Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Jakub Święcicki, Ivona Tautkutė
Nr albumu: 292355, 296836

Wybrane statystyczne metody wykrywania stanów alarmowych na przykładzie danych o zachorowaniach na grypę w Polsce

Praca licencjacka
na kierunku MATEMATYKA

Praca wykonana pod kierunkiem
dra Konrada Furmańczyka
Katedra Zastosowań Matematyki SGGW

Lipiec 2012

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

W pracy przedstawiono algorytmy CUSUM, RKI i Model Farringtona statystycznego monitoringu wykrywania stanów alarmowych. Do ilustracji działania algorytmów wykorzystano dane o zachorowaniach na grypę w Polsce, które przenalizowano za pomocą pakietu statystycznego R.

Słowa kluczowe

stany alarmowe, grypa, CUSUM, epidemia, Model Farringtona, RKI

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.2.Statystyka

Klasyfikacja tematyczna

62P10

Tytuł pracy w języku angielskim

Statistical Surveillance of Influenza Outbreaks with the Use of Statistical Methods in Statistical Package R

Spis treści

Spis rysunków

Spis tabel

Udział w przygotowaniu pracy

Niniejsza praca została napisana przez dwie osoby: Jakuba Święcickiego oraz Ivonę Tautkutė. Praca była tworzona wspólnie, ale można wskazać osoby, które przyczyniły się w większym stopniu do napisania poszczególnych rozdziałów. Dla Jakuba Święcickiego są to podrozdziały 1.2, 1.3, 2.1, 2.3, 2.4, 2.5, 3.2, 3.4, 3.5, 3.6, 3.7; dla Ivony Tautkutė : Wprowadzenie, 1.1, 2.2, 3.1, 3.3, Podsumowanie. Dodatek A był pisany wspólnie.

Wprowadzenie

Niniejsza praca jest poświęcona wykrywaniu stanów alarmowych na przykładzie danych o zachorowaniach na grypę w Polsce, pochodzących z raportów Narodowego Instytutu Zdrowia Publicznego - Panstwowego Zakładu Higieny. System monitoringu epidemiologicznego jest istotną częścią wszystkich instytucji medycznych w związku z potrzebą wczesnego wykrycia zbliżającej się epidemii, które może zezwolić na odpowiednią interwencję i zmniejszenie ryzyka rozprzestrzenienia się choroby, jak również przygotowanie się do zwiększonej aktywności placówek medycznych w celu ochrony zdrowia publicznego.

Zgodnie z terminologią Rządowego Centrum Bezpieczeństwa:

Definicja 0.0.1 (Grypa) *Grypa jest ostrą chorobą zakaźną układu oddechowego, wywołaną zakażeniem wirusem grypy (w tym wyszczególniane są wirusy typu A, B i C). Wirus przenoszony jest drogą kropelkową. Liczba zachorowań wzrasta sezonowo, niekiedy dochodzi do epidemii. Według danych Światowej Organizacji Zdrowia (WHO, ang. World Health Organization), rocznie odnotowywane jest 100 mln przypadków zachorowań na grypę, a umieralność w wyniku powikłań pogrypowych sięga od 500 tys. do 1 000 000 ludzi rocznie.*

Do badania tego typu zjawisk stosowane przez statystyków są liczne metody: parametryczne i semi-parametryczne metody regresji, metody kontroli procesów statystycznych, metody inspirowane procesami Markova i inne.

Do analizy danych o zachorowaniach na grypę w niniejszej pracy wykorzystane zostały trzy metody: CUSUM, Model Farringtona i RKI (niem. *Robert Koch Institut*). Wybór tych konkretnych metod argumentowany jest udokumentowaną stosowalnością do badania chorób zakaźnych, w szczególności chorób sezonowych, jaką jest grypa [17, 13, 8, 18]. Przy wyborze metod statystycznych kierowano się również celem porównania algorytmów, przede wszystkim odmienności podejścia do badanego problemu, a także porównania ich skuteczności w wykrywaniu stanów alarmowych, z uwzględnieniem minimalizacji fałszywych alarmów.

Do testowania algorytmów na danych o zachorowaniach na grypę w Polsce wykorzystano pakiet statystyczny R (www.r-project.org). Skorzystano z funkcji pakietu `surveillance` (www.cran.r-project.org/web/packages/surveillance) [8].

W rozdziale 1 przedstawiona jest teoria dotycząca działania algorytmów CUSUM, RKI i Modelu Farringtona. W kolejnych rozdziałach zamieszczono opis wykorzystanych funkcji pakietu R oraz szczegółową analizę wyników zastosowania algorytmów dla danych.

Rozdział 1

Teoria

1.1. CUSUM

1.1.1. Opis algorytmu

Algorytm CUSUM (ang. cumulative sum) należy do klasy algorytmów bazujących na kontroli procesów statystycznych (SPC) [8]. Po raz pierwszy został przedstawiony przez E. S. Page w 1954 r. [12]. Technika ta wykrywa odstępstwa od bazowych wskaźników w systemie monitoringu. W odróżnieniu od konwencjonalnych metod, CUSUM zezwala na wykrycie niewielkich, nagłych zmian. W minionych dekadach użyto algorytmu CUSUM do wykrywania epidemii salmonellozy [10], grypy [7, 26, 17, 6], a także do zmian w zanieczyszczeniu powietrza [1], zaś inspiracją do zastosowań w badaniach medycznych była analogia do wykrywania aberracji w procesach produkcji przemysłowych [12].

W klasycznym modelu zakładamy, że obserwacje w okresie niealarmowym są w przybliżeniu próbkami iid z rozkładu Poissona $y_1, ..., y_n \sim Pois(m)$, gdzie m jest średnią liczbą obserwowanych zdarzeń (aberracji technicznych, zachorowań, zgonów..., ze względu na temat pracy w dalszej części będziemy się odnosić do zdarzeń jako do zachorowań na grypę).

Warto zaznaczyć, że większość publikacji dotyczących algorytmu CUSUM bierze pod uwagę tylko tzw. algorytm jednostronny, czyli zmianę średniej w jednym kierunku. W praktycznych zastosowaniach interesuje nas zwykle wzrost średniej zachorowań, podczas gdy zmiany w odwrotnym kierunku nie bierzemy pod uwagę. Jeżeli w kolejnych okresach czasu liczba zachorowań ulegnie takiej zmianie, że spowoduje odpowiednio duży wzrost średniej liczby zachorowań, sygnalizowany jest alarm.

Standardowy algorytm CUSUM do wykrywania zmian w średniej rozkładu wykorzystuje statystykę:

$$S_t = max(0, S_{t-1} + (y_t - k)),\ t = 1, .., n, \tag{1.1}$$

gdzie $S_0 = 0$.

Definicja 1.1.1 *Wartość odniesienia k - ustalona wartość referencyjna, o którą są redukowane obserwacje.*

Definicja 1.1.2 *Wartość alarmowa h - ustalona wartość, której przekroczenie przez statystykę $S_t > h$ sygnalizuje stan alarmowy.*

W chwili t kolejna suma kumulacyjna jest obliczana jako suma wartości statystyki w chwili $t-1$ oraz nowych obserwacji y_t, zredukowanych o wartość odniesienia k. Możliwe

jest również zmniejszenie się wartości statystyki, w przypadku gdy liczba zachorowań y_t jest mniejsza od wartości odniesienia. Warunek *max* gwarantuje, że statystyka zawsze przyjmuje wartości nieujemne.

W przypadku liczby zachorowań w populacji nie przekraczającej wartości odniesienia k, wykres statystyk CUSUM w czasie jest krzywą poziomą, zaś dla zwiększonej liczby chorób, przekraczającej k, ma dodatni kąt nachylenia. Po przekroczeniu przez wykres wysokości h sygnalizowany jest alarm.

W celu zbadania skuteczności i czułości algorytmu definiujemy:

Definicja 1.1.3 *ARL (average run length) - średni czas oczekiwania na alarm, gdzie wyróżniamy:*

- *ARL_0 - średni czas oczekiwania na pierwszy alarm, gdy średnia w rozkładzie nie uległa zmianie(fałszywy alarm).*
- *ARL_1 - średni czas oczekiwania na pierwszy alarm w stanie alarmowym (gdy średnia w rozkładzie uległa zmianie).*

Podczas tworzenia modelu odzwierciedlającego badane zjawisko, staramy się dobrać wartości k i h tak, aby minimalizować czas do pierwszego alarmu (ARL_1), jednocześnie starając się uniknąć fałszywego alarmu, czyli maksymalizując ARL_0. Za niskie wartości k i h spowodowałyby zbyteczną wrażliwość algorytmu i bezustanną sygnalizację alarmu, zaś zanadto wysokie wartości skutkowałyby za późnym dostrzeżeniem sytuacji alarmowej, co w konsekwencji braku podjęcia odpowiednich działań mogłoby doprowadzić do sytuacji niekontrolowanej i wybuchu epidemii. Więcej o wyborze parametrów k i h patrz: 1.1.4.

1.1.2. Metoda aproksymacyjna

W klasycznym modelu CUSUM, zaproponowanym przez E. S. Page [12], zakładamy, że obserwacje $y_1, ..., y_n$ są realizacjami zmiennej losowej Y o rozkładzie Poissona i średniej m. Jednak klasyczny model ma liczne wady podczas implementacji algorytmu dla danych rzeczywistych, gdy struktura populacji objętej obserwacją, czy jej liczebność ulegają zmianie.

W niniejszej pracy zostanie wykorzystana modyfikacja algorytmu CUSUM, przy użyciu metody aproksymacyjnej, zaproponowanej przez G. Rossi, L. Lampugnani i M. Marchi (1999), której stosowalność dla danych epidemiologicznych została empirycznie udowodniona przez autorów [14].

Dla uzyskanej próbki obserwacji G. Rossi, L. Lampugnani i M. Marchi [14] proponują następującą transformację zmiennej Y do asymptotycznego rozkładu normalnego $N(0,1)$ (ze względu na m dążące do nieskończoności):

$$X = \frac{Y - 3m + 2\sqrt{mY}}{2\sqrt{m}} \tag{1.2}$$

Aby otrzymać zmienną X, korzystamy z dwóch zmiennych pomocniczych X_1 i X_2, gdzie X_1 ma rozkład asymptotycznie normalny $N(0,1)$, przy m dążącym do nieskończoności, jako transformacja Y o średniej m i odchyleniu standardowym $\sqrt{m}$:

$$X_1 = \frac{(Y - m)}{\sqrt{m}}. \tag{1.3}$$

Ponadto, X_2 jest posiadającą własność stabilizacji wariancji pierwiastkową transformacją rozkładu Poissona [14], ze średnią $\sqrt{m}$ i odchyleniem standardowym $\frac{1}{2}$:

$$X_2 = 2(\sqrt{Y} - \sqrt{m}). \tag{1.4}$$

X otrzymujemy jako zwykłą średnią arytmetyczną X_1 i X_2:

$$X = \frac{X_1 + X_2}{2}. \tag{1.5}$$

Po transformacji statystyka CUSUM przybiera postać:

$$S_t = max(0, S_{t-1} + (x_t - k)). \tag{1.6}$$

1.1.3. Modelowanie sezonowe

Do utworzenia modelu do monitorowania liczby zachorowań potrzebne jest oszacowanie parametru m, średniej w rozkładzie Poissona, z którego w założeniu pochodzą obserwacje. Średnia ta może być oszacowana na podstawie danych empirycznych z obserwacji z lat wcześniejszych, o których zakładamy, że pochodzą z okresu bez epidemii i stanowią model odniesienia. W celu uniknięcia zawyżonej średniej stosuje się np. ucinanie ustalonej liczby obserwacji o największej ilości zachorowań, jednak takie podejście ma liczne wady.

O wiele skuteczniejszym podejściem jest uzmiennienie parametru m, co zezwoli na uwzględnienie specyficznych aspektów populacji objętej badaniem, a w szczególności zmian w strukturze populacji, czy zmiany rozmiaru populacji w czasie, gdy dla różnych obserwacji y_t rozważa się różne m_t.

Do przybliżenia m_t dla algorytmu CUSUM sugeruje się w [9] skorzystanie z modelu regresji Poissona opartego na uogólnionych modelach liniowych (GLM).

Model regresji Poissona

Do modelowania zmian w średniej m_t w czasie, w okresie bez epidemii, proponowany jest następujący sezonowy model Poissona [8]:

$$log(m_t) = \alpha_0 + \sum_{s=1}^{S} \left(\alpha_s cos\left(\frac{2\pi}{T} st\right) + \alpha_s sin\left(\frac{2\pi}{T} st\right)\right), \tag{1.7}$$

gdzie S jest liczbą fal harmonicznych do uwzględnienia w modelu (przy zastosowaniach do zachorowań na grypę zwykle się przyjmuje $S = 1$), T jest okresem pomiaru (np. dla cotygodniowych badań $T = 52$), a α_i parametrami regresji do estymowania. Estymacja parametrów została opisana w podrozdziale: 1.9.

Zastosowanie tej modyfikacji algorytmu jest uwzględnione w pakiecie *surveillance* programu R [8].

1.1.4. Wybór odpowiednich parametrów h i k

Do konstrukcji systemu monitorującego, posługującego się algorytmem CUSUM, oprócz specyfikacji średniej liczby zdarzeń m w okresie bez epidemii, potrzebna jest również specyfikacja dwóch parametrów: wartości odniesienia k oraz wartości alarmowej h.

Wartości k i h są dobrane tak, by minimalizować czas do pierwszego alarmu (ARL_1), jednocześnie starając się uniknąć fałszywego alarmu, czyli maksymalizując ARL_0. Parametr h jest najczęściej empirycznie dopasowywany tak, by gwarantować stały, zdefiniowany przez użytkownika, wskaźnik fałszywych alarmów.

Z powodu braku matematycznych algorytmów na wybór optymalnych wartości parametrów k i h proponowane są alternatywne podejścia. W literaturze są dostępne tabele wartości k i h dla danych ARL_0 i ARL_1 otrzymane za pomocą aproksymacji (patrz: Ewan, Kemp [3], Montgomery [11]). Przykładowo dla $ARL_0 = 500$ i $ARL_1 = 7$ estymowane są wartości $k = 0.6$ i $h = 3.8$.

Rogerson i Yamada [14] sugerują dobór stałej k jako połowy sumy średnich z okresu bezalarmowego i z okresu alarmowego. Jednak w standardowych badaniach monitoringowych oczekujemy, że liczba zachorowań w stanie alarmowym będzie rosła z biegiem czasu, zatem teoria nie określa sposobu optymalnego wyboru parametru k.

W zastosowaniach praktycznych często się wykorzystuje również symulację danych, by otrzymać "optymalne" wartości k i h dla danej choroby (patrz: [13]). Podczas zastosowań modyfikacji aproksymacyjnej algorytmu CUSUM dla wykrywania stanów alarmowych zachorowań na grypę i choroby grypopodobne symulacyjnie wyliczono [16], że optymalną wartością k dla chorób tego typu jest wartość z przedziału 0.5 do 2.5, zaś h z przedziału od 2 do 4. W [6] dla grypy sugerowany jest wybór $k = 0.5$ lub 1.5 dla standardowego algorytmu CUSUM.

1.1.5. Wady algorytmu

Długa regeneracja po stanie alarmowym

Statystyka CUSUM zależy bezpośrednio od wcześniejszych sum skumulowanych, przez co ewentualne wystąpienie stanu alarmowego w zeszłym tygodniu będzie miało wpływ na zawyżoną sumę skumulowaną w bieżącym tygodniu. Tego wynikiem jest wydłużony czas regeneracji, czyli powrotu do stanu normalnego, sumy CUSUM. Jak wynika z przykładu ilustrującego zastosowanie CUSUM przy wykrywaniu stanów alarmowych zachorowań na Hepatitis A, na podstawie danych Szwedzkiego Instytutu Kontroli Chorób Zakaźnych (SMI) [13] [33], okres regeneracji algorytmu wyniósł odpowiednio 10 tygodni i rok dla dwóch zasygnalizowanych alarmów na początku 1997r. i w drugiej połowie 1997r. (patrz: rys. 1.1.5). W czasie regeneracji algorytmu nie było możliwości ocenienia, czy sygnalizowany alarm jest właściwy z powodu zawyżonych wartości statystyki CUSUM.

1.2. Model Farringtona

Model Farringtona został zaproponowany przez Farringtona, Andrewsa, Bealy'ego i Catchpole'a w roku 1996 [5]. Wykorzystywany jest on w praktyce w Wielkiej Brytanii (Anglii i Walii) między innymi przez CDSC (*Communicable Disease Surveillance Center*) w Walii. Ponieważ model Farringtona znalazł praktyczne zastosowanie w Wielkiej Brytanii nazywa się go często modelem angielskim (ang. *English model*).

