([14,15]), co jednak nie rozwiązuje problemu. W obu rozwiązaniach ocena reguł nadal będzie dawać podobne wartości, więc żadne z nich ostatecznie nie rozwiązuje problemu. Rozwiązaniem opracowanym w tej pracy badawczej, jest uznanie regionalnych różnic w danych i odpowiednie dostosowanie zakresów i konstrukcji reguł. W lokalizacji 1 wszystkie pola obszarów danych pomocniczych są niewielkie, więc zakres, w jakim powinny być oceniane, jest inny niż zakres w lokalizacji 2, gdzie obszar danych pomocniczych zajmuje większą powierzchnię komórki danych wejściowych. Jest to z kolei inny przypadek niż w lokalizacji 3, w której obszar nakładania się ma pełny możliwy zakres wartości. W związku z powyższym, opracowano dwie różne metody obliczania odpowiedniego zakresu dla danej zmiennej, a także nowe algorytmy budowy i oceny systemu reguł, w celu uwzględnienia tych lokalnych zakresów (sekcja 4.2.4).

Obliczenia zakresu wartości parametru dla danego segmentu można zdefiniować na kilka sposobów, przynależnych do trzech kategorii zakresów. *Zakres globalny* to nazwa nadana aktualnie używanej metodzie, która definiuje ten sam zakres dla wszystkich komórek wprowadzanych do systemu regułowego. Kolejne kategorie są zdefiniowane w niniejszej pracy, specjalnie dla danych przestrzennych. Nowe kategorie to: *zakres lokalny*, który definiuje zakres wartości parametrów, używając wartości minimalnych i maksymalnych dla parametru, z wybranego zestawu danych oraz *zakres szacunkowy*, w którym minimalne i maksymalne wartości są obliczane osobno dla danej wartości konkretnego parametru. W zakresie lokalnym zaproponowano różne sposoby definiowania zestawów danych do obliczenia zakresu. - ponieważ dane odnoszą się do segmentu siatki, można wykorzystać przestrzenne połączenie między tymi segmentami i innymi danymi. Przykładami są: parametry uwzględniające określoną odległość od bieżącego segmentu, parametry uwzględniające powiązania z innymi segmentami, należącymi do tej samej komórki danych wejściowych lub bardziej skomplikowane predykaty, dotyczące relacji przestrzennych. Podobnie w przypadku zakresu szacunkowego, gdzie zaproponowano i wdrożono różne definicje, określające możliwe wartości minimalne i maksymalne dla rozważanych parametrów. Przykładem definicji takiego zakresu jest następujący zakres: rozważane są wszystkie segmenty nakładające się z jedną komórką siatki wejściowej, minimalną wartość parametru definiuje się jako 0 (co wskazuje, że wartość jest odwzorowana całkowicie w innym segmencie), a maksymalna możliwa wartość jest równa wartości w komórce siatki wejściowej (co wskazuje, że segment ten odwzorowuje całą wartość komórki).

Można podać wiele różnych definicji wartości parametrów i połączyć je z różnymi sposobami określania zakresów, co daje bardzo duży zestaw możliwych parametrów, przy czym niektóre z nich osiągają dobre wyniki, podczas gdy inne nie przyczyniają się do dobrego oszacowania podstawowego rozkładu przestrzennego danych wejściowych. Może to wynikać z faktu, że sam



Rysunek 4: Przykład porównania danych na siatkach.

zestaw danych nie nadaje się do roli danych pomocniczych lub, że rozważane obliczenia dla wartości parametru i / lub jego zakresu nie dają wartości, które wykazują relację z rozkładem którego poszukujemy. Takie parametry nie poprawią wyniku przekształcenia natomiast niepotrzebnie powiększą bazę reguł. Wybór najbardziej odpowiednich zmiennych i definicji zakresów a priori, pozwoliłby wykorzystać najlepsze parametry, a tym samym poprawić wydajność. W związku z tym, należy ocenić, czy kombinacja definicji zmiennej i zakresu jest odpowiednia dla zadanych danych pomocniczych, które mają być uwzględnione w bazie reguł. Aby ocenić zmienną i jej definicję zakresu pod względem jej przydatności, przyjęto następujące podejście: dla wszystkich komórek wyjściowych obliczamy zmienną wartości i zakres; zmienna jest skalowana w swoim zakresie i przypisana do komórki. Wszystkie komórki będą teraz miały wartości w zakresie $[0,1]$. Następnie wartości komórek są skalowane tak, aby całkowita suma wartości w komórkach była równa sumie wartości w siatce wyjściowej zbioru treningowego. Jeśli siatka ta przypomina idealną siatkę wyjściową, to, można przyjąć, że kombinacja zmiennej i zakresu dla tego zestawu danych pomocniczych dobrze odpowiada podstawowemu rozkładowi siatki wyjściowej. Należy zwrócić uwagę, że w tym momencie obie siatki mają tę samą geometrię: istnieje odwzorowanie jeden-do-jeden między komórkami. To jednak powoduje, że powstaje nowy problem: jak określić czy dwa zestawy danych przestrzennych (opisanych na rastrach) odwzorowują ten sam rozkład. Problem można zdefiniować szerzej: jak porównać które zbiory danych przestrzennych odwzorowują dany rozkład lepiej niż inne. Zagadnienie to nie było do tej pory szerzej badane, więc nowe, autorskie rozwiązanie tego zagadnienia jest dokładniej opisane w następnej sekcji.

4.2.3 Porównywanie danych na siatkach

W literaturze podobieństwo zbiorów danych określa się często za pomocą korelacji Pearsona, sumy kwadratów różnic lub innych podobnych miar (np. [16]). Jednak te miary podobieństwa uwzględniają tylko liczbowe aspekty danych, nie uwzględniają natomiast innych, jak na przykład aspektu przestrzennego. Aspekt ten powoduje, że standardowe metody, takie jak korelacja Pearsona dają nieprawidłowe wyniki. Problem ten jest szerzej omówiony w [17], tutaj kwestia ta zostanie przedstawiona na prostym przykładzie. Rysunek 4 pokazuje trzy niewielkie przykłady siatek, z których każda ma tylko trzy komórki. Idealny rozkład przestrzenny jest pokazany na górnej siatce,

gdzie wartości wynoszą (100, 0, 0). Poniżej zaprezentowane są dwie siatki, uważane za podobne do siatki pierwszej, i mające odpowiednio wartości (80, 20, 0) i (40, 30, 30). Korelacja Pearsona pierwszej i drugiej siatki daje wynik 0.97, natomiast korelacja pierwszej i trzeciej siatki wynosi 1.0. Oznacza to, że zgodnie z korelacją Pearsona, trzecia siatka bardziej przypomina pierwszą. Może to być prawdą w niektórych interpretacjach (np. te same dane, reprezentowane w różnych jednostkach), jednak sytuacja zmienia się, gdy uwzględnimy rozkład przestrzenny. Wartości w pierwszej siatce wyraźnie pokazują, że większość danych znajduje się po lewej stronie siatki. Ponadto, gdy trzecia siatka ma również największą wartość w skrajnie lewej komórce, druga siatka wyraźnie stanowi lepszą reprezentację „danych położonych głównie po lewej stronie siatki". Trudnością w określeniu tego rodzaju podobieństwa i powodem, dla którego klasyczne miary podobieństwa, np. korelacja Pearsona, zawodzą, jest to, że nie wszystkie różnice są równoważne. W opisanym przykładzie w przypadku danych (80,0,20), wynik powinien być gorszy, niż ten zaprezentowany w drugim przykładzie, ale tylko dlatego, że wysokie wartości powinny znajdować się w pobliżu lewej strony. Korelacja Pearsona dla wartości (80,0,20) oraz dla wartości (80,20,0) byłaby taka sama, podobnie suma kwadratów różnic.

