# ALGORYTM UZUPEŁNIANIA BRAKUJĄCYCH DANYCH W ZBIORACH REJESTROWANYCH NA STACJACH MONITORINGU POWIETRZA

Szymon HOFFMAN, Rafał JASIŃSKI
Politechnika Częstochowska
Wydział Inżynierii i Ochrony Środowiska
ul. Dąbrowskiego 69, 42-200 Częstochowa
szymon@is.pcz.czest.pl; raphael@is.pcz.czeast.pl

STRESZCZENIE

W pracy zaproponowano algorytm uzupełniania brakujących danych w zbiorach rejestrowanych na stacjach automatycznego monitoringu powietrza. Wykorzystano wyniki analizy możliwości zastosowania różnych metod modelowania w celu aproksymacji stężeń zanieczyszczeń powietrza. Porównano dokładność kilku różnych grup modeli, w tym modeli szeregów czasowych, modeli regresji wielowymiarowej i zwykłych modeli interpolacyjnych. W algorytmie przyjęto zasadę minimalizacji błędów modelowania, którą zrealizowano poprzez specyficzne traktowanie różnych serii pomiarowych. Zastosowane w algorytmie metody predykcji uzależniono od rodzaju zanieczyszczenia, od długości luki pomiarowej i od miejsca modelowanego przypadku w luce pomiarowej, a także od dostępności innych danych, wymaganych w używanych metodach aproksymacji.

## 1. Wprowadzenie

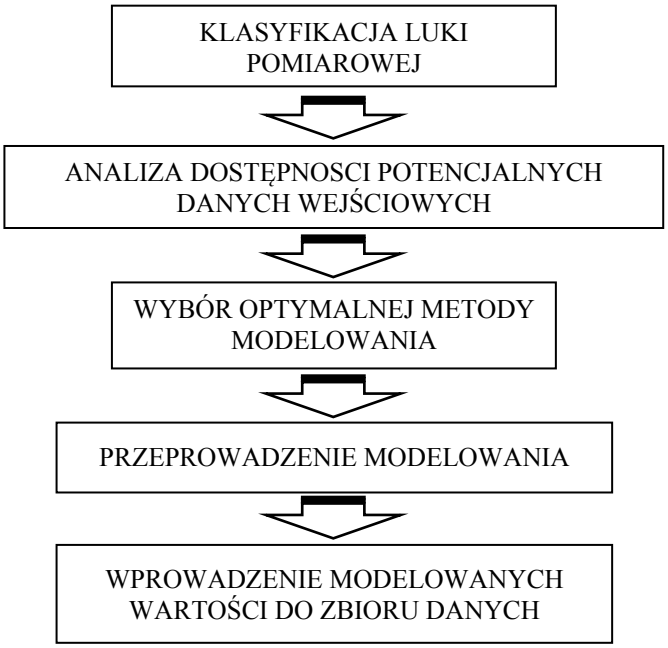
Obowiązujące przepisy prawne dopuszczają możliwość wykorzystania modelowania w celu uzupełnienia brakujących danych w systemach monitoringu powietrza, gdy kompletność serii pomiarowych stężeń poszczególnych zanieczyszczeń jest zbyt mała [1]. Dane pomiarowe gromadzone w systemach monitoringu powietrza są cennym źródłem wiedzy o zależnościach między mierzonymi parametrami. Ukryta w tych danych wiedza o zależnościach regresyjnych i autoregresyjnych umożliwia aproksymację brakujących przypadków [2-3]. Metody modelowania stosowane do predykcji brakujących danych powinny zapewniać możliwie największą dokładność, aby wygenerowane dane były najbardziej zbliżone do rzeczywistości. W zależności od charakteru luki pomiarowej różne metody aproksymacji mogą okazać się najdokładniejsze. W obrębie samej luki pomiarowej do poszczególnych przypadków można rekomendować różne metody predykcji.

Przedstawioną w pracy koncepcję algorytmu uzupełniania brakujących danych w zbiorach rejestrowanych na stacjach automatycznego monitoringu powietrza oparto o dwie podstawowe zasady: pierwszą - korzystania wyłącznie z tzw. modeli autonomicznych i drugą – zapewnienia maksymalnej dokładności predykcji. Modele autonomiczne wykorzystują wiedzę ukrytą w danych historycznych, zgromadzonych w rozpatrywanym systemie monitoringu powietrza [3]. Do predykcji nie są potrzebne żadne dane pochodzące spoza systemu. Algorytm wykorzystujący modele autonomiczne jest możliwy do utworzenia w każdym dowolnym systemie automatycznego monitoringu powietrza. Jednak w poszczególnych stacjach monitoringu dokładność rozpatrywanych metod modelowania może być różna. Dlatego dedykowana dla określonej stacji pomiarowej propozycja algorytmu powinna być poprzedzona wstępną analizą dokładności możliwych do stosowania metod modelowania.

Celem pracy jest prezentacja algorytmu uzupełniania brakujących danych przygotowanego dla stacji monitoringu powietrza w Radomiu. Stacja ta została potraktowana jako element sieci monitoringu powietrza obejmującej w sumie 8 stacji pomiarowych. Do konstrukcji algorytmu wykorzystano rekomendacje wynikające z analizy dokładności modelowania różnymi metodami.

## 2. Metodyka

Ogólna koncepcja algorytmu polega na całościowym przeprowadzeniu procesu uzupełniania brakujących danych w bazie danych zawierającej zarówno wartości stężeń zanieczyszczeń jak i pozostałych parametrów meteorologicznych, przy czym uzupełniane są tylko wartości stężeń zanieczyszczeń. Proces odbywa się automatycznie i polega na wyszukiwaniu przez program luki danych, zakwalifikowaniu jej do odpowiedniej klasy, analizie dostępności potencjalnych danych wejściowych i doborze optymalnej metody modelowania (rys. 1). W miarę działania algorytmu luki danych zastępowane są wartościami modelowanymi w porządku chronologicznym.




Rys. 1. Schemat ideowy algorytmu uzupełniania brakujących danych, wg [4].