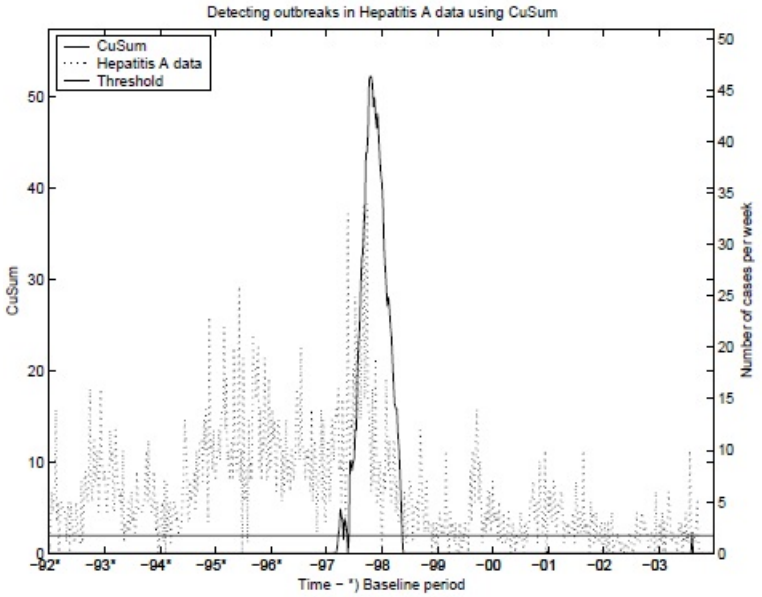
1.2.1. Postać i estymacja współczynników modelu

Postać modelu

Model Farringtona zakłada, że obserwacje są niezależne oraz pochodzą z rozkładu Poissona z parametrem dyspersji ϕ. Przy założeniu, że y_i jest bazową liczbą wystąpień w czasie t_i badanego zdarzenia, postać modelu Farringtona można w ogólności zapisać w sposób następujący:

$$E(y_i) = \mu_i$$

$$\log \mu_i = \alpha + \beta t_i$$



Rysunek 1.1: Wykrywanie stanów alarmowych na podstawie danych o zachorowaniach na Hepatitis A (1992-2003) [13]

$$Var(y_i) = \phi\mu_i, \tag{1.8}$$

gdzie α i β są parametrami modelu zaś ϕ parametrem dyspresji. Zauważyć należy, że rozkład zmiennej y_i należy do wykładniczej rodziny rozkładów. Ponadto funkcja $\log x$ jest funkcją różnowartościową klasy C^∞ na przedziale $(0, \infty)$ oraz związek między $log\mu_i$ jest w modelu 1.8 kombinacją liniową zmiennych objaśniających. W związku z tym powyższy model zaliczyć można do klasy uogólnionych modeli liniowych (GLM, ang. *Generalized Linear Model*). Szczegóły dotyczące uogólnionych modeli liniowych znaleźć można w podręczniku Farawaya [4].

Estymacja współczynników modelu

Do estymacji parametrów uogólnionych modeli liniowych bardzo często wykorzystuje się iteracyjnie ważoną metodę najmniejszych kwadratów IRWLS (IRWLS, ang. *Iteratively Re-Weighted Least Squares*). Stosowana jest w większości pakietów statystycznych. Szczegółowy opis algorytmu w pełnej ogólności można znaleźć w [4] oraz w [19]. W skrócie można go zapisać w sposób następujący. Niech będzie dany model:

$$E(y_i) = \mu_i$$

$$g(\mu_i) = \eta_i = \alpha + \sum_{k=1}^{p-1} \beta_k x_{ik}$$

$$Var(y_i) = \phi v(\mu_i), \tag{1.9}$$

gdzie g jest funkcją monotoniczną i różniczkowalną, x_{ik} to wartość k-tej zmiennej objaśniającej dla i-tej obserwacji, a v pewną funkcją. Ponadto $i = 1, \ldots, n$. Dla powyższego modelu procedura IRWLS wygląda następująco:

1. Obliczenie początkowych wartości $\hat{\boldsymbol{\eta}}_{\mathbf{0}}$ oraz $\hat{\boldsymbol{\mu}}_{\mathbf{0}}$ przy pomocy metody najmniejszych kwadratów.

2. Stworzenie zmiennej pomocniczej $\mathbf{z_0} = (z_{01}, \ldots, z_{0n})$, gdzie $z_{0i} = \hat{\eta}_{0i} + (y_i - \hat{\mu}_{0i})\eta_i'(\hat{\mu}_{0i})$, $\eta_i'(\hat{\mu}_{0i})$ oznacza pochodną w punkcie $\hat{\mu}_{0i}$.

3. Stworzenie wag postaci:
$$\omega_{0i} = \frac{1}{(\eta_i'(\hat{\mu}_{0i}))^2 \hat{\phi} v(\hat{\mu}_{0i})} \tag{1.10}$$
i zdefiniowanie macierzy $\mathbf{W} = diag[\omega_{01}, \ldots, \omega_{0n}]$.

4. Przy pomocy estymatora metody ważonych najmniejszych kwadratów otrzymujemy estymator parametrów:
$$\hat{\boldsymbol{\Gamma}} = (\mathbf{X'WX})^{-\mathbf{1}}\mathbf{X'Wz},$$
gdzie $\mathbf{W}$ jest macierzą diagonalną złożoną z wag, $\boldsymbol{\Gamma}$ wektorem wszystkich parametrów części liniowej modelu 1.9, zaś $\mathbf{X}$ macierzą obserwacji. Stąd wyliczamy $\hat{\boldsymbol{\eta}}$ oraz $\hat{\boldsymbol{\mu}}$ w celu wykorzystania ich w kolejnych iteracjach.

5. Kroki 2, 3 i 4 powtarzamy iteracyjnie z wykorzystaniem $\hat{\boldsymbol{\eta}}$ i $\hat{\boldsymbol{\mu}}$ (w miejscu $\hat{\boldsymbol{\eta}}_{\mathbf{0}}$ i $\hat{\boldsymbol{\mu}}_{\mathbf{0}}$) z poprzedniego punktu aż różnica wyników między kolejnymi iteracjami będzie mała.

W przypadku dodatkowej specyfikacji wag w modelu należy we wzorze 1.10 umieścić je w liczniku [19]. Zauważyć należy, że w celu otrzymania modelu Farringtona we wzorze 1.9 należy wziąć $g(\mu_i) = log\mu_i$, $p = 2$, $x_{i1} = t_i$ oraz $v(\mu_i) = \mu_i$.

We wzorze 1.10 pojawia się także estymator współczynnika dyspersji $\hat{\phi}$. Faraway [4] zaznacza, że nie da się go estymować, wykorzystując podejście oparte na funkcji wiarogodności. W związku z tym proponowanych jest kilka wersji estymatora ϕ. Najczęściej używaną jest [4]:
$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \tag{1.11}$$
gdzie $\hat{\mu}_i = g^{-1}(\alpha + \sum_{i=k}^{p-1}\beta_i x_{ik})$.

W przypadku modelu 1.8, Farrington zakłada możliwość wystąpienia jedynie naddyspersji, w związku z czym estymator ϕ przyjmuje postać:
$$\hat{\phi} = \max\left\{\frac{1}{n-p}\sum_{i=1}^{n}\omega_i\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, 1\right\}, \tag{1.12}$$
gdzie ω_i oznacza wagi, zaś $p = 2$ lub $p = 1$ w zależności od tego, czy estymowany model ma uwzględniać trend liniowy. Ponadto n to nie jest liczba wszystkich obserwacji, a jedynie niektóre z nich. Dokładne wyjaśnienie liczby n oraz wag dokonane zostanie w dalszej części pracy.

Niech teraz t_0 oznacza bieżący moment, zaś y_0 liczbę zdarzeń w okresie. Wtedy oczekiwaną liczbę zdarzeń szacuje się w następujący sposób:
$$\hat{\mu}_0 = \exp(\hat{\alpha} + \hat{\beta}t_0), \tag{1.13}$$
gdzie $\hat{\alpha}$ oraz $\hat{\beta}$ są estymatorami odpowiednich parametrów w modelu 1.8.

1.2.2. Trend i sezonowość w modelu

Model Farringtona umożliwia zastosowanie w modelu liniowego trendu oraz uwzględnienie sezonowości, co jest jednym z głównych warunków przydatności modelu wykrywającego stany alarmowe liczby zachorowań na choroby. W niniejszym podrozdziale obowiązują oznaczenia wykorzystane we wcześniejszej części pracy.

Trend

Liniowy trend względem $\log(\mu_i)$ wynika bezpośrednio z postaci modelu 1.8, a odpowiada za niego współczynnik β. Jeżeli $\beta = 0$, to oznacza, że w specyfikacji modelu nie uwzględniono trendu, jeżeli zaś $\beta \neq 0$, to trend w modelu występuje i w zależności od znaku współczynnika β jest on dodatni lub ujemny.

W modelu Farringtona uwzględnia się liniowy trend tylko w przypadku, gdy zachodzą następujące warunki:

1. długość historycznego szeregu czasowego jest większa niż 3 lata,
2. współczynnik β jest istotny statystycznie,
3. $\mu_0 \leqslant \max\{y_i : i = 1, \ldots, n\}$.

Sezonowość

Model Farringtona pozwala na uwzględnienie sezonowości poprzez szacowanie progów alarmowych na podstawie obserwacji z porównywalnych okresów roku. Dla przykładu niech poszczególne obserwacje będą oznaczały liczbę zdarzeń dla poszczególnych tygodni w roku. Liczbę obserwacji wziętych pod uwagę w trakcie modelowania wyznacza się na podstawie b -liczby poprzednich lat oraz w - tzw. „szerokości okna". Niech teraz bieżącym tygodniem będzie tydzień x roku y. Wtedy w analizie wykorzystane zostaną obserwacje z tygodni od $x - w$ do $x + w$ z lat od $y - b$ do $y - 1$. Daje to łączną liczbę obserwacji:

$$n = b(2w + 1).$$

Pozostaje pytanie, jak dobrać b i w. Zwiększając n, zwiększana jest precyzja modelu, jednakże łączy się to ze zwiększeniem w i/lub b. W pierwszej sytuacji może wystąpić problem z uchwyceniem wahań sezonowych, podczas gdy zbyt duża wartość b zmusza do porównywania bieżących wyników z danymi z odległych lat, co uniemożliwi uwzględnienie pewnych zachodzących tendencji. W związku z tym, powyższe wartości należy ustalać indywidualnie dla każdego badania, biorąc pod uwagę analizowane zjawisko i rodzaj danych poddanych analizie.

1.2.3. Próg

W celu ustabilizowania prawdopodobieństwa fałszywego alarmu Farrington et al.[5] zaproponowali transformację danych, podnosząc zmienne do potęgi $\frac{2}{3}$. W konsekwencji zaproponowano przybliżony przedział ufności dla y_0 na poziomie $1 - 2\alpha$:

$$U = \hat{\mu}_0 \left(1 + \frac{2}{3} z_\alpha \left(\frac{\hat{\tau}}{\hat{\mu}_0}\right)^{1/2}\right)^{3/2} \tag{1.14}$$

$$L = \max\left\{\hat{\mu}_0\left(1 - \frac{2}{3}z_\alpha\left(\frac{\hat{\tau}}{\hat{\mu}_0}\right)^{1/2}\right)^{3/2}, 0\right\}, \tag{1.15}$$

gdzie U i L to odpowiedni górny i dolny koniec przedziału ufności, zaś z_α jest kwantylem rzędu α ze standardowego rozkładu normalnego. Ponadto we wzorach 1.14 i 1.15 $\hat{\tau}$ definiowane jest w sposób następujący:

$$\hat{\tau} = \hat{\phi} + \frac{Var(\hat{\mu}_0)}{\hat{\mu}_0}.$$

Wyprowadzenie powyższych przedziałów znaleźć można w [5].Warto jedynie dodać, że wykonanie powyższej transformacji pozwala na ustabilizowanie prawdopodobieństwa fałszywego alarmu na poziomie α, co zostało potwierdzone empirycznie przez Farringtona et al.

Zaznaczyć należy, że w przypadku analizy danych epidemiologicznych, ograniczyć się można do górnego przedziału ufności. Jednakże w wielu innych dziedzinach (np. przemyśle) nie trudno znaleźć procesy, w których istotne będzie przekorczenie dolnego progu lub którego-kolwiek z opisanych progów.

1.2.4. Wagi

Istotnym zagadnieniem w przedstawionym podejściu jest redukcja wpływu przeszłych obserwacji, dla których wystąpiła epidemia bądź stan alarmowy. W związku z tym w modelu Farringtona postanowiono wprowadzić wagi dla obserwacji ω_i. Przy pomocy pierwotnych estymatorów μ_i oraz współczynnika dyspersji oszacowanego w przypadku gdy $\omega_i = 1$ dla każdego i, zdefiniowane zostały reszty s_i:

$$s_i = \frac{3}{2\hat{\phi}^{1/2}} \frac{y_i^{2/3} - \hat{\mu}_i^{2/3}}{\hat{\mu}_i^{1/6}(1 - h_{ii})^{1/2}} \tag{1.16}$$

Ostatecznie wagi definiowane są w następujący sposób:

$$\omega_i = \begin{cases} \gamma s_i^{-2} & \text{jeżeli } s_i < 1 \\ \gamma & \text{w przeciwnym przypadku,} \end{cases} \tag{1.17}$$

gdzie γ jest stałą dobraną tak by $\sum_{i=1}^{n} \omega_i = n$, zaś h_{ii} elementami diagonalnymi macierzy projekcji:

$$\mathbf{H} = \mathbf{W^{1/2}X(X'WX)^{-1}X'W^{1/2}}. \tag{1.18}$$

W powyższym wzorze $\mathbf{X}$ jest macierzą obserwacji, a $\mathbf{W}$ jest macierzą diagonalną składającą się z wag ω_{ei}wykorzystywanych w algorytmach umożliwiających estymację modelu, np. dla metody *IRWLS* wagi dla omawianego modelu przedstawiono w ogólności poniżej:

$$\omega_{ei} = \frac{\omega_i}{\hat{\phi}\hat{\mu}_i(\log(\hat{\mu}_i)')^2} = \frac{\omega_i\hat{\mu}_i}{\hat{\phi}} \tag{1.19}$$

Dobór wag ω_i i reszt s_i został dokonany przez autora modelu w sposób empiryczny przy założeniu, że obserwacje z dużymi resztami powinny otrzymać bardzo małe wagi.

1.2.5. Algorytm

W poniższym podrozdziale przedstawiona zostanie algorytm wykrywania stanów alarmowych przy pomocy modelu Farringtona.

W pierwszej kolejności szacowany jest wyjściowy model (przy założeniu, że wagi $\omega_i = 1$), dzięki, któremu otrzymuje się wyjściowe wartości estymatorów $\hat{\mu}_i$ oraz $\hat{\phi}$. Następnie obliczane są wagi zgodnie ze wzorami z podrozdziały 1.2.4 oraz model 1.8 jest reestymowany. Stąd otrzymuje się nowy estymator współczynnika dyspersji $\hat{\phi}$. Kolejnym krokiem jest sprawdzenie istotności statystycznej trendu w modelu 1.8. Jeżeli okazałby się on nieistotny wszystkie poprzednie czynności muszą zostać powtórzone, tym razem bez uwzględnienia trendu. Ostatecznie oblicza się wartość progu alarmowego.

W przypadku, gdy liczba wystąpień badanego zjawiska przekracza górną granicę U, mówi się o wystąpieniu alarmu. W przeciwnym przypadku nie odnotowuje się alarmu. Odpowiednioą w przypadku procesów, w których bada się, czy dane zjawisko nie miało miejsca zbyt rzadko powiedzieć można, że alarm wystąpił, jeżeli liczba wystąpień badanego zjawiska nie przekroczyła dolnej granicy L.

1.3. Metoda RKI

Metoda RKI jest bardzo prostą i intuicyjną metodą statystyczną do wykrywania stanów alarmowych. Nazwa RKI wzięła się od Instytutu Roberta Koch (niem. *Robert Koch Institut*), w którym metoda ta jest wykorzystywana [2]. *Robert Koch Institut* jest niemiecką instytucją federalną (częścią Federalnego Ministerstwa Zdrowia) odpowiedzialną za kontrolę i prewencję zachorowań m.in. na choroby zakaźne.

Pierwszym etapem w metodzie RKI jest wyznaczenie obserwacji odniesienia. Często wykorzystuje się (podobnie jak w przypadku modelu Farringtona) obserwacje z przeszłych lat z okresów bliskich analizowanemu okresowi. Granica alarmu szacowana jest na dwa sposoby. Jeżeli średnia z obserwacji odniesienia jest większa niż 20, jako granicę alarmu przyjmuje się średnią plus dwa odchylenia standardowe. W przeciwnym przypadku wykorzystuje się wybrany kwantyl rozkładu Poissona z parametrem równym średniej z obserwacji odniesienia [2]. W związku z tym wzór na granicę alarmu można zapisać w sposób następujący:

$$U = \begin{cases} \bar{x} + 2s & \text{jeżeli } \bar{x} > 20 \\ \delta_\alpha & \text{w przeciwnym przypadku,} \end{cases} \tag{1.20}$$

gdzie $\bar{x}$ jest średnią z obserwacji odniesienia, s odchyleniem standardowym z obserwacji odniesienia, zaś δ_α najmniejszym kwantylem spełniającym warunek:

$$P\{\delta \geqslant \delta_\alpha\} \leqslant \alpha$$

$$\delta \sim Poiss(\bar{x}) \tag{1.21}$$

Alarm jest sygnalizowany w przypadku, gdy bieżąca liczba zachorowań przekracza obliczoną granicę alarmu.

Rozdział 2

Funkcje pakietu R do analizy stanów alarmowych

W poniższym rozdziale przedstawione zostaną funkcje pakietu statystycznego R służące do estymacji modeli oraz wykonywania algorytmów przedstawionych w rozdziale 1. Funkcje te znajdują się w pakiecie `surveillance`. Do niektórych funkcji dodano przykłady. Przy ich konstruowaniu wykorzystano dane wbudowane w pakiet `surveillance` o nazwie *salmonella.agona*. Są to tygodniowe dane opisujące liczbę przypadków salmonelli w Wielkiej Brytanii w latach 1990-1995. Zbiór składa się z 312 obserwacji. W niniejszym rozdziale celowo wykorzystano dane, które nie będą wykorzystane w dalszej analizie, ponieważ główny nacisk tego rozdziału ma być położony na aspektach technicznych modelowania, a nie na uzyskanych wynikach. Szczegółową analizę poruszanego w pracy tematu badawczego zamieszczono w rozdziale 3.

2.1. Tworzenie obiektu klasy `disProg`

W pakiecie `surveillance` podstawową zmienną wejściową funkcji jest obiekt klasy `disProg`. Przechowuje on m.in. dane na temat obserwowanej częstości zdarzeń w poszczególnych okresach, informację o tym, czy w rzeczywistości miał miejsce alarm oraz częstotliwości pojawiania się raportów. Podstawowym narzędziem do tworzenia obiektów klasy `disProg` jest funkcja `create.disProg`. Na podstawie danych wejściowych tworzy ona obiekt wspomnianej klasy, który może być przetwarzany przez funkcje pakietu `surveillance`. Składnia funkcji `create.disProg` przedstawiona jest poniżej, zaś opis ważniejszych argumentów znaleźć można w tabeli 2.1.

```
create.disProg(week, observed, state, start=c(2001,1), freq=52,
neighbourhood=NULL, populationFrac=NULL, epochAsDate=FALSE)
```

2.2. CUSUM

W niniejszym podrozdziale opiszemy funkcję:

`algo.cusum`

występującą w pakiecie programu R `surveillance` do implementacji algorytmu CUSUM, bazującą na metodzie aproksymacyjnej [14].