W pracy tej opracowano nowy algorytm do określania, na ile dwie siatki są podobne do tego samego przestrzennego rozkładu danych. Przedstawiono go poniżej (Algorytm 2).

Stworzenie siatki odniesienia:
WEJŚCIE: Siatka A, Maska M
Siatka R = taka sama geometria jak w A
Dla każdej komórki c w A (lub R)
    L = min(wartości komórek w M)
    U = sum(wartości komórek w M)
    F = zbiór rozmyty zdefiniowany
        używając f(c), L oraz U
    Przypisanie F do komórki c w R
zwróć R

Określenie wartości w rankingu:
WEJŚCIE: Siatka odniesienia R, siatka B
m = 0
Dla każdej komórki w B (także komórki w R)
    m+= Ocena f(b) w F (z R)
zwróć m/(liczba komórek w B)

Algorytm 2: Stworzenie siatki odniesienia i określenie wartości rankingu w opracowanym algorytmie oceny podobieństwa.

Metoda zaczyna się od ustalenia, która z siatek jest siatką odniesienia (w powyższym przykładzie byłaby to siatka górna). Następnie, konieczne jest zdefiniowanie maski, która definiuje sąsiedztwo komórki. W przypadku siatki regularnej, może ona być otoczeniem komórki, dla

nieregularnej siatki należy skonstruować maskę dla danego przypadku, na przykład przez uwzględnienie komórek leżących bliżej niż pewna zdefiniowana odległość. Do maski mogą zostać dodane wagi dla każdego elementu. Dzięki temu, maska może definiować stopień rozproszenia danych na pewnym obszarze i nadal wskazywać na duże podobieństwo. Biorąc minimalną wartość w komórkach L i sumę (lub sumę ważoną) U, otrzymano wstępny zakres wartości $[L,U]$. Zakres ten można uznać za wsparcie trójkątnego zbioru rozmytego, przy czym f(c) jest najbardziej prawdopodobną wartością. Może to jednak być rozkład skośny, który, w algorytmie rankingowym, miałby ten skutek uboczny, że różnice od najbardziej prawdopodobnej wartości nie byłyby traktowane symetrycznie. Ponieważ symetria jest właściwością pożądaną, można to rozwiązać przy pomocy symetrycznego odstępu $[x_1,x_2]$ wokół f(c) postaci

$$[x_1^i,x_2^i]=[f(a_i)-max(f(a_i)-L_i,U_i-f(a_i)),f(a_i)+max(f(a_i)-L_i,U_i-f(a_i))]$$

Przedział ten jest użyty do zdefiniowana funkcji $f_1^i(x)$ oraz $f_2^i(x)$, potrzebnych do zdefiniowania trójkątnego zbioru rozmytego na siatce odniesienia. Funkcje te przyjmują następującą postać

$$f_1^i(x)=\frac{1}{f(c_i)-x_1^i}(x-f(c_i))+1$$

$$f_2^i(x)=\frac{1}{f(c_i)-x_2^i}(x-f(c_i))+1$$

Ponieważ przyjmowane są wartości dodatnie (np. w przypadku stężeń zanieczyszczeń), zbiór rozmyty określony za pomocą powyższych funkcji jest odcięty na poziomie 0, tworząc zbiór rozmyty, względem którego oceniane będą wartości w komórkach rozważanych siatek. Każda komórka siatki odniesienia będzie zatem miała zdefiniowany własny zbiór rozmyty.

Ranking siatek jest otrzymywany poprzez przypisanie każdej siatce wartości oceny. Dla każdej siatki wartość każdej komórki jest oceniana względem zbioru rozmytego komórki siatki odniesienia. Wszystkie te wartości są w zakresie $[0,1]$ i są agregowane przy użyciu średniej, co daje jedną wartość wskazującą podobieństwo siatki. O ile sama wartość jest trudna do zinterpretowania i zależy od wielu czynników, to jest ona bardzo przydatna do określenia, która siatka lepiej przybliża siatkę odniesienia. W celu uzyskania pełnych informacji odnosimy się do [17]. Algorytm porównywania siatek został opracowany dla wsparcia rozwiązywania problemu nakładania siatek, ale może być niezależnie stosowany jako samodzielny algorytm bądź użyty w innych zastosowaniach i dziedzinach badań.

4.2.4 Konstrukcja bazy reguł i ewaluacja

W sercu opracowanego systemu jest system wnioskowania rozmytego Mamdani. Jak wspomniano w sekcji 4.2.1, jest to system z wieloma regułami postaci:

JEŻELI x_1 JEST L_1^k ORAZ x_i JEST L_i^k ORAZ ... x_n JEST L_n^k WTEDY y JEST L_y^k

gdzie wartości liczbowe są oceniane na podstawie wyrażeń lingwistycznych, reprezentowanych przez zbiory rozmyte. Typowym sposobem konstruowania takiego systemu jest użycie zbioru uczącego, czyli zbioru zestawów danych, które zawierają zarówno wartości wejściowe, jak i wartości wyjściowe. Algorytm budowy systemu reguł według Mamdani można znaleźć w [13]. Konstrukcja zazwyczaj rozpoczyna się od określenia zakresu danych wejściowych i wyjściowych,

które są następnie dzielone na obszary rozmyte, z którymi zostają powiązane wyrażenia lingwistyczne. Zakres zmiennej (wejściowej lub wyjściowej) jest ogólnie definiowany poprzez znalezienie najmniejszych i największych wartości, występujących w zbiorze treningowym, dla ustalenia zakresu możliwych przyjmowanych wartości. Jak wyjaśniono w sekcji 4.2.2, uwzględnienie aspektu przestrzennego sprawia, że definiowanie globalnych zakresów wartości parametru może nie być wystarczające. Może to skutkować niemożnością rozróżniania, przez system wnioskowania, podobnych wartości. W sekcji 4.2.2 pokazano, że możliwe jest zdefiniowanie bardziej odpowiedniego zakresu dla każdej zmiennej. Aby możliwe było stworzenie systemu bazy reguł, zarówno algorytm tworzenia reguł, jak i algorytm ich oceny, muszą zostać zmodyfikowane. Jedną z możliwych modyfikacji jest zmiana skali wartości. Zamieszczony w rozdziale 6 pracy [18] przykład pokazuje jednak, że jest to niewystarczające dla zastosowania przy danych przestrzennych, ponieważ ogranicza możliwości defuzyfikacji wyniku.

Opracowane rozwiązanie (Algorytm 3) wprowadza pojęcie przestrzeni zmiennych i oddziela definicję wyrażeń lingwistycznych od tych przestrzeni. W fazie uczenia, zbiory rozmyte związane z wyrażeniami lingwistycznymi służą jedynie do wyboru, które wyrażenia są najodpowiedniejszymi reprezentacjami wartości parametrów. Kiedy zostanie to określone, kolejne operacje oceny i łączenia reguł polegają na dopasowaniu wyrażeń lingwistycznych, zamiast łączenia powiązanych zbiorów rozmytych. To rozłączenie znajduje odzwierciedlenie w powyższym nowo opracowanym algorytmie., gdzie na wstępie, dla każdej zmiennej wybierana jest pewna liczba wyrażeń lingwistycznych, dla których zostają zdefiniowane zbiory rozmyte.