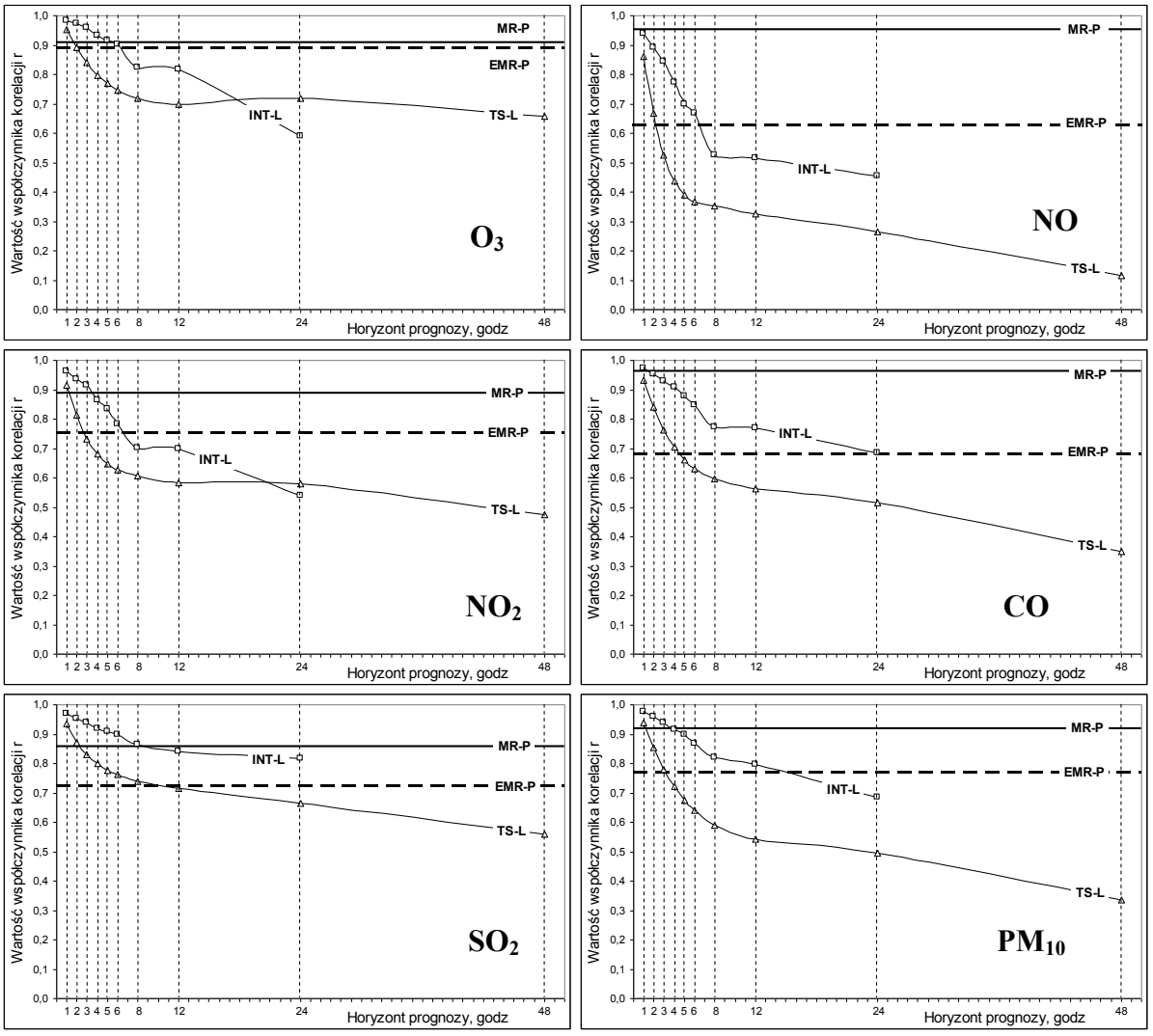
Algorytm uzupełniania brakujących danych w seriach czasowych rejestrowanych na stacjach monitoringu powietrza w Radomiu utworzono korzystając z koncepcji ogólnej algorytmu (rys. 1) i z opublikowanych w pracy [4] wyników analizy błędów różnych metod aproksymacji i wynikających z tej analizy rekomendacji najdokładniejszych modeli. Wykorzystano dane zarejestrowane w latach 2004-2008 na ośmiu stacjach monitoringu powietrza, działających w różnych miejscowościach województw łódzkiego i mazowieckiego. Porównano dokładność kilku różnych grup modeli w tym modeli szeregów czasowych (TS-L), modeli regresji wielowymiarowej bazujących wyłącznie na danych rejestrowanych na stacji w Radomiu (MR-P), modeli regresji wielowymiarowej wykorzystującej dane pochodzące z innych stacji monitoringu (EMR-P) i zwykłych modeli interpolacyjnych (INT-L). Dokładniejszy opis tych modeli znajduje się w pracy [4]. Wykonana analiza wykazała, że dla każdego z zanieczyszczeń powietrza należy

rekomendować inne metody predykcji, ponieważ występują duże różnice w możliwościach modelowania poszczególnych zanieczyszczeń powietrza. Wykazano również, że dla różnych fragmentów luki pomiarowej powinny być rekomendowane odmienne metody aproksymacji.

Uwzględniając wyniki tej analizy zaproponowano szczegółowy algorytm sekwencyjnego uzupełniania brakujących danych, który może być stosowany do dowolnego przypadku luki pomiarowej występującej w danych monitoringowych pochodzących ze stacji w Radomiu. Jako kryterium dokładności predykcji przyjęto wartość współczynnika korelacji.