Kod wywołujący funkcję `algo.cusum` w programie R ma postać:

Tabela 2.1: Opis argumentów funkcji `create.disProg`.

Argument	Opis
`week`	indeks dla macierzy obserwacji (zazwyczaj wyrażony w tygodniach)
`observed`	macierz obserwacji dla poszczególnych jednostek
`state`	macierz stanów alarmowych, przyjmujących wartości 0 i 1
`start`	dwuelementowy wektor: rok i numer pierwszej obserwacji
`freq`	częstotliwość danych w skali roku, np. 52 dla danych tygodniowych, 12 dla danych miesięcznych, itd.
`neighbourhood`	macierz, mówiąca o tym, czy jednostki badania są sąsiadujące, np. regiony kraju

```
algo.cusum(disProgObj, control = list(range = range, k = 1.04, h = 2.26,
m = NULL, trans = "standard"),
```

gdzie podstawowym argumentem jest obiekt klasy `disProgObj` opisany w podrozdziale 2.1, zawierający wektor wszystkich obserwacji. Przy pomocy argumentu `range` ustalamy punkty w czasie, które zamierzamy monitorować. Wybieramy również odpowiednią wartość odniesienia k i wartość alarmową h.

Argument m odpowiada za sposób estymowania średniej zachorowań m w okresie bez epidemii. Przyjmuje wartości:

- `numeric` - wektor ustalonych wartości średniej m dla różnych punktów czasowych. Powinien być tej samej długości co `range`.
- `NULL` - na podstawie wszystkich obserwacji, z wykluczeniem badanych, tzn. obserwacje poprzedzające pierwszą wartość z zakresu `range`, obliczana jest średnia, która się nie zmienia w czasie.
- `"glm"` - do obserwacji dopasowywany jest model GLM, o domyślnej dla funkcji postaci:

$$log(m_t) = \alpha_0 + \sum_{s=1}^{S} \alpha_s(sin\frac{2\pi}{T}st) + \alpha_s cos(\frac{2\pi}{T}st),$$

  z wartością domyślną $S = 1$, gdzie T jest częstością obserwacji, odczytaną z argumentu `freq` obiektu klasy disProgObj.

Możliwy jest również wybór wykorzystywanej transformacji do zmiennej o rozkładzie normalnym, za pomocą argumentu `trans`.

- `trans="standard"` jest standartową transformacją, bazującą na asymptotycznej normalności zmiennej o średniej m i odchyleniu standartowym $\sqrt{m}$.
- `trans="rossi"` jest opisaną wcześniej (podrozdział 1.1.2) transformacją zaproponowaną przez G. Rossi, L. Lampugnani i M. Marchi [14].

Wynikiem działania funkcji jest obiekt klasy `survRes` zawierający wektor wartości alarmowych dla punktów czasowych z zakresu `range` oraz wektor sum skumulowanych w danych punktach czasowych z zakresu. Wartość `upperbound` wskazuje wyliczoną liczbę zachorowań w danym punkcie czasowym, przy której zostałby zasygnalizowany alarm. Po zasygnalizowaniu alarmu statystyka nie jest resetowana, więc alarm jest wciąż sygnalizowany aż do momentu

powrotu wartości CUSUM poniżej granicy `upperbound`. Za pomocą `control$m.glm` możemy wywołać dopasowany model GLM, jeżeli w funkcji użyto argumentu `glm`.

Skrótowy opis argumentów funkcji jest przedstawiony w tabeli 2.2.

Tabela 2.2: Opis argumentów funkcji `algo.cusum`

Argument	Opis
`disProgObj`	obiekt klasy `disProg`.
`control(range)`	ustala punkty czasowe do ewaluacji.
`control(k)`	wartość odniesienia
`control(h)`	wartość alarmowa.
`control(m)`	średnia oczekiwana liczba zachorowań.
`control(trans)`	jedna z możliwych transformacji danych do rozkładu normalnego.

Przykład zastosowania algorytmu CUSUM przy użyciu danych salmonella.agona

Dane dotyczące *salmonella.agona* są cotygodniowe, zatem przyjmujemy w przykładowym modelu `freq=52`, obejmują 6-letni okres czasu, ogółem 312 obserwacji. Do wykrycia epidemii posłużymy się modelem GLM estymowanym na podstawie danych z pierwszych 100 obserwacji. Do graficznej ilustracji danych posłużymy się wbudowaną funkcją `plot(survRes)`.

Kod 2.1: Analiza danych `salmonella.agona` przy pomocy algorytmu CUSUM.

```
data(salmonella.agona)
n <- length(salmonella.agona$observed)
res <- algo.cusum(salmonella.agona, control = list(range = 100:n, k=1.04,
 h=2, m="glm", trans="rossi"))
plot(res)
plot(res$cusum)
control$m.glm
```

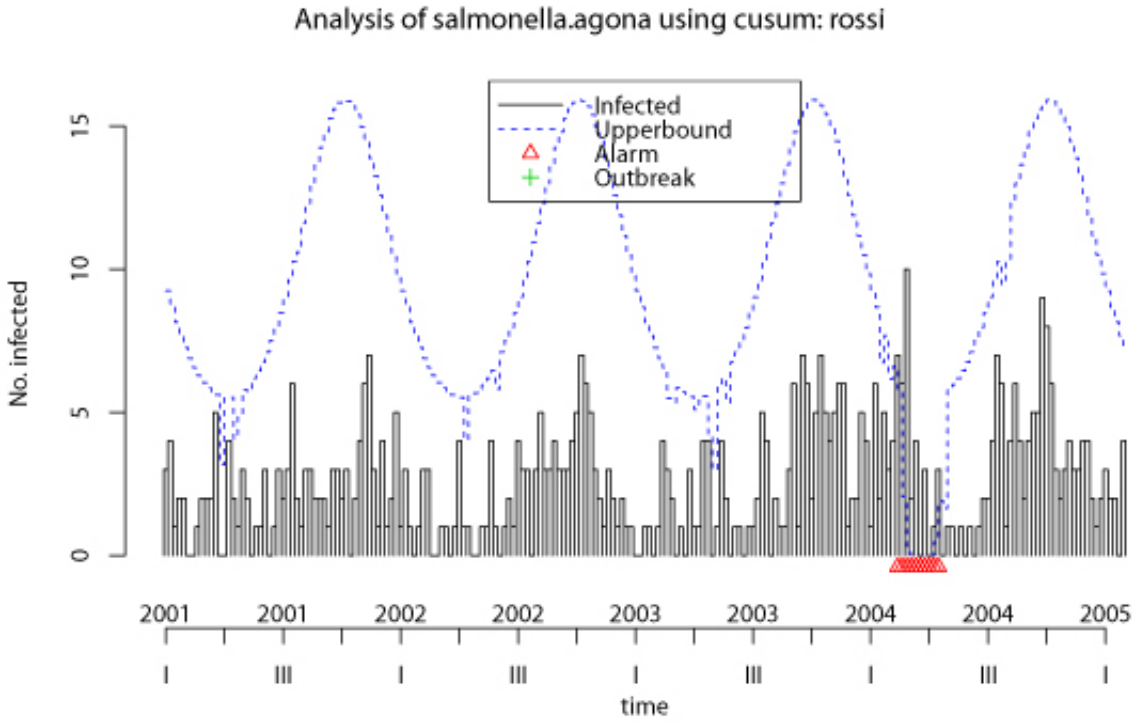
Na rysunku 2.1 widoczne jest zasygnalizowanie alarmu w 2004 roku. Alarm jest sygnalizowany po przekroczeniu przez liczbę zachorowań wartości `upperbound`, która jest wyliczana korzystając ze statystyki CUSUM. W przypadku wystąpienia epidemii są one zaznaczone symbolem `outbreak`.

2.3. Model Farringtona

Do estymacji modelu Farringtona w programie R służy funkcja `algo.farrington` pakietu `surveillance`. Dane wykorzystywane w analizie muszą być zapisane jako obiekt klasy disProg. Pełna specyfikacja funkcji przedstawiona jest poniżej:

```
algo.farrington(disProgObj, control=list(range=NULL, b=3, w=3,
reweight=TRUE, verbose=FALSE, alpha=0.01, trend=TRUE, limit54=c(5,4),
powertrans="2/3", fitFun=c("algo.farrington.fitGLM")))
```

Argumenty funkcji `algo.farrington` pozwalają na dużą kontrolę własności modelu Farringtona. Po pierwsze przy pomocy opcji `b` oraz `w` można dowolnie wybierać zakres obserwacji odniesienia. Ponadto analizie poddać można tylko część dostepnego szeregu czasowego bez



Rysunek 2.1: Wykrywanie stanów alarmowych dla danych `salmonella.agona` przy użyciu algorytmu CUSUM.

konieczności przetwarzania danych. W razie konieczności funkcja `algo.farrington` umożliwia estymację modelu bez przeważania obserwacji (patrz podrozdział 1.2.4), ustawiając `reweight` = `FALSE`. Ważną funkcjonalnością opisywanej funkcji jest możliwość ustawienia dowolnego poziomu istotności dla przedziałów ufności, wyłączenia trendu (`trend` = `FALSE`) oraz ustawienia minimalnej liczby zdarzeń, dla której może wystąpić alarm przy pomocy argumentu `limit54`. Przykładowo, jeżeli uzna się, że alarm nie może wystąpić, gdy łącznie przez ostatnie 4 okresy badania (włącznie z obecnym) wystąpiło mniej niż 5 zdarzeń to deklaracja argumentu `limit54` wyglądałaby w sposób następujący `limit54` $= c(4, 5)$. Zestawienie i krótkie wyjaśnienie wszystkich argumentów funkcji `algo.farrington` przedstawiono w tabeli 2.3.

Wynikiem działania funkcji `algo.farrington` jest obiekt klasy `survRes`. Opis obiektu klasy `survRes` zamieszczono w podrozdziale 2.5.

Przykładowo w celu analizy danych `salmonella.agona` przy pomocy modelu Farringtona wykorzystano kod na podstawie [8].

Kod 2.2: Analiza danych `salmonella.agona` przy pomocy modelu Farringtona.

```
n <- length(salmonella.agona\$observed)
control <- list(b=4,w=3,range=(n-100):n,reweight=TRUE,
                verbose=FALSE,alpha=0.01)
res <- algo.farrington(salmonella.agona,control=control)
plot(res,disease="Salmonella␣Agona",method="Farrington")
```

Druga linia kodu tworzy listę z argumentami wykorzystywanymi w funkcji `algo.farrington`, mianowicie obserwacje odniesienia mają być brane z ostatnich czterech lat, z każdego roku po

Tabela 2.3: Opis argumentów funkcji `algo.farrington`

Argument	Opis
`disProgObj`	obiekt klasy disProg,
`control`	lista elementów kontrolujących,
`range`	indeksy punktów w czasie, dla których przeprowadzana jest analiza,
`b`	liczba lat, z których brane są obserwacje odniesienia,
`w`	szerokość okna (patrz rozdział 1.2),
`reweight`	uwzględnienie przeważenia obserwacji (patrz podrozdział 1.2.4), typ logical,
`trend`	uwzględnienie trendu w modelu (pod warunkiem, że spełnione są odpowiednie założenia (podrozdział 1.2.2)), typ logical,
`verbose`	pokazanie dodatkowych informacji dotyczących bugów,
`powertrans`	rodzaj transformacji użytej do obliczania progu alarmowego,
`alpha`	poziom istotności dla dwustronnego przedziału ufności,
`limit54`	określenie minimalnej liczby zdarzeń w pewnym okresie potrzebnej do wystąpienia alarmu,
`fitFunString`	rodzaj funkcji wykorzystanej do dopasowania modelu.

7 obserwacji $(2w + 1)$, a analizie ma być poddanych 100 ostatnich obserwacji szeregu czasowego. Ponadto algorytm ma wykonywać przeważenie, zaś poziom istotności ustawiony został na $0,01$. W wyniku dwóch kolejnych linii wykonywany jest algorytm, wyniki są przypisywane do zmiennej `res` i ostatecznie uzyskiwany jest wykres z wynikami analizy.

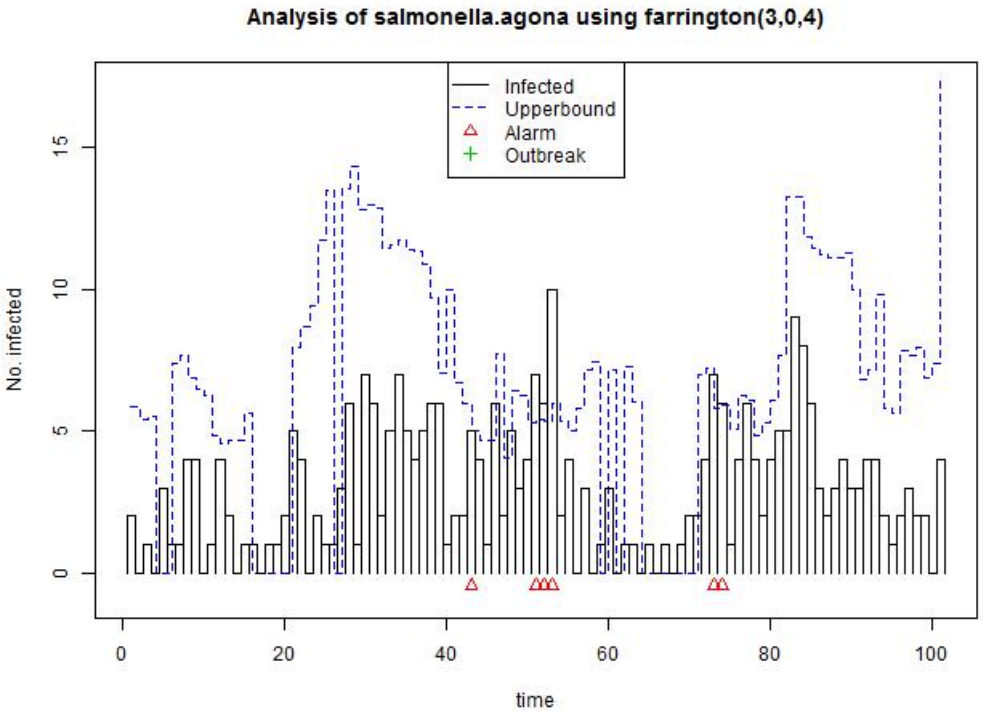
Na rysunku 2.2 zaznaczono liczbę zdarzeń w analizowanych okresach, próg alarmowy oraz momenty, w których wykryto alarm (czerwone trójkąty). Tego rodzaju wykres jest najwygodniejszym sposobem analizy uzyskanych wyników.

2.4. Metoda RKI

Do wykrywania stanów alarmowych przy pomocy metody RKI wykorzystać można funkcję `algo.rki` z pakietu `surveillance`. Dane wykorzystywane w analizie muszą być zapisane jako obiekty klasy `disProg`. Ogólną postać funkcji można zapisać w sposób następujący:

```
algo.rki(disProgObj, control = list(range = range, b = 2, w = 4, actY
= FALSE))
```

Argument `range` odpowiada za wybór obserwacji do analizy. Analogicznie jak w przypadku modelu Farringtona argument `b` odpowiada za liczbę poprzednich lat, z których pobierane są obserwacje odniesienia, zaś argument `w` określa tzw. „szerokość okna", czyli dla analizowanego momentu w czasie liczbę obserwacji przed i po analogicznym momencie we wcześniejszych latach. Przykładowo dla ustalonych wartości b i w liczba obserwacji odniesienia



Rysunek 2.2: Wykrywanie stanów alarmowych dla szeregu `salmonella.agona` przy użyciu modelu Farringtona

wynosić będzie $b*(w+1)$. Ostatnim argumentem funkcji `algo.rki` jest `actY`. Przyjmuje one wartość `TRUE`, gdy do obserwacji odniesienia włączone mają być obserwacje z bieżącego roku oraz wartość `FALSE` w przeciwnym przypadku. Argumenty funkcji `algo.rki` przedstawiono także w formie tabelarycznej w tabeli 2.4.

Wynikiem działania funkcji `algo.rki` jest obiekt klasy `survRes`. Warto dodać, że w pakiecie R zaimplementowane są trzy dodatkowe funkcje `algo.rki1`, `algo.rki2` oraz `algo.rki3`. Są to wersje funkcji `algo.rki` z ustalonymi już parametrami `b` i `w`. Ponieważ są one szczególnymi przypadkami funkcji `algo.rki`, w dalszej części paracy rozważana będzie jedynie ta ostatnia.

Wyniki analizy z wykorzystaniem metody RKI najwygodniej przedstawić w formie wykresu. Sposób jego tworzenia jest analogiczny jak w przypadku modelu Farringtona, dlatego zostanie on pominięty w tym miejscu.

2.5. Obiekt klasy `survRes`

Wynikiem działania przedstawionych funkcji jest obiekt klasy `survRes`. Obiekt tej klasy rozumieć można jako listę zawierającą nie tylko wyniki funkcji, ale także zadeklarowane wartości argumentów wykonanej funkcji oraz obiekt wejściowy klasy `disProg`. Najważniejszymi elementami obiektu klasy `survRes` są:

- wektor o wartościach 1 lub 0 informujący, czy w danym momencie czasu zgodnie z algorytmem wykryto alarm,
- wektor progów alarmowych,

Tabela 2.4: Opis argumentów funkcji `algo.rki`

Argument	Opis
`disProgObj`	obiekt klasy disProg,
`control`	lista elementów kontrolujących,
`range`	indeksy punktów w czasie, dla których przeprowadzana jest analiza,
`b`	liczba lat, z których brane są obserwacje odniesienia,
`w`	„szerokość okna" - analogicznie jak w modelu Farringtona (patrz 1.2),
`actY`	uwzględnienie w zbiorze obserwacji odniesienia obserwacji z bieżącego roku.