W fazie tworzenia bazy reguł, przetwarzany jest każdy warunek i wartość parametru w zbiorze treningowym, ustalane są zakresy dla każdej ze zmiennych (sekcja 4.2.2) oraz zdefiniowane zostają wyrażenia lingwistyczne dla zmiennej, która jest wyskalowana i dopasowania do przyjętego zakresu. Wyrażenia lingwistyczne, odnoszące się do zbiorów rozmytych, najbardziej pasujących do wartości parametrów, zostają wybrane do stworzenia reguł. Każdej regule przypisana jest waga (bazująca na wartości funkcji przynależności), a następnie reguła ta jest dodawana do bazy reguł. Od tego momentu ważne są wyrażenia lingwistyczne, a nie zbiór rozmyty. Wyrażenia lingwistyczne są rozważane w odniesieniu do zakresu. Określenie „wysoki” będzie zatem oznaczać „wysoki w kontekście rozważanej jednostki”. Baza reguł nadal będzie w spójny sposób dopasowywać wyrażenia językowe wartości wejściowych, do określonych wartości wyjściowych. Również późniejsze tworzenie połączonej bazy reguł odbywa się w sposób tradycyjny, ponieważ nie jest wymagana żadna dodatkowa wiedza na temat zbiorów rozmytych.

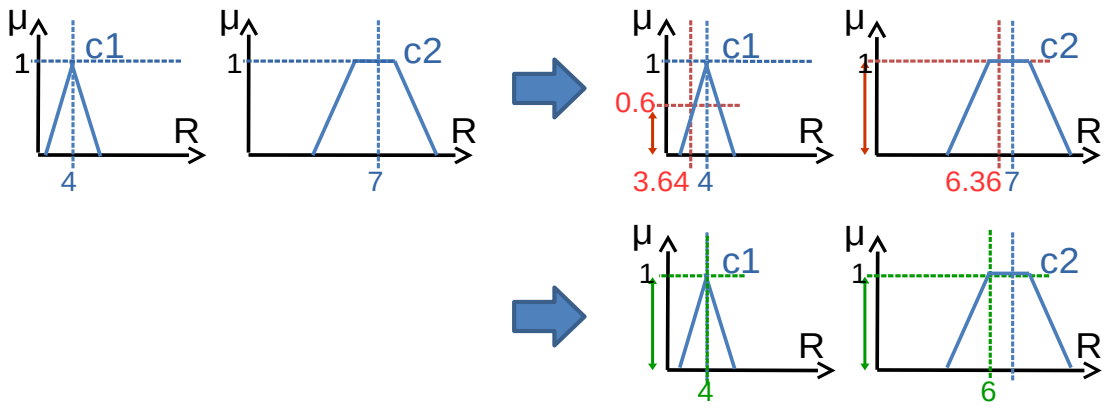
Zastosowanie bazy reguł również wymaga znajomości zakresów dopuszczalnych wartości. Jest to bowiem niezbędne do prawidłowego przeskalowania kształtu zbiorów rozmytych dla wyrażeń lingwistycznych. Zakres zmiennej wyjściowej również musi być określony, aby umożliwić prawidłowe zdefiniowanie powiązanych wyrażeń lingwistycznych. Baza reguł może być następnie zastosowana tak, jak w tradycyjnych systemach wnioskowania Mamdani, czego wynikiem jest zbiór rozmyty zdefiniowany w odpowiedniej domenie.

Przedstawiony poniżej algorytm ma bardzo konkretne zastosowanie oraz wymaga danych, dla których jest możliwe określenie przedziałów wartości dla każdej zmiennej. W rezultacie pozwala on systemowi wnioskowania na operowanie w zakresach bardziej dopasowanych do danych – wyrażenie lingwistyczne „wysoki” może być teraz zależne od danych i interpretowane jako „wysoki dla tego parametru i zakresu”. To nie tylko otwiera nowe możliwości stosowania regułowych systemów wnioskowania, ale także zapewnia dodatkowe narzędzie, które może pomóc w zmniejszeniu rozmiaru systemów reguł bez zmniejszania ekspresji stanów i reguł.

Tradycyjna konstrukcja bazy wiedzy
Podzielenie przestrzeni wejściowej na rozmyte regiony
Podzielenie przestrzeni wyjściowej na rozmyte regiony
Zdefiniowanie wyrażeń lingwistycznych dla tych przestrzeni
Dla każdego parametru i wartości wykonaj
    Wygeneruj regułę rozmytą
    Przypisz wagę do każdej reguły
Stwórz połączoną bazę reguł rozmytych

Nowo opracowana konstrukcja bazy wiedzy
Dla każdego parametru i wartości wykonaj
    Wyznacz przestrzeń wejść
    Wyznacz przestrzeń wyjść
    Zdefiniuj wyrażenia lingwistyczne
        w tych przestrzeniach
    Wygeneruj rozmytą regułę
    Przypisz wagę dla reguły
Stwórz połączoną bazę reguł rozmytych

Algorytm 3: Nowa konstrukcja bazy reguł dla przestrzeni zmiennych przestrzeni



Rysunek 5: Ilustracja defuzyfikacji z ograniczeniami; górne rozwiązanie jest otrzymane przez defuzyfikację każdego zbioru indywidualnie, dolne rozwiązanie pokazuje możliwości wyznaczania wyniku, gdy rozważane są wszystkie zbiory jednocześnie.

4.2.5 Defuzyfikacja z ograniczeniami

Defuzyfikacja to proces wyodrębniania pojedynczej ostrej wartości ze zbioru rozmytego. Jest to operacja bardzo istotna, ponieważ często zbiór rozmyty nie może być użyty do kolejnych operacji i należy określić dla niego najbardziej odpowiednią wartość ostrą. Zwykle odbywa się to poprzez wybranie funkcji defuzyfikującej, przyjmowaną przez zbiór rozmyty, która wyznacza dla niego wartość ostrą. Istnieją różne defuzyfikatory, z których każdy ma inne właściwości ([19]). W opracowanym podejściu do problemu przekształcenia siatki, lepsze wyniki można uzyskać nie defuzyfikując pojedynczych zbiorów rozmytych, ale rozpatrując kilka zbiorów rozmytych łącznie. Pozwala to na wybór lepszych wartości ze zbiorów rozmytych. Istniejące badania nad zniekształcaniem w ramach ograniczeń ([20]). Koncentrują się one na ograniczeniach nałożonych na domenę pojedynczego zbioru rozmytego. W związku z tym, konieczne było opracowanie nowatorskiej metody defuzyfikacji, biorącej pod uwagę wiele zbiorów rozmytych jednocześnie i wyznaczającej wartości ostre dla każdego z tych zbiorów, zapewniając, że nałożone ograniczenia są spełnione i najniższa wartość funkcji przynależności jest maksymalizowana dla wszystkich zbiorów. Rozwinięto nowatorską defuzyfikację. Rozważa ona wiele zestawów rozmytych jednocześnie i generuje błędne wartości dla każdego z nich, upewniając się, że wstępnie zdefiniowane ograniczenie zostało spełnione, a najniższa ocena członkostwa z wyników dla poszczególnych zbiorów jest zmaksymalizowana.

System wnioskowania rozmytego zwraca, w wyniku, zbiór rozmyty, a defuzyfikacja jest ważnym elementem, pozwalającym na prawidłową interpretację wyniku. W większości zastosowań system wnioskowania zwraca ostateczny wynik końcowy. Jednak w przedstawionej w tej pracy sekwencji czynności, system wnioskowania służy do określania wartości segmentu (sekcja 4.2.1). Segmenty dzielą jedną komórkę siatki wejściowej oraz docelowo będą połączone, do uzyskania ostatecznej wartości komórki wyjściowej. Stąd też ograniczenie, nałożone na segmenty należące do jednej komórki siatki wejściowej, które mówi, że ich wartości muszą się sumować do wartości w