## 3. Wyniki

Na rys. 2 porównano błędy predykcji chwilowych stężeń zanieczyszczeń powietrza zarejestrowanych na stacji monitoringu powietrza Radom. Na oddzielnych wykresach dla $O_3$, NO, $NO_2$, CO, $SO_2$, $PM_{10}$ przedstawiono krzywe ilustrujące zmienność wartości współczynnika korelacji r w miarę wydłużania horyzontu prognozy od 1 do 48 godzin. Na każdym wykresie porównano wyniki dla modeli wygenerowanych 4 różnymi technikami predykcji, w tym dla modeli szeregów czasowych (TS-L), dla nieliniowych model regresji wielowymiarowej (MR-P), dla nieliniowych modeli regresji wielowymiarowej eksplorujących dane pochodzące z sąsiednich stacji monitoringu (EMR-P) i dla liniowych modeli interpolacyjnych (INT-L).



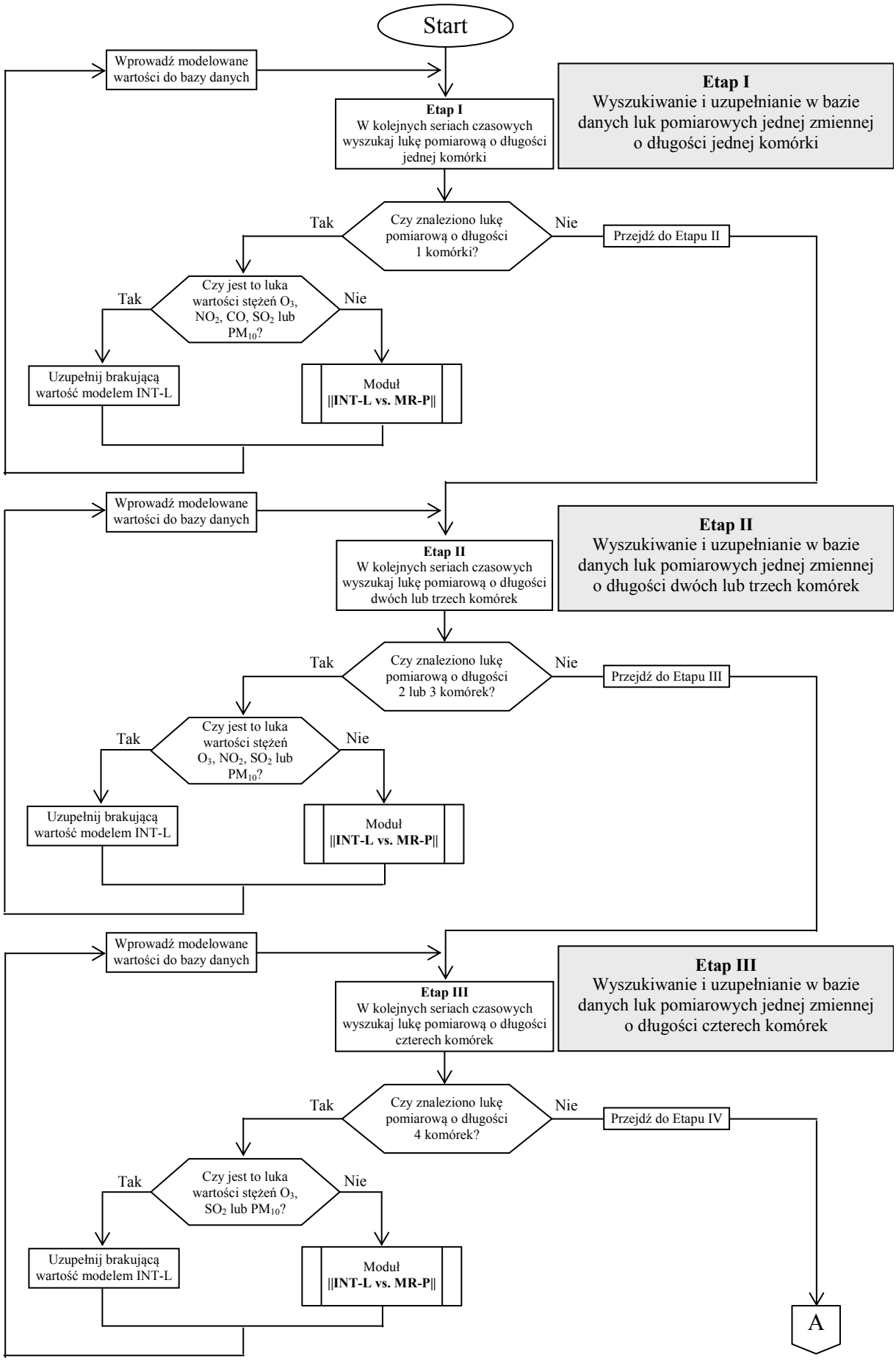
Rys. 2. Zmiany wartości współczynnika korelacji dla różnych metod predykcji w zależności od horyzontu prognozy w luce pomiarowej.

