dr inż. Marcin Luckner
Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska
Warszawa, 16 października 2022

Autoreferat

Posiadane dyplomy, stopnie naukowe

22.10.2010 **Stopień naukowy doktora nauk technicznych**, *Instytutu Badań Systemowych Polskiej Akademii Nauk*, informatyka techniczna
Rozprawa doktorska *Problem eliminacji nieprzystających obiektów w zadaniu rozpoznawania wzorca*. Promotor: prof. dr hab. inż. Władysław Homenda.

1998–2003 **Stopień zawodowy magistra inżyniera**, *Politechnika Warszawska, Fizyka Techniczna i Matematyka Stosowana*, informatyka w zakresie informatyki stosowanej
Praca magisterska *Automatyczna identyfikacja wybranych symboli notacji muzycznej*. Promotor: prof. dr hab. inż. Władysław Homenda. Dyplom magistra inżyniera z oceną celującą.

Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych

14.11.2013 – **dyrektor**, *Politechnika Warszawska, Wydział Matematyki i Nauk Informacyjnych*, Ośrodek Badań dla Biznesu

01.03.2011 – **adiunkt**, *Politechnika Warszawska, Wydział Matematyki i Nauk Informacyjnych*, Zakład Zastosowań Informatyki i Metod Numerycznych

2009 – 2011 **asystent**, *Politechnika Warszawska, Wydział Matematyki i Nauk Informacyjnych*, Zakład Zastosowań Informatyki i Metod Numerycznych

2004 – 2009 **asystent**, *Politechnika Warszawska, Wydział Geodezji i Kartografii*, Instytut Geodezji Wyższej i Astronomii Geodezyjnej

Omówienie osiągnięć, o których mowa w art. 219 ust. 1 pkt 2 Ustawy

Na osiągnięcie składa się cykl powiązanych tematycznie artykułów naukowych, pt.:

Metody analizy przestrzennej wykorzystujące dane z sieci bezprzewodowych (GSM i Wi-Fi)

zawarty w pracach [A1–A11].

Cykl publikacji

[A1] M. Luckner, I. Krzemińska, P. Wawrzyniak i J. Legierski, "Estimating Population Density Without Contravening Citizen's Privacy: Warsaw Use Case," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, t. 52, nr. 7, s. 4494–4506, 2022, **punkty MEiN: 200, IF: 11.471**.

[A2] M. Luckner i R. Górak, "Automatic detection of changes in signal strength characteristics in a Wi-Fi network for an indoor localisation system," *Sensors (Switzerland)*, t. 20, nr. 7, s. 1–13, 2020, **punkty MEiN: 100, IF: 3.847**, ISSN: 14248220.

[A3] R. Górak i M. Luckner, "Automatic detection of missing access points in indoor positioning system," *Sensors (Switzerland)*, t. 18, nr. 11, paź. 2018, **punkty MEiN: 100, IF: 3.847**, ISSN: 14248220.

[A4] M. Luckner, A. Rosłan, I. Krzemińska, J. Legierski i R. Kunicki, "Clustering of Mobile Subscriber's Location Statistics for Travel Demand Zones Diversity," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 40**, t. 10244 LNCS, 2017, s. 315–326, ISBN: 978-3-319-59104-9.

[A5] M. Luckner, B. Topolski i M. Mazurek, "Application of XGboost algorithm in fingerprinting localisation task," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 40**, t. 10244 LNCS, 2017, s. 661–671, ISBN: 9783319591049.

[A6] R. Górak i M. Luckner, "Long term analysis of the localization model based on Wi-Fi network," w *Studies in Computational Intelligence*, **punkty MEiN: 20**, t. 642, 2016, s. 87–96, ISBN: 9783319312767.

[A7] R. Górak i M. Luckner, "Modified random forest algorithm for Wi–Fi indoor localization system," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 9876 LNCS, 2016, s. 147–157.

[A8] R. Górak, M. Luckner, M. Okulewicz, J. Porter-Sobieraj i P. Wawrzyniak, "Indoor Localisation Based on GSM Signals: Multistorey Building Study," *Mobile Information Systems*, t. 2016, 2016, **punkty MEiN: 40, IF: 1.863**, ISSN: 1875905X.

[A9] M. Luckner i R. Górak, "Comparison of floor detection approaches for suburban area," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 9622, 2016, s. 782–791, ISBN: 9783662493892.

[A10] M. Luckner i R. Górak, "Hybrid algorithm for floor detection using GSM signals in indoor localisation task," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 9648, 2016, s. 730–741, ISBN: 9783319320335.

[A11] R. Górak i M. Luckner, "Malfunction immune wi–fi localisation method," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 9329, 2015, s. 328–337.

Suma punktów za cykl, wyliczonych zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 9 lutego 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych, wynosi **630**.

Cykl obejmuje cztery publikacje w czasopismach [A1–A3, A8], których Sumaryczny Impact Factor został wyliczony na podstawie najnowszego wykazu Journal Citation Reports z 2021 roku i wynosi **21.028**.

Prace [A4, A5] zostały zaprezentowane na konferencjach z serii *International Conference on Computer Information Systems and Industrial Management* o randze C w rankingu Australian Computing Research and Education Association of Australasia (CORE).

Prace [A6, A9] zostały zaprezentowane na konferencjach z serii *Asian Conference on Intelligent Information and Database Systems* o randze B4 w rankingu Qualis.

Praca [A10] została zaprezentowana na konferencji z serii *International Conference on Hybrid Artificial Intelligence Systems* randze C w rankingu CORE i B4 w rankingu Qualis.

Prace [A7, A11] zostały zaprezentowane na konferencjach z serii *International Conference on Computational Collective Intelligence: Semantic Web, Social Networks and Multiagent Systems* o randze C w rankingu CORE.

Wstęp

Cykl publikacji, będący osiągnięciem, dotyczy wykorzystania danych z sieci bezprzewodowych (GSM i Wi-Fi) do rozwiązywania problemów analizy przestrzennej. Na cykl składają się głównie prace realizowane w ramach projektów badawczych: projektu NCBiR *LOKKOM: kompleksowe metody wyznaczania lokalizacji terminala sieci telefonii komórkowej przemieszczającego się w terenie otwartym i budynkach*, w którym pełniłem rolę kierownika grupy zadań, oraz europejskiego projektu badawczego *VaVeL: Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors*, którego byłem kierownikiem z ramienia Politechniki Warszawskiej.

Wszystkie prace zawarte w cyklu są pracami współautorskimi, stworzonymi przez osoby pracujące w wyżej wymienionych projektach. W szczególności, prace zrealizowane w ramach projektu LOKKOM, zostały zrealizowane we ścisłej współpracy z doktorem nauk matematycznych Rafałem Górakiem, który pełnił rolę kierownika powiązanej z moją grupy zadań. Podczas omawiania cyklu publikacji będę skupiać się na opisaniu mojego wkładu w poszczególne pozycje.

Przed omówieniem poszczególnych pozycji zarysuję tematykę prowadzonych badań. Analiza przestrzenna, oparta na systemie lokalizacyjnym, jest używana na codzień, gdy chcemy sprawdzić swoją położenie, upewnić się, że podążamy wyznaczoną trasą, czy uzyskać informacje o zatłoczeniu na drodze. O ile usługi oparte na globalnym systemie pozycjonowania GPS (ang. *Global Positioning System*) zapewniają wystarczającą dokładność w środowisku zewnętrznym, o tyle kwestia określenia dokładnej pozycji w pomieszczeniach jest znacznie bardziej skomplikowana. Po pierwsze, wszystkie rozwiązania GPS zawodzą wewnątrz budynków. Po drugie, w pomieszczeniach wymagany jest wyższy poziom dokładności niż na zewnątrz. Wynika to z faktu, że nawet stosunkowo niewielki błąd lokalizacji w środowisku wewnętrznym może oznaczać, że lokalizowane urządzenie znajduje się w innym pomieszczeniu lub na innym piętrze.

Ponieważ lokalizacja oparta na badaniu propagacji sygnałów jest zadaniem skomplikowanym analitycznie , w praktyce do rozwiązania problemów analizy przestrzennej opartych na lokalizacji używa się narzędzi uczenia maszynowego i innych algorytmów analizy danych. Te jednak potrzebują odpowiednich danych do przeprowadzenia procesu uczenia.

W omawianych pracach, dane treningowe algorytmów uczenia maszynowego, są wektorami zawierającymi obserwowaną siłą sygnału RSS (ang. *Received Signal Strength*) pochodzącą z punktów dostępowych APs (ang. *Access Points*) sieci Wi-Fi lub ze stacji bazowych BTS (ang. *Base Transceiver Station*) sieci komórkowej. Analizowane są dwa podejścia do pozyskania danych. Dokładna i kosztowna metoda fingerprintingu, która dostarcza pełnej wiedzy o punkcie pomiarowym oraz metoda crowdsourcingu, stosunkowo tania i dostarczająca dużo, ale ograniczonych danych.

Podejście fingerprintingowe do lokalizacji polega na zbudowaniu – na podstawie zbioru danych treningowych – modelu estymującego, dla wektora RSS, położenie punktu, w którym dokonano pomiarów.

Definicja 1 (Fingerprinting)**.**

(i) $\mathcal{AP}$ *to zbiór wszystkich AP używanych w modelu lokalizacyjnym, wyliczony przez kolejne liczby naturalne* $(1, 2, 3, \dots)$.

(ii) $\mathcal{F} = \mathbf{R}^2 \times \mathbb{Z} \times \mathbf{R} \times \mathbf{R}^n$ *to przestrzeń pomiarowa gdzie* $n = \sharp\mathcal{AP}$.
Dla $F \in \mathcal{F}$
  (a) współrzędne $F_1[m], F_2[m]$ *oznaczające horyzontalne położenie punktu pomiarowego;*
  (b) współrzędna $F_3 \in \mathbb{Z}$ *oznacza piętro, na którym znajduje się punkt pomiarowy;*
  (c) $F_4[s]$ *jest czasem dokonania pomiaru;*
  (d) $F_k[dBm]$*, gdzie* $4 < k \leqslant n + 4$*, jest RSS z* k*-ego źródła z* $\mathcal{AP}$*. Jeśli nie ma sygnału z* k*-ego AP wtedy* $F_{k+4} = \varnothing$*, gdzie* $\varnothing$ *jest specyficzną unikalną wartością spoza zakresu pomiarowego.*

(iii) Zbiór wektorów sygnałów $S \subset \mathcal{F}$ *definiuje serię pomiarową. Zazwyczaj* S *jest zbierany w ciągu jednego lub kilku kolejnych dni na tym samym zestawie punktów pomiarowych w danym obszarze.*

(iv) $\mathcal{L} = (\mathcal{L}_x, \mathcal{L}_y, \mathcal{L}_f) : \mathcal{F} \mapsto \mathbf{R}^2 \times \mathbb{Z}$ *jest rzutem na pierwsze trzy współrzędne zbioru* $\mathcal{F}$ *a* $\pi : \mathcal{F} \mapsto \mathbf{R}^n$ *to rzut na ostatnie* n *współrzędnych. Innymi słowy* $\mathcal{L}(F)$ *zapewnia nam lokalizację punktu pomiaru wektora* F *gdy* $\pi(F)$ *jest wektorem RSS związanym z* F.

(v) Dla $v = (v_1, ..., v_n) \in \mathbf{R}^n$ *oznaczamy* $supp(v) = \{k : \ v_k \neq \varnothing\}$ *który jest zbiorem widocznych punktów dostępowych (APs) dla* v.

W niektórych przypadkach lokalizacji zmienna F_3 jest zmienną rzeczywistą i opisuje wysokość punktu pomiarowego. Dla łatwiejszego zrozumienia opisywanych problemów, w niektórych przypadkach stosuję do opisu punktu pomiarowego $\mathcal{L}(F) = p = (F_1, F_2, F_3)$, notację $p = (x, y, f)$ i analogicznie dla estymacji $\hat{p} = (\hat{x}, \hat{y}, \hat{f})$.

Podejście crowdsourcingowe do lokalizacji polega na zbudowaniu – na podstawie zbioru danych treningowych – modelu wyznaczającego, dla wektora RSS, przybliżonego obszaru, w którym dokonano

pomiarów.

Definicja 2 (Crowdsourcing)**.**
(i) $\mathcal{AP}$ *to zbiór wszystkich AP używanych w modelu lokalizacyjnym, wyliczony przez kolejne liczby naturalne* $(1, 2, 3, \dots)$.
(ii) $\mathcal{F} = \mathcal{Z} \times \mathbf{R} \times \mathbf{R}^n$ *to przestrzeń pomiarowa gdzie* $n = \sharp\mathcal{AP}$.
Dla $F \in \mathcal{F}$
(a) *współrzędna* F_1 *jest obszarem* $F_1 \subset \mathbf{R}^2 \times \mathbb{Z}$ *oznaczającym przybliżone położenie punktu pomiarowego;*
(b) $F_2[s]$ *jest czasem dokonania pomiaru;*
(c) $F_k[dBm]$*, gdzie* $2 < k \leqslant n + 2$*, jest RSS z* k*-ego źródła z* $\mathcal{AP}$*. Jeśli nie ma sygnału z* k*-ego AP wtedy* $F_{k+2} = \varnothing$*, gdzie* $\varnothing$ *jest specyficzną unikalną wartością spoza zakresu pomiarowego.*
(iii) *Zbiór wektorów sygnałów* $S \subset \mathcal{F}$ *definiuje serię pomiarową.*
(iv) $\mathcal{L} : \mathcal{F} \mapsto \{0, 1\}$ *określa czy sygnały pochodzą z obszaru* $\mathcal{Z}$ *a* $\pi : \mathcal{F} \mapsto \mathbf{R}^n$ *to rzut na ostatnie* n *'współrzędnych. Innymi słowy* $\mathcal{L}(F)$ *zapewnia nam przybliżoną lokalizację punktu pomiaru wektora* f *gdy* $\pi(F)$ *jest wektorem RSS związanym z* F.
(v) *Dla* $v = (v_1, ..., v_n) \in \mathbf{R}^n$ *oznaczamy* $supp(v) = \{k : \ v_k \neq \varnothing\}$ *który jest zbiorem widocznych punktów dostępowych (APs) dla* v.

Ponieważ dane z crowdsourcingu da się zrzutować na format przyjęty dla danych z fingerprintingu (np. poprzez użycie centroidu obszaru $\mathcal{Z}$ do wyznaczenia pierwszych dwóch współrzędnych i przyjęcie $F_3 = 0$), dla zachowania przejrzystości, dalsze definicje będą przyjmowały, że dane pochodzą z fingerprintingu.

Aby umożliwić symulację przemieszczania się na obserwowanym obszarze, określmy graf $\mathcal{G}_S$, którego wierzchołkami są punkty pomiarowe serii S. Oznaczmy ten zbiór przez V_S. Zakładając rozmieszczenie punktów pomiarowych w kwadratowej siatce, każdy wierzchołek może być połączony krawędziami z ośmioma najbliższymi punktami pomiarowymi.

Zdefiniujmy następujące typy krawędzi:
przestrzeń krawędź nie przecina żadnej przeszkody i istnieje swobodne przejście wzdłuż krawędzi.
drzwi krawędź przecina drzwi.
ściana wszystkie inne przypadki. To znaczy: krawędź przecina ścianę lub inną nieruchawą przeszkodę.

Będziemy rozważać graf, który składa się tylko z krawędzi typu **przestrzeń** i **drzwi** tylko. Oznaczamy ten zbiór krawędzi przez E_S. Nie uwzględniamy krawędzi typu **ściana**, ponieważ testowania proponowanego rozwiązania lokalizacyjnego chcemy zasymulować ruchy ludzi wewnątrz budynku.

Definicja 3 (Graf pomiarowy)**.** *Graf pomiarowy* $\mathcal{G}_S = (V_S, E_S)$ *dla serii pomiarowej* S*, jest grafem o wierzchołkach umieszczonych w rozłożonych na kwadratowej siatce punktach pomiarowych serii* S *i krawędziach* E_S *łączących każdy punkt z ośmioma najbliższymi punktami siatki, o ile dokonano w nich pomiaru. Krawędzie grafu są etykietowane jako* ***przestrzeń****,* ***drzwi*** *i* ***ściana*** *w zależności od topografii analizowanego obszaru.*

Definicja 4 (Graf przejścia)**.** *Graf przejścia* $\mathcal{G}'_S = (V'_S, E'_S)$ *dla serii pomiarowej* S*, jest największym podgrafem grafu pomiarowego* $\mathcal{G}'_S \subset \mathcal{G}_S$*, nie zawierającym krawędzi typu* ***ściana****.*

Definicja 5 (Trasa)**.** *Trasa jest ścieżką w grafie przejścia* $\mathcal{G}'_S$.

Sformułujmy problemy analizy przestrzennej, które są rozwiązywane w ramach osiągnięcia.

Problem 1 (Lokalizacja)**.** *Na podstawie serii pomiarowej* $\mathcal{S}_L$ *(seria ucząca) skonstruować model lokalizacyjny. Model lokalizacji jest funkcją* $\widehat{\mathcal{L}} : \mathbf{R}^n \mapsto \mathbf{R}^2 \times \mathbb{Z}$ *taką, że biorąc pod uwagę wektor RSS* $v \in \mathbf{R}^n$*,* $\widehat{\mathcal{L}}(v)$ *przewiduje lokalizację, w której pomiar* v *został dokonany.*

Problem 2 (Śledzenie trasy)**.** *Używając modelu lokalizacyjnego* $\widehat{\mathcal{L}}$*, zbadać odstępstwo od zadanej ścieżki* $\mathcal{P}_R = (v_1, v_2, \dots, v_n)$*, przebytej ścieżki* $\mathcal{P}_T = \left(\widehat{\mathcal{L}}(F^1), \widehat{\mathcal{L}}(F^2), \dots, \widehat{\mathcal{L}}(F^m)\right)$ *w przestrzeni* $\mathbf{R}^2 \times \mathbb{Z}$.

Problem 3 (Estymacja zagęszczenia)**.** *Używając modelu lokalizacyjnego $\widehat{\mathcal{L}}$, na podstawie pomiarów $\mathcal{F}$, takich, że $F_4 \in [t_b, t_e]$ i $\widehat{\mathcal{L}}(\pi(F)) \in \mathcal{Z}$, gdzie $\mathcal{Z}$ jest podzbiorem obserwowanego obszaru, określić liczbę terminali przebywających w okresie $[t_b, t_e]$ na obszarze $\mathcal{Z}$.*

Jak widać z określenia problemów analizy przestrzennej, jakość uzyskanych rozwiązań jest ściśle powiązana z jakością modelu lokalizacyjnego $\widehat{\mathcal{L}}$. Dlatego należy precyzyjnie zdefiniować miary oceny jakości tego modelu.

Definicja 6. *Niech $\mathcal{S}_T$ (seria testowa) będzie serią pomiarową, a $\widehat{\mathcal{L}}$ modelem lokalizacyjnym. Dla elementu $s \in \mathcal{S}_T$ wprowadzamy następujące pojęcia:*
błąd horyzontalny:

$$\mathcal{E}_h(\widehat{\mathcal{L}}, s) = \sqrt{(\hat{x} - x)^2 + (\hat{y} - y)^2}$$

błąd rozpoznania piętra:

$$\mathcal{E}_f(\widehat{\mathcal{L}}, s) = |\hat{f} - f|,$$

gdzie $\widehat{\mathcal{L}}(\pi(s)) = (\hat{x}, \hat{y}, \hat{f})$ i $\mathcal{L}(s) = (x, y, f)$. Innymi słowy, (x, y, f) to prawdziwa pozycja dokonania pomiaru s podczas gdy $(\hat{x}, \hat{y}, \hat{f})$ to przewidywana pozycja na podstawie wektora RSS $\pi(s)$.