komórce siatki wejściowej (w przypadku gdy wartość ta jest podzielna). System wnioskowania rozmytego w wyniku zwraca zbiór rozmyty, który zawiera wszystkie możliwe wartości, jakie może przyjąć dany segment, wraz z prawdopodobieństwem ich wystąpienia. Po defuzyfikacji każdego segmentu (wybrania najbardziej prawdopodobnej wartości) nie ma żadnej gwarancji, że ograniczenie na sumę wartości segmentów będzie spełnione. Rozważmy następujący przykład. Mamy dwa zbiory rozmyte, jak pokazano na rysunku 5, oraz ograniczenie, że suma zdefuzyfikowanych wartości musi być równa danej wartości (np. C=10). Defuzyfikacja pierwszego zbioru rozmytego c1, przy użyciu na przykład metody Centroid lub MeanOfMax, daje wartość 4, podczas gdy defuzyfikacja drugiego zbioru rozmytego c2 daje wartość 7. Ograniczenie, że suma powinna być równa 10, nie jest spełnione, ponieważ obie wartości sumują się do 11. Najbardziej oczywistym rozwiązaniem dla spełnienia ograniczenia byłoby przeskalowanie wartości, ale na prostym przykładzie można wykazać, że nie jest to dobre podejście. W tym prostym przykładzie przeskalowanie oznacza, że wartości będą $10\times\frac{4}{4+7}=3.64$ dla pierwszego zbioru rozmytego i $10\times\frac{7}{4+7}=6.36$ dla drugiego. Biorąc pod uwagę zbiory rozmyte dla obu komórek, widzimy, że $\mu_{c_1}(3.64)=0.6$ i $\mu_{c_2}(6.36)=1$. W zbiorze możliwych wartości, dla obu komórek można jednak znaleźć wartości w c1 i w c2, które sumują się do 10, i dla których zachodzi $\mu_{c_1}(4)=\mu_{c_2}(6)=1.0$. Ogólnie oznacza to, że 4 i 6 stanowią rozwiązanie, całkowicie zgodnie z oboma zestawami rozmytymi, a jednocześnie spełniające wymagane ograniczenie. Dlatego rozwiązanie to jest lepsze, niż rozwiązanie otrzymane przez skalowanie wartości.

Opracowany algorytm jest zoptymalizowany dla wypukłych, odcinkami liniowych zbiorów rozmytych i wyznacza on wartości ostre wszystkich zbiorów rozmytych równocześnie. Algorytm rozpoczyna od określenia najwyższego poziomu wartości alfa, dla którego wszystkie zbiory mają niepuste przedziały. Począwszy od znanego defuzyfikatora (w tym przypadku MeanOfMax), algorytm przeszukuje te przedziały w poszukiwaniu rozwiązania. Ważnym kryterium jest to, że rozwiązanie powinno minimalizować *względną odległość* od środka przedziału. *Względna odległość* oznacza, że przy większych przedziałach, dozwolone odchylenia także są większe. Kryterium to zostało wprowadzone, aby umożliwić wybór pomiędzy wieloma równorzędnymi rozwiązaniami. W powyższym przykładzie, na tym etapie, algorytm wyznaczy wartości 4 i 6. Jeśli jednak nie znaleziono rozwiązania w przedziałach, algorytm kontynuuje szukanie rozwiązania. Dla lepszej czytelności przykładu przyjmijmy, że wartość ograniczająca C jest mniejsza, niż suma niższych granic przedziałów (ograniczenie z drugiej strony byłoby symetryczne). Jeżeli suma dolnych granic obu przedziałów jest wciąż zbyt duża, to następnym krokiem jest znalezienie brakującej różnicy. Odbywa się to poprzez obniżenie wartości alfa i przyjęcie jej nowego poziomu dla wszystkich zbiorów rozmytych, ograniczając przedziały do części po lewej stronie, od wcześniej rozważanych niższych ograniczeń. Daje to przedziały, które mają nowe dolne ograniczenie, podczas gdy górne ograniczenie jest równe poprzedniemu ograniczeniu dolnemu. W nowo powstałych przedziałach, szukamy wartości, które sumują się do brakującej różnicy, tym razem minimalizując względną odległość od ograniczenia górnego. Opisane kroki są powtarzane, dopóki nie zostanie znaleziony poziom alfa, dla którego rozwiązanie istnieje. Powstaje pytanie, o ile

powinno się obniżać poziom alfa w każdej iteracji. Dzięki zastosowaniu odcinkowo liniowych zbiory rozmyte, można do tego celu wykorzystać punkty zmiany nachylenia funkcji.

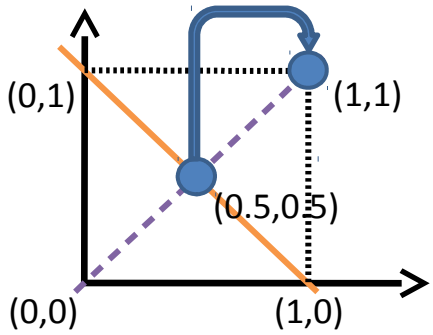
Defuzyfikacja z ograniczeniami musiała zostać zdefiniowana przy problemie zmiany siatek, jednakże algorytm może znaleźć zastosowanie wszędzie tam, gdzie poszczególne wartości ostre (wynikające z defuzyfikacji) są ze sobą w jakiś sposób połączone lub od siebie zależne. Metoda ta jest ogólna i działa niezależnie od innych czynności w sekwencji. Jest to bardzo przyszłościowe zagadnienie badawcze, szczególnie biorąc pod uwagę dużą popularność systemów rozmytych.

4.2.6 Rozwiązanie w przypadku braku danych treningowych

Do tej pory zakładano, że dane szkoleniowe są dostępne i są to dane, dla których znane jest idealne rozwiązanie. Dane szkoleniowe powinny zawierać nie tylko te same rodzaje modelowanych danych, ale dane te powinny być również reprezentowane na takiej samej siatce, jak rozważany problem. Ponieważ są to duże wymagania, takie dane mogą w wielu przypadkach być niedostępne, co ogranicza stosowalność zaprezentowanego podejścia. Dane treningowe i reguły mogą być przygotowane przez eksperta, ale wówczas taki system byłby raczej uproszczony, ze względu na wielki nakład pracy, jaki ekspert musiałby włożyć w przygotowanie danych. Ekspert narzuciłby też własne założenia o danych i rozpatrywanych zjawiskach. W związku z tym, dla zachowania ogólności podejścia, potrzebny był algorytm do generowania danych treningowych dla danego problemu. Opracowany algorytm pobiera geometrię z siatek danych wejściowych i zapełnia ją przy użyciu wygenerowanego podstawowego rozkładu pseudolosowego.

Zaprojektowanie metody generacji danych treningowych było niemałym wyzwaniem. Aby je zrealizować, konieczne było przeprowadzenie wiele eksperymentów, zanim udało się stwierdzić, które dane są odpowiednie. Eksperymenty te wykazały, że w przypadku zestawów danych, zbudowanych na tych samych geometriach, te same parametry zostały ocenione jako dobre. Ta właściwość doprowadziła do powstania następującego algorytmu. Rozważmy siatkę wejściową A, docelową siatkę B i dane pomocnicze, reprezentowane na siatce C. Stosując te same definicje siatki, rozważamy odpowiednio siatki A', B' i C'. Następnie wygenerowano wzór punktów, pokrywający obszar zainteresowania. Aby zmaksymalizować spójność metody względem różnych wywołań algorytmu, wzór punktów nie został wygenerowany w pełni losowo: w każdej komórce wyjściowej siatki B' ustawiono losową liczbę punktów (0, 1, 4, 5 lub 9). Punkty te zostały umieszczone według wstępnie zdefiniowanych wzorców w komórce. Nie użyto jednego wzorca, ponieważ mogłoby to wygenerować niewłaściwe rozmieszczenie punktów na siatce wejściowej lub siatce pomocniczej (na przykład zbyt wiele punktów na krawędziach komórek siatki). Punktom tym przypisano losowe wartości, których możliwy zakres został określony przez możliwy zakres wartości w komórkach w siatce wejściowej. Powodem doboru takiego zakresu jest zapewnienie, że zbiór uczący będzie miał wartości takiego samego rzędu, jak oryginalne dane. Nie jest to bardzo istotne wymaganie, ale zostało uznane za bezpieczny wybór w kontekście błędów zaokrąglania. Zbiór punktów został następnie spróbkowany, z wykorzystaniem siatek A ', B' i C ', czego rezultatem są siatki stanowiące dane treningowe. W razie potrzeby, można wygenerować wiele takich pseudolosowych zestawów punktów, co spowoduje wygenerowanie większego zestawu danych treningowych.