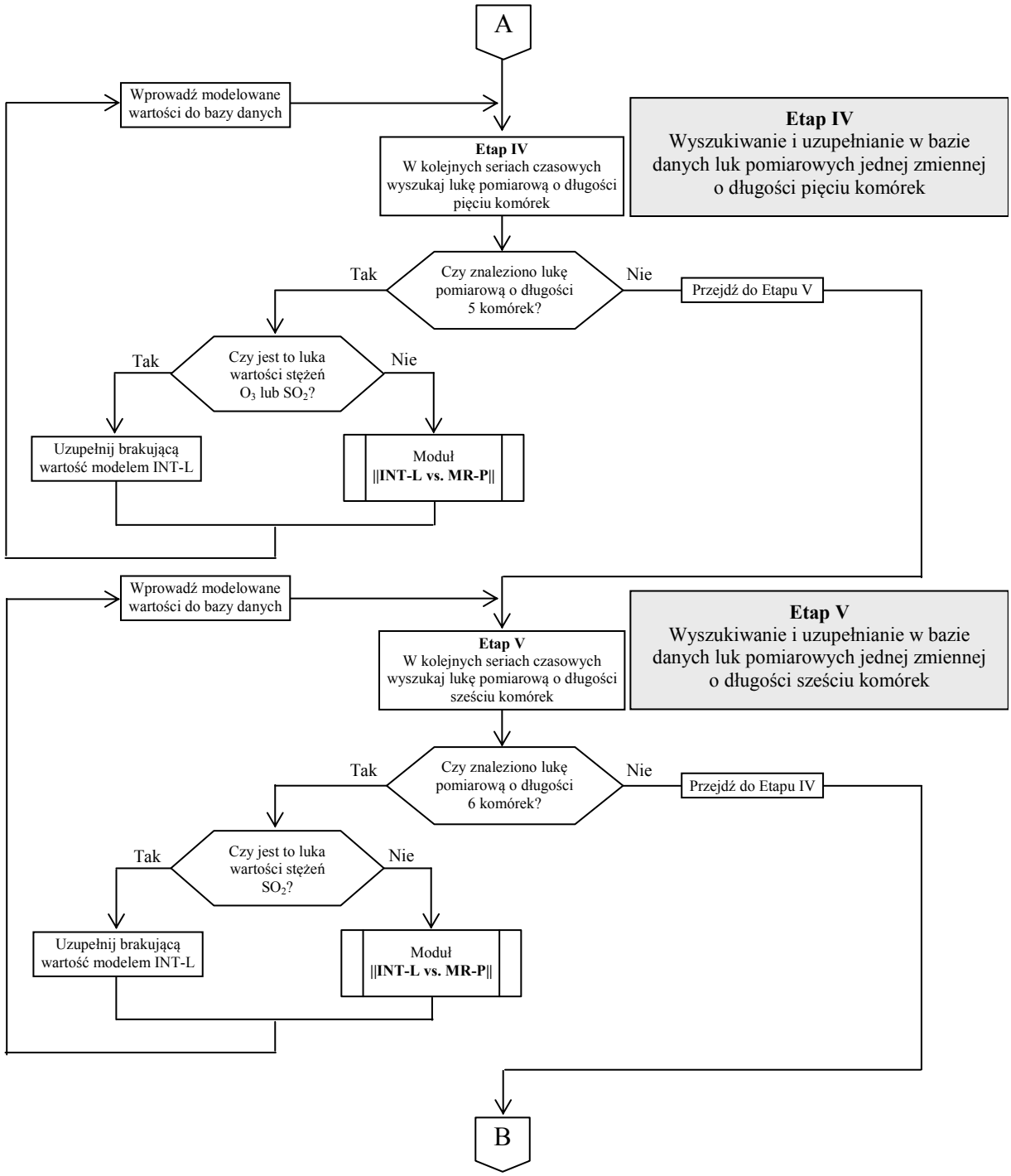
Zaproponowany w pracy algorytm został przygotowany w oparciu o porównania wartości błędu modelowania dla kolejnych przypadków w ewentualnych lukach pomiarowych poszczególnych serii czasowych. Na przykład, analiza wykresów prowadzi do wniosku, że dla krótkich horyzontów prognozy najdokładniejszymi modelami aproksymacyjnymi wszystkich zanieczyszczeń, z wyjątkiem NO, są modele interpolacyjne INT-L. Jednak dokładność tych modeli szybko maleje w miarę wydłużania horyzontów prognozy. Oznacza to, że tylko dla krótkich luk pomiarowych modele INT-L są dokładniejsze od pozostałych. Algorytm został tak zbudowany, że wstępnie klasyfikuje długość luki pomiarowej w serii czasowej, a następnie w zależności od klasyfikacji narzuca określony (najdokładniejszy) sposób modelowania. W krótkich lukach pomiarowych do predykcji stosowane będą modele INT-L, natomiast w dłuższych modele szeregów czasowych TS-L albo modele regresyjne MR-P, w zależności od przypadku i od rodzaju zanieczyszczenia.

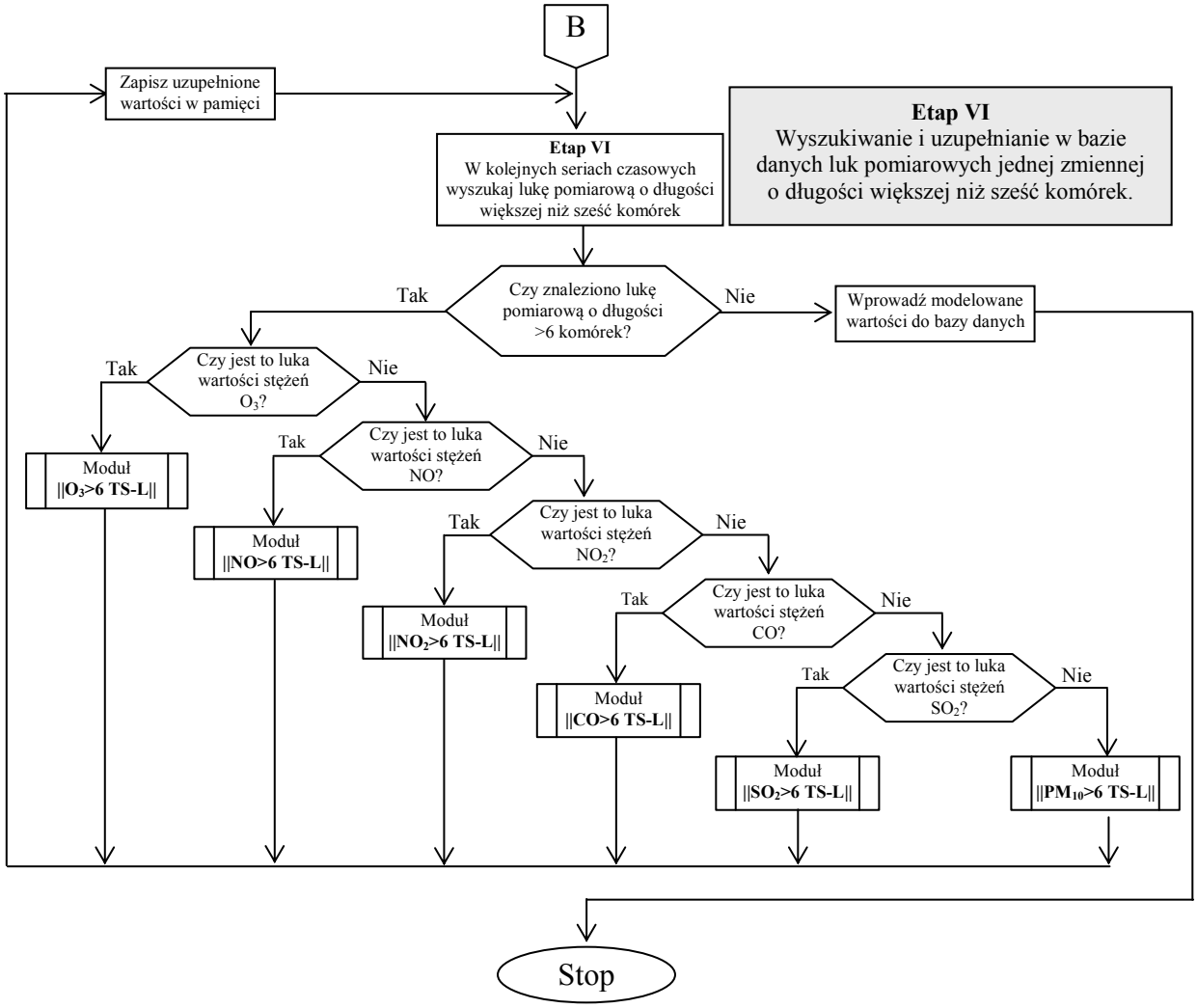
Schemat algorytmu przedstawiono na rys. 3. Algorytm podzielono na VI etapów, uwzględniających różne długości luki pomiarowej. W I etapie następuje wyszukiwanie i uzupełnianie w bazie danych luk pomiarowych o długości jednej komórki dla kolejnych zmiennych. W etapie II następuje wyszukiwanie i uzupełnianie w kolejnych seriach czasowych stężeń luk pomiarowych o długości dwóch i trzech komórek. W etapach III, IV i V następuje wyszukiwanie i uzupełnianie luk pomiarowych o długości odpowiednio czterech, pięciu i sześciu komórek. W omawianych etapach I-V wszystkie modelowane wartości są automatycznie wpisywane w luki pomiarowe i w następnych krokach traktowane przez algorytm tak jak inne wartości rzeczywiste. W VI etapie następuje wyszukiwanie i modelowanie brakujących wartości stężeń poszczególnych zanieczyszczeń, występujących w lukach o długości większej niż sześć komórek. W tym etapie wartości modelowane zachowywane są w pamięci i dopiero po zakończeniu całego etapu są wpisywane we właściwe miejsce w lukach pomiarowych.