Definicja 7. *Dla serii testowej $\mathcal{S}_T$ i modelu lokalizacyjnego $\widehat{\mathcal{L}}$ zdefiniujmy:*
(i) Średni błąd horyzontalny

$$\textbf{\textit{HME}}(\widehat{\mathcal{L}}, \mathcal{S}_T)\ [m] = \frac{\sharp\{\mathcal{E}_h(\widehat{\mathcal{L}}, s) :\ s \in S\}}{\sharp\mathcal{S}_T};$$

(ii) Medianę błędu horyzontalnego

$$\textbf{\textit{HMED}}(\widehat{\mathcal{L}}, \mathcal{S}_T)\ [m] = \textit{mediana}\{\mathcal{E}_h(\widehat{\mathcal{L}}, s) :\ s \in S\};$$

(iii) Percentyl błędu horyzontalnego

$$\textbf{\textit{HPERC}}(\widehat{\mathcal{L}}, \mathcal{S}_T)\ [m] = \textit{Perc}_{80}\{\mathcal{E}_h(\widehat{\mathcal{L}}, s) :\ s \in S\};$$

(iv) Skuteczność rozpoznania piętra

$$\textbf{\textit{ACC}}(\widehat{\mathcal{L}}, \mathcal{S}_T) = \frac{\sharp\{s \in \mathcal{S}_T :\ \mathcal{E}_f(\widehat{\mathcal{L}}, s) \neq 0\}}{\sharp\mathcal{S}_T};$$

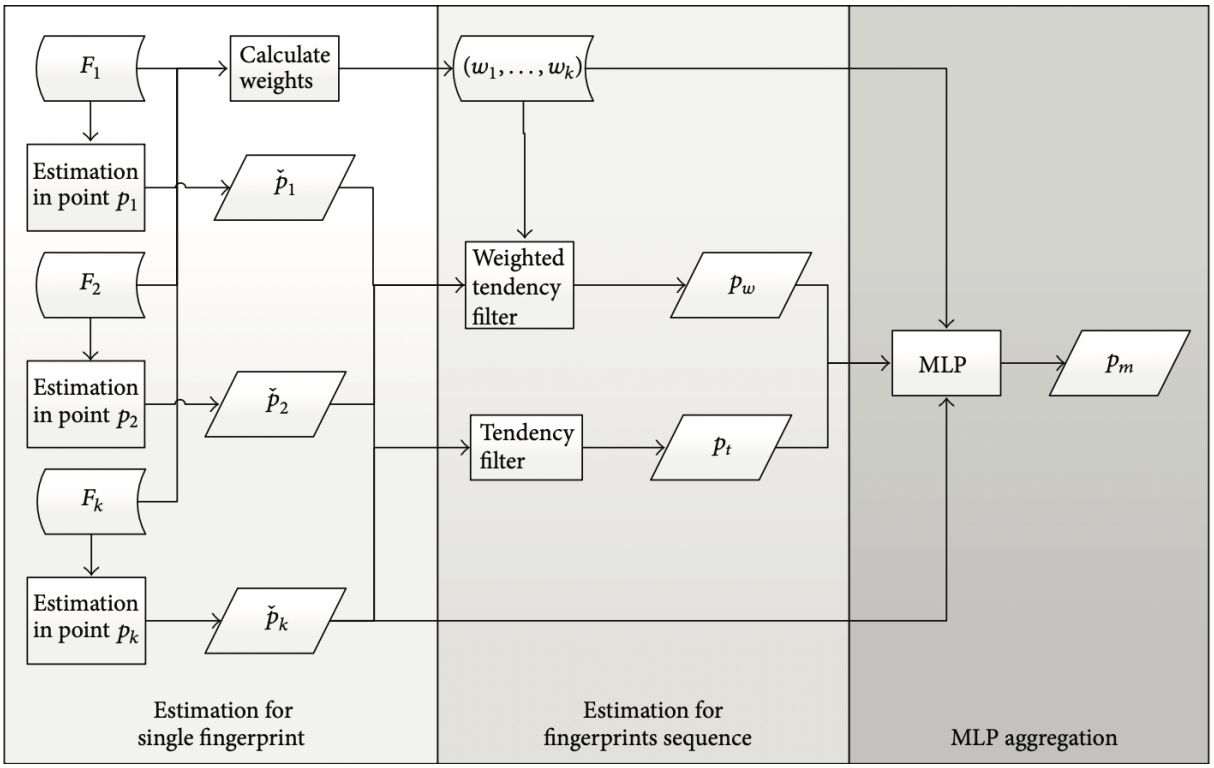
*Gdy z kontekstu jasno wynika, jaki model jest testowany i jaka seria testowa została użyta, symbol modelu i serii testowej zostanie pominięty, tzn: **HME** $= \textbf{\textit{HME}}(\widehat{\mathcal{L}}, \mathcal{S}_T)$, **ACC** $= \textbf{\textit{ACC}}(\widehat{\mathcal{L}}, \mathcal{S}_T)$ itp..*

Lokalizacja w oparciu o sygnały GSM

R. Górak, M. Luckner, M. Okulewicz, J. Porter-Sobieraj i P. Wawrzyniak, "Indoor Localisation Based on GSM Signals: Multistorey Building Study," *Mobile Information Systems*, t. 2016, 2016, **punkty MEiN: 40, IF: 1.863**, ISSN: 1875905X

W pracy [A8] podjęliśmy problem wykrycia aktualnej lokalizacji $p = (x, y, f)$ śledzonego obiektu na podstawie sekwencji pomiarów. System lokalizacyjny przyjmował jako dane wejściowe sekwencję sygnałów GSM $(F^1, F^2, \ldots, F^k)$, gdzie $\pi(F^i) \in \mathbf{R}^{39}$. Sygnały były zbierane w kolejnych punktach ścieżki ruchu, gdzie wektor F^k jest ostatnim zebranym wektorem sygnałów.

Zaproponowałem algorytm lokalizacji, przedstawiony na Rys.1. Schemat przebiega osobno dla każdej współrzędnej p (czyli dla $p^1 = x$, $p^2 = y$ i $p^3 = f$). Proces jest hierarchiczny i ma trzy kroki dla każdej $j$tej współrzędnej ($j \in \{1, 2, 3\}$).



Rysunek 1: Wieloetapowa estymacja punktu p. Od lewej estymacja pojedynczego punktu z użyciem uczenia maszynowego, estymacja sekwencji punktów stosując funkcję trendu, agregacja cząstkowych estymacji. Źródło [A8].

Krok 1 Estymacja współrzędnych dla pojedynczego wektora sygnałów F^i jest wykonywana dla każdego $i \leqslant k$. W ten sposób otrzymujemy k estymacji punktu $\{\hat{p}_i^j\}_{i \leqslant k}$ reprezentujących sekwencję przemieszczania się lokalizowanego terminala.

Krok 2 Dla ostatniego punktu w sekwencji wyliczane są estymacje $\hat{p}_t = (\hat{x}_t, \hat{y}_t, \hat{f}_t)$, jako średnia z estymacji $\{\hat{p_1}, \ldots, \hat{p_k}\}$, oraz $\hat{p}_w = (\hat{x}_w, \hat{y}_w, \hat{f}_w)$, która jest estymacją wyliczaną jako tendencja centralna za pomocą ważonych estymacji.

Można założyć, że dla niewielkich długości sekwencji (k), średnia estymacja pozycji ($\hat{p}_t$), w której pobrano F^k będzie dokładniejsza niż estymacja na podstawie pojedynczego odczytu. Dodatkowo, w estymacji ważonej ($\hat{p}_w$), wagi są dobrane tak, aby wzmocnić wpływ punktów obarczonych niskim błędem lokalizacji. W pracy wykazano, że istnieje korelacja między błędem lokalizacji, a liczbą źródeł sygnałów zarejestrowanych w wektorze F^i ($|supp(v)|$).

Krok 3 Ostateczną estymację wylicza funkcja $F^k \rightarrow \hat{p}_k^j$, która opiera się ona na agregacji trzech wyliczonych dotąd estymacji położenia: $\hat{p}_k$, $\hat{p}_t$ oraz $\hat{p}_w$. Do tego celu stosujemy perceptron wielowarstwowy, który oblicza ostateczną estymację $\hat{p}_m = (\hat{x}_m, \hat{y}_m, \hat{f}_m)$.

Tab. 1 podsumowuje błędy lokalizacji na każdym etapie procesu lokalizacji. Zaproponowany proces lokalizacji powoduje zmniejszenie błędu dla każdej z zastosowanych miar.

Tabela 1: Podsumowanie błędów lokalizacji w poszczególnych etapach procesu. Źródło [A8].

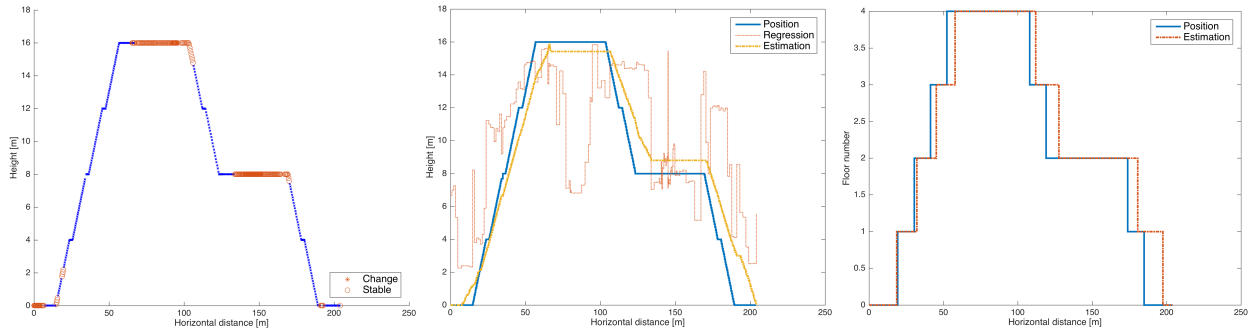
Etap	HME	HMED	HPERC	ACC
Estymacja z punktu	6.75	5.66	9.86	42.93
Estymacja z sekwencji	5.32	4.67	7.51	37.24
Estymacja z agregacji	5.18	4.39	7.47	35.79

Identyfikacja piętra na trasie w oparciu o sygnały GSM

M. Luckner i R. Górak, "Hybrid algorithm for floor detection using GSM signals in indoor localisation task," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 9648, 2016, s. 730–741, ISBN: 9783319320335

Wyniki przedstawione w pracy [A8] pokazały, że główną trudnością lokalizacji wewnątrz budynków, na podstawie sygnałów GSM, jest wykrycie na którym piętrze znajduje się lokalizowany obiekt.

W pracy [A10] zaproponowałem algorytm off-line, który etykietuje wektory RSS numerem bieżącego piętra. Algorytm używa do nauki jednego przejścia wyznaczoną trasą. Następnie działa wielofazowo. W pierwszej fazie wykrywa punkty potencjalnej zmiany piętra. W drugiej fazie, funkcja regresji normalizuje wysokość zmiany i oblicza jej kierunek. W ostatniej fazie, estymowana wysokość jest mapowana na numer piętra. Produkty poszczególnych faz przedstawiono na Rys. 2.



Rysunek 2: Hybrydowe wykrywanie bieżącego piętra. Od lewej: wykrycie punktów potencjalnej zmiany piętra, estymacja kierunku i wielkości zmiany, końcowa estymacja piętra. Źródło [A10].

Do inicjacji algorytm potrzebuje zapisu przejścia wyznaczoną trasą w postaci uporządkowanego zbioru wektorów sygnałów mapowanych na wysokość $\left((\pi(F^1), F_3^1), (\pi(F^2), F_3^2), \ldots, (\pi(F^n), F_3^n)\right)$. W procesie wstępnego przetwarzania danych porównywana jest wysokość pomiędzy dwoma kolejnymi punktami. Jeżeli różnica wysokości $|F_3^k - F_3^{k+1}|$ jest większa niż $0,1$ metra to oba punkty są oznaczane jako punkty zmiany piętra. W innym przypadku punkty te są oznaczane jako punkty stabilne. Należy podkreślić, że ścieżka uczenia się jest jedyną wiedzą o infrastrukturze budynku wykorzystywaną przez algorytm.

Następnie, punkty (oznaczone jako punkty zmiany lub stabilne) tworzą zbiór uczący dla metody klasyfikacji binarnej. Przykład wykrytych punków różnych typów przedstawia pierwszy wykres na Rys. 2.

W kolejnej fazie, zapis przejścia jest wykorzystywany ponownie, aby wytrenować funkcję regresji, estymującą wysokość na podstawie wektora sygnałów. Środkowy wykres na Rys. 2 pokazuje, że wyniki działania takiej funkcji potrafią być dość chaotyczne, i dlatego sama w sobie nie nadaje się ona do rozwiązania problemu określenia bieżącego piętra.

Po wytrenowaniu metody klasyfikacji i funkcji regresji na trasie uczącej możliwe jest zastosowanie algorytmu estymacji piętra do innych tras, które przechodzą przez ten sam lub podobny zestaw punktów. Jeśli początek trasy znajduje się na tym samym piętrze co koniec, to w procedurze estymacji przyszłości nie są potrzebne żadne dodatkowe informacje.

Na trasach, punkty zmiany wysokości są wykrywane przez metodę klasyfikacji. Zmianę wysokości można oszacować przy założeniu, że każdy punkt zmiany oznacza zmianę wysokości o stałą d. W tym kroku pojawiają się dwa problemy. Pierwszym z nich jest wykrycie kierunku lokalnej zmiany. Drugim jest oszacowanie bezwzględnej wartości d dla całej trasy.

W celu rozwiązania pierwszego problem wysokość bazowa jest ustawiona jako zero. Gdy algorytm napotyka na sekwencję punktów zmiany, regresja wysokości jest obliczana dla wszystkich punktów tej sekwencji. Jeżeli większość z nich jest zlokalizowana powyżej poziomu bazowego to sekwencja jest wznosząca. W przeciwnym przypadku sekwencja jest malejąca. Następnie poziom bazowy jest zmieniany

o $\pm d \times n$, a procedura jest powtarzana dla następnej sekwencji punktów zmiany.

Zmianę wysokości dla wysokości bazowej f i ciągu punktów n można obliczyć jako.

$$\delta f = n \times d \times \text{sgn}\left(\sum_{1}^{n} \text{sgn}(f - \hat{f}_i)\right), \tag{1}$$

gdzie f_i jest wartością funkcji regresji dla punktu i, a sgn jest funkcją signum.

Drugi problem rozwiązuje się następująco. Najpierw ustalono, że d wynosi 16 cm, co jest częstym rozmiarem stopnia na klatce schodowej. Następnie utworzona estymacja jest normalizowana. Normalizacja dokonywana jest na podstawie funkcji regresji, która szacuje wysokość na całej ścieżce.

Normalizacja jest wykonywana osobno dla części wstępującej i części zstępującej ścieżki. Najpierw lokalizowany jest ostatni punkt sekwencji zstępującej, a jego położenie oznaczane jest jako h. Dla wszystkich punktów od 1 do h normalizowana wysokość $\bar{f}_i$ jest obliczana jako:

$$\bar{f}_i = \frac{max\hat{f}(f_i - \min_i(f_i))}{\max_i(f_i) - \min_i(f_i)} \text{ dla } i = 1, \ldots, h, \tag{2}$$

gdzie $\hat{f}$ jest funkcją regresji, a f jest wysokością oszacowaną za pomocą wzoru (1).

Druga część punktów jest normalizowana przy użyciu następującego wzoru:

$$\bar{f}_i = \frac{\bar{f}_h(f_i - f_h)}{f_h - \min_i(f_i)} + \bar{f}_h \text{ dla } i = h+1, \ldots, n. \tag{3}$$

.

Na koniec uzyskana funkcja wysokości $\bar{f}$ jest przekształcana na wartości dyskretne w celu wykrycia aktualnego piętra.

Proponowany algorytm rozwiązuje problem określenia aktualnego piętra dla ustalonej trasy. Mimo tego ograniczenia, może być zastosowany w praktyce np. do analizy realizacji zadań patrolowych dla pracowników ochrony. Dzięki połączeniu detekcji punktów zmiany piętra z estymacją wysokości otrzymujemy wyniki o 40 procent lepsze od wyników uzyskanych przy zastosowaniu samej regresji.

Lokalizacja w oparciu o sygnały Wi-Fi

R. Górak i M. Luckner, "Modified random forest algorithm for Wi–Fi indoor localization system," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 9876 LNCS, 2016, s. 147–157

W pracy [A7] zaproponowaliśmy zmodyfikowany algorytm lokalizacyjny oparty o zastosowanie lasów losowych. Modyfikacja polega na stworzeniu kombinacji modeli lokalizacyjnych zbudowanych dla każdego punktu dostępowego z sieci budynku z osobna. Zaproponowana modyfikacja daje nam znaczną poprawę dokładności w porównaniu do bezpośredniego zastosowania lasu losowego.

W ramach pracy porównywałem stworzony model z innymi rozwiązaniami. Przyrost dokładności poziomej **HME**, dla zaproponowanej metody $\widehat{\mathcal{L}}_{mRF}$, wynosi od 5 do 9 procent w porównaniu do lasu losowego. W pracy (Karwowski, Okulewicz i Legierski 2013) autorzy testowali, na tych samych danych, lokalizator wykorzystujący perceptron wielowarstwowy. Najlepsze wyniki, jakie uzyskali, to błąd 5.18m, dla 90 percentyla, przy estymacji $\hat{x}$ i 5.82m, dla 90 percentyla, przy estymacji $\hat{y}$. Testując $\widehat{\mathcal{L}}_{mRF}$ otrzymujemy błąd 4.11m, dla 90 percentyla, przy estymacji $\hat{x}$ i 4.72m, dla 90 percentyla, przy estymacji $\hat{y}$.

M. Luckner, B. Topolski i M. Mazurek, "Application of XGboost algorithm in fingerprinting localisation task," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 40**, t. 10244 LNCS, 2017, s. 661–671, ISBN: 9783319591049

W pracy [A5] przedstawiliśmy zastosowanie, nie używanego wcześniej do lokalizacji, algorytmu XGBoost (Chen i Guestrin 2016). Lokalizacja była oparta na danych z sieci Wi-Fi.

Zaproponowałem nowe podejście do zastosowania algorytmu w zadaniu lokalizacji. Aplikacja bezpośrednia oznaczała by stworzenie trzech oddzielnych modeli do estymacji każdej ze współrzędnych osobno (co było np. punktem wyjścia w pracy [A7]). W zaproponowanym podejściu, zamiast używać dwóch modeli do estymacji współrzędnych horyzontalnych w całym budynku, budujemy dwa modele dla każdego piętra budynku. Poprzez uczynienie każdego modelu bardziej specyficznym niż model ogólny możemy dokonać dokładniejszych predykcji na każdym piętrze.

Zastosowanie modeli XGBoost odbyło się z wykorzystaniem następującego schematu. Pierwszy z modeli $\widehat{\mathcal{L}}_f$ estymował numer bieżącego piętra:

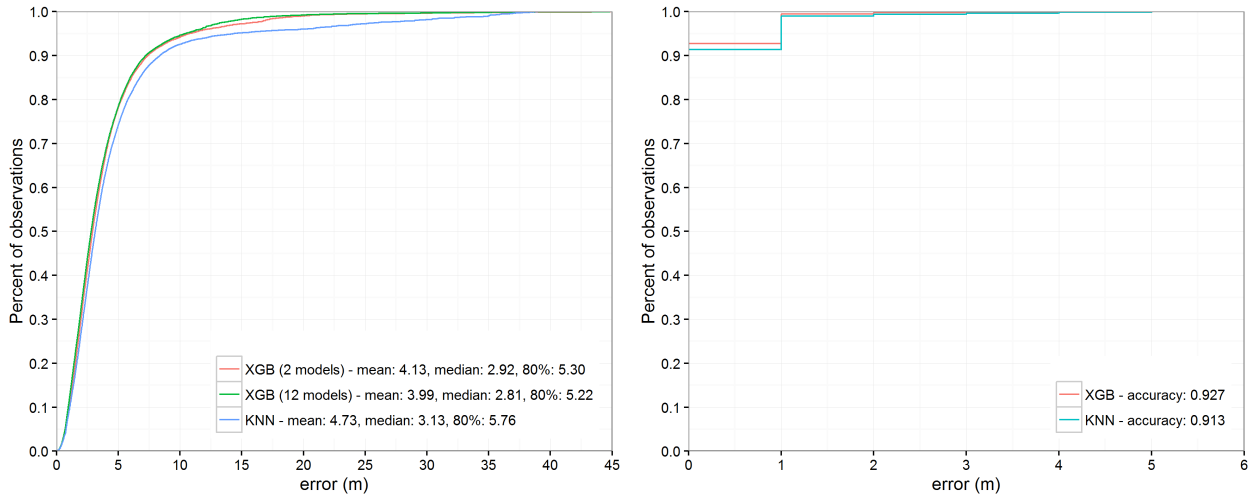
$$\hat{f}_m = \widehat{\mathcal{L}}_f(m). \tag{4}$$

Dla każdego piętra stworzono osobne modele do estymacji współrzędnych poziomych. Współrzędne $\hat{x}_m$ i $\hat{y}_m$ są określane odpowiednio przez modele $\widehat{\mathcal{L}}_x^i$ i $\widehat{\mathcal{L}}_y^i$ gdzie $i \in [\min(\hat{f}_m), \ldots, \max(\hat{f}_m)]$.