Zaletą takiego podejścia jest to, że nawet dla małych sieci, gdzie nie ma możliwości wydzielenia danych uczących, można taki zbiór wygenerować. W obecnej implementacji, tak



Rysunek 6: Przykład błędów wynikających z ograniczeń reprezentacji liczb w formacie zmiennoprzecinkowym na przykładzie obliczeń geometrycznych.

wygenerowane dane wykazują proporcjonalne lub odwrotnie proporcjonalne relacje z danymi wejściowymi, dla parametrów wykorzystujących zachodzenie lub ważone zachodzenie na siebie danych. Możliwe jest generowanie bardziej skomplikowanych siatek danych pomocniczych, ale wówczas należy podać więcej informacji na temat tego, w jaki sposób rozkład danych pomocniczych odnosi się do danych wejściowych.

4.2.7 Odporność obliczeń geometrii

Opracowane rozwiązanie, wykorzystujące metody sztucznej inteligencji do rozwiązania problemu nakładania map, dla danych opisanych na siatkach, zostało zaimplementowane jako prototyp oraz dowód poprawności (*proof-of-concept*). Podczas implementacji pojawił się problem niezawodności obliczeń geometrii ([21]), który wpłynął na prawidłowość obliczeń przestrzennych zależności między komórkami siatki (np. przejście ze zmiany siatki do przestrzennej dezagregacji) oraz między komórkami i cechami. Zastosowana przy implementacji biblioteka przestrzenna (JTS) ma metody dedykowane tym problemom, aczkolwiek były one w tym przypadku niewystarczające, gdyż nadal powodowały otrzymanie zbyt wielu fałszywie pozytywnych wyników testów przecięcia.

Problem odporności obliczeń geometrii przedstawiono na rysunku 6. Przyczyną problemu jest ograniczona reprezentacja liczb rzeczywistych w systemach komputerowych. Rysunek 6 pokazuje dwie linie – jedną łączącą (0,0) z (1,1) i drugą, łączącą (0,1) z (1,0). Przy założeniu, że dla współrzędnych możliwe są tylko wartości całkowite, punkt przecięcia (0.5, 0.5) nie może być poprawnie odwzorowany i będzie reprezentowany jako punkt (1,1), który nie znajduje się na żadnej z linii. To samo dzieje się przy obliczeniach komputerowych, z powodu ograniczonej reprezentacji wartości zmiennoprzecinkowych: niektóre współrzędne nie mogą być reprezentowane w dokładności oferowanej przez wewnętrzną reprezentację wartości rzeczywistych więc są zaokrąglane. To oczywiście stwarza problemy i skutkuje efektami ubocznymi, takimi jak punkty przecięcia dwóch linii, które nie znajdują się na żadnej z tych linii, geometrie, które przecinają się, pomimo, że nie powinny, lub na odwrót, a nawet niespójności między różnymi algorytmami (np. stwierdzenie istnienia przecięcia kontra obliczanie przecięcia). Podczas eksperymentów, problemy z odpornością obliczeń ujawniły się jako powtarzające się wzorce w rozwiązaniach. Biblioteka przestrzenna JTS zapewnia funkcjonalność radzenia sobie z problemem, poprzez obniżenie dokładności reprezentacji współrzędnych. Obliczenia są wykonywane wewnętrznie z pełną dokładnością, co pozwala na wykrycie problemu podczas wykonywania obliczeń. Chociaż takie

podejście zmniejsza prawdopodobieństwo problemów, to okazuje się niewystarczające, aby zapobiec negatywnym efektom ubocznych. W związku z tym, opracowano inną metodę rozwiązania problemu.

Głównym problemem, wynikającym z niedokładności obliczeń, były zaobserwowane fałszywie pozytywne wyniki testów przecięcia geometrii i niespójności między różnymi algorytmami. Pierwszy ze wspomnianych problemów został rozwiązany, poprzez opracowanie algorytmu, który dla dwóch argumentów (geometrii) określa rozmiar ich przecięcia i jeśli jest on mniejszy niż pewien zadany epsilon, zastępuje daną geometrię geometrią niższego wymiaru. Na przykład, jeżeli wielokąt jest wystarczająco mały, jest on zastąpiony przez linię, wystarczająco krótka linia zostaje zastąpiona przez punkt, itd.. Epsilon jest określany automatycznie przy użyciu względnego rozmiaru użytej geometrii. Usuwa to zależność od skali danych i sprawia, że zachowanie jest spójne, niezależnie od jednostki reprezentacji.

Drugi ze wspomnianych problemów został rozwiązany przez zmianę sposobu obliczania relacji przestrzennych. Relacje przestrzenne między dwiema geometriami są definiowane przy użyciu macierzy intersekcyjnej wymiaru 3×3 , która uwzględnia wszystkie możliwe kombinacje przecięć wewnętrznych, granicznych i zewnętrznych obu geometrii. Ze względu na wydajność, obliczenie tej macierzy zazwyczaj wykorzystuje inne algorytmy od tych używanych do obliczania różnych przecięć. Zostało to zmienione w niniejszej implementacji- nie tylko wykorzystano tu te same algorytmy do wykrywania przecięć i ich obliczania, ale także wyżej wspomniane podejście zostało uwzględnione w tych obliczeniach. W związku z tym macierz intersekcyjna stała się w pełni zgodna z wynikami innych obliczeń. Stało się to jednak kosztem wydajności.

Opracowane podejście tylko częściowo rozwiązuje problemy związane z odpornością, ponieważ nie rozwiązuje problemu fałszywych wyników negatywnych w testach przecięcia: dwie przecinające się geometrie, dla których test przecięcia wychodzi negatywnie ze względu na zaokrąglenia współrzędnych, nie jest korygowany przez zaprezentowaną metodę. Jednak w algorytmie zmiany siatki takie sytuacje nie miałyby dużego wpływu na wyniki, a przeprowadzone eksperymenty pokazały, że nie stanowi to problemu. Metodologia obniżania wymiaru przecięć ma uniwersalne zastosowanie w sytuacjach, w których należy unikać fałszywych pozytywnych wyników.

4.2.8 Eksperymenty i wnioski

W celu zweryfikowania wyników badań, przeprowadzono wiele eksperymentów na różnych zestawach danych. Dla wstępnych eksperymentów zostały wygenerowane uproszczone sztuczne przykłady, aby można było lepiej zrozumieć działanie algorytmów. Dzięki tym eksperymentom zauważono wiele problemów, opisanych w poprzednich rozdziałach. Wielokrotne testy pozwoliły na dopracowanie opisanych algorytmów i, w efekcie, stworzenie całej sekwencji czynności. Następnie przeprowadzono bardziej zaawansowane eksperymenty, z wykorzystaniem rzeczywistych zbiorów danych, dotyczących zanieczyszczenia powietrza w rejonie Warszawy. Pewne elementy tych zestawów danych zostały sztucznie wygenerowanych, aby mieć możliwość kontrolowania sytuacji i generowania przypadków szczególnych. Wszystko po to, aby przetestować możliwości i ograniczenia nowego podejścia w różnych warunkach. Ponadto, pozwoliło to na wyliczenie rozwiązania idealnego, którego użyto do weryfikacji jakości wyniku.