W etapach I-V algorytmu wprowadzono dodatkowy moduł wyboru modelu predykcyjnego ||**INT-L vs. MR-P**||. Moduł ten jest zdefiniowanym algorytmem wyboru metody uzupełniania brakujących danych dla luki pomiarowej w serii czasowej, dla której rekomendowany jest nieliniowy model regresji wielowymiarowej MR-P, pod warunkiem, że w wierszu danej komórki luki pomiarowej występuje komplet pozostałych danych. W przypadku, nie spełnienia tego warunku, moduł ||**INT-L vs. MR-P**|| zarekomenduje liniowy model interpolacyjny INT-L. Dla każdej kolejnej brakującej wartości luki zostanie sprawdzony warunek kompletności pozostałych wartości w wierszu, i w zależności od spełnienia tego warunku, zarekomendowany jeden z dwóch modeli MR-P lub INT-L.

W etapie VI zastosowano moduły algorytmów postępowania, dedykowane dla każdego zanieczyszczenia indywidualnie: ||**$O_3$ >6 TS-L**||, ||**NO >6 TS-L**||, ||**$NO_2$ >6 TS-L**||, ||**CO >6 TS-L**|| **i** ||**$SO_2$ >6 TS-L**||, ||**$PM_{10}$ >6 TS-L**||. W modułach tych nie uwzględniono liniowego modelu interpolacyjnego INT-L, z uwagi na jego niską dokładność modelowania dla dłuższych luk pomiarowych. Moduły pozwalają wykorzystać do modelowania początkowych i końcowych komórek w luce pomiarowej modele szeregów czasowych TS-L w zależności od spełnienia warunku kompletności 24 wartości danej zmiennej przed lub po luce danych. Ten warunek jest wymagany modelach szeregów czasowych TS-L, w których wykorzystuje się 24 wartości opóźnione jako zmienne wejściowe. Przykładową procedurę postępowania w ramach modułu dedykowanego dla poszczególnych zanieczyszczeń przedstawiono na przykładzie modułu ||**$NO_2$ >6 TS-L**||, dedykowanego dla uzupełnienia brakujących wartości stężeń $NO_2$.