W zależności od wykrytego piętra współrzędne $\hat{x}_m$ i $\hat{y}_m$ są wyliczane jako:

$$\hat{x}_m = \widehat{\mathcal{L}}_x^{\hat{f}_m}(m), \tag{5}$$

$$\hat{y}_m = \widehat{\mathcal{L}}_y^{\hat{f}_m}(m). \tag{6}$$



Rysunek 3: Porównanie wyników implementacji XGBoost dla dwóch i 12 modeli z wynikami kNN. Po lewej rozkład błędów lokalizacji horyzontalnej, po prawej detekcji piętra. Źródło [A5].

Zaproponowane podejście zapewniło uzyskanie średniego błędu lokalizacji mniejszego niż trzy metry i 93 procentową skuteczność wykrywania piętra przy ograniczeniu się do analizy punktów dostępowych należących do infrastruktury budynku. Okazało się też statystycznie lepsze od wcześniejszych rozwiązań działających na tych samych danych kNN (Rys. 3) i lasów losowych [A7].

Testy systemów lokalizacyjnych

R. Górak i M. Luckner, "Long term analysis of the localization model based on Wi-Fi network," w *Studies in Computational Intelligence*, **punkty MEiN: 20**, t. 642, 2016, s. 87–96, ISBN: 9783319312767

W pracy [A6] przeprowadziliśmy analizę starzenia się systemu lokalizacyjnego. Analizowano pięć serii pomiarów. System lokalizacyjny, trenowany na danych z pierwszej serii, był testowany na danych z pozostałych serii. Testowany system był oparty na lasie losowym [A7].

Tabela 3: Analiza zmiany błędów lokalizacji w czasie. Źródło [A6].

Seria	rok	**ACC**	**HME** [m]	**HMED** [m]	**HPERC** [m]
1.	2012	0.99	1.32	0.18	1.19
2.	2012	0.97	3.61	2.85	5.20
3.	2012	0.98	3.52	2.95	5.09
4.	2014	0.92	4.39	3.70	6.49
5.	2014	0.92	4.52	3.79	6.48

Uporządkowane chronologiczne serie pokazują starzenie się systemu (patrz Tab. 2). Analiza wskazuje, że choć wydajność raz zbudowanego modelu maleje z czasem, to nadal jest on użyteczny. Przykładowo, po dwóch latach od zbudowania systemu, poziomy błąd średni wynosi około 4.5 m, a ACC 92%.

M. Luckner i R. Górak, "Comparison of floor detection approaches for suburban area," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 9622, 2016, s. 782–791, ISBN: 9783662493892

W pracy [A9] badałem możliwość wykorzystania stworzonych systemów lokalizacyjnych w budynkach prywatnych. Podczas gdy większość testów przeprowadza się w rozległych budynkach publicznych, testy przeprowadzono w podmiejskim 3-piętrowym budynku ze słabo rozwiniętą lokalną infrastrukturą Wi-Fi i komórkową. Dane były zbierane w procesie crowdsourcingowym przez telefon komórkowy przyczepiony do samobieżnego robota (patrz Rys. 4). Dzięki temu znana jest współrzędna f_4 (numer piętra), ale nie pozostałe współrzędne lokalizacyjne.

Rysunek 4: Obszar pomiarowy i jednostka mierząca. Źródło [A9].

Jednakże, w opisywanym przypadku, lokalizacja horyzontalna nie jest tak istotna, jak określenie na którym piętrze znajduje się namierzane urządzenie (telefon komórkowy). Porównałem więc metody detekcji piętra korzystając z sygnałów telefonii komórkowej drugiej (2G) i trzeciej (3G) generacji (zbiory oznaczone odpowiednio jako GSM i UMTS), z sygnałów lokalnych sieci Wi-Fi i z czujnika ciśnieniowego.

Ze względu na różne typy sygnałów, w celu rzetelnego porównania ich przydatności w zadaniu lokalizacji, przeprowadzono trzy rodzaje testów. Pierwszy test zakładał wykorzystanie wszystkich źródeł sygnałów danego typu i wszystkich wykonanych pomiarów.

Jednakże, wyniki takiego testu mogą być tendencyjne, gdyż pomiar ciśnieniowy jest pojedynczym sygnałem, a i sieci różnych typów różnią się liczbą punktów dostępowych ($\max(|supp(v)|)$). Z tego powodu przeprowadzono drugi test, w którym system używa tylko jednego źródła sygnału, najistotniejszego dla dyskryminacji, podczas poprzedniego testu.

Drugim czynnikiem, który może wpływać na wyniki detekcji, jest rozmiar zbioru danych. Telefon zbiera

dane z różnych czujników z różną częstotliwością, więc każdy test odbywał się na zbiorze danych o innej liczności. W celu wyeliminowania tego czynnika, przeprowadzono trzeci test, w którym na każde piętro przypadała taka sama liczba pomiarów per czujnik. Wyniki testów przedstawiono w Tab. 3.

Tabela 4: Uzyskane **ACC** dla różnych metod detekcji piętra. Testy przeprowadzono w trzech wariantach: stosując wszystkie dane danego czujnika, tylko najbardziej znaczące źródło sygnałów, taką samą liczbę pomiarów. Źródło [A9].

Metoda	Pełne dane	Pojedyncze źródło	Identyczna liczba pomiarów
Ciśnienie	0.74	0.74	0.72
GSM	0.98	0.78	0.97
Wi–Fi	0.99	0.84	0.99
UMTS	1.00	0.89	1.00

Wyniki pokazują, że opisane czynniki, aczkolwiek mogą wpływać na ogólną wartość **ACC**, nie wpływają na ranking metod. W rezultacie, badanie pokazuje, że w przeciwieństwie do wysokich budynków publicznych [A8], sygnały GSM mogą być skutecznie wykorzystywane przy rozpoznawaniu bieżącego piętra.

Lokalizacja odporna na błędy infrastruktury

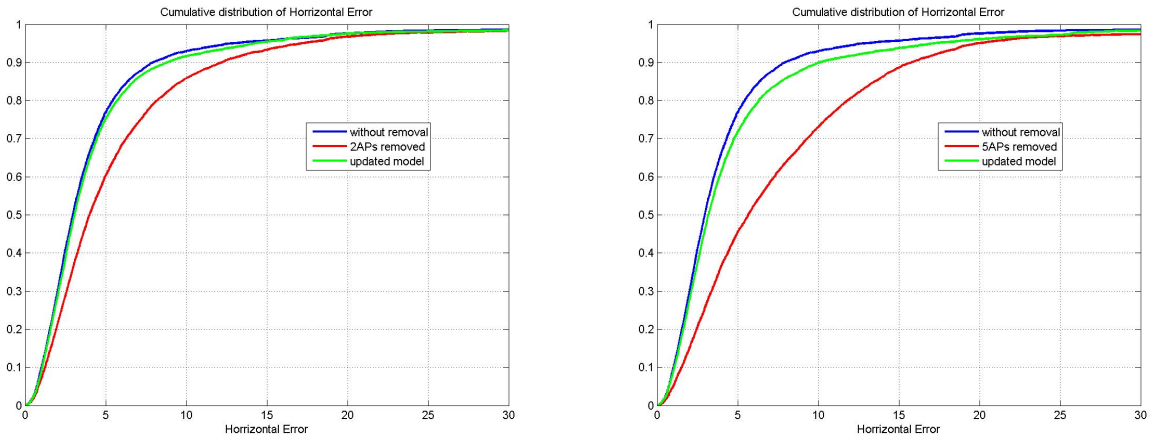
R. Górak i M. Luckner, "Malfunction immune wi–fi localisation method," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 9329, 2015, s. 328–337

W pracy [A11] przedstawiamy jak ograniczyć liczbę obserwowanych AP przez system lokalizacji oraz jak uodpornić metodę lokalizacji na awarie. Proponowane rozwiązanie, oparte na lasie losowym, ogranicza wzrost błędu lokalizacji.

W ramach tej pracy zajmowałem się zagadnieniem selekcji AP i analizą wyników. Selekcja AP zostanie omówiona szczegółowo w ramach streszczenia pracy [A3].

Rys. 5 przedstawia dwa przypadki usunięcia dwóch i pięciu punktów dostępowych. Skumulowana empiryczna dystrybucja błędu **HME** pokazuje, że ich usunięcie znacznie pogarsza jakość lokalizacji horyzontalnej.

.

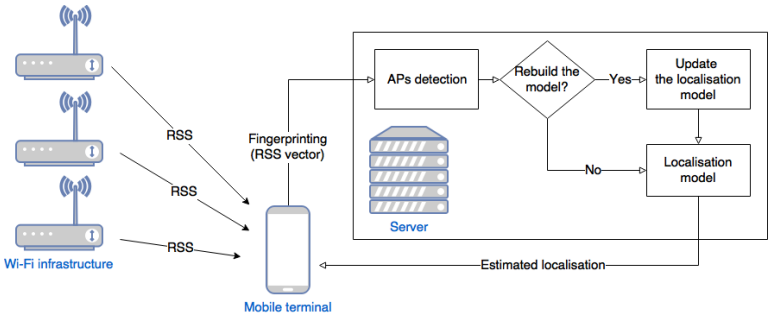


Rysunek 5: Wpływ awarii punktów dostępowych na wyniki lokalizacji. Skutki usunięcia dwóch (po lewej) i pięciu AP (po prawej). Źródło [A11].

Równocześnie widzimy, że wprowadzenie poprawki, polegającej na wytrenowaniu modelu lokalizacyjnego, który nie wykorzystuje uszkodzonych punktów dostępowych do lokalizacji, znacząco zmniejsza skutki awarii. System rozwijany jest w pracach [A2, A3].

R. Górak i M. Luckner, "Automatic detection of missing access points in indoor positioning system," *Sensors (Switzerland)*, t. 18, nr. 11, paź. 2018, **punkty MEiN: 100, IF: 3.847**, ISSN: 14248220

W pracy [A3] przedstawiliśmy system lokalizacji oparty o technologię Wi-Fi. Składa się on z dwóch głównych części, modelu lokalizacyjnego oraz modułu wykrywania punktów dostępowych. System wykorzystuje wektory RSS zbierane przez wiele terminali mobilnych do wykrycia, który AP powinien zostać włączony do modelu lokalizacji i czy model ten wymaga aktualizacji (przebudowy). Przebudowa modelu lokalizacyjnego zabezpiecza system lokalizacyjny przed znaczną utratą dokładności (patrz [A11]). Proponowana automatyczna detekcja brakujących AP ma charakter uniwersalny i może być zastosowana do dowolnego modelu lokalizacyjnego Wi-Fi, który został utworzony metodą fingerprintingu. Rys. 6 prezentuje schemat zaproponowanego rozwiązania.



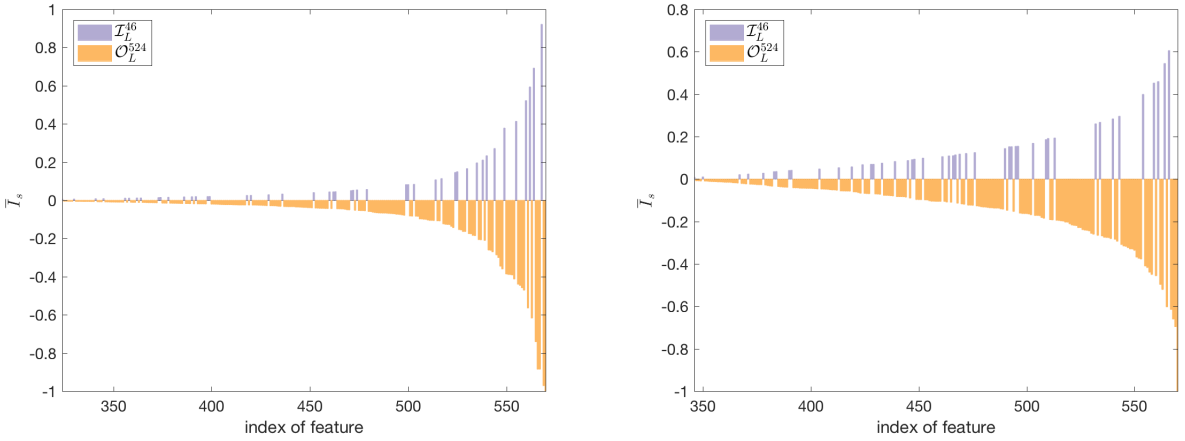
Rysunek 6: Schemat systemu lokalizacyjnego. Źródło [A3].

W ramach pracy [A3] przeprowadziłem analizę punktów dostępowych i stworzyłem rozwiązanie ograniczające liczbę AP, które są brane pod uwagę przy budowie modelu lokalizacji. Możemy wyróżnić dwa rodzaje AP dla serii uczącej $\mathcal{S}_L$. Pierwsza grupa $\mathcal{I}_L \subset \mathcal{S}_L$ należy do znanej infrastruktury Wi-Fi i przeważnie zarządca budynku może zdiagnozować stan AP z tej grupy. Druga grupa $\mathcal{O}_L \subset \mathcal{S}_L$ zawiera AP, które nie należą do infrastruktury. Ta grupa zmienia się znacznie bardziej dynamicznie i zawiera obserwowane stacjonarne AP z sąsiedztwa, jak również AP mobilne.

W budynku, w którym przeprowadziliśmy badania, zaobserwowaliśmy 46 AP z infrastruktury. Reszta - ponad pięćset - to były zewnętrzne AP. O ile bezpieczniej jest ograniczyć system lokalizacji do AP z infrastruktury, to pozostałe AP mogą wnieść do systemu dodatkową wiedzę.

Do szacowania ważności AP wykorzystywana jest metoda stosowana w drzewach klasyfikacyjnych CART. Istotność cech jest pochodną jakości podziału dokonywanego przez daną cechę na zbiorze uczącym, mierzoną odpowiednią miarą błędu (Rokach i Maimon 2010).

Rys. 7 przedstawia istotność wykorzystywanych AP w zadaniu lokalizacji horyzontalnej i detekcji piętra. W przypadku lokalizacji horyzontalnej ważność została obliczona jako średnia ważności dla współrzędnych x i y. Dane dla AP z infrastruktury zostały oznaczone wartościami dodatnimi, a dla zewnętrznych ujemnymi.

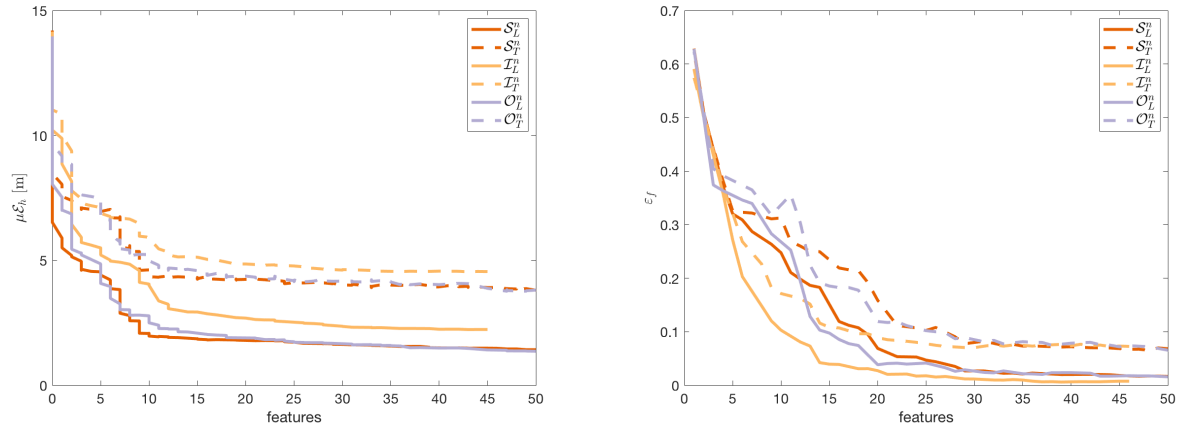


Rysunek 7: Znaczenie punktów dostępowych sieci Wi-Fi obliczane osobno dla lokalizacji horyzontalnej (po lewej) i detekcji piętra (po prawej). Źródło [A3].

Wykresy pokazują, że sygnały zewnętrzne są ważne dla stworzonego modelu lokalizacyjnego. Widzimy również, że liczba ważnych AP - o ważności powyżej 0.1 - jest większa dla zadania lokalizacji horyzontalnej niż dla zadania detekcji piętra. Jednakże powyższa wizualizacja nie daje nam jeszcze informacji o wpływie cech na błąd lokalizacji.

Uporządkujmy zaobserwowane AP w zbiorze uczącym sL według ważności. Zdefiniujmy trzy podzbiory iL^n, oL^n, sL^n, które składają się z sygnałów od n AP o największej ważności. W pierwszym zestawie znalazły się tylko AP z infrastruktury. Drugi zestaw składa się z zewnętrznych AP. Ostatni zestaw jest połączeniem obu poprzednich zestawów i zawiera wszystkie AP.

Na Rys. 8 pokazano, jak zmniejszenie liczby cech poprzez odcięcie AP o najmniejszym znaczeniu wpływa na miary błędu **HME** i **ACC**. Porównanie przeprowadzono dla zbiorów uczących iL^n, oL^n, sL^n i testowych iT^n, oT^n, sT^n. Dla czytelności wykres ogranicza się do 50 najważniejszych cech.



Rysunek 8: Zależność pomiędzy liczbą cech a miarami błędu **HME** (po lewej) i **ACC** (po prawej). Źródło [A3].

Widać, że możemy zredukować błędy **HME** i **ACC** poprzez zwiększenie liczby cech uwzględnianych w zadaniu lokalizacji. Jednak system lokalizacji będzie mniej wydajny, gdy liczba cech będzie bardzo duża.

Dla obu błędów charakterystyka krzywej dla zbiorów uczących i testowych jest zbliżona do siebie. Możemy zatem przyjąć, że zmniejszenie liczby cech przyniesie podobne rezultaty dla zbioru testowego jak dla zbioru uczącego i błąd lokalizacji nie wzrośnie po zmniejszeniu liczby cech w porównaniu z wynikiem uzyskanym dla pełnego zestawu cech.

Pewne ciekawe spostrzeżenia można poczynić, gdy porównamy uzyskane wyniki na różnych zbiorach

danych, dla tej samej liczby cech. W przypadku lokalizacji horyzontalnej wyniki uzyskane na zestawach uczących wykorzystujących wszystkie 46 AP z infrastruktury są gorsze od wyników uzyskanych na pozostałych zestawach AP. Tymczasem wyniki uzyskane na zbiorach testowych dla 46 cech są niemal takie same niezależnie od użytego zestawu cech. Jednak w przypadku wykrywania piętra obserwacje są odmienne. Po pierwsze, AP infrastrukturalne dają najlepsze rezultaty na zbiorze uczącym, chociaż wyniki są bardzo podobne. Po drugie, wyniki uzyskane na zbiorze testowym dla AP z infrastruktury są zauważalnie gorsze od wyników uzyskanych na pozostałych zbiorach. Można zatem stwierdzić, że lepiej jest włączyć do systemu lokalizacji punkty zewnętrzne.