Zaprezentowana sekwencja czynności to zupełnie nowe podejście do problemu, nie tylko jako metoda, ale również na poziomie koncepcyjnym: opracowane podejście analizuje problem pod zupełnie innym kątem niż tradycyjne i może być wykorzystane w szeregu innych problemów. Podczas pracy nad sekwencją czynności, opracowano rozwiązania pewnych problemów, związanych z systemami wnioskowania, których zastosowanie wykracza poza kontekst przedstawionej pracy. Chodzi przede wszystkim o zmiany w algorytmach budowy i stosowania bazy reguł, pozwalające na nowatorską, względną interpretację wyrażeń lingwistycznych, a także opracowanie nowego algorytmu defuzyfikacji (wyostrzania danych), który pozwala wybrać najlepsze rozwiązanie jednocześnie dla szeregu zbiorów rozmytych, mających wspólne ograniczenie. Nowatorski algorytm porównywania danych na siatkach, w kontekście podobieństwa przestrzennego, prezentuje rozwiązanie, do tej pory bagatelizowanego, acz bardzo istotnego problemu. Metoda obliczania podobieństwa także ma szansę znaleźć liczne zastosowania, poza problematyką przedstawioną w niniejszej pracy.

Cytowana bibliografia

[1] Tainio, M., Sofiev, M., Hujo, M., Tuomisto, J., Loh, M., Jantunen, M., Karppinen, A., Kangas, L., Karvosenoja, N., Kupiainen, K., Porvari, P. and Kukkonen, J. Evaluation of the European population intake fractions for European and Finnish anthropogenic primary fine particulate matter emissions. Atmospheric Environment, 2009, Vol. 43 (19), pp. 3052-3059

[2] Woodcock, J., Tainio, M., Cheshire, J., O'Brien, O. and Goodman, A. *Health effects of the London bicycle sharing system: health impact modelling study.* British Medical Journal, 2014, Vol. 348

[3] Martinez-Urtaza J., Bowers J.C., Trinanes J., DePaola A. C*limate anomalies and the increasing risk of Vibrio parahaemolyticus and Vibrio vulnificus illnesses*, Food Research International, Volume 43, Issue 7, 2010, pp. 1780-1790.

[4] Rigaux, P., Scholl, M. and Voisard, A. *Spatial databases with applications to GIS.* Morgan Kaufman Publishers, 2002.

[5] Shekhar, S. and Chawla, S. *Spatial databases: a tour.* Pearson Education, 2003.

[6] Holnicki, P. *Advanced Air Pollution; eds. Nejadkoorki, F.* Chapter 14. Uncertainty in Integrated Modelling of Air Quality. INTECH, 2011, pp. 239 -– 260.

[7] Tomlin, C. *Special Issue Landscape Planning: Expanding the Tool Kit Map algebra: one perspective.* Landscape and Urban Planning, 1994, Vol. 30(1), pp. 3 - 12

[8] Gotway, C. A. and Young, L. J. *Combining incompatible spatial data.* Journal of the American Statistical Association, 2002, Vol. 97(458), pp. 632-648.

[9] Mugglin, A. S., Carlin, B. P.., Zhu, L. and Conlon, E. Bayesian areal interpolation, estimation, and smoothing: An inferential approach for geographic information systems. Environment and Planning A, 1999, Vol. 31(8), pp. 1337-1352.

[10] Mugglin, A. S., Carlin, B. P. and Gelfand Alan, E. *Fully Model-Based Approaches for Spatially Misaligned Data.* Journal of the American Statistical Association, 2000, Vol. 95(451), pp. 877-887.

[11] Volker, W. and Fritsch, D. *Matching spatial datasets: a statistical approach.* International Journal of Geographical Information Science, 1999, Vol. 13 (5), pp. 445-473.

[12] Duckham, M. and Worboys, M. *An algebraic approach to automated information fusion.* International Journal of Geographic Information Systems, 2005, Vol. 19, pp. 537-558.

[13] Wang, L.-X. and Mendel, J. M. *Generating fuzzy rules by learning from examples.* IEEE Transactions on systems, man and cybernetics, 1992, Vol. 22(6), pp. 1414-1427.

[14] Murata, T. and Ishibuchi, H. *Adjusting membership functions of fuzzy classification rules by genetic algorithms.* Proceedings of 1995 IEEE International Conference on Fuzzy Systems. 1995, Vol. 4, pp. 1819-1824 vol.4

[15] Nomura, H., Hayashi, I. and Wakami, N. *A self-tuning method of fuzzy reasoning by genetic algorithm.* Proc. of the 1992 International Fuzzy Systems and Intelligent Control Conference,1992, pp. 236-245.

[16] Aczel, *Complete Business Statistics,* McGraw-Hill Education (India) Pvt Limited, 2007.

[17] Verstraete, J. *Fuzzy quality assessment of gridded approximations.* Applied Soft Computing, 2017, Vol. 55, pp. 319 – 330.

[18] Verstraete J. *Artificial intelligent methods for handling spatial data - Fuzzy rulebase systems and gridded data problems*, 2018.

[19] Van Leekwijck, W. and Kerre, E. E. *Defuzzification: criteria and classification.* Fuzzy Sets and Systems, 1999, Vol. 108, pp. 159-178.

[20] Yager, R. R. and Filev, D. P. *Defuzzification with Constraints.* Fuzzy Logic and its Applications to Engineering, Information Sciences, and Intelligent Systems - Theory and Decision Library, 1996, Vol. 16, pp. 157-166.

[21] Hoffmann, C. M. *Geometric and Solid Modeling: An Introduction.* Morgan Kaufmann Publishers Inc., 1992.

5 Inne wyniki

5.1 Operacje oraz topologia obszarów rozmytych (Operations and topology on fuzzy regions)

Po ukończeniu doktoratu oraz po krótkiej przerwie w karierze naukowej dr Verstraete kontynuował swoje badania nad modelowaniem operacji na danych przestrzennych, obarczonych niepewnościami i danymi nieprecyzyjnymi. Nie są to zagadnienia ściśle związane z tematyką habilitacji, aczkolwiek pośrednie kroki metody używają struktur i przekształceń bardzo podobnych do tych, opracowanych podczas pracy nad doktoratem, niejako łącząc prowadzone badania. Badania były kontynuowane w tematyce rozmytego modelu opartego na cechach geometrycznych, w szczególności przy opracowaniu operacji i topologii, połączenia niepewnych i nieprecyzyjnych danych, jak również reprezentację modeli dostępną dla systemów komputerowych. Te badania zaowocowały trzema publikacjami w czasopismach:

1) Verstraete J. (100%) (2012). Deriving topological concepts for fuzzy regions: from properties to definitions. Control and Cybernetics, vol. 41, eds. Zbigniew, Nahorski; Jan Owsiński, Polish Academy of Sciences, no. 1, pp. 113-143.

   Jestem jedynym autorem tej pracy.

2) Verstraete J. (100%) (2009). Fuzzy Regions: interpretations of surface area and distance. Control and Cybernetics, vol. 38, eds. Zbigniew, Nahorski; Jan Owsiński, Polish Academy of Sciences, no. 2, pp. 509-526.

   Jestem jedynym autorem tej pracy.

3) Verstraete J. (70%), De Tré G. (10%), Hallez A. (10%), De Caluwe R. (10%) (2007). Using TIN-Based Structures for the Modelling of Fuzzy GIS Objects in a Database. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 15, 1-20.

   Wkład własny w powstanie tej publikacji szacuje na 70%. Obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowanie jej koncepcji i układu oraz sformułowanie celu badań 70%; b) projekt i implementacje algorytmów oraz wyprowadzenie wyników 70%; c) analizę i interpretacje wyników 70%; d) przygotowanie publikacji do druku 70%.

ośmioma publikacjami konferencyjnymi indeksowanymi w WoS:

1) Verstraete J. (100%) (2012). Surface area of level-2 fuzzy regions. In: Rutkowski, Leszek; Korytkowski, Marcin; Scherer, Rafał; Tadeusiewicz, Ryszard; Zadeh, Lotfi; Żurada, Jacek (Eds.) Artificial Intelligence and Soft Computing - Artificial intelligence and soft computing: 11th International Conference (ICAISC 2012, Zakopane, Poland, April 29-May 03, 2012), series Lecture Notes in Computer Science, 7267, Springer-Verlag, pp. 342-349.