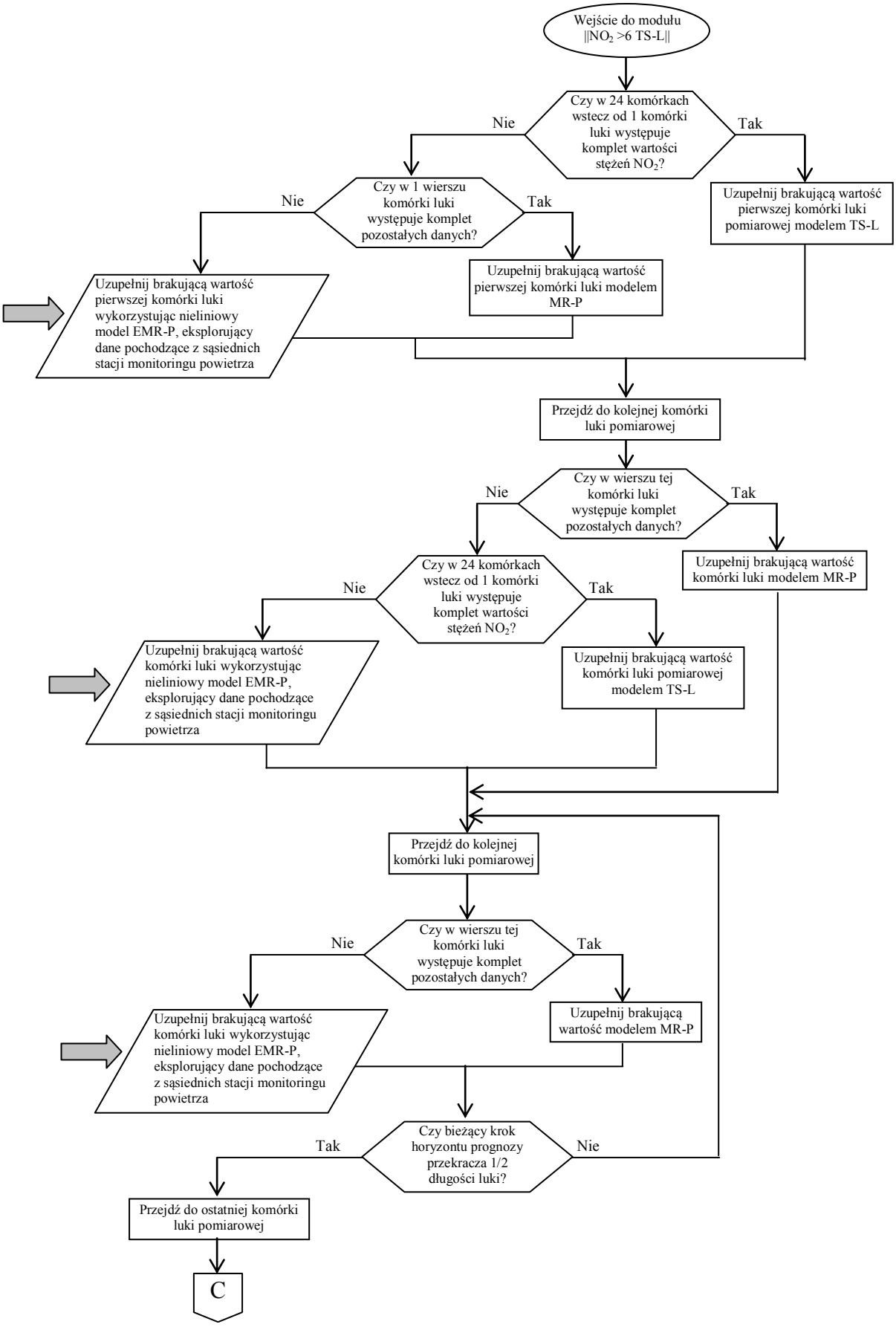

Start
Wprowadź modelowane wartości do bazy danych
Etap I
W kolejnych seriach czasowych wyszukaj lukę pomiarową o długości jednej komórki
Etap I
Wyszukiwanie i uzupełnianie w bazie danych luk pomiarowych jednej zmiennej o długości jednej komórki
Czy znaleziono lukę pomiarową o długości 1 komórki?
Tak
Nie
Przejdź do Etapu II
Czy jest to luka wartości stężeń $O_3$, $NO_2$, CO, $SO_2$ lub $PM_{10}$?
Tak
Nie
Uzupełnij brakującą wartość modelem INT-L
Moduł ||INT-L vs. MR-P||
Wprowadź modelowane wartości do bazy danych
Etap II
W kolejnych seriach czasowych wyszukaj lukę pomiarową o długości dwóch lub trzech komórek
Etap II
Wyszukiwanie i uzupełnianie w bazie danych luk pomiarowych jednej zmiennej o długości dwóch lub trzech komórek
Czy znaleziono lukę pomiarową o długości 2 lub 3 komórek?
Tak
Nie
Przejdź do Etapu III
Czy jest to luka wartości stężeń $O_3$, $NO_2$, $SO_2$ lub $PM_{10}$?
Tak
Nie
Uzupełnij brakującą wartość modelem INT-L
Moduł ||INT-L vs. MR-P||
Wprowadź modelowane wartości do bazy danych
Etap III
W kolejnych seriach czasowych wyszukaj lukę pomiarową o długości czterech komórek
Etap III
Wyszukiwanie i uzupełnianie w bazie danych luk pomiarowych jednej zmiennej o długości czterech komórek
Czy znaleziono lukę pomiarową o długości 4 komórek?
Tak
Nie
Przejdź do Etapu IV
Czy jest to luka wartości stężeń $O_3$, $SO_2$ lub $PM_{10}$?
Tak
Nie
Uzupełnij brakującą wartość modelem INT-L
Moduł ||INT-L vs. MR-P||
A