- wszystkie informacje dotyczące wejściowego obiektu klasy `disProg`,
- wartość argumentów wykonanych funkcji `algo.farrington` w liście `control`.

W przypadku funkcji `algo.farringtona` ważnym elementem obiektu `survRes` jest wektor o wartościach 1 lub 0 informujący, czy dla danego momentu czasu algorytm uwzględnił trend.

Obiekt klasy `survRes` jest bardzo wygodnym sposobem przedstawiania wyników funkcji. Pozwala on na stworzenie odpowiednich rysunków ilustrujących wynik funkcji, o czym wspomniano już w poprzednich podrozdziałach. Ponadto umożliwia on pracę na wynikach, np. dodawania warunków koniecznych wystąpienia alarmu, tak jak zrobiono to w podrozdziałach 3.3 i 3.5.

Rozdział 3

Analiza danych rzeczywistych

W niniejszej pracy przyjęte są następujące definicje:

3.1. Charakterystyka epidemii

Definicja 3.1.1 (Epidemia) *Epidemia (gr. epi - nad, demos - lud) - duży i gwałtowny wzrost zachorowalności na daną chorobę w porównaniu do danych historycznych, często cechujący się rozprzestrzeniem w różnych krajach. [35]*

Definicja 3.1.2 (Pandemia) *Pandemia (gr. pan - wszyscy, demos - lud) - epidemia choroby zakaźnej rozprzestrzeniająca na dużym obszarze na kilku kontynentach w tym samym czasie. [35]*

Największe pandemie grypy w XX i XXI wieku odnotowane zostały w 1918r. i 2009r. Pierwsza z nich została spowodowana nasileniem migracji po I wojnie światowej i złymi warunkami sanitarnymi. Zgony skutkowane pandemią są szacowane na 50-100 mln ludzi. Kolejna pandemia została ogłoszona w 2009 r., nazywana też grypą meksykańską lub świnską grypą, przenoszona przez wirus A/H1N1. Odnotowano około 850 tys. zachorowań, w tym ok. 12 tys. zgonów. [35].

Przy wykrywaniu epidemii i pandemii grypy istotna jest jak najwcześniejsza interwencja, ważne jest zatem uwzględnienie w monitoringu epidemiologicznym wyróżnianych przez WHO sześciu faz pandemii:

1. Nie występuje zakażenie człowieka, wirus występuje u zwierząt.
2. Następuje zakażenie człowieka.
3. Początek okresu alarmu pandemicznego. Bardzo ograniczona transmisja z człowieka na człowieka.
4. Zwiększone ryzyko transmisji wirusa z człowieka na człowieka.
5. Poważne ryzyko pandemii.
6. Zwiększona i trwała transmisja wirusa w populacji.

W odróżnieniu od okresów pandemicznych, co roku występuje grypa sezonowa, którą należy uwzględnić przy modelowaniu układów odniesienia. Za sezon grypy w Polsce uważany jest okres od przełomu września i października do końca listopada oraz, po okresie stabilizacji w grudniu i styczniu, okres od lutego do końca marca [35].

3.2. Opis zbioru danych

Dane wykorzystane na potrzeby analizy pochodzą z raportów Narodowego Instytutu Zdrowia Publicznego - Państwowego Zakładu Higieny [27] dotyczących liczby zachorowań na grypę w Polsce. Zaznaczyć należy, że zostały one pozyskane ze strony internetowej Wrocławskiego Zlotu Użytkowników R [34].

Dane obejmują okres od stycznia 2000 roku do połowy maja 2010. Wyniki raportów Państwowego Zakładu Higieny publikowane były czterokrotnie w miesiącu dla okresu od stycznia do kwietnia i od października do grudnia. W przypadku pozostałych miesięcy raporty publikowane były dwukrotnie w miesiącu. Zwiększona częstotliwość pojawiania się wyników od października do kwietnia związana jest z typową strukturą liczby zachorowań na grypę w roku. Mianowicie na półkuli północnej sezon grypowy przypada na okres późnej jesieni, zimy oraz wczesnej wiosny, co wynika np. z raportów CDC (*Centers for Disease Control and Prevention*). W związku z tym zwiększona liczba raportów w tym okresie pozwala na pełniejsza analizę badanego zjawiska.

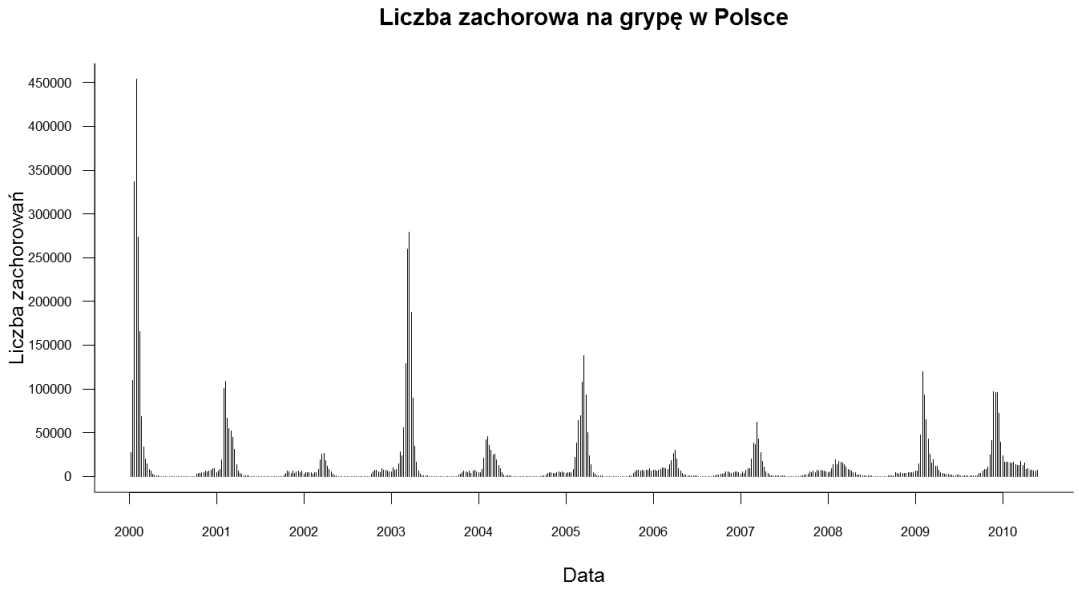
Ponieważ w opisanych w rozdziale 1 metodach i ich implementacjach wymagany jest podział roku na równe okresy, zdecydowano się na przetwarzanie danych. Motywacją do tego była struktura liczby zachorowań na grypę w ciągu roku. W okresach rzadszego raportowania liczba zachorowań jest nieporównywalnie mała względem liczby zachorowań z okresów częstszego raportowania. Ponadto prewencja epidemiologiczna wymaga jak największej dokładności zwłaszcza w okresach grypy sezonowej. W związku z powyższym zdecydowano się na podział każdego z raportów z okresu maj-wrzesień na dwa oraz przyjęto, że w każdym z tych dwóch okresów liczba zachorowań jest równa. Oznacza to, że uznano, iż korzyści wynikające z dokładniejszej analizy danych z okresu grypy sezonowej są większe niż koszty związane ze wspomnianym założeniem.

Ostatecznie dla każdego województwa sporządzony został szereg czasowy składający się z 499 obserwacji, tzn. po 48 obserwacji dla lat 2000-2009 oraz 19 obserwacji z roku 2010. Dodatkowo skonstruowano analogiczny szereg czasowy dla liczby zachorowań w całej Polsce, jako sumą obserwacji z poszczególnych województw.

Rys. 3.1 potwierdza, że grypa w Polsce jest chorobą sezonową. Większość zachorowań na grypę notuje się w miesiącach październik-kwiecień, zaś w sezonie letnim jest ich nieporównywalnie mało. Zauważyć także można wyraźnie zwiększoną liczbę zachorowań na początku roku 2000 oraz na przełomie lat 2002-2003 i 2004-2005. Ponadto rys. 3.1 sugeruje pewną typową struturę liczby zachorowań na grypę w Polsce w ciągu roku. Mianowicie w okolicy października następuje wzrost liczby zachorowań na grypę. Następnie zachorowalność na grypę stabilizuje się na pewnym poziomie i zauważyć można nieznaczne wahania (często nieznaczny spadek liczby zachorowań). Na początku roku kalendarzowego (najczęściej w lutym-marcu) odnotowuje się znaczący wzrost liczby zachorowań, zaś po kilku tygodniach zaobserwować można wyraźny jej spadek. W dalszej części pracy często spotkać będzie można odwołania do opisanej struktury liczby zachorowań na grypę.

W niniejszej pracy analizie poddano dane o zachorowaniach na grypę z 16 województw w Polsce z lat 2000-2010. Do wykrywania stanów alarmowych w latach 2005-2010 posłużono się modelem estymowanym na podstawie 240 obserwacji z lat 2000-2005 (48 obserwacji rocznie). Analizie występowania stanów alarmowych poddano 259 ostatnich obserwacji, co oznacza lata 2005-2009 oraz 19 pierwszych raportów z roku 2010. W celu zachowania porównywalności wyników okres ten został przyjęty dla wszystkich rozważanych metod.

W zbiorze danych odnotowano braki danych dla niektórych województw w przypadku pierwszego raportu z roku 2000. Zdecydowano się na niewykorzystywanie tego raportu w kolejnych analizach, ponieważ mógłby on zaburzyć ich wyniki.



Rysunek 3.1: Liczba zachorowań na grypę w Polsce w latach 2000-2010.

3.3. Algorytm CUSUM

Algorytm CUSUM jest często stosowany w wykrywaniu stanów alarmowych zachorowań na grypę. Jednymi z przykładów wykorzystania CUSUM są prace Ministerstwa Zdrowia Dystryktu Kolumbii [6], Centrum Chorób Zakaźnych w Wielkiej Brytanii [17] czy Szwedzkiego Instytutu Kontroli Chorób Zakaźnych [13].

3.3.1. Modyfikacje funkcji algo.cusum

Do lepszej ilustracji sytuacji epidemiologicznej w Polsce przy pomocy algorytmu CUSUM zastosowano pewne modyfikacje funkcji `algo.cusum` pakietu `surveillance` programu R, opisanej w rozdziale 2.2. Do redakcji skryptu użyto funkcji `fix()`. Skuteczność niżej opisanych modyfikacji została zbadana empirycznie.

Model GLM

Aby uwzględnić wahania sezonowe zachorowań na grypę wykorzystano model GLM. Domyślną postacią modelu GLM w funkcji `algo.cusum`, przy ustawieniu parametru `m="glm"` jest:

```
glm(x  ~ 1 + cos(2 * pi/p * t) + sin(2 *pi/p * t)), family = poisson()
```

gdzie p jest częstością obserwacji.

Dla danych o zachorowaniach na grypę w Polsce w latach 2000-2010, lepsza okazała się jednak być postać uwzględniająca większą liczbę składników odpowiadających za odzwierciedlenie okresowości w algorytmie:

```
control$m.glm <- glm(x ~ 1 + cos(2 * pi/p * t) + sin(2 *pi/p * t)+
    sin(2*pi/p * 2*t) + cos (2*pi/p * 2*t), family = poisson())
```

Resetowanie algorytmu

W podrozdziale 1.1.5 zwróciliśmy uwagę na jedną z wad CUSUM, którą jest długa regeneracja algorytmu po wykryciu stanu alarmowego. Podczas zastosowania CUSUM do danych o zachorowaniach na grypę w Polsce, z powodu dużej ilości zachorowań w okresie zimowym, wystąpił wzrost statystyki CUSUM, który skutkował brakiem lub zbyt długą regeneracją algorytmu, w wyniku którego statystyka nie była w stanie się zregenerować w okresie letnim i rosła zbyt szybko tworząc nowe fałszywe alarmy.

Jednym ze sposobów uniknięcia zawyżonych wartości statystyki CUSUM jest cykliczne jej wyzerowanie. W lipcu aktywność grypy jest odpowiednio niska, by merytorycznie uzasadnić zresetowanie CUSUM właśnie w tym okresie.

Zastosowano modyfikację:

```
if ((t+23)%% 48 == 0) cusum[t]<-0
```

Ustalenie dolnej granicy alarmowej

Zasygnalizowanie alarmu w funkcji `algo.cusum` następuje po przekroczeniu przez statystykę granicy alarmu (ang. `upperbound`), która jest wyliczana na podstawie zdefiniowanych parametrów k i h oraz średniej m, estymowanej (w niniejszej pracy) odpowiednim modelem glm. Podczas modelowania sezonowego oczekujemy mało wystąpień grypy w okresie letnim, zatem granica alarmu jest odpowiednio niższa w tym okresie.

Z drugiej strony, podczas przekroczenia przez statystykę granicy alarmowej mimo relatywnie małej ilości zachorowań, nie definiujemy jako sytuacji alarmowej. Rozsądnym wyjściem jest zatem zdefiniowanie minimalnej ilości zachorowań potrzebnej do zasygnalizowania alarmu.

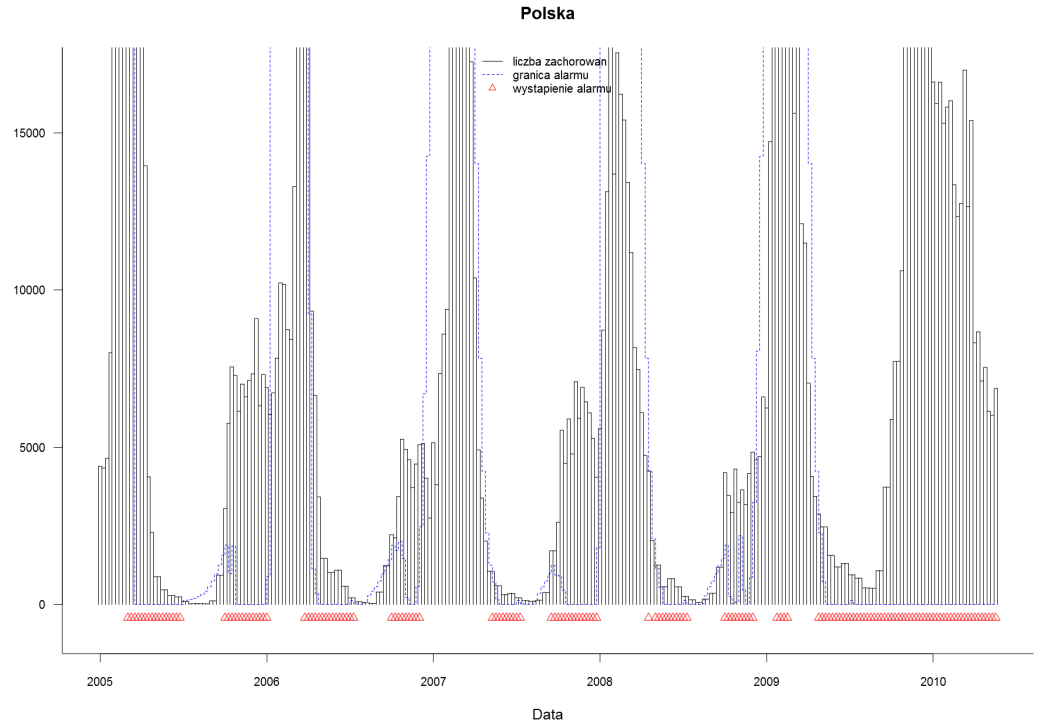
W niniejszej pracy zdefiniowano 200 jako minimalną liczbę zachorowań w jednym województwie potrzebną do zasygnalizowania alarmu. Do zastosowania modyfikacji użyto prostej pętli:

```
for (i in 1:259) {
  if (grypapod$observed[i+240] <  200 )
    { (result$alarm[i] <- 0 ) }
    }
```

3.3.2. Analiza zachorowań na grypę w Polsce

Do analizy stanów alarmowych zachorowań na grypę w Polsce zbudowano model CUSUM, bazujący na wspomnianych w podrozdziale 3.3.1 modyfikacjach. Do budowy modelu wykorzystano standaryzację danych do rozkładu normalnego opisaną przez G. Rossi i L. Lampugnani [14], średnią estymowano za pomocą modelu GLM, a jako parametry wybrano $k = 2.5$ i $h = 4$, które empirycznie zostały potwierdzone jako odpowiednie dla badania grypy [6] [16]. Do ustalenia parametrów modelu CUSUM w wyżej wymienionych publikacjach posłużono się badaniami symulacyjnymi, pozwalającymi na wybór modelu najlepiej odzwierciedlającego specyfikę chorób grypopodobnych, zazwyczaj przy uwzględnieniu trzech kryteriów: czułości algorytmu, częstości fałszywych alarmów oraz skuteczności w szybkim zasygnalizowaniu nadchodzącej epidemii. Z powodu dużej ilości wyprodukowanych przez algorytm fałszywych alarmów, nie rozpatrywano mniejszych wartości k i h (które by skutkowały jeszcze większą ilością fałszywych alarmów).

Jak wynika z wykresu 3.2, w latach 2005-2009 algorytm CUSUM zasygnalizował niezwykle dużą ilość alarmów. W porównaniu z sytuacją rzeczywistą odnotowaną w raportach Głównego



Rysunek 3.2: Analiza stanów alarmowych liczby zachorowań na grypę w Polsce przy wykorzystaniu algorytmu CUSUM

Inspektoratu Sanitarnego([28],[29],[30]), większość tych alarmów była fałszywa. W następnym podrozdziale analizie podane zostaną wyniki z poszczególnych województw w Polsce.