   Jestem jedynym autorem tej pracy.

2) Verstraete J. (100%) (2011) Higher reasoning with level-2 fuzzy regions. In: Christiansen, Henning; De Tré, Guy; Yazici, Adnan; Zadrozny, Slawomir; Andreasen, Troels and Larsen, Hendrik Legind (Eds.). Flexible Query Answering Systems - 9th International Conference,

(FQAS 2011, Ghent, Belgium, October 26-28, 2011), series Lecture Notes in Computer Science - Lecture Notes in Artificial Intelligence, 7022, Springer-Verlag, pages 49-59.

Jestem jedynym autorem tej pracy.

3) Verstraete J. (100%) (2011), Using level-2 fuzzy sets to combine uncertainty and imprecision in fuzzy regions. In: Mugellini, E.; Szczepaniak, P.S.; Chiara Pettenati, M.; Sokhn, M. (Eds.). Advances in Intelligent Web Mastering - 3, Springer-Verlag, Berlin Heidelberg, 2011, Series: Advances in Intelligent and Soft Computing, Vol. 86, pp. 163-172.

   Jestem jedynym autorem tej pracy.

4) Verstraete J. (100%) (2010). A quantitative approach to topology for fuzzy regions. Rutkowski, L.; Scherer, R.; Tadeusiewicz, R.; Zadeh, L.A.; Zurada, J.M. (Eds.). Artificial Intelligence and Soft Computing, Part I, Lecture Notes in Artificial Intelligence vol. 6113. ("10th International Conference on Artificial Intelligence and Soft Computing", ICAISC 2010, Zakopane, Poland, June 13-17), 248-255.

   Jestem jedynym autorem tej pracy.

5) Verstraete J. (100%) (2010). Fuzzy regions: adding subregions and the impact on surface and distance calculation. In Hüllermeier E; Kruse, R; Hoffmann, F (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems, Communications in Computer and Information Science, Vol. 80, Part 1. (13th International Conference, IPMU 2010, Dortmund, Germany, June 28–July 2, 2010), 561-570.

   Jestem jedynym autorem tej pracy.

6) Verstraete J. (80%), Hallez A. (10%), De Tré G. (10%) (2007). Fuzzy Regions: Theory and Applications. In A. Morris, S. Kokhan (Eds.), Geographic Uncertainty in Environmental Security (pp. 1-17). Dordrecht, The Netherlands: Springer.

   Wkład własny w powstanie tej publikacji szacuje na 80%. Obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowanie jej koncepcji i układu oraz sformułowanie celu badań 80%; b) projekt i implementacje algorytmów oraz wyprowadzenie wyników 80%; c) analizę i interpretacje wyników 80%; d) przygotowanie publikacji do druku 80%.

7) Charlier N. (50%), De Tré G. (30%), Gautama S. (20%), Verstraete J. (20%) (2007). Measuring Quality of Geographical Information Using Imperfect Reference Data. 11th IASTED International Conference Artificial Intelligence and Soft Computing, 2007, 80-85.

   Wkład własny w powstanie tej publikacji szacuje na 20%. Obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowanie jej koncepcji i układu oraz sformułowanie celu badań 10%; b) projekt i implementacje algorytmów oraz wyprowadzenie wyników 30%; c) analizę i interpretacje wyników 20%; d) przygotowanie publikacji do druku 20%.

8) Verstraete J. (90%), Hallez A. (10%) (2007). Numerical Properties of Fuzzy Regions: Surface Area. In P. Melin, O. Castillo, L.T. Aguilar, J. Kacprzyck, W. Pedrycz (eds.). Foundations of Fuzzy Logic and Soft Computing, Lecture Notes in Artificial Intelligence 4529, 155-161.

   Wkład własny w powstanie tej publikacji szacuje na 90%. Obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowanie jej koncepcji i układu oraz sformułowanie celu badań 90%; b) projekt i

implementacje algorytmów oraz wyprowadzenie wyników 90%; c) analizę i interpretacje wyników 90%; d) przygotowanie publikacji do druku 90%.

jednym rozdziałem w książce:

1) Verstraete J. (70%), Hallez A. (10%), De Tré G. (10%), Matthé T. (10%) (2008). Topological relations on fuzzy regions: an extended application of intersection matrices. In B. Bouchon-Meunier, R.R. Yager, C. Marsala, and M. Rifqi eds., Uncertainty and Intelligent Information Systems (pp. 487-500). World Scientific.

   Wkład własny w powstanie tej publikacji szacuje na 70%. Obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowanie jej koncepcji i układu oraz sformułowanie celu badań 70%; b) projekt i implementacje algorytmów oraz wyprowadzenie wyników 70%; c) analizę i interpretacje wyników 70%; d) przygotowanie publikacji do druku 70%.

oraz dwoma innymi artykułami konferencyjnymi:

1) Verstraete J. (100%) (2012). Implementable representations of level-2 fuzzy regions for use in databases and GIS. In Greco, Salvatore; Bouchon-Meunier, Bernadette; Coletti, Giulianella; Fedrizzi, Mario; Matarazzo, Benedetto; Yager, Ronald R. (Eds.) Advances on Computational Intelligence - 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2012), series: Communications in computer and information science; Vol. 297, Springer-Verlag, pp. 361-370.

   Jestem jedynym autorem tej pracy.

2) Verstraete J. (100%) (2011). Union and intersection of Level-2 fuzzy regions. Yager, Ronald R.; Reformat, Marek Z.; Shahbazova, Shahnaz N. and Ovchinnikov, Sergei (eds.). World Conference on Soft Computing (WConSC 2011), May 23-26 2011, San Francisco, CA, USA. paper nr. 123.

   Jestem jedynym autorem tej pracy.

5.2 Wieloagentowe systemy zarządzania energią elektryczną

Poza skupieniem się na zagadnieniach przestrzennych, dr Verstraete był włączony w badania nad obliczeniem kosztów działania systemu, w celu oceny wydajności rozproszonego wieloagentowego systemu zarządzania energią elektryczną w mikrosieci. Rezultatem tych badań jest publikacja konferencyjna na FedCSIS:

1) Radziszewska W. (60%), Verstraete J. (30%), Wasilewski J. (10%) (2015) Evaluation of distributed multi-agent Energy Management System – cost calculation. Proceedings of Federated Conference on Computer Science and Information Systems.

   Wkład własny w powstanie tej publikacji szacuje na 30%. Obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowanie jej koncepcji i układu oraz sformułowanie celu badań 20%; b) projekt i implementacje algorytmów oraz wyprowadzenie wyników 30%; c) analizę i interpretacje wyników 40%; d) przygotowanie publikacji do druku 30%.

5.3 Historia publikacji – podsumowanie

Dr Jörg Verstraete jest autorem lub współautorem 55 publikacji oraz jednego artykułu w czasopiśmie, przyjętym do publikacji (nie wliczonym w poniższe statystyki). Ze względu na różne literowanie imienia (Jörg, Joerg, Jorg, J.) oraz na fakt, że są inni autorzy o takim samym nazwisku i inicjale, wyszukiwarki on-line nie zawsze wyszukują wszystkie publikacje dra Verstraete. Pełna lista publikacji jest zawarta w Załączniku 4 do dokumentacji habilitacyjnej. Unikalny identyfikator autora w ResearcherID oraz Web Of Science to: **F-9175-2013**, natomiast w Orcid oraz Scopus: **0000-0002-7772-984X**.