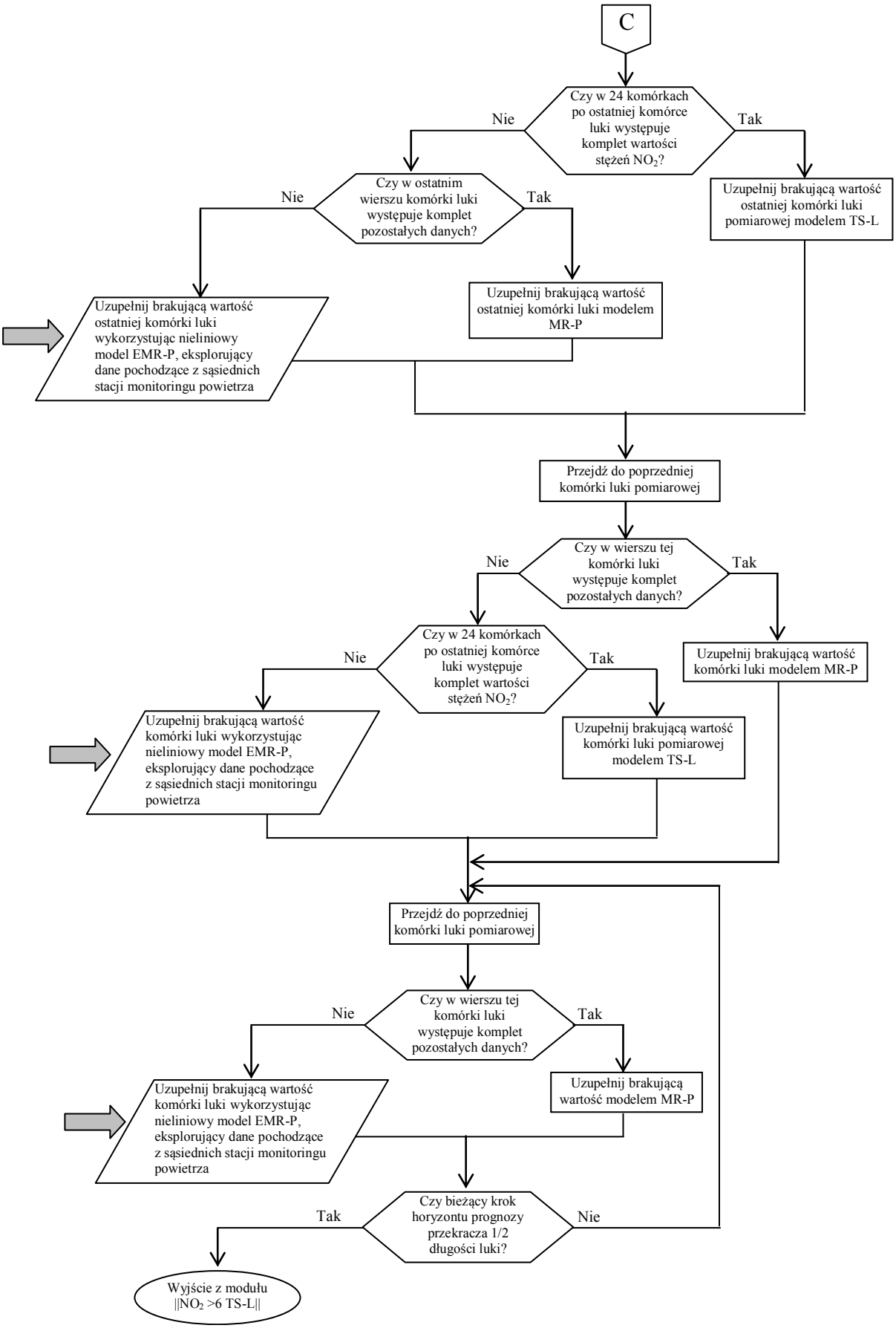

A
Wprowadź modelowane wartości do bazy danych
Etap IV
W kolejnych seriach czasowych wyszukaj lukę pomiarową o długości pięciu komórek
Etap IV
Wyszukiwanie i uzupełnianie w bazie danych luk pomiarowych jednej zmiennej o długości pięciu komórek
Czy znaleziono lukę pomiarową o długości 5 komórek?
Tak
Nie
Przejdź do Etapu V
Czy jest to luka wartości stężeń $O_3$ lub $SO_2$?
Tak
Nie
Uzupełnij brakującą wartość modelem INT-L
Moduł ||INT-L vs. MR-P||
Wprowadź modelowane wartości do bazy danych
Etap V
W kolejnych seriach czasowych wyszukaj lukę pomiarową o długości sześciu komórek
Etap V
Wyszukiwanie i uzupełnianie w bazie danych luk pomiarowych jednej zmiennej o długości sześciu komórek
Czy znaleziono lukę pomiarową o długości 6 komórek?
Tak
Nie
Przejdź do Etapu IV
Czy jest to luka wartości stężeń $SO_2$?
Tak
Nie
Uzupełnij brakującą wartość modelem INT-L
Moduł ||INT-L vs. MR-P||
B




Rys. 3. Przykład algorytmu postępowania z niekompletną bazą danych z automatycznej stacji monitoringu powietrza, w celu uzupełnienia brakujących wartości stężeń $O_3$, NO, $NO_2$, CO, $SO_2$, $PM_{10}$.

Na rys. 4 przedstawiono moduł **||$NO_2$ >6 TS-L||** - zdefiniowany algorytm wyboru metody uzupełniania brakujących wartości stężeń $NO_2$, dla luki o długości większej niż 6 komórek. Algorytm modułu opracowano w oparciu o wartości współczynnika korelacji Pearsona r, jako wskaźnika oceny błędu modelowania dla stężeń $NO_2$ (rys. 2). Dla pierwszego kroku prognozy najdokładniejszą metodą modelowania jest liniowy model szeregów czasowych TS-L. W przypadku spełnienia warunku występowania kompletu 24 wartości danej zmiennej przed luką, moduł **||$NO_2$ >6 TS-L||** wybierze model TS-L do uzupełnienia brakującej wartości. Jeśli ten warunek nie będzie spełniony moduł **||$NO_2$ >6 TS-L||** wybierze kolejną metodę pod względem dokładności modelowania - model regresyjny MR-P. W przypadku niespełnienia warunku występowania kompletu pozostałych danych w wierszu modelowanej komórki, moduł zarekomenduje zastosowanie modelu EMR-P, eksplorującego dane pochodzące z sąsiednich stacji monitoringowych. Dla drugiego kroku prognozy moduł **||$NO_2$ >6 TS-L||** zarekomenduje kolejno modele M-RP, TS-L lub EMR-P sprawdzając, czy warunki zastosowania tych modeli są spełnione.


Wejście do modułu
||NO2 >6 TS-L||
Czy w 24 komórkach wstecz od 1 komórki luki występuje komplet wartości stężeń NO2?
Nie
Tak
Czy w 1 wierszu komórki luki występuje komplet pozostałych danych?
Nie
Tak
Uzupełnij brakującą wartość pierwszej komórki luki pomiarowej modelem TS-L
Uzupełnij brakującą wartość pierwszej komórki luki modelem MR-P
Uzupełnij brakującą wartość pierwszej komórki luki wykorzystując nieliniowy model EMR-P, eksplorujący dane pochodzące z sąsiednich stacji monitoringu powietrza
Przejdź do kolejnej komórki luki pomiarowej
Czy w wierszu tej komórki luki występuje komplet pozostałych danych?
Nie
Tak
Czy w 24 komórkach wstecz od 1 komórki luki występuje komplet wartości stężeń NO2?
Nie
Tak
Uzupełnij brakującą wartość komórki luki modelem MR-P
Uzupełnij brakującą wartość komórki luki wykorzystując nieliniowy model EMR-P, eksplorujący dane pochodzące z sąsiednich stacji monitoringu powietrza
Uzupełnij brakującą wartość komórki luki pomiarowej modelem TS-L
Przejdź do kolejnej komórki luki pomiarowej
Czy w wierszu tej komórki luki występuje komplet pozostałych danych?
Nie
Tak
Uzupełnij brakującą wartość komórki luki wykorzystując nieliniowy model EMR-P, eksplorujący dane pochodzące z sąsiednich stacji monitoringu powietrza
Uzupełnij brakującą wartość modelem MR-P
Czy bieżący krok horyzontu prognozy przekracza 1/2 długości luki?
Tak
Nie
Przejdź do ostatniej komórki luki pomiarowej
C