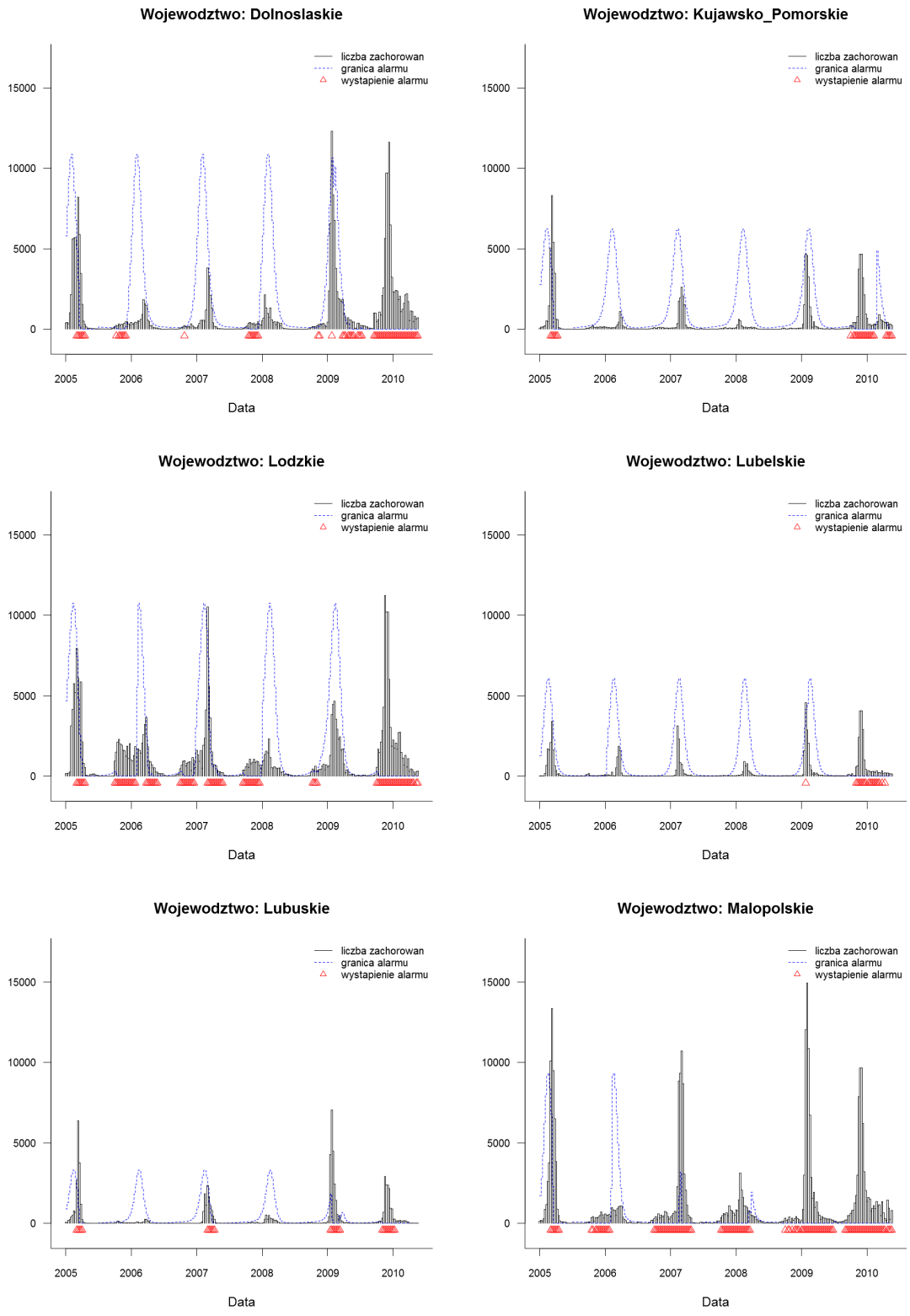
3.3.3. Analiza poszczególnych województw

Mimo tendencji algorytmu do produkowania fałszywych alarmów, nie we wszystkich województwach ta tendencja była jednakowo silna. W województwach Lubelskim i Świętokrzyskim w ciągu badanych pięciu lat, alarmy odnotowano tylko w 2010 roku, zaś w województwie Małopolskim zarejestrowano największą liczbę alarmów.

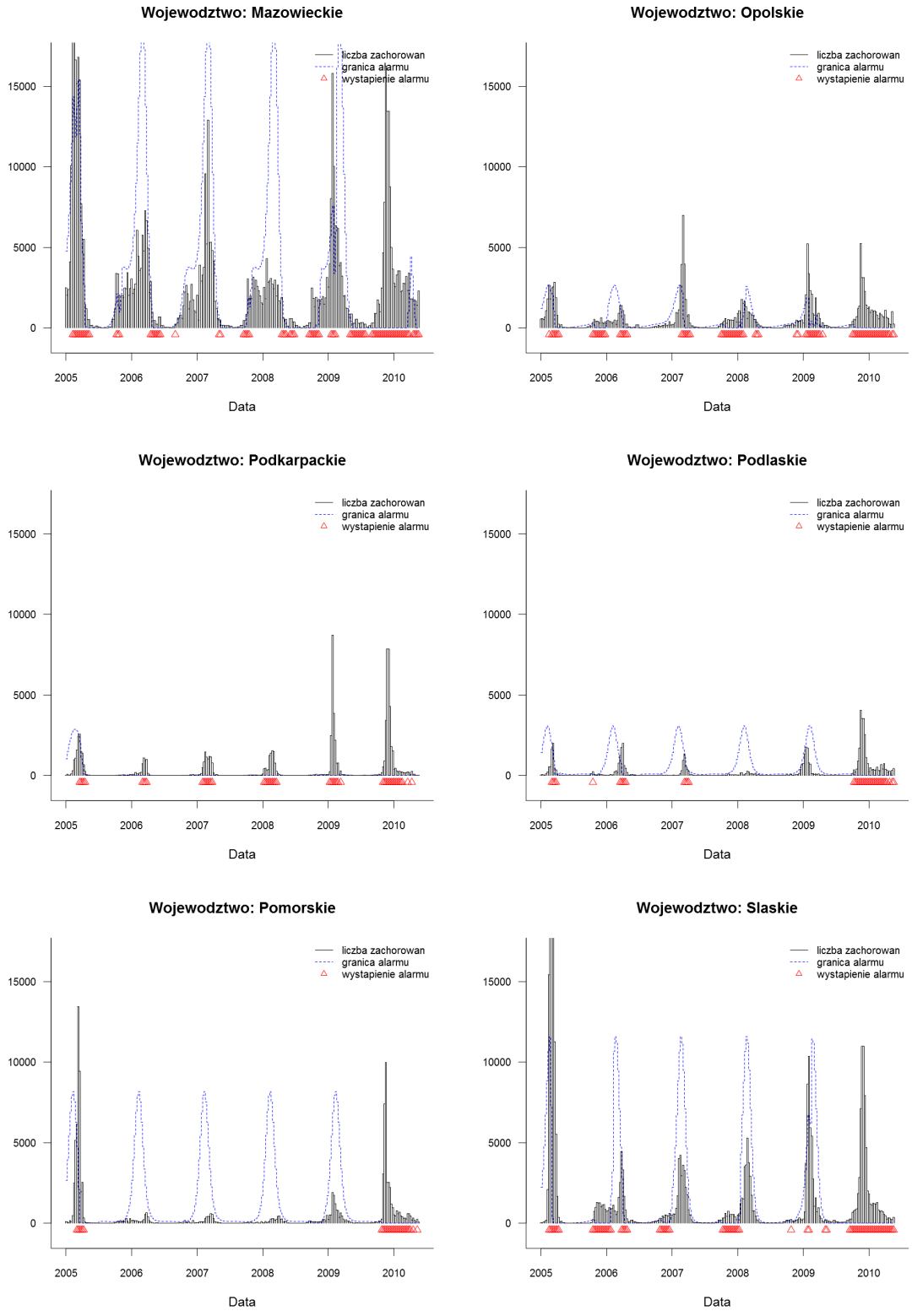
W województwie Dolnośląskim odnotowano alarmy w każdym z badanych lat, przy czym w 2007 roku tylko jeden alarm. Alarmy z 2006, 2007 i 2008 roku były wynikiem wcześniejszego niż w latach wcześniejszych rozpoczęcia "sezonu" grypy. W 2009 roku nastąpił gwałtowny wzrost zachorowań, związany ze światową pandemią grypy. Zachorowania te zawyżyły statystykę CUSUM skutkując długą regeneracją algorytmu i sygnalizowaniem alarmów również w okresie wiosennym i letnim.

W województwie Kujawsko Pomorskim sytuacja alarmowa się zdarzyła w 2005 roku. Liczba zachorowań tam wzrosła gwałtownie, w okresie późniejszym niż estymowany na podstawie wahań sezonowych. W 2010 roku również nastąpiło błędne modelowanie sezonowe. Mimo ilości zachorowań niewiele większej niż w 2009 roku, sygnalizowane były alarmy aż do zresetowania statystyki.

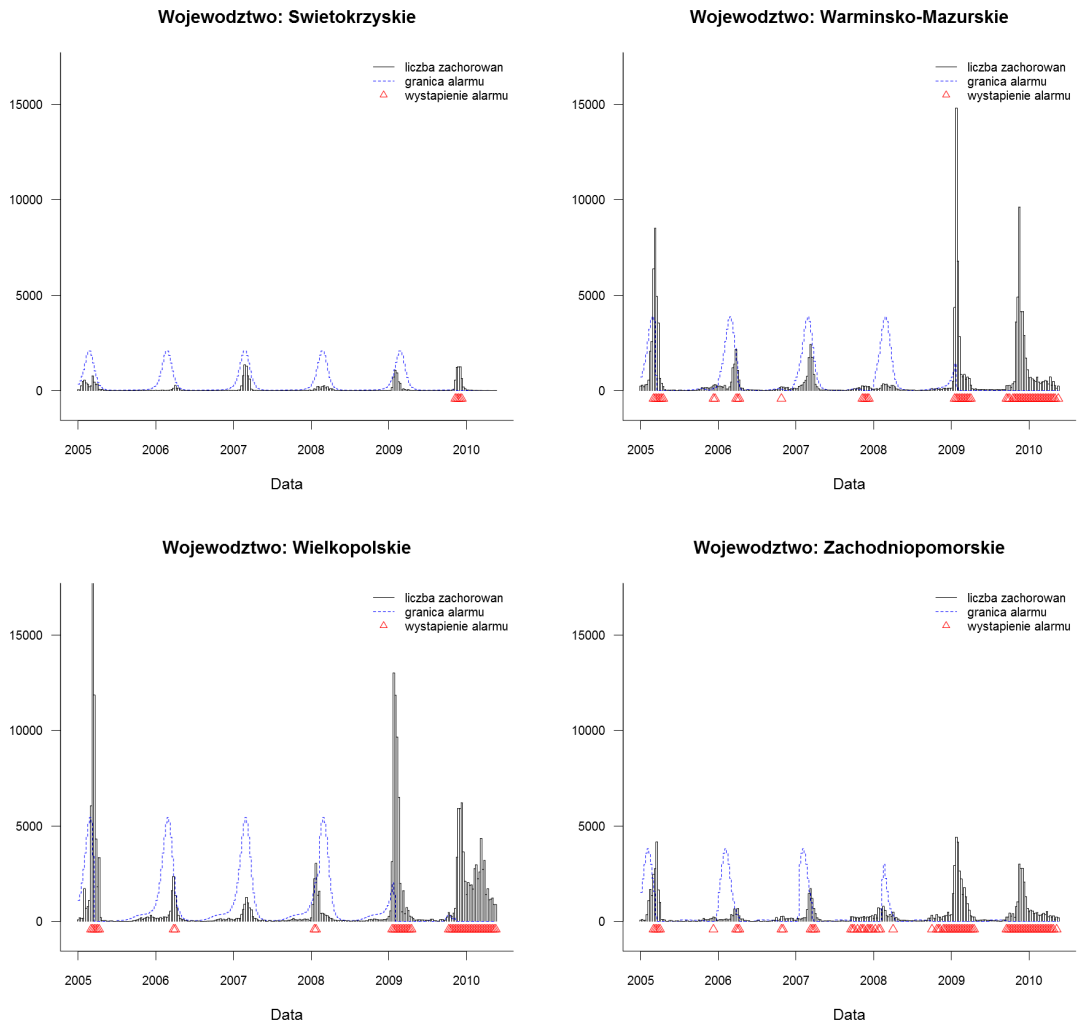
W województwie Lubelskim w latach 2005-2008 zachorowania na grypę nie przekroczyły granicy `upperbound`, w 2009 roku odnotowano jeden alarm, a w 2010 roku statystyka CUSUM skoczyła do góry zaniżając granicę `upperbound` i skutkując bezustanną sygnalizacją alarmu.



Rysunek 3.3: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu algorytmu CUSUM, cz.1.



Rysunek 3.4: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu algorytmu CUSUM, cz.2.

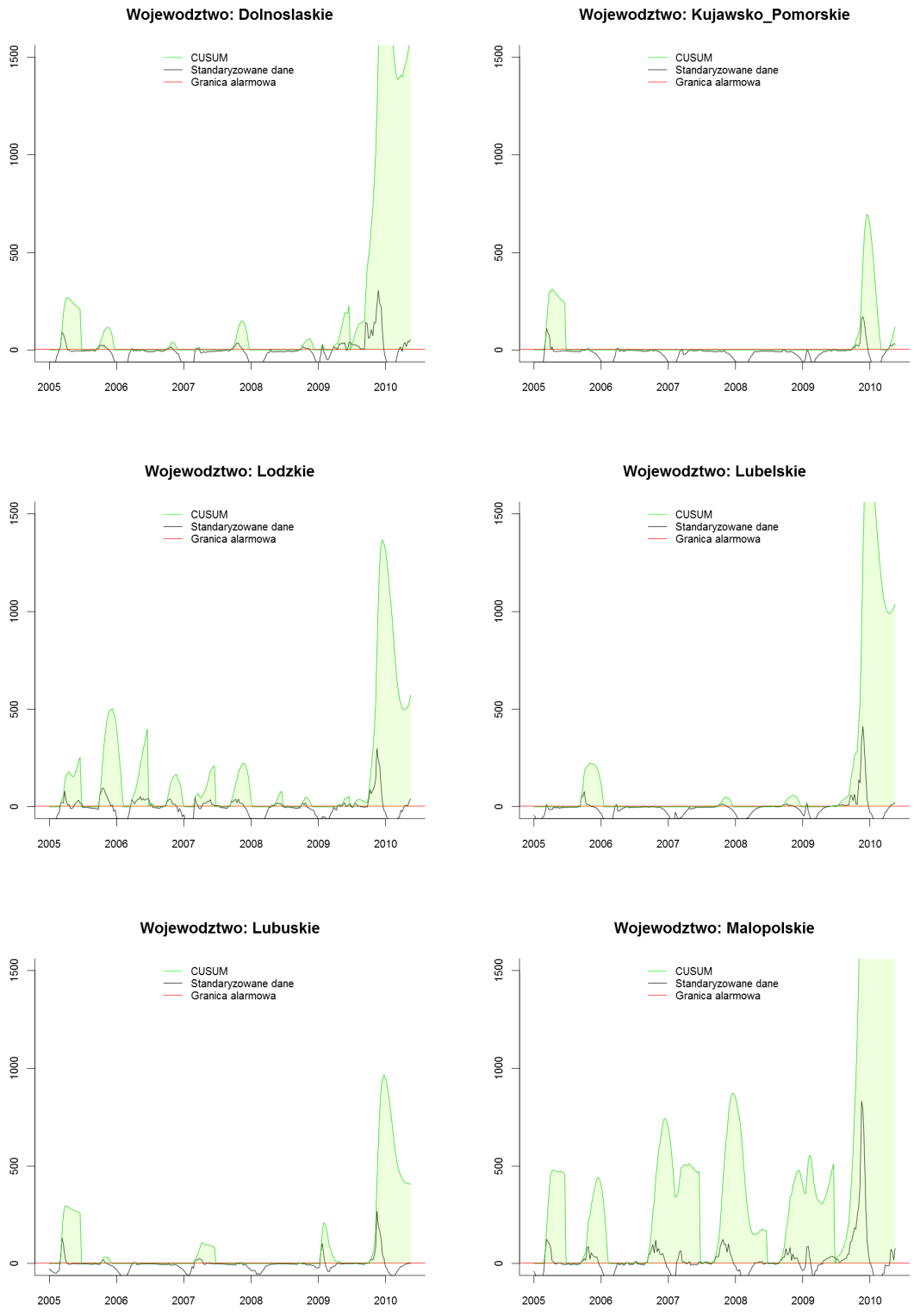


Rysunek 3.5: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu algorytmu CUSUM, cz.3.