W omawianym przypadku liczba AP z infrastruktury budynkowej jest stosunkowo niewielka i można używać ich wszystkich. Jednak stosując zewnętrzne AP, stajemy przed problemem doboru progu odcięcia. Wybór punktu odcięcia może być dokonany na podstawie obserwacji redukcji błędu na zbiorze uczącym, w zależności od liczby użytych cech w zadaniu lokalizacyjnym. W pierwszym podejściu wybieramy taką liczbę cech, która daje minimalny błąd na zbiorze uczącym według następujących wzorów:

$$m_h = \arg\min_{n \leqslant N}(\textbf{HME}(\widehat{\mathcal{L}}, \mathcal{S}_L^n)), \tag{7}$$

$$m_f = \arg\min_{n \leqslant N}(\textbf{ACC}(\widehat{\mathcal{L}}, \mathcal{S}_L^n)). \tag{8}$$

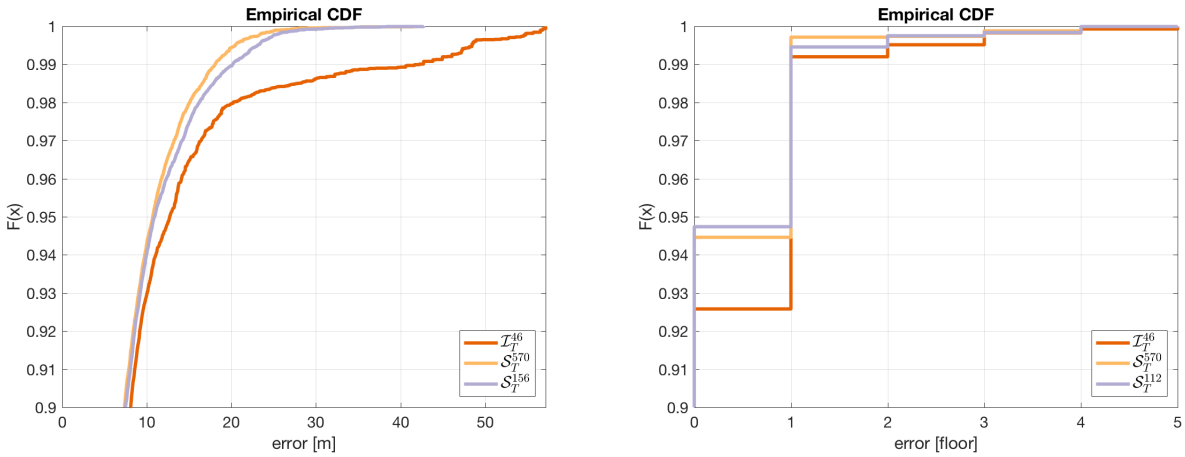
Jednak dla zstępujących funkcji błędu **HME** i **ACC** można się spodziewać, że wartości m_h i m_f będą bliskie maksymalnej liczbie cech N. Podejście alternatywne dopuszcza obranie mniejszej liczby cech kosztem zwiększenia błędu w ramach założonej tolerancji (w naszym przypadku $\theta_f = 5$cm i $\theta_h = 0.5\%$). Wartość odcięcia oblicza się za pomocą następujących wzorów:

$$n_h = \min\{n : \textbf{HME}(\widehat{\mathcal{L}}, \mathcal{S}_L^n) - \textbf{HME}(\widehat{\mathcal{L}}, \mathcal{S}_L^{m_h}) < \theta_h \wedge n \leqslant m_h\}, \tag{9}$$

$$n_f = \min\{n : \textbf{ACC}(\widehat{\mathcal{L}}, \mathcal{S}_L^n) - \textbf{ACC}(\widehat{\mathcal{L}}, \mathcal{S}_L^{m_f}) < \theta_f \wedge n \leqslant m_f\}. \tag{10}$$

Aby podjąć decyzję, który zestaw cech powinien być użyty jako ostateczny, należy porównać błędy **HME** i **ACC** uzyskane na zbiorach $\mathcal{S}_T^N$,$\mathcal{S}_T^{n_f}$, $\mathcal{S}_T^{n_h}$ i $\mathcal{I}_T^{46}$. Zestawy $\mathcal{S}_T^{n_f}$, $\mathcal{S}_T^{n_h}$ zawierają sygnały ze wszystkich rodzajów AP, zostały jednak zredukowane do zestawu najważniejszych AP, za pomocą wzorów (9) i (10). Uzyskana liczba cech to $n_h = 156$ dla zadania lokalizacji horyzontalnej oraz $n_f = 112$ dla detekcji piętra.

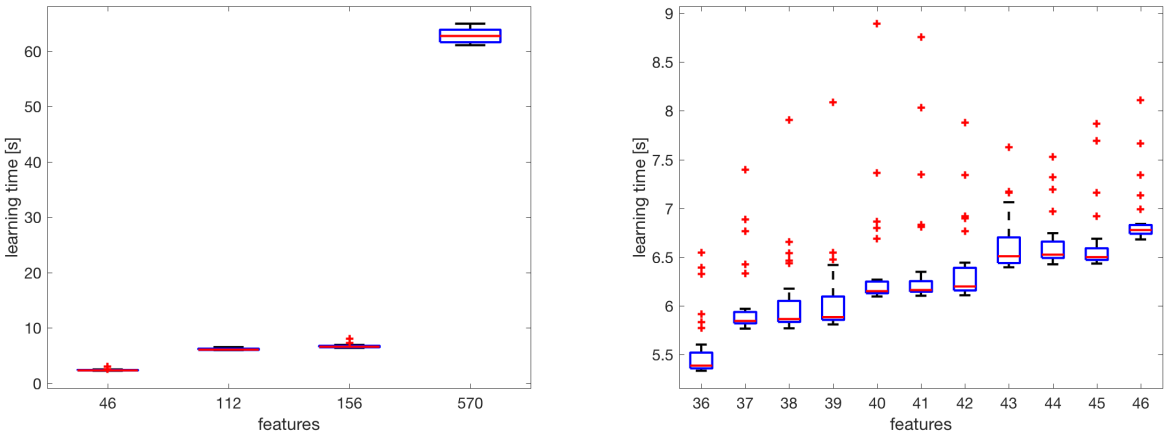
Na Rys 9 porównano wyniki uzyskane przez omawiane zestawy cech z wykorzystaniem empirycznej funkcji dystrybucji. Ze względu na duże podobieństwo wykresów wykres zaczyna się od 0.9.



Rysunek 9: Empiryczna funkcji dystrybucji błędu **HME** (po lewej) i **ACC** (po prawej). Źródło [A3].

Najlepsze wyniki uzyskano na zbiorze zawierającym wszystkie cechy. Możemy jednak ograniczyć liczbę obserwowanych AP do 27 procent używając zredukowanego zestawu cech, co spowoduje wprawdzie zwiększenie błędów grubych, w lokalizacji horyzontalnej, ale jest znaczące dla przyszłej konserwacji systemu i jego wydajności.

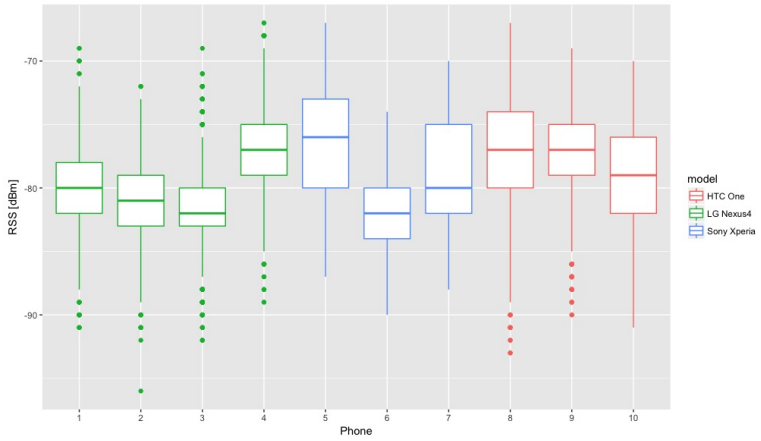
W prezentowanym rozwiązaniu sygnały RSS są analizowane w celu wykrycia brakujących AP, więc dla każdego AP musi zostać wdrożona detekcja awarii. W celu sprawdzenia praktycznej możliwości wdrożenia systemu, przeprowadziłem testy wydajnościowe rozwiązania, na komputerze z procesorem Intel Core i5 2,9 GHz i pamięcią DDR3 16 GB 1867 MHz. Oszacowanie, wykonane na podstawie 30 testów dla każdej rozważanej liczby cech, zostało przedstawione na Rys. 10.



Rysunek 10: Czas budowy klasyfikatora (po lewej), czas przebudowy systemu (po prawej). Źródło [A3].

Porównałem czas uczenia w zależności od liczby obserwowanych AP. Przedstawiony czas jest całkowitym czasem potrzebnym do stworzenia modelu lokalizacyjnego $\widehat{\mathcal{L}}_{RF}$. Średni czas waha się od ponad jednej minuty do mniej niż pięciu sekund w zależności od liczby cech. Porównałem także czas uczenia się w przypadku, gdy algorytm lokalizacji musi zostać zaktualizowany z powodu wykrycia brakującego AP. Jako początkową liczbę obserwowanych AP przyjęto liczbę AP z infrastruktury. Uzyskane wyniki pokazują, że algorytm może być zastosowany jako algorytm czasu rzeczywistego do lokalizacji, ponieważ średni czas potrzebny na stworzenie nowego modelu lokalizacyjnego po, usunięciu brakującego AP nie przekracza siedmiu sekund.

Dodatkowo, przeprowadziłem testy mające pokazać jak zmienia się RSS z tego samego AP dla różnych telefonów i modeli telefonów użytych do zbierania danych. Test został wykonany przy użyciu dziesięciu różnych telefonów jednocześnie zbierających dane.



Rysunek 11: Różnice w rejestrowanej sile sygnału dla różnych egzemplarzy i modeli telefonów. Źródło [A3].

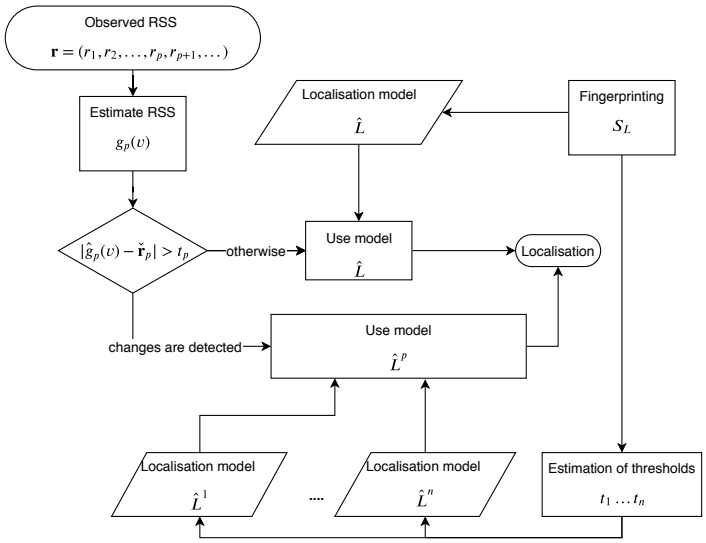
Rys. 11 pokazuje, że różnice pomiędzy RRS są niewielkie. Największa różnica między średnimi wynosi 2 dBm i 6 dBm odpowiednio dla modeli i poszczególnych telefonów. Oznacza to, że system nie stawia specjalnych wymagań dla sprzętu, którym mają być zbierane dane.

M. Luckner i R. Górak, "Automatic detection of changes in signal strength characteristics in a Wi-Fi network for an indoor localisation system," *Sensors (Switzerland)*, t. 20, nr. 7, s. 1–13, 2020, **punkty MEiN: 100, IF: 3.847**, ISSN: 14248220

W pracy [A2] omówiliśmy zagadnienie zmienności siły sygnału odbieranego z obserwowanego punktu dostępowego. Zagadnienie to jest rozwinięciem tematyki pracy [A3], gdzie rozważane było wykrywanie wyłączenia AP. Analizę charakteru RSS pozwala wykryć także inne zmiany w infrastrukturze, np. przesuniecie AP. Taka zmiana może obniżyć jakość usług systemu lokalizacji wewnętrznej opartego na technologii Wi-Fi.

Zaproponowaliśmy dynamiczny system oparty na estymatorze RSS wykorzystującym odczyty z innych AP. Algorytm rozpoznaje AP, który zmienił swoją charakterystykę. Następnie system przebudowuje model lokalizacyjny z pominięciem zmienionego AP, aby zachować jakość lokalizacji.

Zaproponowałem metodę wykrywania, że charakterystyka RSS (dla punktu dostępowego AP_p) uległa zmianie. Rys. 12 przedstawia ogólny schemat metody.



Rysunek 12: Schemat modelu lokalizacji z uwzględnieniem wykrywania zmian charakterystyki sygnału RSS. Źródło [A2].

Metoda rozpoczyna od wyznaczania, dla wektora $\mathbf{r} \in \mathbf{R}^n$, wektora $\breve{\mathbf{r}}_p = (r_1, r_2, ..., r_{p-1}, r_{p+1}, ...) = (r_i)_{i \neq p} \in \mathbf{R}^{n-1}$, czyli wektora $\mathbf{r}$ z usuniętą współrzędną p. Dla danych uczących $\mathcal{S}_L^{\mathcal{AP}}$, definiuje zbiór wektorów $\mathcal{RSS}_p = \{\breve{\mathbf{r}}_p \mid \exists (t, x, y, z) : (\mathbf{r}, t, x, y, z) \in \mathcal{S}_L^{\mathcal{AP}}\}$.

Stąd mamy funkcję (ewentualnie multifunkcję) g_p, która dla danego wektora $\breve{\mathbf{r}}_p \in \mathcal{RSS}_p$ zwraca brakującą współrzędną $r_p \in \mathbf{R}$ usuniętego AP_p. Następnie, na podstawie g_p i wykorzystując las losowy, tworzymy estymator $\widehat{g}_p : \mathbf{R}^{n-1} \mapsto \mathbf{R}$, który przewiduje RSS z usuniętego AP_p, na podstawie RSS z pozostałych AP. Błąd estymacji predykcji RSS definiujemy jako $|\hat{g}_p(\breve{\mathbf{r}}_p) - r_p|$ gdzie $(\mathbf{r}, t, x, y, z) \in \mathcal{S}_L^{\mathcal{AP}}$, a r_p jest p-tą współrzędną $\mathbf{r}$.

Następnie obliczamy próg t, który oddziela istotne zmiany od pozostałych. Ze względu na występowanie w rzeczywistym środowisku, różnych zakłóceń błędy estymacji będą większe od zera. Zakładamy jednak, że błędy wprowadzone przez zmianę (np. zmianę lokalizacji AP) będą znacznie wyższe od błędów typowych. Jeśli nie, to możemy założyć, że wpływ zmiany na jakość usługi lokalizacyjnej nie będzie znaczący.

Próg t należy do zbioru unikalnych wartości błędów $\mathbf{E}$ obliczonych na zbiorze danych uczących $\mathcal{S}_L^E$. Ze względów obliczeniowych można zmniejszyć rozmiar zbioru $\mathbf{E}$ zmniejszając precyzję zapisanych w nim wartości.

Próg obliczany jest według wzoru:

$$t_p = \operatorname{perc}_{80} \left\{ \sum_{(\mathbf{r},t,x,y,z) \in \mathcal{S}_L} \sigma_t \left(|\hat{g}_p(\breve{\mathbf{r}}_p) - r_p| \right) \ : t \in \mathbf{E} \right\}, \tag{11}$$

$$\sigma_t(x) = \begin{cases} 0 & \text{jeżeli } x < t \\ 1 & \text{jeżeli } x \geqslant t. \end{cases} \tag{12}$$

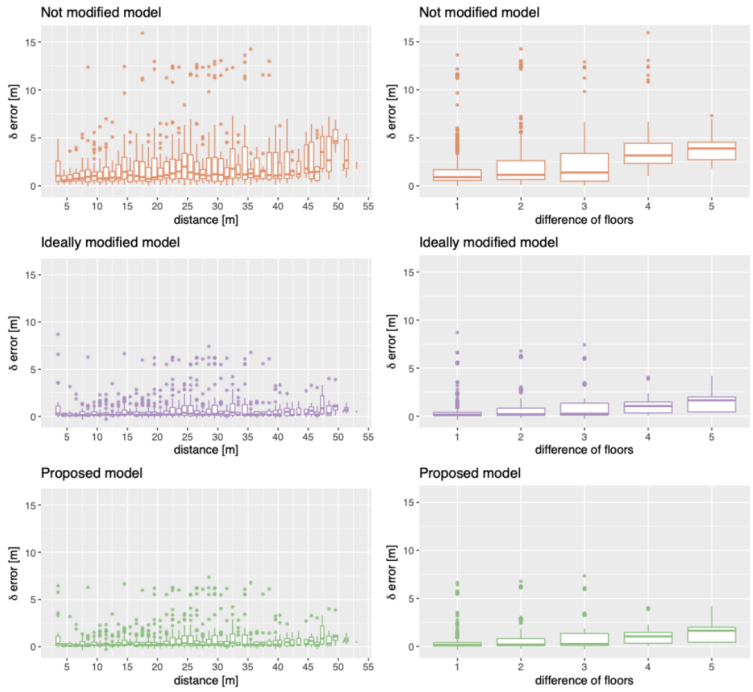
AP_p klasyfikujemy jako punkt dostępowy, który zmienił swoją charakterystykę (być może z powodu zmiany położenia), jeśli dla wektora RSS $\mathbf{r}$, $|\widehat{g}_p(\breve{\mathbf{r}}_p) - r_p| > t_p$. W takim przypadku tworzony jest nowy model lokalizacyjny $\widehat{\mathcal{L}}^p$. Tym razem jednak zbiór danych uczących to $\mathcal{S}_L^{\mathcal{AP}-\{p\}}$. Innymi słowy, usuwamy odczyty AP_p z odczytów serii uczącej $\mathcal{S}_L^{\mathcal{AP}}$. Stąd możemy zdefiniować zmodyfikowany model lokalizacyjny $m\widehat{\mathcal{L}}$ jako:

$$m\widehat{\mathcal{L}}(\mathbf{r}, p) = \begin{cases} \widehat{\mathcal{L}}(\mathbf{r}) & \text{jeżeli system nie wykrył zmiany charakterystyki } AP_p, \\ \widehat{\mathcal{L}}^p(\mathbf{r}) & \text{w p.p..} \end{cases} \tag{13}$$

Aby oszacować jakość modelu $m\widehat{\mathcal{L}}$, porównuje się go z systemem, który idealnie wykrywa, czy AP_p zmienił swoją charakterystykę. W związku z tym wprowadzamy $i\widehat{\mathcal{L}}$, takie, że:

$$i\widehat{\mathcal{L}}(\mathbf{r}, p) = \begin{cases} \widehat{\mathcal{L}}(\mathbf{r}) & \text{jeżeli } AP_p \text{ nie zmienił swojej charakterystyki (np. położenia),} \\ \widehat{\mathcal{L}}^p(\mathbf{r}) & \text{w p.p..} \end{cases} \tag{14}$$

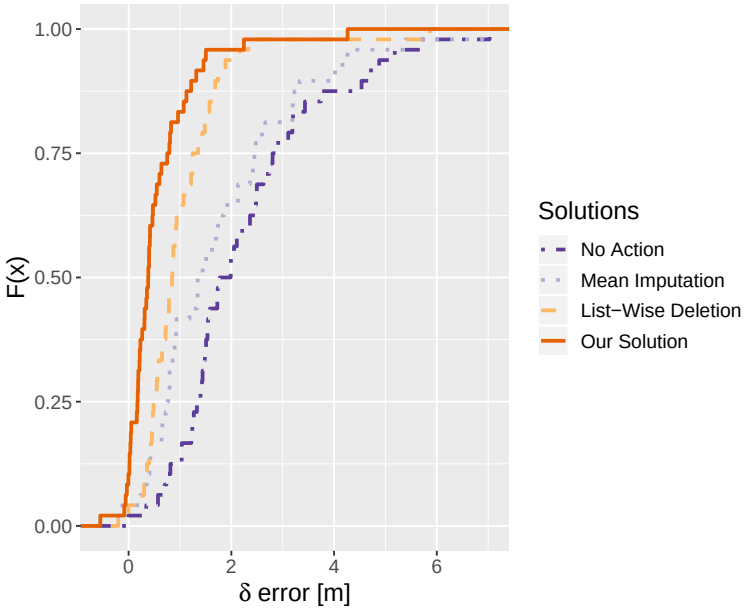
Metody przetestowano symulując zmianę położenia AP. Rysunek 13 Pokazuje jak kształtuje się błąd lokalizacji po przesunięciu AP horyzontalnie i między piętrami. Wyniki pokazano dla niezmodyfikowanego modelu, modelu idealnie wykrywającego zmiany sygnału i dla zaproponowanego rozwiązania.



Rysunek 13: Wpływ przemieszczenia AP na lokalizację. Od góry, model niezmodyfikowany, model idealnie zmodyfikowany, proponowane rozwiązanie. Źródło [A2].

Widać, że zaproponowane rozwiązanie daje wyniki bliskie wynikom idealnego rozwiązania i zdecydowanie lepsze niż niezmodyfikowany system lokalizacji.