C
Czy w 24 komórkach po ostatniej komórce luki występuje komplet wartości stężeń $NO_2$?
Nie
Tak
Czy w ostatnim wierszu komórki luki występuje komplet pozostałych danych?
Nie
Tak
Uzupełnij brakującą wartość ostatniej komórki luki pomiarowej modelem TS-L
Uzupełnij brakującą wartość ostatniej komórki luki modelem MR-P
Uzupełnij brakującą wartość ostatniej komórki luki wykorzystując nieliniowy model EMR-P, eksplorujący dane pochodzące z sąsiednich stacji monitoringu powietrza
Przejdź do poprzedniej komórki luki pomiarowej
Czy w wierszu tej komórki luki występuje komplet pozostałych danych?
Nie
Tak
Czy w 24 komórkach po ostatniej komórce luki występuje komplet wartości stężeń $NO_2$?
Nie
Tak
Uzupełnij brakującą wartość komórki luki modelem MR-P
Uzupełnij brakującą wartość komórki luki wykorzystując nieliniowy model EMR-P, eksplorujący dane pochodzące z sąsiednich stacji monitoringu powietrza
Uzupełnij brakującą wartość komórki luki pomiarowej modelem TS-L
Przejdź do poprzedniej komórki luki pomiarowej
Czy w wierszu tej komórki luki występuje komplet pozostałych danych?
Nie
Tak
Uzupełnij brakującą wartość komórki luki wykorzystując nieliniowy model EMR-P, eksplorujący dane pochodzące z sąsiednich stacji monitoringu powietrza
Uzupełnij brakującą wartość modelem MR-P
Czy bieżący krok horyzontu prognozy przekracza 1/2 długości luki?
Tak
Nie
Wyjście z modułu ||$NO_2$ >6 TS-L||

Rys. 4. Algorytm modułu ||**$NO_2$ >6 TS-L**|| uzupełniania brakujących wartości stężeń $NO_2$ dla luki pomiarowej o długości >6 godz.

Dla kolejnych kroków prognozy model predykcyjny EMR-P daje lepsze rezultaty modelowania niż TS-L, zatem moduł ||**$NO_2$ >6 TS-L**|| wybierze model regresyjny MR-P lub EMR-P, w zależności od spełnienia warunku kompletności wszystkich pozostałych danych w wierszu modelowanej komórki. Przedstawiony powyżej cykl uzupełniania brakujących wartości będzie się powtarzał, aż do momentu uzupełnienia połowy luki danych, wówczas modelowanie kontynuowane będzie przy wykorzystaniu prognozy wstecznej od ostatniej komórki do połowy luki danych przy zastosowaniu tych samych warunków modelowania jak w pierwszej części algorytmu.

## 4. Wnioski

Na podstawie przeprowadzonych badań można sformułować następujące wnioski końcowe:

1. Problem brakujących danych w systemach monitoringu powietrza można rozwiązać za pomocą algorytmu wypełniającego luki w zbiorach danych.
2. Porównując różne metody modelowania można wyznaczyć najdokładniejsze z nich i rekomendować je do algorytmu uzupełniającego dane. Wybór rekomendowanej do algorytmu metody zależy od rodzaju zanieczyszczenia, od długości luki pomiarowej i od miejsca w luce pomiarowej, a także od dostępności innych danych, wymaganych w rozpatrywanych metodach predykcji.
3. Stężenia zanieczyszczeń wszystkich zanieczyszczeń, z wyjątkiem NO, można efektywnie modelować metodą liniowej interpolacji, ale tylko do pewnej długości luki pomiarowej, powyżej której regresyjne metody modelowania okazują się dokładniejsze od interpolacji.
4. Dla rozpatrywanej stacji monitoringu powietrza modele szeregów czasowych są mniej dokładne od metody interpolacyjnej, ale mogą one zapewniać najwyższą dokładność modelowania skrajnych przypadków (pierwszych i ostatnich) w długich lukach pomiarowych.
5. W algorytmie służącym do wypełniania brakujących danych należy przyjąć tylko jedną z miar błędów jako kryterium wyboru metody modelowania. Wtedy wybór dla kolejnych miejsc luki pomiarowej będzie jednoznaczny.

*Praca została wykonana w ramach badań własnych Politechniki Częstochowskiej BW 402-201/06. W analizie wykorzystano wyniki uzyskane w projekcie badawczym nr 1 T09D 037 30.*

## Literatura

1. Rozporządzenie Ministra Środowiska z dnia 17 grudnia 2008 r. w sprawie dokonywania oceny poziomów substancji w powietrzu (Dz. U. Nr 5, poz. 31).
2. Gentili S., Magnaterra L., Passerini G.: An introduction to the statistical filling of environmental data time series. In Latini G., Passerini G. (Eds.): Handling Missing Data: Applications to Environmental Analysis. Wit Press, Southampton, Boston 2006.
3. Hoffman S.: Treating missing data at air monitoring stations. In Pawłowski L., Dudzińska M. R., Pawłowski A. (Eds.): Environmental Engineering, Taylor & Francis Group, London 2007, 349-353.
4. Hoffman S., Jasiński R.: Uzupełnianie brakujących danych w systemach monitoringu powietrza. Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2009.