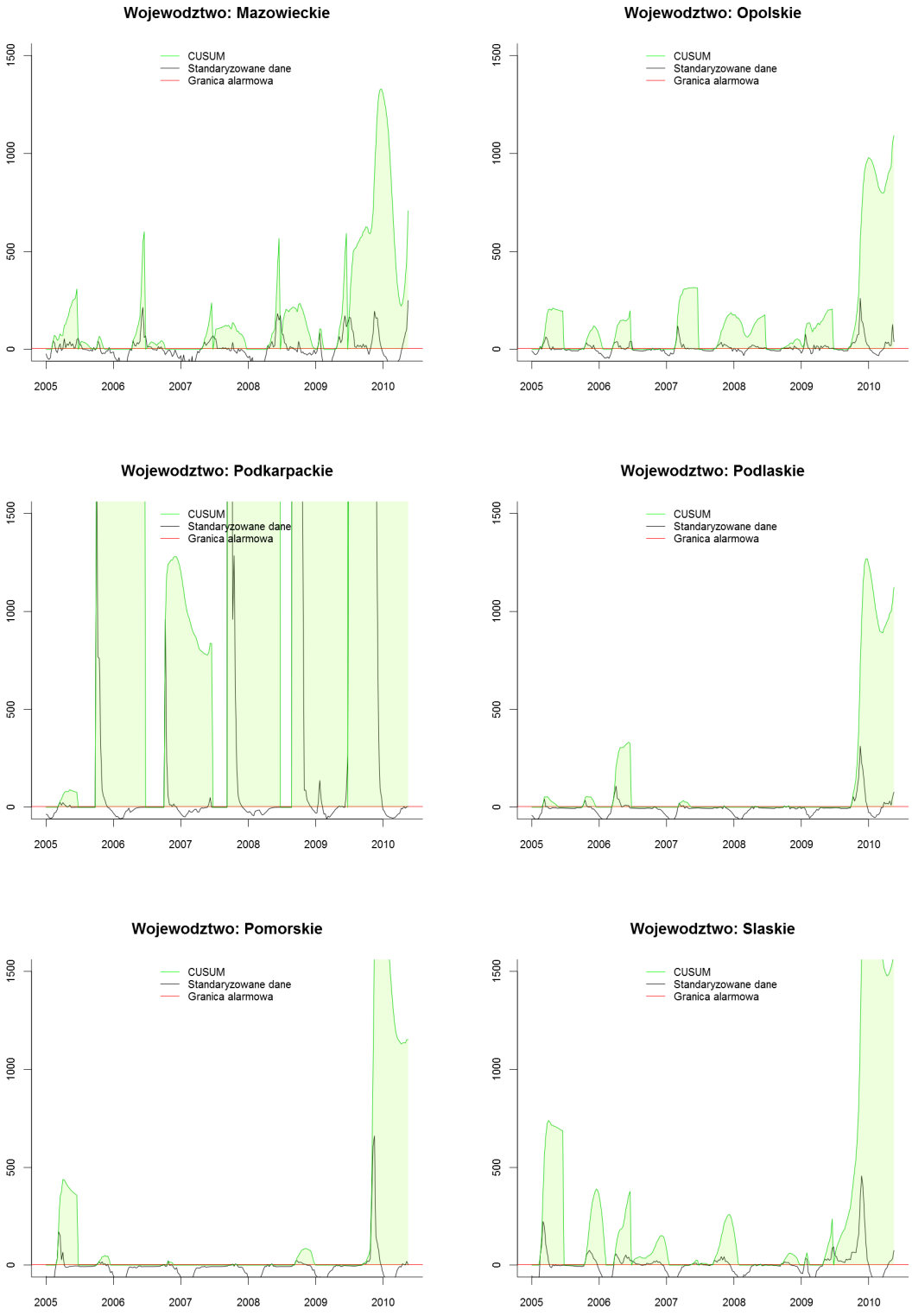
Nieprawidłowe "zniknięcie" granicy alarmowej (`upperbound`) nastąpiło w większości województw w 2010 roku. Wyjątkiem było tylko województwo Kujawsko Pomorskie. W województwach Zachodniopomorskim i Małopolskim podobna sytuacja się zdarzyła również w 2009 roku, a w woj. Podkarpackim we wszystkich latach z zakresu 2006-2010.

3.3.4. Analiza problemu

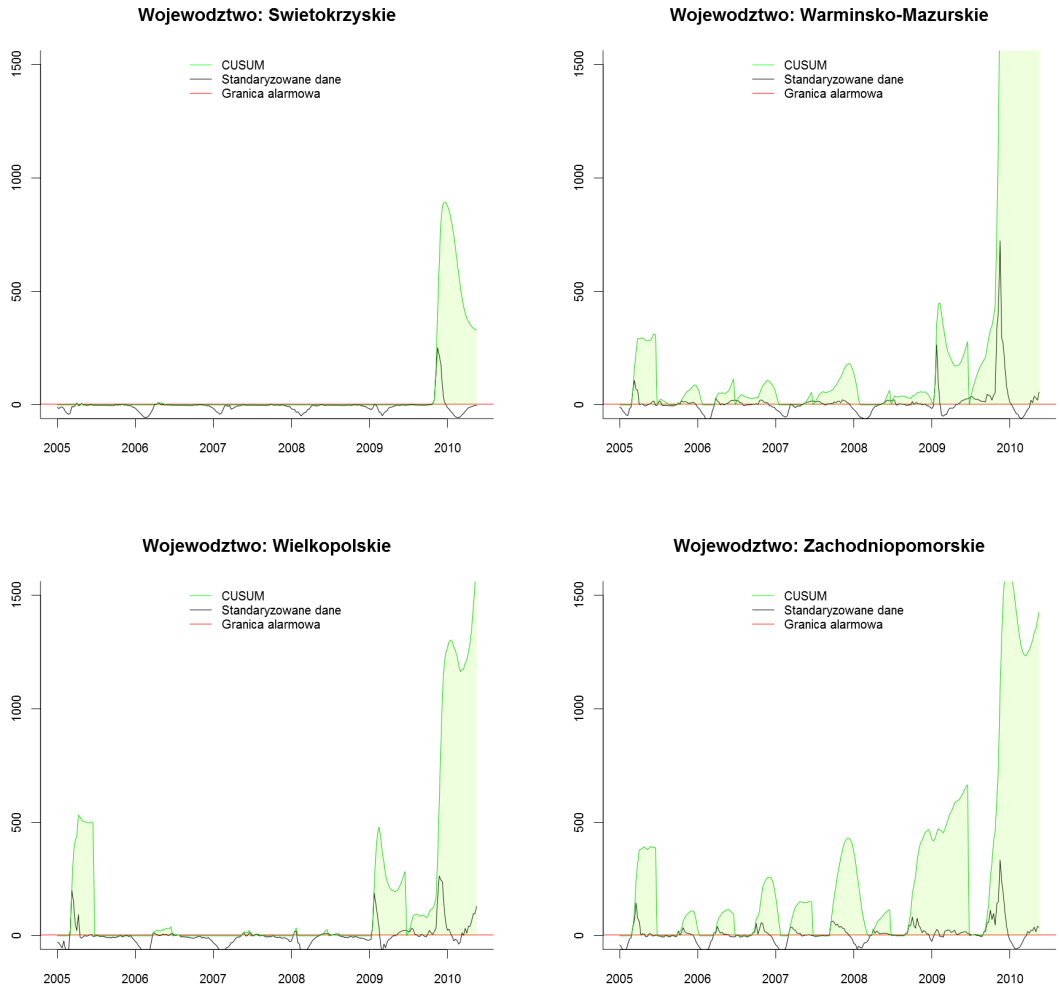
Z wykresów przedstawiających zastosowanie algorytmu CUSUM dla danych o zachorowaniach na grypę w Polsce, wyraźnie wynika nieprawidłowe funkcjonowanie algorytmu. Nawet po zastosowaniu modyfikacji funkcji `algo.cusum`, zmniejszających liczbę fałszywych alarmów, w większości województw na przeciągu pięciu lat odnotowano ogromną liczbę alarmów. Zastosowanie takich wyników analizy do kontroli procesów epidemiologicznych ma znikomą wartość. Z powodu licznego zastosowania algorytmu CUSUM do badania grypy ([13],[17],[16]) przyczyną problemu nie jest specyfika badanej choroby.



Rysunek 3.6: Analiza danych standaryzowanych i statystyki CUSUM w poszczególnych województwach, cz.1.



Rysunek 3.7: Analiza danych standaryzowanych i statystyki CUSUM w poszczególnych województwach, cz.2.



Rysunek 3.8: Analiza danych standaryzowanych i statystyki CUSUM w poszczególnych województwach, cz.3.

Aby lepiej zilustrować przyczyny błędnego wykrywania stanów alarmowych przez algorytm, dokładniej przeanalizujemy i porównamy sytuację w poszczególnych województwach na podstawie wykresów przedstawiających dane o zachorowaniach, po zastosowaniu standaryzacji opisanej przez G. Rossi i L. Lampugnani (patrz: 1.1.2) oraz wykresu statystyki CUSUM. Opisane niżej wykresy są przedstawione na rysunkach 3.6, 3.7, 3.8.

Jak wynika z wykresów, mimo standaryzacji danych, obserwacje w 2009-2010 roku, wysoce odbiegają od średniej rozkładu $N(0,1)$, niektóre z nich nawet o więcej niż 1500 jednostek (dla lepszej czytelności wykresów, została podana skala od 0 do 1500). Szczególnie widoczne jest to w woj. Podkarpackim, gdzie każdego roku w okresie jesiennym standaryzowane ilości zachorowań "wybuchały" do góry.

Standaryzowane obserwacje są redukowane o parametr k, który jest znikomie mały w porównaniu z tak wielkimi wartościami x_t. W konsekwencji, statystyka CUSUM kumuluje co raz większe wartości, kilkanaście razy większe od wartości alarmowej h, zatem sygnalizowany

jest alarm. CUSUM przyjmuje tak duże wartości, że obserwacje w następnych punktach czasowych przestają odgrywać rolę, gdyż nawet jeżeli w kolejnej chwili czasu zaobserwujemy zero zachorowań, to statystyka CUSUM zredukuje się tylko o małą wartość k i wciąż będzie sygnalizowany alarm. Granica alarmowa, która wskazuje, przy jakiej liczbie zachorowań alarm byłby sygnalizowany, spłaszcza się i przyjmuje wartość zero.

W sytuacji tak wielkich wychyleń standaryzowanych danych od średniej rozkładu $N(0,1)$ resetowanie statystyki ratuje sytuację tylko na chwilę i następuje kolejny gwałtowny wzrost CUSUM.

W woj. Dolnośląskim dane po standaryzacji miały duży wzrost w 2005 roku, który spowodował duży wzrost statystyki CUSUM i sygnalizowanie alarmu. Statystyka przyjęła wartości rzędu sto razy większe niż $h = 4$, po swym piku wiosną 2005 zaczęła się regenerować, jednak ze względu na długą regenerację algorytmu i tak ogromne wartości statystyki, regeneracja ta potrwałaby o wiele dłużej i spowodowałaby zwiększone wartości CUSUM w kolejnym sezonie, gdyby nie zastosowanie wyzerowania statystyki latem 2005 roku. Pod koniec 2005 roku nastąpiło kolejne alarmowe zwiększenie się ilości zachorowań, po którym statystyka CUSUM naturalnie się zregenerowała w wyniku dużych dewiacji od zera standaryzowanych danych w kierunku ujemnym. Pod koniec 2007 roku również nastąpił duży wzrost CUSUM z regeneracją statystyki na początku 2008 roku. W 2009 roku wartości standaryzowanych danych alarmowo wzrastały powodując duży wzrost CUSUM aż do lipca, po czym statystyka została zresetowana, jednak standaryzowane ilości zachorowań były o wiele większe od średniej 0, przez co statystyka CUSUM znów zaczęła gwałtownie rosnąć, przyjmując kolosalne wartości.

W województwie Kujawsko Pomorskim standaryzowane dane miałe ujemne skoki na poczatku lat 2006, 2007 i 2008, a duże wzrosty standaryzowancych danych nastąpiły w 2005 r. i pod koniec 2009 r., powodując stany alarmowe w tych dwóch okresach.

W woj. Łódzkim duże skoki standaryzowanych danych pojawiały się każdego roku, przy czym największy z nich pod koniec 2009 roku, który spowodował skumulowanie się ponadtysięcznych sum, przez co regeneracja algorytmu wydłużyła się na tyle, że przy zerowej obserwalności zachorowań w 2010 r., statystyka CUSUM byłaby w kolejnych odstępach czasu zmniejszana o wartość $k = 2.5$, zatem jakby w algorytmie nie zastosowano resetowania statystyki latem, dopiero po około pięciu latach przyjęłaby ona wartości mniejsze niż $h = 4$

W woj. Lubelskim duże dewiacje standaryzowanych danych od średniej wystąpiły pod koniec 2005 roku, powodując wzrost statystyki CUSUM. Jak wynika jednak z wykresu 3.3 nie został zasygnalizowany wówczas alarm, z powodu wprowadzonej modyfikacji algorytmu `algo.cusum`, która usuwała alarmy spowodowane mniejszą niż 200 ilością obserwowanych zachorowań.

W woj. Małopolskim standazyzacja danych do rozkładu $N(0,1)$ była jeszcze mniej skuteczna. Co roku występowały przypadki odskoku od średniej zero w górę o sto jednostek, przez co alarm nie występował generalnie tylko w okresach lipcowego resetowania statystyki, a pod koniec 2009 roku wystąpił ogromny skok standaryzowanych danych, powodując wzrost statystyki CUSUM na więcej niż 400 razy większą od wartości alarmowej h

W woj. Mazowieckim i Opolskim standaryzowane dane miały również co roku dewiacje od średniej zero rzędu 100 jednostek większe, z największym skokiem pod koniec 2009 roku, który powodował bezustanną sygnalizację alarmu.

Najgorsza sytuacja miała miejsce w woj. Podkarpackim, gdzie dane po standaryzacji tak dalece odbiegały od średniej rozkładu $N(0,1)$, że prawie co roku występowały kolejne przypadki wartości wiele razy większych nawet od 1500. Resetowanie algorytmu latem w takiej sytuacji minimalnie ratowało sprawę, co ilustruje lipiec 2009 roku, gdzie po zresetowaniu staystyki CUSUM do zera, kolejna standaryzowana obserwacja była na tyle ogromna, że spowodowała wzrost statystyki do wartości nie mieszczących się w skali (0,1500).

W porównaniu do wcześniejszych województw, sytuacja w woj. Świętokrzyskim do 2009 roku, dane po standaryzacji nie miały tak wielkich wychyleń, przez co algorytm CUSUM nie zasygnalizował żadnego alarmu w latach 2005-2009, a z wykresu 3.4 widać również skuteczną predykcję wahań sezonowych. Pod koniec 2009 roku nastąpił duży wzrost standaryzowanych danych, na który wpływ miał również duży wpływ z pewnością rozwój pandemii spowodowanej wirusem A/H1N1. Można wnioskować zatem, że w woj. Świętokrzyskim algorytm CUSUM działał poprawnie.

3.3.5. Podsumowanie

Z przedstawionych wyników i analizy działania algorytmu wynika, że algorytm CUSUM nie jest odpowiedni do wykrywania stanów alarmowych na podstawie danych o zachorowaniach na grypę z raportów Narodowego Instytutu Zdrowia Publicznego - Państwowego Zakładu Higieny. Wyprodukowana przez algorytm ilość alarmów jest zbyt wielka, aby miała jakiekolwiek zastosowanie praktyczne.

Jedną z najbardziej ewidentnych przyczyn nieprawidłowego działania algorytmu dla danych o zachorowaniach na grypę w Polsce jest nieskuteczna standaryzacja danych. Uwzględniając wahania sezonowe na podstawie odpowiedniego modelu GLM estymowane są średnie liczby zachorowań m w odpowiednich punktach czasowych, przez co po standaryzacji danych nie powinna występować już sezonowość w obserwacjach. Dane po standaryzacji powinny mieć średnią zero, z niewielkimi odchyleniami od średniej, by spełniały założenia aproksymacyjnego modelu CUSUM. W wypadku zastosowania metody aproksymacyjnej zaproponowanej przez G. Rossi i L. Lampugnani do danych o zachorowaniach na grypę w Polsce obserwowane są jednak zbyt wielkie i zbyt częste odchylenia standaryzowanych danych od średniej zero, które powodują zwiększoną liczbę fałszywych alarmów.

Dla tak wielkich odchyleń, stosowalne parametry $k = 2.5$ i $h = 4$ tracą na funkcjonalności, gdyż w wypadku osiągnięcia przez statystykę CUSUM wartości rzędu paru tysięcy, tak mała wartość parametru odniesienia k skutkuje zbyt długą regeneracją algorytmu, a tak trywialne mała wartość h powoduje bezustanną sygnalizację alarmu.

Możliwym rozwiązaniem problemu mogłoby być podwyższenie parametrów k i h, by wyeliminować możliwie dużo fałszywych alarmów, rejestrując jedynie istotne z nich. Empirycznie sprawdzono, że dla przykładowych wartości $k = 100$ i $h = 10$ eliminowana jest większość fałszywych alarmów bez straty wczesnego wykrycia epidemii z 2009 roku. Jednak jedną z wad takiego podejścia jest brak merytorycznego uzasadnienia dużych wartości parametrów k i h, a także brak uzasadnienia do korzystania z tranformacji danych, która nie powoduje standaryzacji danych do rozkładu $N(0, 1)$.

Z powodu niestosowalności zaproponowanej metody aproksymacyjnej, innym ewentualnym rozwiązaniem byłoby użycie oryginalnej metody CUSUM bez standaryzowania danych (ale z uzmiennieniem średniej m w czasie, na przykład za pomocą modelu GLM). Takie zmiany wymagałyby jednak całkowicie odmiennej implementacji algorytmu CUSUM w programie R i wcale nie gwarantują lepszych wyników.

3.4. Model Farringtona

W poniższym podrozdziale przedstawiona zostanie analiza danych rzeczywistych z wykorzystaniem modelu Farringtona. W pierwszej kolejności zaprezentowano analizę dla danych dla całego kraju, następnie dokonano uszczegółowienia wyników na poszczególne województwa. Wszystkie kody programu R wykorzystane w niniejszym podrozdziale znaleźć można w

załączniku. W kolejnych podrozdziałach przedstawiono specyfikację modelu, analizę stanów alarmowych dla Polski oraz analizę stanów alarmowych w poszczególnych województwach.

3.4.1. Specyfikacja modelu

Przy specyfikacji modelu Farringtona zdecydowano się na wykorzystanie obserwacji z czterech poprzedzających lat oraz ustalono „szerokość okna" na 4, tzn. $b = 4$ i $w = 4$ (patrz podrozdziały: 1.2.2 i 2.3). Parametr w został wybrany w ten sposób, by z jednej uchwycić ewentualne przesunięcia apogeum liczby zachorowań na grupę, ale z drugiej strony, by uwzględnić wahania sezonowe. Uzasadnieniem dla wielkości parametru b są wielkość próby oraz ciągłe zmiany w wykrywaniu, badaniu i zapobieganiu grypie. Oznacza to, że parametr b nie może być zbyt duży, ponieważ wspomniane zmiany zaburzałyby znacznie wyniki analizy.

W trakcie analizy badanego zjawiska przy wykorzystaniu modelu Farringtona zauważono, że bardzo często zdarzała się sytuacja, że trend (podrozdział 1.2.2) dla jednego okresu był statystycznie istotny, zaś dla kolejnego nie. Nie udało się nawet znaleźć, żadnych zależności pomiędzy częścią roku kalendarzowego a istotnością statystyczną trendu. Ponieważ nie znaleziono merytorycznego uzasadnienia tak częstych zmian istotności trendu, zdecydowano się nie uwzględniać trendu w dalszych analizach.