Model detekcji zmian wykorzystałem także do wykrywania wyłączenia AP. Model porównano z innymi podejściami, stosowanymi w przypadku braków odczytów z AP. Rys. 14 porównuje przyrost błędów w sytuacji gdy nie zastosuje się metod zapobiegających (No Action), odzyskuje brakujące wartości za pomocą średniej (Mean Imputation), wykrywa brakujące sygnały w celu wyeliminowania ich źródła z modelu (List-Wise Deletion) oraz dla proponowanego systemu wykrywania zmian (Our Solution).



Rysunek 14: Porównanie przyrostów błędów otrzymanych dla różnych metod zaradczych dla stosowanych w przypadku braków odczytów z AP. Źródło [A2].

Wszystkie metody są skuteczne i uzyskują lepszy wynik niż ignorowanie zmian. Jednak najlepsze wyniki uzyskuje zaproponowany system. W przypadku zaproponowanego rozwiązania oraz modelu List-Wise Deletion obserwujemy pewne ujemne wartości. Oznacza to, że zaktualizowany system daje w niektórych przypadkach lepsze wyniki niż system przed zmianą, a przyrost błędu jest ujemny.

Estymacja zagęszczenia w oparciu o sygnały GSM pochodzące z crowdsourcingu

M. Luckner, A. Rosłan, I. Krzemińska, J. Legierski i R. Kunicki, "Clustering of Mobile Subscriber's Location Statistics for Travel Demand Zones Diversity," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 40**, t. 10244 LNCS, 2017, s. 315–326, ISBN: 978-3-319-59104-9

W pracy [A4] przedstawiliśmy metodę analizy danych crowdsourcingowych, opisujących przybliżoną, poprzez przypisanie do stacji bazowej, lokalizację abonentów telefonii komórkowej. Obserwując dzienny rozkład zdarzeń zarejestrowanych w stacji bazowej można utworzyć charakterystykę obszaru objętego sygnałem BTS. Podobne charakterystyki są grupowane, tworząc obszar wspólnej charakterystyki dobowej. Obszary te reprezentują różne zagęszczenie dobowe użytkowników sieci komórkowej i mogą być wykorzystane do charakterystyki stref zapotrzebowania transportowego. Co ważne, metoda chroni prywatność użytkowników sieci. Dane zbierane są wyłącznie w postaci statystyki, czyli sumy zdarzeń dla danej komórki (ang. *Cell*) sieci komórkowej i nie ma możliwości odwrócenia tego procesu. Przetwarzane dane nigdy nie pochodzą bezpośrednio od poszczególnych, prywatnych, terminali. Statystyki opisują raczej obciążenie BTS-a niż jakiekolwiek cechy poszczególnych użytkowników (co miało miejsce w pracach [A8, A10]).

W ramach pracy [A4] zaproponowałem sposoby przekształcania danych ze stacji bazowych i ich transformacji do agregatu obejmującego dany obszar. Każda stacja bazowa BTS_i jest określona przez jej położenie (szerokość BTS_i^{lat} i długość BTS^{long} geograficzną), oraz szacowany zasięg wyrażony jako promień okręgu BTS_i^r. Liczba zdarzeń, zarejestrowanych przez stację, jest agregacją dla pewnego okresu czasu (w tym wypadku godziny) i jest określona jako $BTS_i^e(t)$. Wektor $BTS_i^e(t)$ dla $t \in [1, \ldots, 24]$ określa dobowy rozkład zdarzeń zarejestrowanych przez stację BTS_i.

Posiadając zbiór zawierający nienormalizowane wartości bezwzględne zdarzeń $BTS_i^e(t)$, zaproponowa-

łem dwie metody ich normalizacji. Pierwsza normalizacja uwzględnia zasięg stacji bazowej posługując się przekształceniem:

$$|BTS_i^e(t)| = \frac{BTS_i^e(t)}{\pi(BTS_i^r)^2} \tag{15}$$

Druga normalizacja uwzględnia dzienny rozkład danych. Normalizacja może być zastosowana do surowych jak i już znormalizowanych danych (względem obszaru). Przekształca dane do zakresu $[0, 1]$ zgodnie z formułami:

$$\begin{aligned} \overline{BTS_i^e(t)} &= \frac{BTS_i^e(t)}{\max_{t=1}^{24}(BTS_i^e(t))} \\ \overline{|BTS_i^e(t)|} &= \frac{|BTS_i^e(t)|}{\max_{t=1}^{24}(|BTS_i^e(t)|)} \end{aligned} \tag{16}$$

Ostatnia modyfikacja to powiązanie danych ze stacji bazowych (surowych lub znormalizowanych) z wyznaczonymi obszarami. Każdy obszar Z_i jest określony przez wielokąt Z_i^P.

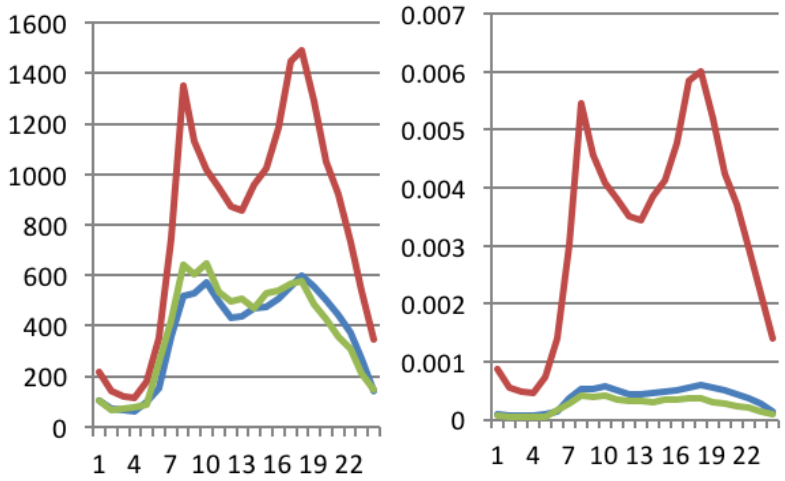
Zdefiniujmy funkcję przynależności:

$$\mu(i,j) = \begin{cases} 1 & \text{jeżeli } (BTS_j^{lat}, BTS_j^{long}) \in Z_i^P \\ 0 & \text{w.p.p.} \end{cases}, \tag{17}$$

wtedy wektor dobowych zdarzeń w obszarze Z_i jest określany jako:

$$\vec{Z_i} = \sum_{j=1}^{|\mathbb{BTS}|} \mu(i,j)\vec{BTS_j}. \tag{18}$$

W pracy porównano możliwości wykorzystania danych znormalizowanych na różne sposoby. Przykładowo na Rys. 15 porównano wyniki normalizacji dla danych z trzech BTS-ów.



Rysunek 15: Charakterystyki przykładowych BTS-ów. Po lewej dane surowe, po prawej dane znormalizowane względem zasięgu stacji bazowej. Źródło [A4].

Uzyskane wykresy dobowej aktywności są rozwiązaniem problemu estymacji zagęszczenia. Widać na nich, że normalizacja może prowadzić do spłaszczenia funkcji dobowej aktywności, co utrudnia rozróżnienie obszarów o różnej charakterystyce i może ograniczyć możliwości analizy przestrzenno-czasowej danych.

Zastosowanymi obszarami agregacji były strefy zapotrzebowania transportowego dla Miasta Stołecznego Warszawy, a dane, po agregacji, posłużyły jako wejście do sieci Kohonena, aby odnaleźć obszary o podobnej dziennej aktywności użytkowników sieci komórkowej. Analizy uzyskanych wyników wykraczają poza zakres tego autoreferatu. Natomiast opracowane metody były kluczowe dla uzyskania wyników zaprezentowanych w pracy [A1].

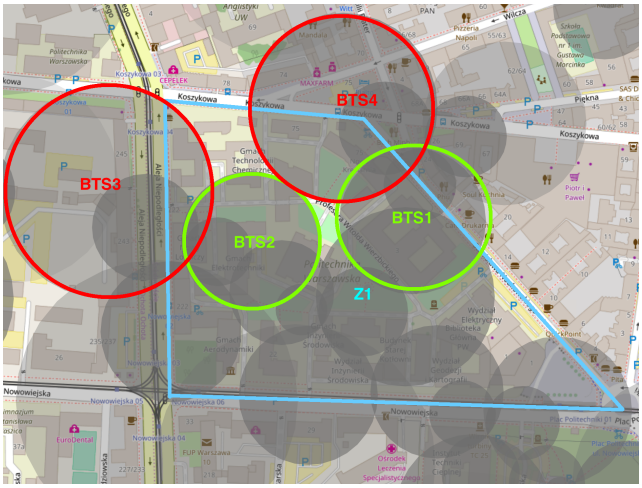
M. Luckner, I. Krzemińska, P. Wawrzyniak i J. Legierski, "Estimating Population Density Without Contravening Citizen's Privacy: Warsaw Use Case," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, t. 52, nr. 7, s. 4494–4506, 2022, **punkty MEiN: 200, IF: 11.471**

W pracy [A1] zaproponowaliśmy wykorzystanie do analiz przestrzennych danych poddanych uprzednio agregacji w poszczególnych komórkach publicznej naziemnej sieci komórkowej. Dzięki temu, w analizach, nie dochodzi do śledzenia pojedynczego urządzenia mobilnego i nie jest naruszane prawa użytkowników do prywatności.

Aby udowodnić przydatność proponowanej metody zbierania danych, porównaliśmy uzyskane wyniki z systemem telewizji przemysłowej służącym do szacowania liczby osób przebywających na obszarze centrum handlowego. Zaproponowany system jest wystarczająco czuły, aby wykryć nietypowe globalne wydarzenia w obszarze miejskim i wyróżnić strefy różnego typu.

Na potrzeby systemu, zaproponowałem modyfikację metody agregacji przedstawionej w pracy [A4], gdyż ma ona pewne wady. Ignorowany jest rzeczywisty zasięg BTS, przez co agregacja zachowuje się tak, jakby wszyscy abonenci byli w centrum obszaru zasięgu BTS.

Rys. 17 przedstawia obszar kampusu głównego Politechniki Warszawskiej. Kółka reprezentują przybliżone zasięgi BTS. Wykorzystując funkcję przynależności (17) ograniczamy liczbę zdarzeń przypisanych do strefy Z_1 do zdarzeń zarejestrowanych przez BTS-y zlokalizowane wewnątrz strefy np. BTS_1 i BTS_2. Metoda ta nie przypisuje jednak do obszaru zdarzeń pochodzących z BTSów BTS_3 i BTS_4.



Rysunek 16: Przykład zależności między analizowanym obszarem, a stacjami bazowymi. Źródło [A1].

W celu eliminacji tego problem, wprowadzamy parametr η_{ij}. Parametr opisuje stosunek zdarzeń, które powinny być przypisane do danego obszaru. Stosunek ten obliczany jest według powierzchni BTS_j ($BTS_j^s = \pi(BTS_j^r)^2$), która obejmuje obszar Z_i^P.

$$\eta_{ij} = \begin{cases} \frac{BTS_j^s \cap Z_i^P}{BTS_j^s} & BTS_j^s \cap Z_i^P \neq \varnothing \\ 0 & \text{w p.p.} \end{cases}. \tag{19}$$

Dodatkowo definiujemy parametr ρ_i opisujący gęstość zaludnienia na danym obszarze. Załóżmy, że liczba mieszkańców w obszarze Z_i jest dana jako Z_i^p. Obliczamy lokalną gęstość ρ_{ij} dla obszaru Z_i

wykorzystując również dane z pozostałych stref objętych zasięgiem BTS_j:

$$\rho_{ij} = \frac{Z_i^p}{\sum_{k=1}^{|\mathbb{Z}|} \delta(\eta_{kj}) Z_k^p}, \tag{20}$$

gdzie

$$\delta(x) = \begin{cases} 1 & x > 0 \\ 0 & \text{w p.p.} \end{cases}. \tag{21}$$

W przypadku braku danych o liczbie ludności w danym obszarze, parametr ρ_{ij}przyjmuje wartość jeden. Teraz, wykorzystując (19) i (20), definiujemy wektor zdarzeń dziennych dla strefy Z_i jako:

$$\vec{Z}_i = \sum_{j=1}^{|\mathbb{BTS}|} \rho_{ij} \eta_{ij} B\vec{T}S_j. \tag{22}$$

Rys. 17 porównuje dwa podejścia do agregacji zdarzeń wewnątrz obszarów. Pierwsze podejście (włączenie) wykorzystuje formułę (18). Drugie podejście (procentowe) wykorzystuje formułę (22) z $\rho_{ij} = 1$. Do testów wykorzystano strefy zapotrzebowania transportowego. Prezentowane dane zostały zebrane w ciągu 12 dni.

Rysunek 17: Porównanie dwóch podejść do agregacji danych sieci komórkowych wewnątrz obszaru. Rozkład zdarzeń w poszczególnych strefach. Źródło [A1].

Włączenie pokazuje więcej zdarzeń w strefach niż rozkład procentowy. Średnia liczba zdarzeń dla podejścia procentowego sięga 273 tysięcy, gdy średnia dla podejścia włączającego przekracza 291 tysięcy. Wynika to z tego, że nie każda strefa w mieście zawiera stację BTS. Gdy podejście procentowe rozdziela zdarzenia pomiędzy 895 stref z 896 stref w mieście, podejście włączające obejmuje tylko 843 strefy. Stąd, wskazane jest stosowanie podejście procentowe, zwłaszcza dla mniejszych obszarów agregacji.

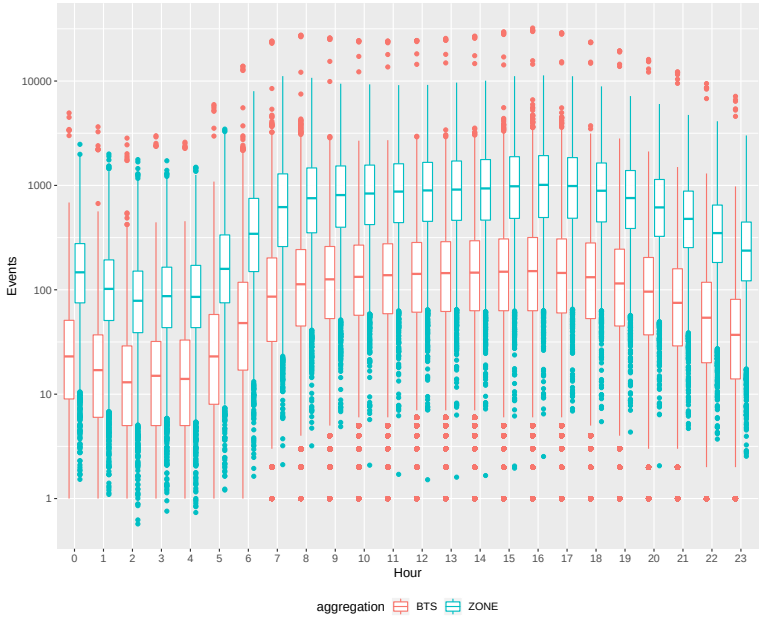
Dodatkowym aspektem agregacji jest wzrost prywatności. Załóżmy, że obserwujemy stacje bazowe o małym zasięgu na przedmieściach (podobnie jak to miało miejsce w pracy [A10]), gdzie występuje tylko jedno lub kilka urządzeń. Jeśli urządzenie, przed zmianą jego identyfikatora w systemie operatora, przeniesie się w zasięg innej stacji bazowej, to dwie lokalizacje tego urządzenia stworzą tworzyć trajektorię. Trajektoria będzie reprezentowała pewne wrażliwe informacje o mieszkańcu obszaru na którym występuje (np. kiedy opuszcza swoje mieszkanie). Z tego powodu, algorytm wyliczający statystyki dla BTS, eksponuje

wszystkie małe odczyty jako wartość dziesięć. Zagrożenie naruszenia prywatności przez małe odczyty może być też niwelowane poprzez agregację do obszaru.

Rys. 18 przedstawia rozkład minimalnej liczby zdarzeń w ciągu jednej godziny zagregowany według stacji bazowych i stref miejskich. Ze względu na duże różnice pomiędzy minimalną liczbą zarejestrowanych zdarzeń dla różnych stacji zastosowano skalę logarytmiczną. W związku z tym prezentowana liczba zdarzeń jest obliczana dla stacji i stref jako:

$$events_{BTS}=\log_{10}\left(\min_{\mathbb{BTS}}\left(BTS^{e}\right)\right), \tag{23}$$

$$events_{Zone}=\log_{10}\left(\min_{\mathbb{Z}}\left(Z\right)\right). \tag{24}$$



Rysunek 18: Rozkład minimalnej liczby agregowanych zdarzeń. Źródło [A1].

Dla agregacji danych na stacjach bazowych, małe odczyty (mniejsze niż 5) stanowią 6.8% wszystkich dobowych obserwacji. Gdy porównamy wyniki z agregacją przy użyciu stref, liczba małych odczytów gwałtownie spada (do 0.29%). Co więcej, dane te nie mogą być wykorzystane do stworzenia trajektorii. Zgodnie ze wzorem (22) małe odczyty uzyskaliśmy, gdy iloczyn liczby zdarzeń i procentowego pokrycia obszaru przez BTS jest mały. Dlatego też odczyty nie mogą być już interpretowane jako pojedyncze urządzenia, o czym dobitnie wskazują odczyty mniejsze od jedności obserwowane np. o 2 rano (Rys. 18).

Aby udowodnić, że dane po agregacji są nie tylko bezpieczne dla prywatności użytkowników, ale też użyteczne, zidentyfikowałem charakterystyczne sygnatury, które są związane z konkretnymi obiektami miejskimi: uniwersytetami, centrami handlowymi i węzłami transportowymi.

W pracy (Furno, Fiore, Stanica, Ziemlicki i Smoreda 2017) zidentyfikowano wiele uniwersalnych wzorców, powtrzających się w miastach z różnych krajów. Uzyskano je przetwarzając dane lokalizacyjne poszczególnych telefonów komórkowych. Powtórzenie tego eksperymentu dla danych zagregowanych będzie świadczyć o ich użyteczności w zaawansowanych analizach przestrzennych.

W przeprowadzonym eksperymencie wybrałem specjalne obszary obejmujące centra handlowe, stacje kolejowe i kampusy uniwersyteckie Miasta Stołecznego Warszawy (po dziesięć obiektów w każdej kategorii). Dla tych obszarów zbieraliśmy dane z jednego tygodnia (16.01.2017-22.01.2017).

Tab. 4 przedstawia informacje o obserwowanych obiektach. Dla przejrzystości prezentacji ograniczyłem ją do obiektów, w których odnotowano średnio co najmniej tysiąc zdarzeń na godzinę. Wyliczone statystyki

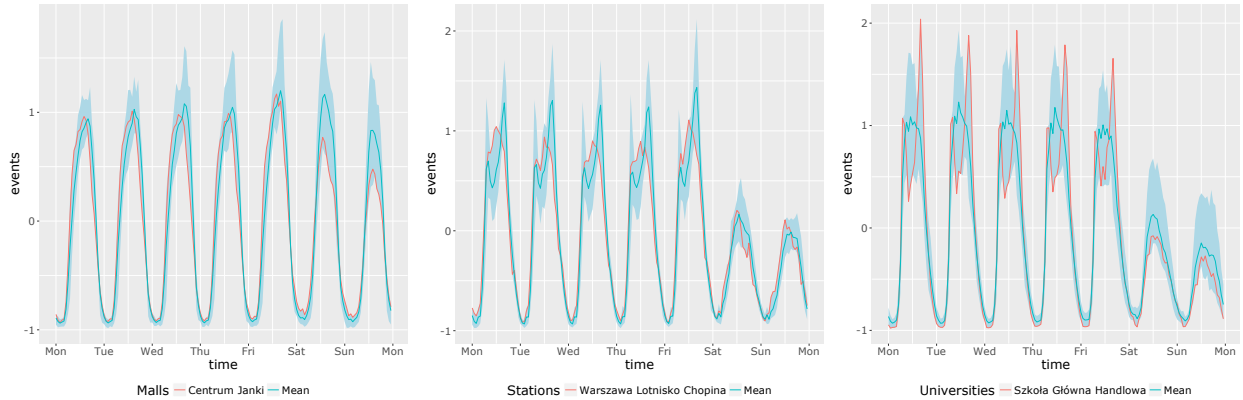
pokazują minimalne, maksymalne i średnie liczby zdarzeń obserwowanych w ciągu godziny.