Kategoria	< PhD	≥ PhD	Razem
Monografie naukowe	0	2	2
Artykuły w czasopismach	1	7	8
Publikacje konferencyjne publikowane w WoS	4	14	18
Rozdziały w książkach indeksowanych w WoS	2	0	2
Publikacje konferencyjne w WoS	11	6	17
Rozdziały w książkach nie w WoS	3	3	6
Prace nierecenzowane (plakaty, abstrakty)	1	1	2
Razem	22	33	55

Impact factor oraz punkty MNiSW:

Kategoria	< PhD	≥ PhD	Razem
Impact factor	0	18.204	18.204
Punkty MNiSW	148	395 (385)	543 (533)
Punkty z listy A	0	230 (195)	230 (195)
Punkty z listy B	14	13 (28)	27 (42)
Punkty z WoS[a]	110	110 (120)	220 (230)
Punkty z monografii	24	25	49
Punkty za rozdziały w książkach	0	17	17

[a]) LNCS, LNAI indeksowane w WoS są liczone za 15 punktów w lub po 2013 roku oraz liczone jako 10 punktów przed 2013.
Uwaga: Wartości są obliczone używając klasyfikacji z roku publikacji: Control and Cybernetics był na liście A z 15 punktami w 2009 roku i na liście A z 20 punktami w 2012 roku, w latach 2013-2016 był na liście ministerialnej B z 14 punktami. Publikacja z ICAISC 2010 została sklasyfikowana jako rozdział w 2010 roku i otrzymała 13 punktów, używając reguł przyjętych w WoS, publikacje do 2013 roku mają 10 punktów. Wartości w nawiasach odnoszą się do wartości z listy MNiSW 2013-2016 oraz obecnych reguł przyjętych w Web Of Science.

Hirsh h-index:

Źródło	Liczba publikacji	Bez autocytowań	Całkowita
Google Scholar	54	5[a]	9
Web of Science	26 (28[b])	3[c]	4
Scopus	30	4	6

[a]) ręczne zliczenie cytowań z Google Scholar, jako że system nie podaje wartości indeksu Hirsha bez autocytowań.
[b]) 26 w Core Collection, 28 w All Databases
[c]) ręczne zliczenie cytowań z Web Of Science, jako że system nie podaje wartości indeksu Hirsha bez autocytowań; na końcu sekcji znajduje się wyjaśnienie, dotyczące możliwych błędów w wynikach z Web Of Science.

Liczba cytowań:

Źródło	Liczba publikacji	Bez autocytowań	Całkowita
Google Scholar	54		224
ResearcherID	53 (26[a])		64
Web of Science	26 (28[b])	39	65
Scopus	30	56	124

[a]) 53 w sumie, 26 z danymi cytowań.
[b]) 26 w Core Collection, 28 w All Databases.
Uwaga: 1 monografii oraz 1 abstraktu brakuje w Google Scholar oraz ResearcherID.
Uwaga 2: na końcu sekcji jest wyjaśnienie dotyczące możliwych błędów w wynikach z Web Of Science.

Rozkład cytowań:

	Przed doktoratem (≤2007)		Po doktoracie (≥2007)	
Źródło	Publikacje	Cytowania	Publikacje	Cytowania
Google Scholar	27	150	31	99
ResearcherID		7		59
Web of Science	12[a]	38 (38[b])	20[a]	42 (22[b])
Scopus	12	80 (47[b])	22	60 (20[b])

[a]) biorąc pod uwagę wszystkie bazy danych
[b]) z wyłączeniem autocytowań
Uwaga 1: 1 monografii oraz 1 abstraktu brakuje w Google Scholar oraz ResearcherID.
Uwaga 2: Publikacje z 2007 są liczone zarówno w sekcji “Przed doktoratem” jak i “Po doktoracie”.
Uwaga 3: Liczba bez autocytowań zwrócona przez WoS wydaje się mało wiarygodna biorąc pod uwagę liczbę całkowitą.

Uwaga odnosząca się do rezultatów zwracanych przez Web Of Science:

Liczba nie autocytowań, zwracanych przez Web Of Science pokazuje nieregularność, która wydaje się mało realistyczna i może wpływać na indeks Hirsha. Problem jest widoczny, gdy ograniczymy zakres czasowy w zapytaniu, jak zaprezentowano w tabeli poniżej, która pokazuje liczbę cytowań według WoS od początku do określonego roku.

Zakres czasowy publikacji	Liczba cytowań	Z wyłączeniem auto-cytowań
do roku (i włącznie z) 2018	65	39
do roku (i włącznie z) 2017	65	39
do roku (i włącznie z) 2016	64	43
do roku (i włącznie z) 2015	64	43
do roku (i włącznie z) 2014	63	45
do roku (i włącznie z) 2013	54	37

Wzrost w liczbie cytowań jest logiczny biorąc pod uwagę upływ czasu, aczkolwiek spadek liczby cytowań bez auto-cytowań, zauważony od 2014 roku już nie jest. Problem był zgłoszony do obsługi klienta Web Of Science (Clarivate Analytics) 8 lutego 2018 roku. Sprawie nadano numer #TS-04495548 i potwierdzono, że jest spowodowany błędem systemu. Niestety problem nie został jeszcze naprawiony.

5.4 Inne osiągnięcia i działalność dydaktyczna

5.4.1 Badania powiązane

Dr Jörg Verstraete jest autorem szeregu recenzji dla czasopism takich jak IEEE Transactions on Fuzzy Systems, Fuzzy Sets and Systems, Advances in Fuzzy Systems, IEEE Emerging Topics in Computational Intelligence, IEEE Computer and Control and Cybernetics. Dodatkowo był także recenzentem wielu artykułów konferencyjnych dla konferencji takich jak IFSA, IPMU, itd., a także rozdziałów kilku książek naukowych.

Dr Jörg Verstraete uczetniczył w ponad 22 konferencjach i prezentował swoje osiągnięcia na 21 z nich. Prowadził także seminaria w IBS PAN, na Politechnice Warszawskiej oraz w CiTIUS.

Dodatkowo prowadził sesje na różnych konferencjach oraz organizował sesje specjalne na IEEE IS'14 oraz na FedCIS 2016. Był członkiem komitetu organizacyjnego *4th International workshop on Uncertainty in atmospheric emissions* oraz *Summer School in Fuzzy Logic* w CiTIUS.

Przed rozpoczęciem pracy w Instytucie Badań Systemowych PAN odbył miesięczny staż na zaproszenie prof. Janusza Kacprzyka. Dodatkowo odbył krótkie staże: przygotowanie projektu w IIASA w Austrii (4 tygodnie) oraz w CiTIUS w Hiszpanii (1 tydzień). Od lutego 2017 roku jest na dwuletnim kontrakcie typu post-doc w CiTIUS i zajmuje się aplikacyjną stroną swoich badań oraz pracą nad przestrzennymi indeksami.

5.4.2 Działalność dydaktyczna

Po ukończeniu doktoratu dr Jörg Verstraete był członkiem komisji doktorskiej pod tytułem *Generic methods for object comparison* (tłumaczenie z oryginalnego flamandzkiego tytułu: Generieke methoden voor objectvergelijking), bronioną przez Axel Hallez na uniwersytecie w Ghent w 2010 roku. Dr Verstraete uczył także przedmiotu Bazy Danych na Politechnice Warszawskiej na wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej w latach 2011-2016 (wykład oraz laboratoria).

Podczas wcześniejszej kariery na uniwersytecie w Ghent dr Verstraete, w latach 2000-2007, prowadził szereg przedmiotów takich jak np. Bazy danych, Struktury danych oraz algorytmy, Kompilatory itp. Dodatkowo był też promotorem oraz członkiem komisji wielu prac magisterskich.