Kolejnymi argumentami koniecznymi do specyfikacji są poziom istoności α oraz minimalna liczba obserwacji konieczna do podniesienia alarmu - parametr `limit54` (patrz podrozdział 2.3). W przypadku poziomu istotności zdecydowano się na wartość $0,01$, ponieważ założono, że prawdopodobieństwo popełnienia błędu pierwszego rodzaju w przypadku badanego zjawiska powinno być jak najmniejsze. Inaczej wywołanie niepotrzebnego alarmu dla całego kraju może nieść za sobą poważne konsekwencje. W przypadku parametru `limit54` przyjęto arbitralnie wartość 600 zachorowań w ostatnich trzech okresach dla województw oraz 4000 zachorowań w ostatnich trzech okresach w przypadku danych dla Polski.

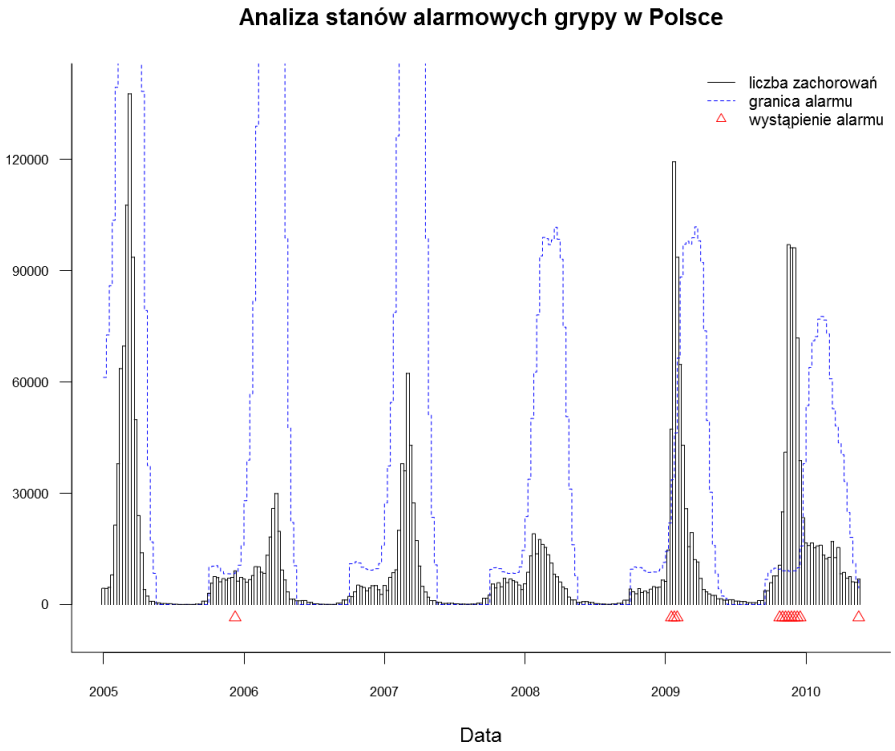
3.4.2. Analiza stanów alarmowych grypy dla Polski

Do przeprowadzenia analizy stanów alarmowych liczby zachorowań na grypę w Polsce wykorzystano specyfikację modelu Farringtona opisaną w podrozdziale 3.4.1. Wyniki analizy zaprezentowano na rys. 3.9.

Wyniki analizy przedstawione na rys. 3.9 pokazują, że wykorzystując do wykrywania stanów alarmowych liczby zachorowań na grypę model Farringtona w niektórych okresach odnotowano alarm. Po pierwsze pojedynczy alarm został zauważony w listopadzie 2006. Z perspektywy czasu widać, że w kontekście całego kraju alarm ten mógłby zostać zignorowany, jednakże wskazywać on może na podwyższone stany zachorowań w pewnych województwach, dlatego w tym przypadku uzasadniona jest bardziej szczegółowa analiza.

Alarmy odnotowano także na początku roku 2009. W tym przypadku sytuacja jest bardziej niepokojąca, ponieważ 3 zauważone alarmy występują po sobie oraz każdy z nich znacznie przekracza wyznaczoną z modelu granicę alarmu.

Kolejnym okresem z zauważonymi alarmami są dwa ostatnie miesiące 2009 roku. Wystąpienie alarmów wynika ze znacznego przesunięcia struktury liczby zachorowań w roku, tzn. w przypadku przełomu 2009-2010, apogeum zachorowań wystąpiło dużo wcześniej niż w przypadku wcześniejszych lat. Ponadto struktura liczby zachorowań na grypę także uległa zmianie. W niemal wszystkich wcześniejszych sezonach grypowych zauważyć można wzrost liczby zachorowań na grypę do pewnego poziomu pod koniec roku, następnie stabilizację liczby zachorowań na zbliżonym poziomie (lub nieznaczny jej spadek) oraz gwałtowny wzrost liczby zachorowań na początku roku kalendarzowego. W przypadku sezonu grypowego 2009-2010



Rysunek 3.9: Analiza stanów alarmowych liczby zachorowań na grypę w Polsce przy wykorzystaniu modelu Farringtona.

gwałtowny wzrost wystąpił już na jesieni 2009 roku, a następnie liczba zachorowań systematycznie malała. Na koniec okresu badawczego odnotowano dodatkowo jeden pojedynczy alarm.

Z rysunku 3.9 wynika także, że model Farringtona dość dobrze odzwierciedla strukturę liczby zachorowań na grypę, o której wspomniano w poprzednim paragrafie. Wskazuje na to kształt granicy alarmu.

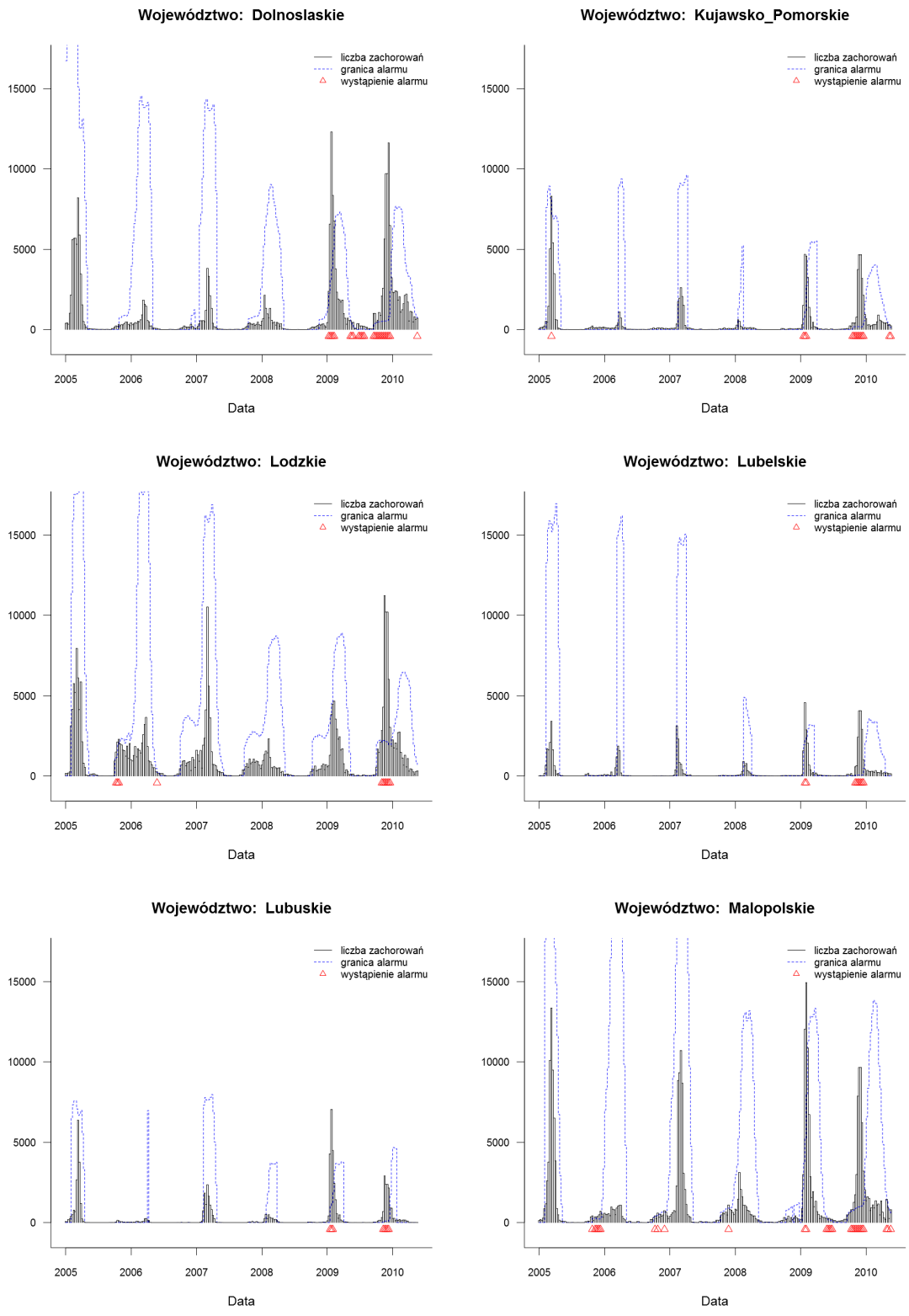
Podsumowując, przy wykorzystaniu modelu Farringtona odnotować można pojedynczy alarm pod koniec roku 2006, sekwencję alarmów na początku roku 2009 oraz sekwencję alarmów pod koniec 2009 roku. Dodatkowo alarm wystąpił w maju 2010 roku, jednakże jest to koniec analizowanego okresu. W celu uszczegółowienia analizy zdecydowano się na analizę stanów alarmowych liczby zachorowań w poszczególnych województwach.

3.4.3. Analiza stanów alarmowych grypy dla poszczególnych województw

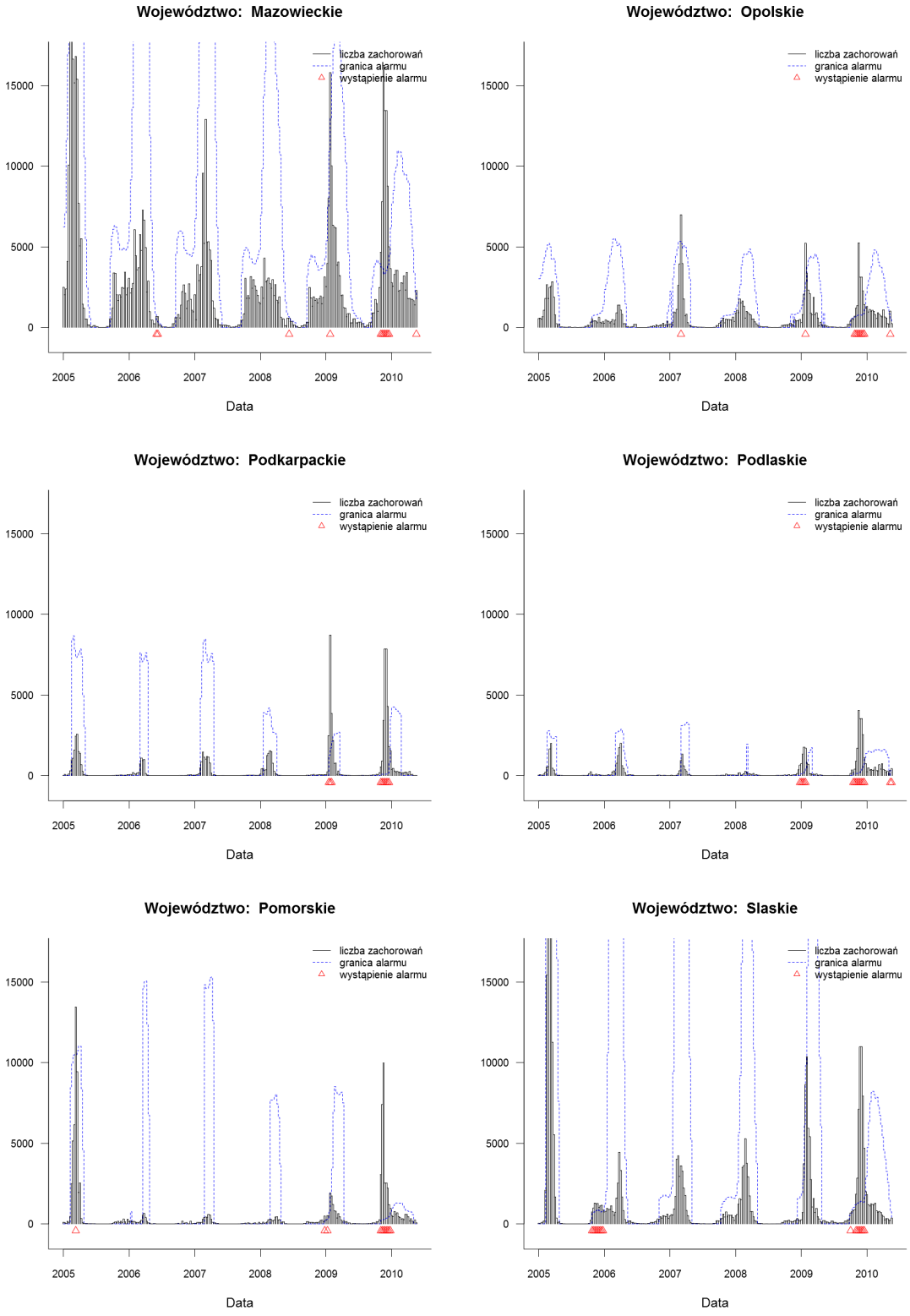
W niniejszym rozdziale wyniki z podrozdziału 3.4.2 zostaną rozwinięte, uwzględniając analizę w poszczególnych województwach. We wszystkich analizach wykorzystano specyfikację modelu Farringtona opisaną w podrozdziale 3.4.1. Wyniki analiz przedstawiono na rys. 3.10, 3.11 oraz 3.12.

Alarmy w poszczególnych województwach występowały w różnych okresach. Wielokrotnie można zauważyć alarmy w sezonach grypowych 2008-2009 oraz 2009-2010, rzadziej w latach wcześniejszych. Alarmy w poszczególnych województwach przedstawiono poniżej:

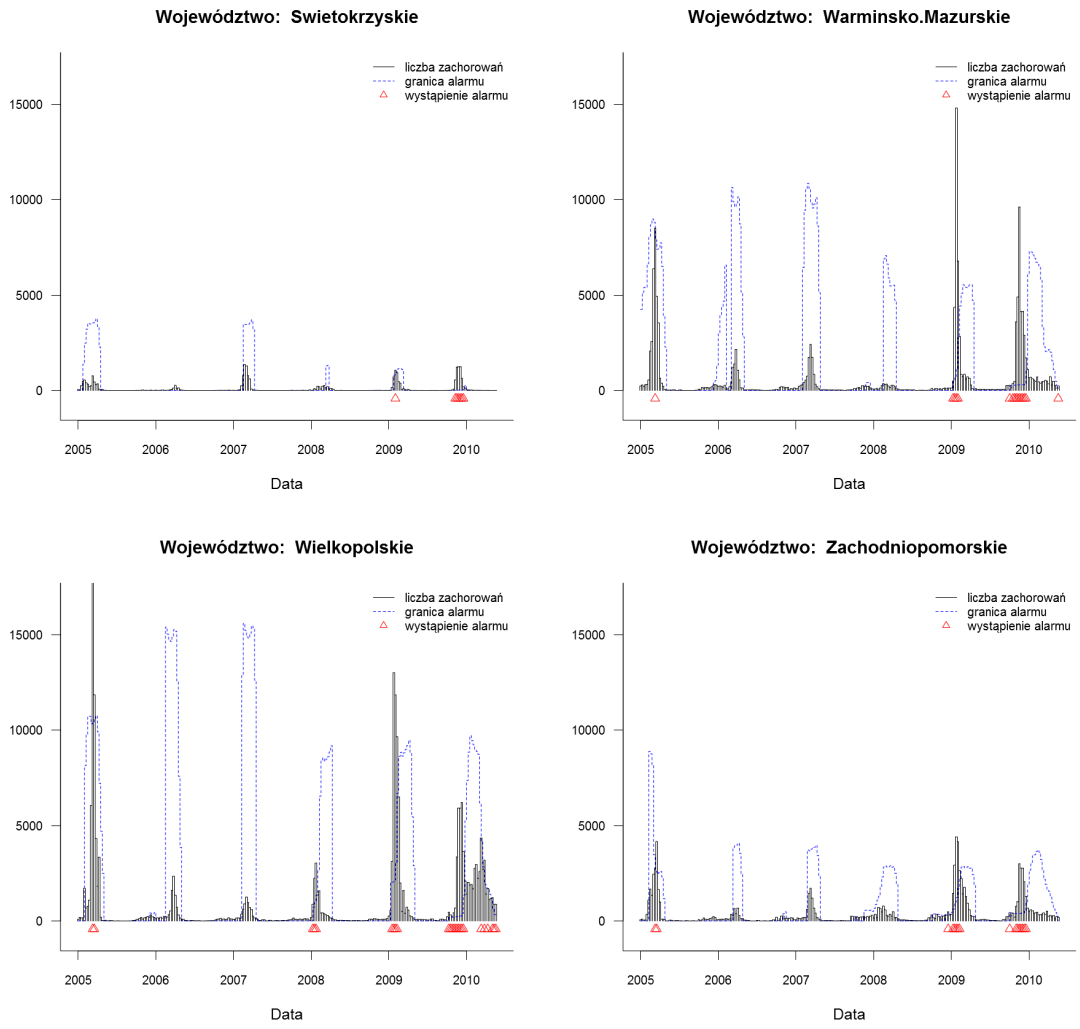
- Województwo Dolnośląskie: liczne sekwencje następujących po sobie alarmów w sezo-



Rysunek 3.10: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu modelu Farringtona, cz.1.



Rysunek 3.11: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu modelu Farringtona, cz.2.