Tabela 18: Liczba zdarzeń obserwowanych w ciągu godziny w wybranych strefach różnego typu. Przedstawiono minimalne, maksymalne i średnie liczby zdarzeń. W szczycie pokazano dzień tygodnia i godzinę z maksymalną liczbą zdarzeń. Źródło [A1].

lp	nazwa	typ	min	średnia	max	szczyt
1	Arkadia	Centrum handlowe	159	3753	9004	Sb. 14
2	Centrum Janki	Centrum handlowe	260	3969	8606	Pt. 14
3	Galeria Mokotów	Centrum handlowe	122	3091	7192	Wt. 15
4	Blue City	Centrum handlowe	161	2716	6012	Śr. 15
5	Złote Tarasy	Centrum handlowe	29	1986	5674	Pt. 17
6	Atrium Targówek	Centrum handlowe	56	1224	3059	Sb. 13
7	Wola Park	Centrum handlowe	70	1059	2899	Sb. 14
8	Atrium Promenada	Centrum handlowe	63	1225	2795	Sb. 16
9	Warszawa Centralna	Stacja kolejowa	147	3163	10277	Pt. 16
10	Warszawa Lotnisko Chopina	Stacja kolejowa	377	4448	9378	Pt. 12
11	Warszawa Wileńska	Stacja kolejowa	37	1147	3163	Pt. 16
12	Warszawa Zachodnia	Stacja kolejowa	70	1155	3001	Pt. 15
13	Warszawa Śródmieście	Stacja kolejowa	83	1215	2919	Pt. 16
14	Warszawa Gdańska	Stacja kolejowa	22	1006	2891	Śr. 16
15	Warszawa Wschodnia	Stacja kolejowa	79	1004	2270	Pt. 15
16	Uniwersytet Medyczny	Uczelnia	82	1722	4700	Wt. 11
17	Szkoła Główna Handlowa	Uczelnia	20	1249	3795	Pon. 16
18	Politechnika Warszawska	Uczelnia	41	1291	3780	Wt. 11
19	Akademia Sztuki Wojennej	Uczelnia	181	1903	3526	Czw. 11
20	Wojskowa Akademia Techniczna	Uczelnia	127	1476	3395	Czw. 11

W centrach handlowych zgrupowano najwięcej zdarzeń, jednak dwoma obiektami o największej liczbie zaobserwowanych zdarzeń były dworce kolejowe. Dodatkowo, przedstawiono dzień tygodnia i godzinę szczytu liczby zdarzeń. Dla większości stacji szczyt można zaobserwować w piątek po południu, kiedy osoby pracujące lub studiujące w Warszawie opuszczają miasto. Dla większości uczelni szczyt obserwowany jest w godzinach porannych, kiedy studenci powinni być na uczelniach. Wreszcie, najpopularniejszymi dniami odwiedzin większości centrów handlowych są weekendy.

Na Rys. 19 porównano tygodniowy rozkład zdarzeń dla obiektów różnych typów. Ze względu na różnice w wielkości zdarzeń dokonano standaryzacji danych.



Rysunek 19: Zestandaryzowane profile obiektów. Od lewej: centra handlowe, stacje kolejowe, uczelnie wyższe. Źródło [A1].

Galerie handlowe mają podobny profil dla dni pracujących, ze szczytem w godzinach popołudniowych oraz szczyt poranny w weekendy. W przypadku dworców obserwujemy dwa szczyty w dni robocze i znaczne zmniejszenie liczby wydarzeń w weekend. Wreszcie uczelnie mają najbardziej chaotyczny profil zdarzeń dziennych, z podobnym rozkładem zdarzeń w dni robocze. Tak jak dla stacji, obserwuje się redukcję liczby zdarzeń w czasie weekendu.

Dla porównania na wszystkich rysunkach uwzględniono obserwacje dla obiektów nietypowych. Na przykład obszar obejmujący Centrum Janki zawiera także część drogi wylotowej z miasta. Stacja kolejowa Warszawa Lotnisko Chopina jest stacją podziemną, która leży pod lotniskiem i pełni inną rolę niż pozostałe stacje. Podobnie uczelnia Szkoła Główna Handlowa znajduje się w pobliżu stacji metra, co może wpływać na jej charakterystykę.

Podsumowując, agregacja danych z urządzeń mobilnych może być wykorzystana do zdefiniowania profili dla różnych typów obiektów podobnie jak w pracy (Furno, Fiore, Stanica, Ziemlicki i Smoreda 2017). Uzyskano wizualnie różne tygodniowe sygnatury dla uczelni wyższych, centrów handlowych i dworców kolejowych. Ponadto wykryto, kiedy te obiekty były najczęściej odwiedzane. Wyniki analiz można wykorzystać w celach urbanistycznych np. aby dostosować transport publiczny do lokalnych potrzeb.

Podsumowanie

W ramach podsumowania, podkreślę najważniejsze wyniki zawarte w cyklu.

- Opracowanie algorytmu do wyznaczania lokalizacji wewnątrz budynków na podstawie sygnałów GSM z dokładnością około pięciu metrów [A8].
- Budowa lokalizatorów działających wewnątrz budynków, które na podstawie sygnałów Wi-Fi, osiągają błąd średni poniżej trzech metrów i rozpoznają piętro ze skutecznością 93 procent [A5, A7].
- Opracowanie algorytmu lokalizacyjnego, odpornego na zaniki sygnałów Wi-Fi i zmiany w infrastrukturze sieci [A2, A3, A11].
- Opracowanie algorytmu przetwarzającego dane crowdsourcingowe, zbierane na stacjach bazowych sieci GSM, do postaci pozwalającej na analizę zagęszczenia na zadanym obszarze, bez naruszania prywatności użytkowników sieci [A1, A4].
- Opracowanie algorytmu selekcji punktów dostępowych, do budowy modelu lokalizacyjnego, aby zwiększyć wydajność modelu bez przekraczania tolerancji błędu [A3].
- Opracowanie algorytmu wykrywania zmian w infrastrukturze Wi-Fi [A2].
- Opracowanie algorytmu do wykrywania bieżącego piętra, w zadaniu śledzenia trasy, na podstawie sygnałów GSM, osiągającego wynik o 40 procent lepszy niż model regresyjny [A10].
- Wykazanie, doświadczalnie, że wdrożone modele lokalizacyjne starzeją się w niskim stopniu, charakteryzują się niskim kosztem obliczeniowym, i mogą być zasilane danymi z różnych typów urządzeń [A3, A6].
- Wykazanie, doświadczalnie, że możliwe jest wdrożenie modelu lokalizacyjnego, ograniczającego się do detekcji piętra, przy użyciu w słabo rozwiniętej infrastruktury sieciowej i nisko-kosztowego sposobu zbierania danych [A9].
- Wykazanie, doświadczalnie, że wdrożone algorytmy agregacji danych z crowdsourcingu zapewniają prywatność użytkowników sieci komórkowej [A1].
- Wykonanie analizy dziennego zatłoczenia głównych budynków i kompleksów budynków użyteczności publicznej w Warszawie, wykorzystując przybliżone dane crowdsourcingowe [A1].

Omówienie pozostałych publikacji naukowo-badawczych

Dorobek po uzyskaniu stopnia doktora

[B1] A. Guerra-Manzanares, H. Bahsi i M. Luckner, "Leveraging the first line of defense: a study on the evolution and usage of android security permissions for enhanced android malware detection," *Journal of Computer Virology and Hacking Techniques*, 2022, **punkty MEiN: 70**.

[B2] A. Guerra-Manzanares, M. Luckner i H. Bahsi, "Android malware concept drift using system calls: Detection, characterization and challenges," *Expert Systems with Applications*, t. 206, s. 117 200, 2022, **punkty MEiN: 140, IF: 8.665**, ISSN: 0957-4174.

[B3] A. Guerra-Manzanares, M. Luckner i H. Bahsi, "Concept drift and cross-device behavior: Challenges and implications for effective android malware detection," *Computers & Security*, t. 120, s. 102 757, 2022, **punkty MEiN: 140, IF: 5.105**, ISSN: 0167-4048.

[B4] P. Wrona, M. Grzenda i M. Luckner, "Streaming Detection of Significant Delay Changes in Public Transport Systems," Clélia i in., red., **punkty MEiN: 140**, Springer International Publishing, 2022, s. 486–499, ISBN: 978-3-031-08760-8.

[B5] M. Wachulec i M. Luckner, "Fault detection of jet engine heat sensor," *Procedia Computer Science*, t. 192, s. 844–852, 2021, **punkty MEiN: 70**, ISSN: 18770509.

[B6] M. Luckner, M. Grzenda, R. Kunicki i J. Legierski, "IoT Architecture for Urban Data-Centric Services and Applications," *ACM Transactions on Internet Technology*, t. 20, nr. 3, s. 1–30, 2020, **punkty MEiN: 140, IF: 3.75**, ISSN: 1533-5399.

[B7] M. Bukowski, M. Luckner i R. Kunicki, "Estimation of Free Space on Car Park Using Computer Vision Algorithms," *Advances in Intelligent Systems and Computing*, t. 920, s. 316–325, 2019, ISSN: 21945357.

[B8] M. Luckner, "Practical web spam lifelong machine learning system with automatic adjustment to current lifecycle phase," *Security and Communication Networks*, t. 2019, 2019, **punkty MEiN 40, IF: 1.968**, ISSN: 19390122.

[B9] M. Luckner, M. Gad i P. Sobkowiak, "Antyscam-Practical web spam classifier," *International Journal of Electronics and Telecommunications*, t. 65, nr. 4, s. 713–722, 2019, **punkty MEiN: 40**, ISSN: 23001933.

[B10] A. Wilkowski, I. Mykhalevych i M. Luckner, "City Bus Monitoring Supported by Computer Vision and Machine Learning Algorithms," *Advances in Intelligent Systems and Computing*, t. 920, s. 326–336, 2019, ISSN: 21945357.

[B11] K. Breński, M. Chołuj i M. Luckner, "Evil-AP - Mobile man-in-the-middle threat," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 40**, t. 10244 LNCS, 2017, s. 617–627, ISBN: 9783319591049.

[B12] M. Luckner i J. Karwowski, "Estimation of Delays for Individual Trams to Monitor Issues in Public Transport Infrastructure," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 10448 LNAI, 2017, s. 518–527, ISBN: 9783319670737.

[B13] M. Luckner, P. Kobojek i P. Zawistowski, "Public transport stops state detection and propagation: Warsaw use case," w *SMARTGREENS 2017 - Proceedings of the 6th International Conference on Smart Cities and Green ICT Systems*, 2017, s. 235–241, ISBN: 9789897582417.

[B14] W. Homenda, M. Luckner i W. Pedrycz, "Classification with rejection: Concepts and evaluations," w *Advances in Intelligent Systems and Computing*, t. 364, 2016, s. 413–425, ISBN: 9783319190891.

[B15] A. Wilkowski i M. Luckner, "Low-cost canoe counting system for application in a natural environment," w *Advances in Intelligent Systems and Computing*, t. 440, 2016, s. 705–715, ISBN: 9783319293561.

[B16] M. Luckner, "Conversion of decision tree into deterministic finite automaton for high accuracy online SYN flood detection," w *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, **punkty MEiN: 20**, 2015, s. 75–82, ISBN: 9781479975600.

[B17] R. Filasiak, M. Grzenda, M. Luckner i P. Zawistowski, "On the testing of network cyber threat detection methods on spam example," English, *Annales des Telecommunications/Annals of Telecommunications*, t. 69, nr. 7-8, s. 363–377, 2014, **punkty MEiN: 40, IF: 1.901**, ISSN: 19589395.

[B18] W. Homenda i M. Luckner, "Pattern recognition with rejection: Application to handwritten digits," w *2014 4th World Congress on Information and Communication Technologies (WICT 2014)*, IEEE, grud. 2014, s. 326–331, ISBN: 978-1-4799-8115-1.

[B19] W. Homenda, M. Luckner i W. Pedrycz, "Classification with rejection based on various SVM techniques," w *Proceedings of the International Joint Conference on Neural Networks*, **punkty MEiN: 140**, 2014, s. 3480–3487, ISBN: 9781479914845.

[B20] M. Luckner, "Global and local rejection option in multi-classification task," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 70**, t. 8681 LNCS, 2014, s. 483–490, ISBN: 9783319111780.

[B21] M. Luckner, M. Gad i P. Sobkowiak, "Stable web spam detection using features based on lexical items," *Computers & Security*, t. 46, s. 79–93, 2014, **punkty MEiN: 140, IF: 5.105**, ISSN: 01674048.

[B22] K. Rzążewska i M. Luckner, "3D model reconstruction and evaluation using a collection of points extracted from the series of photographs," w *2014 Federated Conference on Computer Science and Information Systems, FedCSIS 2014*, wrz. 2014, s. 669–677, ISBN: 9788360810583.

[B23] J. Furtak i in., "Frontiers in network applications, network systems and web services," w *2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, 2013, ISBN: 9781467344715.

[B24] W. Homenda, M. Luckner i W. Pedrycz, "Classification with rejection : concepts and formal evaluations," w *8th International Conference on Knowledge, Information and Creativity Support*, Kraków, 2013, s. 161–172, ISBN: 9781479914845.

[B25] M. Luckner i R. Filasiak, "Flow-level Spam Modelling using separate data sources," w *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, IEEE, 2013, s. 91–98.

[B26] M. Luckner i R. Filasiak, "Reference data sets for spam detection: Creation, analysis, propagation," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 20**, t. 8073 LNAI, 2013, s. 212–221, ISBN: 9783642408458.

[B27] M. Luckner i K. Szyszko, "RBF ensemble based on reduction of DAG structure," w *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, Kraków: IEEE, 2013, s. 99–105.

[B28] J. Rudziński i M. Luckner, "Low-cost computer vision based automatic scoring of shooting targets," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, t. 7828 LNAI, 2013, s. 185–195, ISBN: 9783642373428.

[B29] A. Sroka i M. Luckner, "Tree Symbols Detection for Green Space Estimation," w *Advanced Concepts for Intelligent Vision Systems, Acivs 2013*, **punkty MEiN: 70**, 2013, s. 526–537.

[B30] G. Bagrowski i M. Luckner, "Comparison of Corner Detectors for Revolving Objects Matching Task," w *Artificial Intelligence and Soft Computing Lecture Notes in Computer Science*, **punkty MEiN: 20**, Springer Berlin Heidelberg, 2012, s. 459–467.

[B31] M. Luckner, "Problem eliminacji nieprzystających elementów w zadaniu rozpoznania wzorca," w *Zastosowania metod statystycznych w badaniach naukowych IV*, StatSoft, 2012, s. 283–294.

[B32] M. Luckner i W. Izdebski, "Publication of Geodetic Documentation Center Resources on Internet," w *Advanced Information Systems Engineering Lecture Notes in Computer Science Volume 7328*, **punkty MEiN: 140**, Springer Berlin Heidelberg, 2012, s. 533–548.

[B33] J. Rudziński i M. Luckner, "Automatic scoring of shooting targets with tournament precision," w *Frontiers in Artificial Intelligence and Applications*, t. 243, 2012, s. 324–334, ISBN: 9781614991045.

[B34] M. Grzenda, K. Kaczmarski, M. Kobos i M. Luckner, "Geospatial presentation of purchase transactions data," w *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2011, s. 291–296, ISBN: 978-83-60810-35-4.

[B35] M. Luckner, "Multiclass SVM classification using graphs calibrated by similarity between classes," w *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **punkty MEiN: 70**, t. 6884 LNAI, 2011, s. 435–444, ISBN: 9783642238659.

[B36] M. Luckner, "Reducing Number of Classifiers in DAGSVM Based on Class Similarity," w *Image Analysis and Processing – ICIAP 2011 Lecture Notes in Computer Science*, **punkty MEiN: 70**, Springer Berlin Heidelberg, 2011, s. 514–523.

Dorobek po uzyskaniu stopnia doktora, z wyłączeniem cyklu obejmującego osiągnięcie, obejmuje 36 pozycji [B1–B36]. Suma punktów dorobku po doktoracie, wyliczonych zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 9 lutego 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych, wynosi **1490**. Dorobek obejmuje osiem publikacji w czasopismach [B1–B3, B6, B8, B9, B17, B21], których Sumaryczny Impact Factor został wyliczony na podstawie najnowszego wykazu Journal Citation Reports z 2021 roku i wynosi **26.494**. W ramach dorobku można wyodrębnić kilka obszarów badawczych, omówionych pokrótce w kolejnych sekcjach.

Zagadnienie klasyfikacji

Prace zgrupowane w tej sekcji podejmują zagadnienia algorytmiczne i skupiają się na modyfikacji i optymalizacji klasyfikatorów. Część proc dotyczy zagadnienia optymalizacji budowy zestawów klasyfikatorów binarnych do rozwiązywania problemów wieloklasowych, część podejmuje zagadnienie odrzucania w problemie rozpoznawania wzorca.

Prace [B27, B35, B36] dotyczą wykorzystania struktur drzewiastych (acyklicznych grafów skierowanych) oraz funkcji podobieństwa do optymalizacji struktury zestawu klasyfikatorów binarnych, stosowanej do rozwiązywania problemów wieloklasowych.

W pracy [B35] przedstawiono nowe struktury uczące, oparte na drzewach podobieństwa między klasami i grafie acyklicznym skierowanym. Proponowane struktury oparte są na rozkładzie rozpoznanych klas w przestrzeni danych. Struktury tworzone są poprzez grupowanie podobnych klas. Podobieństwo między klasami jest szacowane przez odległość między klasami. Strategia Jeden-Kontra-Jeden (JKJ) jest realizowana dla klas najbliższych. W pozostałych przypadkach stosowana jest strategia Jeden-Kontra-Wszyscy (JKW). Metoda ta pozwala na redukcję czasu klasyfikacji bez znaczącego wzrostu błędu klasyfikacji.

Praca [B36] analizuje użycie maszyn wektorów nośnych w przypadku problemów klasyfikacji wieloklasowej. Poszczególne klasyfikatory mogą być zebrane w strukturę skierowanego grafu acyklicznego DAGSVM. Struktura taka implementuje strategię JKJ. W tej strategii dla każdej pary klas tworzony jest podział. Jednak, ze względu na hierarchiczną strukturę, tylko część z klasyfikatorów jest wykorzystywana w pojedynczym procesie klasyfikacji. Liczba klasyfikatorów może zostać zmniejszona, jeśli ich zadania klasyfikacyjne zostaną zmienione z rozdzielania pojedynczych klas na rozdzielanie grup klas. Jak w poprzedniej pracy, proponowana metoda opiera się na podobieństwie klas. Dla klas bliskich struktura DAG pozostaje niezmienna. Dla klas odległych więcej niż jedna jest rozdzielana za pomocą jednego klasyfikatora, co zmniejsza koszt klasyfikacji.

W pracy [B27] zaproponowana struktura grafu skierowanego jest testowana w wariantach opartych na różnych metrykach podobieństwa. Do testów wykorzystano zbiory danych z repozytorium UCI, a wyniki porównano z opublikowanymi pracami. Testy dowiodły, że grupowanie radialnych funkcji bazowych, w zaproponowanej strukturze, zmniejsza koszt klasyfikacji, a dokładność rozpoznania nie ulega znacznemu zmniejszeniu.

Podczas gdy praca [B31] podsumowuje, rozpatrywany w mojej rozprawie doktorskiej, problem eliminacji nieprzystających elementów w zadaniu rozpoznania wzorca, kolejne prace [B14, B18–B20, B24] rozwijają ten koncept, częściowo w ramach realizacji projektu NCN *Zagadnienie odrzucania w problemie rozpoznawania wzorca*.

Klasyfikacja, standardowo, przyporządkowuje wszystkie przetwarzane elementy do znanych klas. Zakładamy, że istnieją tylko elementy rodzime i nie występują żadne obce, czyli wszystkie przetwarzane elementy są zaliczane do znanych klas. Jakość klasyfikacji można mierzyć dwoma czynnikami: liczbą elementów prawidłowo i błędnie rozpoznanych. W opisanym podejściu, dopuszczenie elementów obcych zwiększa liczbę błędnie zakwalifikowanych elementów i tym samym pogarsza jakość klasyfikacji. W tym kontekście pożądane jest odrzucanie elementów obcych, czyli nieprzypisywanie ich do żadnej ze znanych klas.

Odrzucanie elementów obcych zmniejszy liczbę elementów błędnie zakwalifikowanych, ale może również odrzucić elementy rodzime zmniejszając, jako efekt uboczny, skuteczność klasyfikatora. Dlatego ważne jest zbudowanie dobrze zaprojektowanych metod, które odrzucą znaczną część obcych i tylko nieliczną rodzimych elementów.

W pracach [B14, B24] przedstawiono koncepcje oceny klasyfikacji z odrzucaniem. Przedstawiono trzy główne modele: klasyfikację bez odrzucania, klasyfikację z odrzucaniem oraz klasyfikację z reklasyfikacją. Koncepcje te są zilustrowane elastycznymi zespołami klasyfikatorów binarnych wraz z teoretycznymi ocenami każdego z modeli. Proponowane modele mogą być stosowane w szczególności jako klasyfikatory pracujące z danymi zaszumionymi, gdzie rozpoznawane dane wejściowe nie są ograniczone do elementów znanych klas.

W pracy [B19] zastosowano zespoły maszyn wektorów wspierających rozwiązujące problemy dwuklasowe i jednoklasowe jako narzędzia klasyfikacyjne oraz jako podstawowe narzędzia do odrzucania elementów obcych. W artykule zaproponowano ocenę jakości metod klasyfikacji i odrzucania, a na koniec przeprowadzono eksperymenty w celu zilustrowania zaproponowanych pojęć i metod. W pracy [B18] przeanalizowano kilka prób odrzucenia elementów obcych w odniesieniu do rozpoznawania pisma ręcznego.

W pracy [B20] przedstawiono dwie opcje odrzucania. Opcja globalnego odrzucania oddziela obserwacje obce - niezdefiniowane w zadaniu klasyfikacyjnym - od obserwacji rodzimych. Opcja lokalnego odrzucania działa po procesie klasyfikacji i oddziela obserwacje indywidualnie dla każdej klasy. Przedstawiamy implementację obu metod dla klasyfikatorów binarnych zgrupowanych w strukturze grafu (drzewo lub skierowany graf acykliczny).

Cyberbezpieczeństwo

Zakres prac dotyczących cyberbezpieczeństwa, przedstawionych w tej sekcji, obejmuje analizę cech niskopoziomowych do wykrywania spamu i ataków DDoS, wykrywania web-spamu i ochrony urządzeń mobilnych z systemem Android.

Pierwsze cztery prace [B16, B17, B25, B26] dotyczą analizy danych niskopoziomowych (rekordów NetFlow) do analizy zagrożeń w ruchu sieciowym. Jedną z głównych przeszkód utrudniających rozwój i ocenę takich metod jest brak referencyjnych zbiorów danych.

W pracach [B17, B26] zaproponowano sposób testowania metod wykrywających zagrożenia sieciowe. Obejmuje on procedurę tworzenia realistycznych zbiorów danych referencyjnych opisujących zagrożenia sieciowe oraz przetwarzanie i wykorzystanie tych zbiorów danych w środowiskach testowych. Proponowane podejście zostało zilustrowane i ocenione na podstawie problemu detekcji spamu. Referencyjne zestawy danych są opracowywane, analizowane i wykorzystywane zarówno do generowania żądanego wolumenu symulowanego ruchu, jak i jego analizy z wykorzystaniem algorytmów uczenia maszynowego. Testy uwzględniają zarówno dokładność, jak i wydajność metod wykrywania zagrożeń w warunkach rzeczywistego obciążenia i ograniczonych zasobów obliczeniowych. Ponadto zaprezentowano hybrydowy klasyfikator, który wykrywa spam ze stosunkowo wysoką skutecznością.

W pracy [B25] przedstawiono niskopoziomowy model spamu. Model ten opisuje podklasy spamu i dostarcza informacji o głównych cechach zadania detekcji spamu. Model ten jest podstawą do budowy drzew

decyzyjnych wykrywających spam. W wyniku analizy detektorów, które zostały wyuczone na podstawie danych zebranych z różnych serwerów pocztowych, uzyskano uniwersalny opis spamu składający się z najistotniejszych cech. Przepływy opisane przez wybrane cechy i zebrane na serwerze Broadband Remote Access (BRAS) były analizowane przez zespół utworzonych klasyfikatorów. Zespół ten wykrył główne źródła spamu wśród adresów IP nadawców.

W pracy [B16] zaproponowano, jak przekształcić drzewo decyzyjne w deterministyczny automat skończony. Drzewo decyzyjne uczy się rozpoznawać zagrożenia wykorzystując zebrane dane. Stworzony w konsekwencji jego uczenia zbiór reguł decyzyjnych jest przekształcany w automat skończony, który może wykrywać zdarzenia zanim zostanie uzyskana pełna komplementarność danych. Metoda ta jest ograniczona do małych drzew, ale może rozwiązywać rzeczywiste problemy. Jako przykład przedstawiono wykrywanie ataku TCP SYN flood. Automat, stworzony dla tego zagadnienia, ma tak samo wysoki współczynnik dokładności jak drzewo decyzyjne, ale może podejmować decyzje ponad trzykrotnie szybciej.

Prace [B8, B9, B21] dotyczą metod i systemów wykrywania web-spamu. Web-spam to metoda manipulowania wynikami wyszukiwarek poprzez podnoszenie rangi stron spamowych. Przybiera różne formy i brakuje mu spójnej definicji. Detektory web-spamu wykorzystują techniki uczenia maszynowego do wykrywania spamu. Jednak ich weryfikacja odbywa się najczęściej na zbiorach danych pochodzących z tego samego okresu, co zbiory uczące. Doprowadza to do tendencyjności otrzymanych wyników.

W pracy [B21] porównaliśmy maszyny wektorów wspierających trenowane i testowane na zestawach danych WEBSPAM–UK z różnych lat. Aby uzyskać stabilne wyniki zaproponowaliśmy nowe cechy oparte na leksyce. Dokument HTML przekształcono w tekst bez znaczników HTML, zestaw widocznych symboli i listę linków, w tym tych ze znaczników. Następnie wyliczono cechy zawierające informacje o dziwnych kombinacjach liter, zbitkach spółgłosek, statystykach dotyczących sylab, słów i zdań oraz Gunning Fog Index. Używając danych zebranych w 2006 roku jako zbioru uczącego, uzyskaliśmy bardzo stabilną dokładność przy transferze rozwiązania pomiędzy latami. Wreszcie, udowodniliśmy, że równowaga pomiędzy czułością i specyficznością mierzoną przez miarę AUC jest poprawiona przez zaproponowane cechy.

Praca [B9] przedstawia system rozpoznawania web-spamu, który okresowo odświeża zbiór uczący, aby stworzyć adekwatny klasyfikator. Nowy klasyfikator jest uczony wyłącznie na danych z ostatniego okresu. W pracy pokazano, że strategia ta sprawdza się lepiej niż inkrementacja zbioru uczącego. Dodatkowo rozwiązano problem rozruchu systemu poprzez skorzystanie ze zbiorów zewnętrznych.

Standardem w detekcji spamu jest stosowanie uczenia maszynowego. Jednakże opieranie się na starych zbiorach treningowych skutkuje słabymi wynikami klasyfikacji. Dlatego potrzebne są stale uczące się systemy detekcji. W pracy [B8] zaproponowałem system do rozpoznawania web-spamu, który automatycznie przebudowuje zbiór uczący, aby uniknąć uczenia się na przestarzałych danych. Co więcej, system może, całkowicie automatycznie, budować zbiór uczący korzystając z zewnętrznych źródeł danych. Testy na prawdziwych danych z serwisów Quora, Reddit i Stack Overflow pokazały wysoką skuteczność systemu. Zarówno skuteczność jak i F1 wyniosły 0.98 i 0.96 odpowiednio dla semi-automatycznego i w pełni automatycznego systemu.

Kolejne prace skupiają się na zagrożeniach związanych z korzystaniem z urządzeń mobilnych, w szczególności obsługiwanych przez system operacyjny Android.

Praca [B11] przedstawia testy dotyczące przeprowadzania ataków man-in-the-middle na publiczne sieci internetowe. Stworzone narzędzie pozwala na przeprowadzenie ataku z telefonu komórkowego, aby później przejmować i przekierowywać ruch internetowy. Praca przedstawia także dobre praktyki pozwalające zapobiec tego typu zagrożeniom.

Prace [B1–B3] są poświęcone analizie złośliwego oprogramowania działającego na systemie Android.

W pracy [B2] zaproponowaliśmy nową metodę wykrywania i skutecznego rozwiązywania problemu dryfu pojęć w zadaniu wykrywaniu złośliwego oprogramowania. Proponowane rozwiązanie pozwala zachować wysoki poziom metryk jakości w długim okresie czasu i minimalizuje wysiłki związane z przekwalifikowaniem modelu.

W pracy [B3] oceniliśmy i porównaliśmy znaczniki czasowe stosowane do modelowania ewolucji

złośliwego oprogramowania. Nasze wyniki pokazują, że powszechnie stosowany w literaturze znacznik daje, po pewnym czasie, słabe wyniki i że lepsze wyniki są osiągane, gdy używany jest wewnętrzny znacznik czasu aplikacji. Dodatkowo, badanie rzuca światło na wykorzystanie różnych źródeł danych i ich wpływ na modelowanie. Stwierdziliśmy, że cechy dynamiczne uzyskane dla poszczególnych aplikacji z różnych źródeł danych (tj. emulatora i urządzenia rzeczywistego) wykazują znaczące różnice, które mogą zniekształcić wyniki modelowania. Dlatego należy unikać fuzji różnych typów podczas tworzenia zbiorów treningowych i testowych.

W pracy [B1] analizowaliśmy uprawnienia bezpieczeństwa (ang. *Security Permissions*) systemu Android w kontekście wykrywania złośliwego oprogramowania. Pokazujemy, że gdy dryf pojęć zostanie uwzględniony, uprawnienia mogą generować długotrwałe i skuteczne systemy wykrywania złośliwego oprogramowania. Ponadto, testując są zdolności dyskryminacyjne różnych zestawów cech. Stwierdziliśmy, że początkowy zestaw uprawnień, zdefiniowany w systemie Android 1.0 (poziom API 1), jest wystarczający do zbudowania skutecznego modelu detekcji, zapewniając średni wynik 0.93 F1 dla danych obejmujących siedem lat.

Przetwarzanie i analiza obrazu

W tej sekcji znajdują się zróżnicowane prace dotyczące przetwarzania i analizy obrazu, związane z widzeniem komputerowym (ang. *Computer Vision*). Zakres prac obejmuje analizę dokumentów drukowanych, zdjęć, wideo, a także automatyczne tworzenie chmur punktów z serii zdjęć.

Pierwsze dwie prace [B22, B30] dotyczą automatycznego tworzeniea chmur punktów z serii zdjęć. Proces ten wymaga sparowania ze sobą par zdjęć, co jest związane z wyszukiwaniem tożsamych punktów na różnych zdjęciach np. narożników analizowanych obiektów.

Praca [B30] zawiera test detektorów narożników stosowanych do znajdowania punktów charakterystycznych obracających się trójwymiarowych obiektów. Przedstawiono pięć różnych algorytmów, począwszy od historycznego detektora Moraveca, a skończywszy na ówcześnie najnowszych, takich jak SUSAN i Trajkovic. Ponieważ algorytmy są porównywane z punktu widzenia wykorzystania do modelowania 3D, porównanie dotyczy serii zdjęć i wymaga znalezienia projekcji punktu 3D na dwa lub trzy kolejne zdjęcia. Jakość algorytmów jest dyskutowana na podstawie zdolności do wykrywania narożników modelowanych obiektów, odporności na szumy, oraz koszu obliczeniowego.

Praca [B22] opisuje cały proces rekonstrukcji modelu 3D. Rozpoczyna się od przedstawienia metody, która służy do znalezienia dopasowania pomiędzy zdjęciami oraz metodologii wykorzystania danych do utworzenia wstępnej struktury zrekonstruowanego modelu, reprezentowanego przez chmurę punktów. Jako kolejny etap wykonywany jest proces rafinacji, wykorzystujący metodę dopasowania wiązek. W dalszej kolejności wykorzystywany jest zestaw metod stereowizji w celu znalezienia bardziej szczegółowego rozwiązania. Algorytmy te wykorzystują pary obrazów, dlatego badany jest zestaw procedur agregujących wyniki. Pracę kończy opis sposobu przetwarzania chmury punktów, w tym rekonstrukcji powierzchni, w celu utworzenia wyniku. Opisana metodologia została zilustrowana rekonstrukcjami trzech serii profesjonalnych zdjęć z publicznego repozytorium oraz jednej serii zdjęć amatorskich stworzonych specjalnie na potrzeby tej pracy. Wyniki zostały ocenione za pomocą zaproponowanych miar dopasowania powierzchniowego i dopasowania konturowego.

W pracach [B28, B33] przedstawiono algorytm automatycznego punktowania celów strzeleckich oparty na technikach widzenia komputerowego. W przeciwieństwie do profesjonalnych rozwiązań, proponowany system nie wymaga dodatkowego sprzętu i opiera się wyłącznie na prostych metodach przetwarzania obrazu, takich jak algorytm detekcji krawędzi Prewitta i transformacja Hougha. Błąd estymacji dla ponad 91 procent przestrzelin jest niższy niż próg punktacji turniejowej (do jednej dziesiątej pola). Dlatego system może być odpowiedni dla strzelców amatorów zainteresowanych turniejową dokładnością.

Kolejna praca [B29] dotyczy analizy geodezyjnych map zasadniczych, które są bardzo szczegółowym źródłem informacji. Jednak takie mapy są tworzone dla specjalistów i niezrozumiałe dla nieprofesjonalistów. Przykładem informacji, która może być przydatna dla obywatela są zmiany w miejskich terenach zielonych. Te, cenne dla lokalnej społeczności przestrzenie, mogą być niszczone przez deweloperów lub władze lokalne. Dlatego ważnym zadaniem jest monitoring terenów zielonych, który można prowadzić na podstawie map

z Ośrodków Dokumentacji Geodezyjnej. W pracy [B29] przedstawiono studium wykonalności estymacji terenów zielonych z zeskanowanych map. Rozwiązanie opiera się na detekcji symboli. Dwa rodzaje symboli (drzewa iglaste i liściaste) są rozpoznawane przez następujący algorytm. Wykrywane są kropki z centrów symboli i ekstrahowane jest ich sąsiedztwo. Określone cechy są obliczane jako wejście dla sieci neuronowych, które wykrywają symbole drzew. Dokładność detekcji wynosi 90 procent, co jest wystarczające do oszacowania obszaru terenów zielonych.

Praca [B15] przedstawia niskokosztowy system zliczania kajaków i kajakarzy w celu kontroli szlaków turystycznych. Stworzony system został zaimplementowany na Raspberry Pi 2, a całkowity koszt urządzenia śledzącego jest mniejszy niż 200 dolarów. Zaproponowany algorytm wykorzystuje odejmowanie tła i maszynę wektorów wspierających do śledzenia jednostek pływających i rozpoznawania wśród nich kajaków. Uzyskane wyniki są satysfakcjonujące jak na niskokosztowe rozwiązanie. W zależności od rozważanej grupy obiektów dokładność algorytmu osiąga 84, 89.5 i 96 procent odpowiednio dla kajaków, łodzi i pozostałych obiektów.

Kolejne dwie prace, z zakresu analizy obrazów w przestrzeni miejskiej, były realizowane w ramach europejskiego projektu badawczego *Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors (VaVeL)*.

W pracy [B7] zaproponowano system do monitoringu wolnych miejsc parkingowych używając widzenia maszynowego. Zastosowano trzy ogólnie znane metody przetwarzania obrazu i autorskie rozwiązanie łączące te podejścia poprzez perceptron wielowarstwowy. System osiągnął 95 procentową skuteczność na rzeczywistych danych.

W pracy [B10] opisano eksperyment dotyczący automatycznego rozpoznania autobusów komunikacji miejskiej na zapisie wideo. Problem polegał na wyodrębnieniu pojazdów z tła, a następnie na odróżnieniu autobusów miejskich od samochodów osobowych i innych autobusów. Osiągnięto skuteczność 85 procent w rozpoznawaniu autobusów przy 15 procentach błędnych rozpoznań innych pojazdów.

Analiza danych przestrzennych i miejskich

W tej sekcji znajdują się prace dotyczące analizy danych przestrzennych, jak i analizy danych wieloskalowych (ang. *Big Data*) pochodzących z miejskich systemów informacyjnych.

W pracy [B34] przedstawiono automatyczny system dla małych i średnich firm internetowych sprzedających towary. System łączy czasowe dane o sprzedaży z jej lokalizacją geograficzną i prezentuje uzyskane informacje na mapie. Takie podejście do prezentacji danych ma ułatwić zrozumienie struktury sprzedaży i może być pomocne w generowaniu pomysłów na poprawę strategii sprzedaży, a w konsekwencji na zwiększenie przychodów firmy. System może być dostosowany do przetwarzania i prezentacji danych w obrębie różnych poziomów podziału administracyjnego, z wykorzystaniem różnych hierarchii sprzedawanych towarów. Opisując system, przedstawiamy również jego prototyp, który w sposób interaktywny wizualizuje dane na trójwymiarowej mapie.

W pracy [B32] omówiono formę publikacji zasobów geodezyjnych i kartograficznych zgodną z dyrektywą Unii Europejskiej INSPIRE na podstawie aplikacji iGeoMap. Aplikacja internetowa iGeoMap, której byłem współautorem, może publikować dane przestrzenne z plików (tekstowych lub binarnych), baz danych wyspecjalizowanych w obsłudze danych przestrzennych (PostgreSQL, ORACLE) lub usług internetowych (Web Map Service, Web Feature Service). W ramach pracy omówiono wyszukiwanie najbardziej popularnych danych (działki, punkty adresowe, punkty kontrolne).

Kolejne prace, z zakresu analizy danych miejskich, były realizowane w ramach europejskiego projektu badawczego *Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors (VaVeL)*.

Praca [B12] przedstawia wyniki analiz strumieniowych danych opisujących lokalizację środków transportów publicznego w Warszawie. Dzięki zbudowanemu rozwiązaniu do analizy danych masowych można mierzyć aktualne opóźnienia poszczególnych autobusów i tramwajów, a dzięki opracowanym miarom także przewidywać globalne awarie w mieście.

Praca [B13] przedstawia system do obserwacji oficjalnych kanałów komunikacji dotyczących transportu publicznego (Twitter, strony internetowe). Przychodzące informacje są analizowane, aby uzyskać obraz aktualnego statusu infrastruktury komunikacji miejskiej. Uzyskane informacje są przetwarzane do postaci

danych przestrzennych, aby prezentować informacje w sposób niezależny językowo.

W pracy [B6] opisano architekturę miejskiego internetu rzeczy (ang. *Internet of Things*) ugruntowaną we wzorcach big data i ukierunkowaną na potrzeby miast i ich kluczowych interesariuszy. Zaproponowano architekturę platformy USE4IoT (Urban Service Environment for the Internet of Things), która gromadzi i przetwarza miejskie dane oraz rozszerza architekturę Lambda. Przedstawiono, jak platforma została wykorzystana do uczynienia IoT technologią wspomagającą inteligentne planowanie transportu. Ponadto zdefiniowano kluczowe komponenty przetwarzania danych niezbędne do zapewnienia wysokiej jakości strumieni danych IoT w czasie zbliżonym do rzeczywistego. Dodatkowo zamieszczono testy pokazujące, jak opisana platforma IoT zapewnia środowisko analityczne o niskim opóźnieniu dla inteligentnych miast.

Praca [B4] została zrealizowana w ramach europejskiego projektu badawczego *Co-designing Inclusive Mobility (CoMobility)*. W pracy zaproponowano zarówno metodę wykrywania znaczących opóźnień transportu publicznego, jak i architekturę referencyjną, bazującą na silnikach przetwarzania strumieniowego, w których metoda ta jest zaimplementowana. Zapewnia to zarówno identyfikację online istotnych i powtarzalnych opóźnień, jak i odporność na ograniczoną jakość danych lokalizacyjnych. Proponowana metoda może być wykorzystana z różnymi detektorami zmian, takimi jak ADWIN, zastosowanymi do strumienia danych lokalizacyjnych obserwowanych na poszczególnych krawędziach grafu transportowego. Pozwala ona w trybie on-line wykryć, na których krawędziach obserwowane są statystycznie istotne opóźnienia oraz na których krawędziach opóźnienia powstają i są redukowane. Detekcja taka może być wykorzystana do modelowania wyborów transportowych i ilościowego określenia wpływu regularnych, a nie losowych zakłóceń na planowane i odbyte podróże przy użyciu multimodalnych silników modelowania podróży.

Pozostałe prace

W zakresie nakreślonych obszarów badawczych nie mieści się praca [B5] przedstawiająca algorytm przewidujący awarię czujnika poziomu i temperatury oleju (OLTS) w celu jego wymiany zanim awaria pociągnie za sobą poważne koszty. Czujnik OLTS, wskazujący zbyt wysoką temperaturę oleju, jest przyczyną poważnych konsekwencji związanych z zawróceniem samolotu i nakazem wyłączenia silnika w locie. Przewidywanie awarii czujnika jest możliwe, ale operator potrzebuje co najmniej 11 miesięcy danych historycznych. Opracowany algorytm automatyzuje proces wykrywania potencjalnych awarii za pomocą opartego na danych modelu bazującego na niepodobieństwie. Oblicza on średnią kroczącą różnicy temperatur oleju pomiędzy silnikami siostrzanymi dla okresów krótkoterminowych i długoterminowych (liczonych w lotach). Jeżeli różnica między średnią krótkoterminową a długoterminową jest większa od ustalonego progu ogłaszany jest alarm. Proponowany model zmniejszył zakres danych wymagany do wykrycia usterki do trzech miesięcy.

Dodatkowo, w zakres wymienionych prac, wchodzi edytorial *Frontiers in network applications, network systems and web services* [B23] do materiałów konferencji *Federated Conference on Computer Science and Information Systems, FedCSIS 2013*.

Dorobek przed uzyskaniem stopnia doktora

[C1] M. Luckner, "Comparison of hierarchical SVM structures in letters recognition task," w *IEEE CIS-Poland Chapter Edited Volume*, Warsaw, Poland, 2008.

[C2] M. Luckner, "Distances Tree as SVM Ensemble in Digits Recognition Task," w *Proceedings of the 11th Joint Conference on Information Sciences*, 2008.

[C3] M. Luckner, "Recognition of Noised Patterns Using Non–Disruptive Learning Set," *Journal of Digital Information Management*, t. 5, nr. 3, 2007.

[C4] W. Homenda i M. Luckner, "Automatic Knowledge Acquisition: Recognizing Music Notation with Methods of Centroids and Classifications Trees," w *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, IEEE, 2006, s. 3382–3388, ISBN: 0-7803-9490-9.

[C5] M. Luckner, "Recognition of Noised Patterns Using Non-Disruption Learning Set," w *Sixth International Conference on Intelligent Systems Design and Applications*, t. 1, IEEE, paź. 2006, s. 557–562, ISBN: 0-7695-2528-8.

[C6] M. Luckner i W. Homenda, "Braille Score," w *Sixth International Conference on Intelligent Systems Design and Applications*, t. 1, IEEE, paź. 2006, s. 775–780, ISBN: 0-7695-2528-8.

[C7] W. Homenda i M. Luckner, "Hierarchical OCR system for texts in musical scores," w *Eleventh International Fuzzy Systems Association World Congress,*, Beijing, China, 2005.

[C8] W. Homenda i M. Luckner, "Automatic Recognition of Music Notation Using Neural Networks," w *International Conference on AI and Systems*, Divnormorkoye, 2004.

[C9] W. Homenda i M. Luckner, "Optical Music Recognition : A Niche of Research and Technology," w *WISIS 2004 First Warsaw International Seminar on Intelligent Systems*, 2004, s. 1–17.

Dorobek przed uzyskaniem stopnia doktora obejmuje dziewięć pozycji [C1–C9], w tym jedną publikację w czasopiśmie [C3]. Suma punktów dorobku przed doktoratem, wyliczonych zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 9 lutego 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych, wynosi **200**. Dorobek powstały przed doktoratem, stworzony został w głównej mierze pod opieką promotora prof. hab. inż. Władysława Homendy.

W pracach [C4, C7–C9] przedstawiono badania nad automatycznym rozpoznawaniem wybranych symboli notacji muzycznej z wykorzystaniem różnych metod uczenia maszynowego, a także hierarchicznym systemem OCR wyspecjalizowany w rozpoznawaniu tekstów zawartych na partyturze.

W pracy [C6] przedstawiono program komputerowy, który pomaga osobom niewidomym pracującym z notacją muzyczną. Program umożliwia przetwarzania muzyki od wydrukowanej partytury do pliku MIDI, który może być wykonany przez instrument elektroniczny. Motorem systemu jest moduł rozpoznawania oparty na zaawansowanej technologii sztucznej inteligencji. Rozpoznane partytury są konwertowane na specjalną wewnętrzną reprezentację, która pozwala na przekazanie wszystkich niuansów muzyki. Zapis może być również przetwarzany za pomocą edytora, który jest przeznaczony dla osób niewidomych.

Prace [C3, C5] omawiają rozpoznawanie silnie zaszumionych symboli na podstawie wzorców bez zakłóceń. Przedstawiono model systemu rozpoznawania nietypowych symboli muzycznych. W opisywanym modelu symbole nietypowe są wykorzystywane do generowania zbioru uczącego, który umożliwia poprawę rozpoznawania, co zostało zaprezentowane na rzeczywistym przykładzie rozpoznawania wybranych symboli muzycznych. Przedstawiono kilka technik o różnej szybkości rozpoznawania i czasie obliczeniowym, w tym nadzorowane i nienadzorowane.

W pracach [C1, C2] przedstawiono badanie zależności pomiędzy strukturą hierarchicznego rozpoznawania wzorców a dokładnością klasyfikacji. Porównano różne zespoły maszyn wektorów wspierających uporządkowane według struktury drzewa binarnego w zadaniu rozpoznawania liter z repozytorium UCI. Przedstawiono również praktyczne testy rozpoznawania tekstów map.

Informacja o wykazywaniu się istotną aktywnością naukową realizowaną w więcej niż jednej uczelni lub instytucji naukowej, w szczególności zagranicznej

Poniższe informacje zostały uporządkowane według instytucji, z którymi współpracuję lub współpracowałem, z wyłączeniem Wydziału Matematyki i Nauk Informacyjnych Politechniki Warszawskiej, w którym jestem zatrudniony w wymiarze pełnego etatu i jestem w grupie pracowników zaliczonych do liczby N – jest to moje podstawowe miejsce pracy, do którego afiliuję mój dorobek. Wykaz obejmuje tylko instytucje, które pojawiły się w afiliacjach publikacji współautorskich.

University of Alberta (UoA), Edmonton, Kanada

W roku 2006, za pracę zespołu w skład którego wchodził prof. Witold Pedrycz z UoA, otrzymaliśmy nagrodę zespołową I stopnia JM Rektora PW, za osiągnięcia naukowe w latach 2004-2005.

W latach 2011-2014 kontynuowałem współpracę z profesorem, gdy realizowaliśmy, jako główni wykonawcy, projekt NCN *Zagadnienie odrzucania w problemie rozpoznawania wzorca*. Projekt dotyczył badania metod i algorytmów rozpoznawania wzorców w połączeniu z nową funkcjonalnością odrzucania elementów

pomylonych i obcych. Wyniki tej współpracy zostały przedstawione w publikacjach współautorskich z prof. Witoldem Pedryczem [B14, B19, B24].

Tallinn University of Technology (TalTech), Talin, Estonia

Od 2020 roku współpracuję z Alejandro Guerra-Manzanaresem i Hayretdinem Bahsi z TalTech. Nasza współpraca dotyczy analizy wieloletniego dryftu konceptualnego cech opisujących złośliwe oprogramowanie atakujące urządzenia z systemem Android. Wyniki naszej współpracy przedstawiono w pracach [B1–B3]. Współpraca jest kontynuowana i planowane są kolejne publikacje.

W okresie 20.04.2022-21.05.2022 przebywałem na stażu w Centre for Digital Forensics and Cyber Security w TalTech, gdzie współpracowałem z Alejandro Guerra-Manzanaresem i Hayretdinem Bahsi nad analizą ewolucji różnych rodzin złośliwego oprogramowania.

Informacja o osiągnięciach dydaktycznych oraz organizacyjnych

Wkład w rozwój kadry naukowej

W ramach mojej pracy naukowej, współpracowałem z doktorantami Politechniki Warszawskiej i innych instytucji naukowych. Współpraca dotyczyła głównie realizacji projektów badawczych, ale przejawiała się też w postaci wspólnych publikacji. Doktoranci, z którymi dzielę dorobek w postaci wymienionych publikacji to Izabella Krzemińska (Uniwersytet Ekonomiczny w Poznaniu) [A1, A4], Robert Kunicki (Politechnika Warszawska) [B6, B7, A4] i Przemysław Wrona (Politechnika Warszawska) [B4].

Jestem także współpromotorem rozprawy doktorskiej Alejandro Guerra-Manzanaresa pod tytułem *Machine Learning-Based Detection and Characterization of Evolving Threats in Mobile and IoT Systems* (Manzanares 2022). Pozostali promotorzy to profesor Hayretdin Bahsi i Sven Nõmm z Tallinn University of Technology. Rozprawa, została obroniona na TallTech w dniu 01.08.2022. W dorobku uwzględniono nasze prace współautorskie [B1–B3].

Pełnienie roli promotora prac magisterskich i inżynierskich

W latach 2011-2022 byłem promotorem:

- 22 obronionych prac inżynierskich (łącznie 47 studentów, ze względu na prace wieloosobowe),
- 23 obronionych prac magisterskich (brak prac wieloosobowych).

Następujące prace, w dorobku, zostały napisane wraz z moimi dyplomantami jako rozwinięcie prac inżynierskich i magisterskich lub wykorzystują częściowo uzyskane w nich wyniki [B5, B7, B11, B22, B27–B30, B33].

Działalność dydaktyczna

Przedmioty prowadzone na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej:

2014/2015 - 2014/2015 — **Algorithms and Computability**, *przedmiot obowiązkowy*, computer science, II stopień
ćwiczenia, laboratoria

2010/2011 - 2020/2021 — **Teoria automatów i lingwistyka matematyczna**, *przedmiot obowiązkowy*, matematyka, specjalność: Matematyka w naukach informacyjnych
ćwiczenia

2009/2010 - — **Automata Theory and Languages**, *przedmiot obowiązkowy*, computer science, I stopień
ćwiczenia

2009/2010 - — **Teoria algorytmów i obliczeń**, *przedmiot obowiązkowy*, informatyka, informatyka i systemy informacyjne, II stopień
ćwiczenia, laboratoria

2009/2010 - 2015/2016 — **Teoria automatów i języków**, *przedmiot obowiązkowy*, informatyka, informatyka i systemy informacyjne, I stopień
ćwiczenia

2009/2010 - 2009/2010 **Data bases**, *przedmiot obowiązkowy*, computer science, I stopień
laboratoria

2010/2011 - 2012/2013 **Bazy danych**, *przedmiot obowiązkowy*, informatyka, I stopień
laboratoria

2009/2010 - 2013/2014 **Programowanie komponentowe w technologii Java Enterprise Edition**, *przedmiot obieralny*, informatyka, II stopień
laboratoria

Przedmioty prowadzone i koordynowane na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej.

2021/2022 - **Data Science Workshop**, *przedmiot obowiązkowy*, data science, II stopień
wykład

2019/2020 - **Projekt interdyscyplinarny**, *przedmiot obowiązkowy*, inżynieria i analiza danych, I stopień
wykład

2019/2020 - **Podstawy przetwarzania danych**, *przedmiot obowiązkowy*, informatyka i systemy informacyjne, specjalność: Metody sztucznej inteligencji
wykład, laboratoria

2019/2020 - 2020/2021 **Warsztaty Badawcze**, *przedmiot obowiązkowy*, inżynieria i analiza danych, II stopień
wykład

2018/2019 - **Zaawansowane programowanie obiektowe i funkcyjne**, *przedmiot obowiązkowy*, inżynieria i analiza danych, I stopień
wykład

2017/2018 - **Programowanie obiektowe**, *przedmiot obowiązkowy*, inżynieria i analiza danych, I stopień
wykład

2016/2017 - **Kreatywne rozwiązywanie problemów**, *przedmiot obowiązkowy*, inżynieria i analiza danych, I stopień
wykład, ćwiczenia

2015/2016 - 2017/2018 **Business Analytics Programming**, *przedmiot obowiązkowy*, computer science, II stopień
wykład, projekt

2014/2015 - 2018/2019 **Android Application Development**, *przedmiot obieralny*, computer science, I-II stopień
wykład, laboratoria

2014/2015 - 2018/2019 **Aplikacje mobilne: Android**, *przedmiot obieralny*, informatyka, I-II stopień
wykład, laboratoria

2014/2015 - 2017/2018 **Pracownia projektowa**, *przedmiot obieralny*, informatyka/matematyka, I-II stopień
projekt

2013/2014 - 2015/2016 **Java SE**, *przedmiot obieralny*, informatyka/matematyka, I-II stopień
wykład, laboratoria

2013/2014 - 2017/2018 **Programming 3 (Advanced, Java)**, *przedmiot obowiązkowy*, computer science, I stopień
wykład, laboratoria

Przedmioty prowadzone na studiach podyplomowych na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej.

2021/2022 - **Metody analizy danych: eksploracja danych i sztuczna inteligencja**, *przedmiot obowiązkowy*, data science, studia podyplomowe
wykład

2019/2020 - **Metody Sztucznej Inteligencji**, *przedmiot obowiązkowy*, inżynieria i analiza danych, I stopień
wykład

Przedmioty prowadzone na studiach MBA Szkoły Biznesu Politechniki Warszawskiej.

2021/2022 - **Data Science**, *przedmiot obowiązkowy*, data science, MBA Digital Transformation
ćwiczenia

Ponadto, w latach 2004-2009 prowadziłem, na Wydziale Geodezji i Kartografii, zajęcia z Metod Numerycznych i Informatyki.

Dodatkowa działalność dydaktyczna, wykraczająca poza prowadzenie zajęć dydaktycznych.

2022 **Nagroda zespołowa Ministra Edukacji i Nauki**, *za znaczące osiągnięcia w zakresie działalności dydaktycznej*
Za uruchomienie na Wydziale Matematyki i Nauk Informacyjnych innowacyjnego programu studiów drugiego stopnia o profilu ogólnoakademickim na kierunku Data Science.

2020 - 2021 **Członek zespołu wykonawczego POWER NERW 2**, *10 - Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych*
Prace nad zmianami w programie studiów kierunku Informatyka i Systemy Informacyjne

2018 - 2019 **Członek zespołu wykonawczego POWER NERW**, *10 - Przygotowanie i uruchomienie nowego kierunku studiów na studiach II stopnia - Inżynieria i Analiza Danych (IAD)*
Przygotowanie materiałów dydaktycznych dla nowych zajęć na II stopniu kierunku Inżynieria i Analiza Danych

2017 - 2020 **Koordynator zadania projektu POWER Kadra/NERW**, *Kompetentny wykładowca - wysoki poziom nauczania*
Szkolenia dla kadry akademickiej Politechniki Warszawskiej

2014 - **Członek Zespołu Rektorskiego ds. Innowacyjnych Form Kształcenia (INFOX)**
Uczestnik szeregu projektów z zakresu opracowania nowych metodyk, pilotażowych wdrożeń, współpracy z otoczeniem społeczno-gospodarczym w ramach innowacyjnych form zajęć ze studentami.

Działalność organizacyjna

2020 - **Członek Komisji Egzaminu Dyplomowego**, *Kierunek Inżynieria i Analiza Danych*

2020 - **Członek Dziekańskiej Komisji ds. Finansowania Badań Naukowych**

2020 - **Przewodniczący Podkomisji Informatyki Dziekańskiej Komisji ds. Finansowania Badań Naukowych**

2021 - **Członek komisji Oceny Śródokresowej**, *Szkoła Doktorska nr 3, Politechnika Warszawska*

2021 - **Sekretarz Podkomisji Oceny Śródokresowej**, *Szkoła Doktorska nr 3, Politechnika Warszawska*

2020- **Członek Rady Naukowej**, *Centrum Badawczego POB*, Fizyka wysokich energii i technika eksperymentu.

2020 - **Członek Komisji Rady Wydziału ds. Nagród i Odznaczeń**

2019 - **Członek Rady Dyscypliny**, *Informatyka Techniczna i Telekomunikacja*

2018 - **Członek Komisji Programowych**, *Kierunki Inżynieria i Analiza Danych i Informatyka i Systemy Informacyjne*

2018 - **Członek Rady Wydziału**, *Wydział Matematyki i Nauk Informacyjnych*

2017 **Nagroda zespołowa II stopnia JM Rektora PW**, *za osiągnięcia organizacyjne w roku akademickim 2016/2017*
Organizacja międzywydziałowych, interdyscyplinarnych zajęć i projektów.

2014 - **Członek Zespołu Rektorskiego ds. Innowacyjnych Form Kształcenia (INFOX)**
Organizacja międzywydziałowych, interdyscyplinarnych zajęć i projektów, współpraca z międzynarodowymi organizacjami akademickimi.

2011 - **Dyrektor Ośrodka Badań dla Biznesu**
Organizacja i prowadzenie Ośrodka, którego celem jest współpraca z biznesem i realizacja projektów rozwojowo badawczych.

Działalność popularyzatorska

2020 **Coronathon**
Juror Hackathonu dotyczącego walki z konsekwencjami pandemii COVID-19.

2019 - 2020 **Universities of the Future**
Publikacja Handbook for Industry - publikacji opisującej możliwości związane z Przemysłem 4.0 dla przedsiębiorstw.

2019 **Centrum Unijnych Projektów Transportowych**
Publikacja artykułu *Big Data w analizie funkcjonowania systemu komunikacji miejskiej* promującego wykorzystanie infrastruktury Big Data do analizy danych miejskich.

2018 **Data Science Summit**
Prowadzenie Hackathonu *Conquer Urban Big Data hackathon*.

2018 **Red Bull Tech Lab**
Mentor opiekujący się zespołem startującym w konkursie.

2018 **Campus App Challenge**
Juror Hackathonu dotyczącego lokalizacji wewnątrzbudynkowej

2018 **Design Thinking Week**
Prowadzenie warsztatu *Po co gonić ten tramwaj?*.

2017 **Design Thinking Week**
Prowadzenie warsztatu o komunikacji między pokoleniowej *W czasie deszczu dzieci się nudzą*.

2017 **Środa z DRIMn Ścieżka Przemysł 4.0**
Wykład *Uczenie maszynowe i sztuczna inteligencja – szansa czy zagrożenie?*.

2017 **Klub Informatyka**, *Polskie Towarzystwo Informatyczne*
Udział w debacie *Dane masowe za a nawet przeciw*.

2017 **Kreatywny Projekt Zespołowy**
Prowadzenie projektu edukacyjnego mającego na celu rozbudowanie umiejętności i zainteresowań z zakresu nauki i działań zespołowych wśród młodzieży w wieku ponadgimnazjalnym.

2016 **TedX Warsaw**
Prowadzenie warsztatu Design Thinking.

2016 **Design Thinking Week**
Prowadzenie warsztatu *Nauczycielu akademicki - naucz się nauczać kreatywnie*.

Bibliografia

T. Chen i C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *CoRR*, t. abs/1603.02754, 2016.

A. Furno, M. Fiore, R. Stanica, C. Ziemlicki i Z. Smoreda, "A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas," *IEEE Trans. Mob. Comput.*, t. 16, nr. 10, s. 2682–2696, 2017.

J. Karwowski, M. Okulewicz i J. Legierski, "Application of Particle Swarm Optimization Algorithm to Neural Network Training Process in the Localization of the Mobile Terminal," w *Engineering Applications of Neural Networks - 14th International Conference, EANN 2013, Halkidiki, Greece, September 13-16, 2013 Proceedings, Part I*, 2013, s. 122–131.

A. G. Manzanares, "Machine Learning-Based Detection and Characterization of Evolving Threats in Mobile and IoT Systems," 2022, ISBN: 9789949838981.

L. Rokach i O. Maimon, "Classification Trees," w *Data Mining and Knowledge Discovery Handbook*, O. Maimon i L. Rokach, red. Boston, MA: Springer US, 2010, s. 149–174, ISBN: 978-0-387-09823-4.