Rysunek 3.12: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu modelu Farringtona, cz.3.

nach 2008-2009 i 2009-2010,

- Województwo Kujawsko-Pomorskie: pojedynczy alarm na początku roku 2005, dwa alarmy na początku roku 2009, sekwencja następujących po sobie alarmów w sezonie grypowym 2009-2010,
- Województwo Łódzkie: alarmy w sezonie grypowym 2005-2006 związane z licznymi zachorowaniami pod koniec roku 2005, liczne alarmy w sezonie grypowym 2009-2010,
- Województwo Lubelskie: pojedynczy alarm na początku roku 2009 oraz liczne alarmy w sezonie 2009-2010,
- Województwo Lubuskie: dwa alarmy na początku roku 2009 oraz liczne alarmy w sezonie 2009-2010,
- Województwo Małopolskie: liczne alarmy w sezonie 2005-2006 związane z wypłaszczeniem struktury liczby zachorowań na grypę, pojedyncze alarmy w kolejnych dwóch

sezonach, liczne alarmy w sezonach 2008-2009 oraz 2009-2010,

- Województwo Mazowieckie: pojedyncze alarmy od roku 2006 do roku 2009, liczne alarmy w sezonie 2009-2010,
- Województwo Opolskie: pojedyncze alarmy na początku roku 2007 oraz na początku roku 2009, liczne alarmy w sezonie 2009-2010,
- Województwo Podkarpackie: 3 alarmy na początku roku 2009, liczne alarmy w sezonie 2009-2010,
- Województwo Podlaskie: liczne alarmy w sezonach grypowych 2008-2009 oraz 2009-2010,
- Województwo Pomorskie: pojedynczy alarm na początku roku 2005, 2 alarmy pod koniec roku 2008, liczne alarmy pod koniec roku 2009,
- Województwo Śląskie: liczne alarmy pod koniec roku 2005 oraz liczne alarmy pod koniec roku 2009,
- Województwo Świętokrzyskie: pojedynczy alarm na początku roku 2009 oraz liczne alarmy pod koniec roku 2009,
- Województwo Warmińsko-Mazurskie: pojedynczy alarm na początku roku 2005, liczne alarmy w sezonach 2008-2009 i 2009-2010,
- Województwo Wielkopolskie: dwa alarmy na początku 2005 roku, 3 alarmy na początku roku 2008, liczne alarmy w sezonach grypowych 2008-2009 oraz 2009-2010,
- Województwo Zachodniopomorskie: dwa alarmy na początku roku 2005, liczne alarmy w sezonach grypowych 2008-2009 oraz 2009-2010.

Po pierwsze w przypadku wszystkich województw wykryto alarmy w ostatnim analizowanym sezonie grypowym, tzn. 2009-2010. Charakterystyczne jest też to, że we wszystkich tych przypadkach alarmy wynikają ze zmiany struktury liczby zachorowań, podobnie jak to miało miejsce dla analizy całego kraju (podrozdział 3.4.2). Mianowicie sezon grypowy zaczął się wcześniej i maksimum zachorowań miało miejsce także wcześniej niż zwykle. Uznać, więc można, że zaobserwowana zmiana była bardzo podobna w skali całego kraju. W przypadku wcześniejszego sezonu grypowego sytuacja w Polsce wystąpień alarmów w poszczególnych województwach była bardziej zróżnicowana.

Bardziej szczegółowy opis sytuacji wyłaniającej się z analizy liczby zachorowań na grypę przy wykorzystaniu modelu Farringtona znajduje się w kolejnym podrozdziale (3.4.4). Podjęta w nim została próba integracji wyników dla Polski i województw oraz oceny przydatności modelu Farringtona.

3.4.4. Uogólnienie wyników analizy i ocena przydatności modelu

W niniejszym podrozdziale dokonano integracji wyników z poprzednich dwóch rozdziałów (3.4.2 i 3.4.3). Mianowicie starano się pokazać, które województwa miały wpływ na powstawanie alarmów dla danych dotyczących całego kraju oraz zaprezentowano wyniki w ujęciu przestrzenno-czasowym.

W celu zwiększenia przejrzystości zaprezentowanych wyników postanowiono sporządzić mapy Polski, na których zaprezentowano liczbę alarmów dla poszczególnych województw.

Na rys. 3.13 przedstawiono pięć map - dla każdego z pięciu pełnych sezonów grypowych poddanych analizie (od sezonu 2005-2006 do sezonu 2009-2010).

Na początek analizy zaznaczyć należy, że alarmy w poszczególnych województwach niekoniecznie muszą się przekładać na alarmy w całym kraju, tzn. z alarmów w poszczególnych województwach nie musi wynikać alarm w całym kraju (zauważyć to można w sezonie grypowym 2006-2007). Jest to oczywisty fakt wynikający z agregacji danych w przypadku analizy dla Polski.

Zgodnie z podrozdziałem 3.4.2 w sezonie grypowym 2005-2006 odnotowany został pojedynczy alarm pod koniec roku 2005. Z rys. 3.13 wynika, że za podniesiony alarm odpowiadać może sytuacja w województwach Śląskim, Małopolskim i Łódzkim. Stosunkowo dużą liczbę zachorowań, jak na porę roku (późna jesień) odnotowano w Województwach Śląskim i Łódzkim. Dodatkowo we wszystkich trzech województwach zauważyć można zmianę typowej struktury liczby zachorowań na grypę w trakcie sezonu grypowego.

Mimo, że w sezonach grypowych 2006-2007 oraz 2007-2008 nie odnotowano stanu alarmowego w skali całego kraju, to pojawiały się one w poszczególnych województwach. W sezonie 2006-2007 alarmy pojawiły się w województwach Małopolskim i Opolskim. W przypadku pierwszego z nich odnotowano wzrost liczby zachorowań późną jesienią, w przypadku drugiego, przekroczenie granicy alarmu nastąpiło tylko raz, jednakże miało to miejsce w momencie największej zachorowalności, co sugerowałoby podjęcie działań zapobiegających dalszemu rozprzestrzenianiu się choroby. W sezonie 2007-2008 alarmy odnotowano w województwach Wielkopolskim i Małopolskim. W województwie Wielkopolskim nastąpiło przesunięcie apogeum liczby zachorowań o około 3 tygodnie. W przypadku województwa Małopolskiego powtórzyła się sytuacja z poprzedniego sezonu.

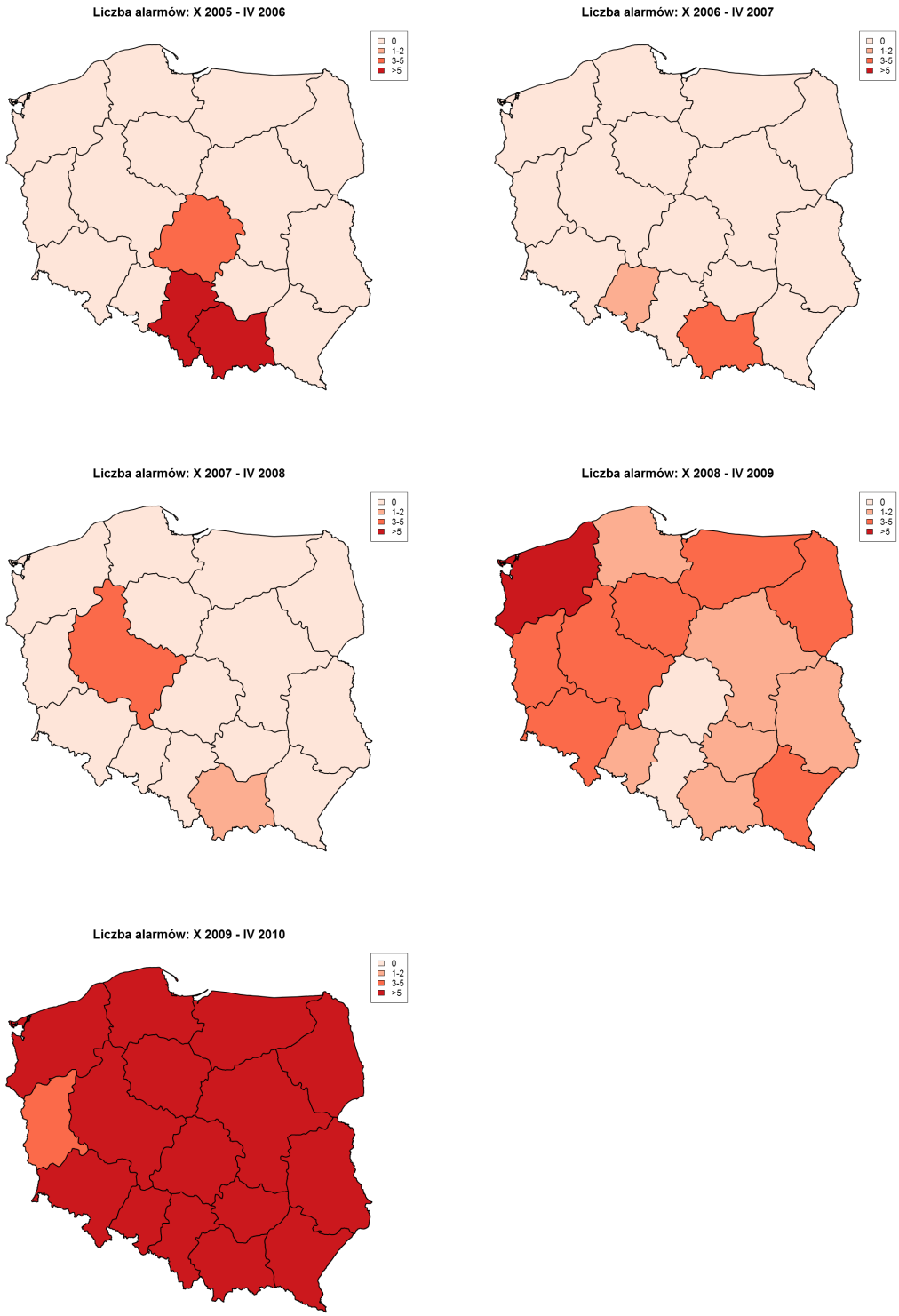
W sezonie grypowym 2008-2009 odnotowano alarmy zarówno dla całego kraju, jak i dla większości województw. Województwa, w których wykryto alarm oraz stopień natężenia alarmów przedstawiono na rys. 3.13. Większość alarmów nie wynikała z przesunięcia struktury liczby zachorowań, a z bardzo dużej liczby przypadków zachorowań. Sugeruje to niebezpieczną sytuację epidemiologiczną w tamtym okresie. Warto dodać, że w przypadku większości województw, w których odnotowano alarm związany z przesunięciem struktury, maksimum liczby zachorowań i tak przewyższało maksimum granicy alarmu w sezonie. Daje to jednoznaczny sygnał do zintesyfikowania przeciwdziałaniu grypie.

W ostatnim analizowanym sezonie (2009-2010) okres największej liczby zachorowań na grypę nastąpił dużo wcześniej niż zazwyczaj. W tej sytuacji model Farringtona zareagował bardzo szybko, wykazując liczne alarmy w województwach oraz w całym kraju. Mimo wszystko z rys. 3.10, 3.11 oraz 3.12 wynika także znaczne zwiększenie liczby zachorowań (wielokrotne przekroczenie maksimum granicy alarmu dla sezonu).

Z przeprowadzonej analizy wydaje się, że model Farringtona dość dobrze sprawdza się w przypadku danych dotyczących zachorowań na grypę w Polsce. Bardzo dobrze wykrywa przede wszystkim przesunięcia struktury liczby zachorowań oraz znaczne zwiększenie liczby zachorowań.

Ponieważ grypa jest chorobą zakaźną, spodziewać się można, że jeżeli alarmy występują w kilku województwach, to województwa te powinny ze sobą sąsiadować lub znajdować się we względnie niedużej odległości. Z map 3.13 wynika, że model Farringtona dosyć dobrze oddaje zakaźny charakter choroby, tzn. we wszystkich badanych sezonach grypowych (oprócz sezonu 2007-2008) województwa, w których wykryto alarmy leżą względnie blisko siebie.

Kolejnym atutem modelu Farringtona jest bardzo dobre odzwierciedlenie wahań sezonowych. Widać to zarówno na wykresie dla Polski (rys. 3.9), jak i na wykresach dla poszczególnych województw (np. Mazowieckiego, rys. 3.11). Postać modelu Farringtona pozwala zatem na modelowanie zjawisk z różnymi wahaniami sezonowymi.



Rysunek 3.13: Liczba odnotowanych alarmów w poszczególnych województwach w wybranych sezonach grypowych - model Farringtona.

Oprócz wspomnianych atutów, model Farringtona posiada również kilka wad. W kontekście analizowanego w pracy zjawiska, zauważyć można, że zdarzało się, że model zbyt pochopnie sygnalizował alarmy związane z przesunięciem struktury zachorowań (np. w przypadku województwa Małopolskiego w sezonie grypowym 2005-2006). Zaznaczyć jednak należy, że wniosek taki wysnuć można jedynie z perspektywy czasu, a nie w momencie zaistnienia alarmu. Można uniknąć, ustawiając bardziej restrykcyjny warunek konieczny powstania alarmu (w funkcji programu R parametr `limit54`), jednakże należy wtedy uważać na pominięcie ważnych alarmów.

Inną wadą modelu Farringtona jest konieczność kompromisowego wyboru liczby obserwacji zbioru uczącego. Chcąc zwiększyć precyzję modelu konieczne jest zwiększenie obserwacji w zbiorze uczącym. Implikuje to zwiększenie liczby lat odniesienia (parametr `b` w specyfikacji modelu) lub zwiększenie liczby uwzględnionych w zbiorze uczącym obserwacji z każdego z wybranych lat odniesienia [5]. W pierwszym przypadku zbyt duże `b` prowadziłoby zbyt powolnego dostosowywania modelu do zmian w takich obszarach jak: profilaktyka i zapobieganie zachorowaniom na grypę. W drugim przypadku wahania sezonowe nie byłyby już tak dobrze odzwierciedlane [5]. Zaznaczyć jednak należy, że wspomniany dylemat wyboru obserwacji odniesienia jest mankamentem wielu algorytmów i modeli do wykrywania stanów alarmowych.

Podsumowując, mimo kilku wad model Farringtona wydaje się dość dobrze wykrywać stany alarmowe liczby zachorowań na grypę w Polsce. Potwierdza to powyższa analiza. W związku z tym model ten mógłby być stosowany jako narzędzie wspomagające wykrywanie podwyższonej liczby zachorowań na grypę.

3.5. Metoda RKI

W poniższym podrozdziale przedstawiona zostanie analiza danych rzeczywistych z wykorzystaniem metody RKI. Na początek zaprezentowana zostanie specyfikacja modelu, następnie analiza dla Polski i dla poszczególnych województw wraz z interpretacją wyników.

3.5.1. Specyfikacja

Podobnie jak w przypadku analizy z wykorzystaniem modelu Farringtona postanowiono włączyć do zbioru uczącego obserwacje z czterech ostatnich lat ($b = 4$) oraz ustawić szerokość okna w na 4 ($w = 4$). Uzasadnienie wyboru takich właśnie wartości znaleźć można w rozdziale 3.4.1. Ponadto zdecydowano się nie wykorzystywać obserwacji z bieżącego roku ($\texttt{actY} = FALSE$) w celu uniknięcia sytuacji, w której sygnalizowany alarm w bieżącym roku implikowałby brak alarmu w analizowanym momencie.

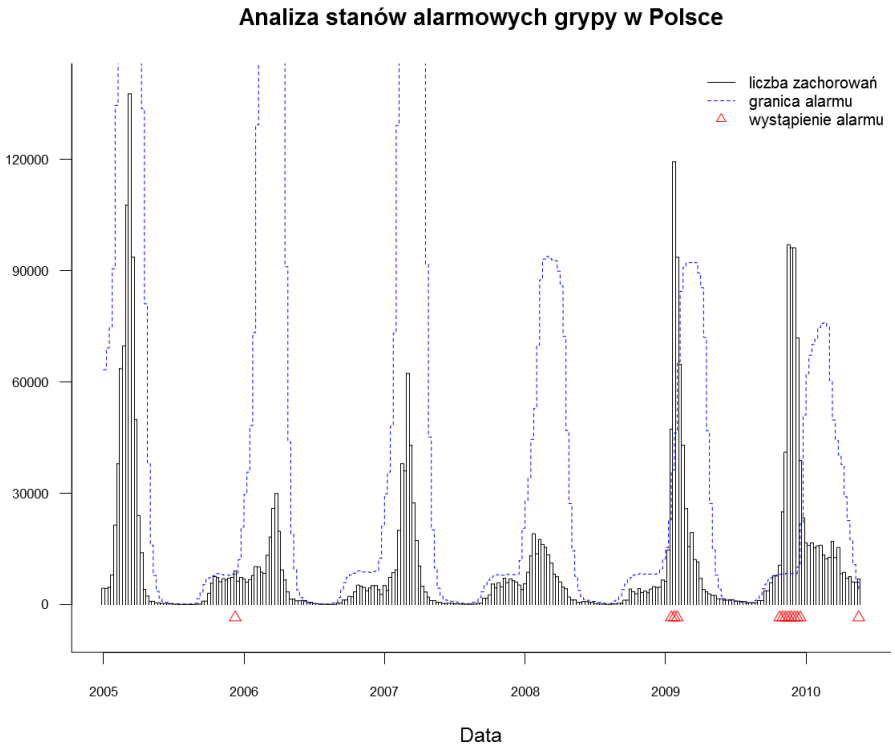
W celu uzyskania porównywalności wyników uzyskanych metodą RKI i przy pomocy modelu Farringtona, wprowadzono analogiczny warunek konieczny zasygnalizowania alarmu. Mianowicie alarm może być wywołany pod warunkiem, że w ciągu ostatnich trzech okresów odnotowano odpowiednio 600 zachorowań dla województw i 4000 dla Polski. Pozwoli to także na uniknięcie fałszywych alarmów w sezonie letnim. Odpowiednią linię kodu znaleźć można w załączniku (str. 66).

3.5.2. Analiza danych dla Polski i dla poszczególnych województw

W poniższym rozdziale przedstawiona zostanie analiza danych rzeczywistych z wykorzystaniem metody RKI ze specyfikacją z podrozdziału 3.5.1. Część ta nie będzie tak rozbudo-

wana jak w przypadku modelu Farringtona, ponieważ w dużej części wyniki uzyskane tymi dwiema metodami dają podobne rezultaty.

Na początek na rys. 3.14 przedstawiono wyniki dla całego kraju.



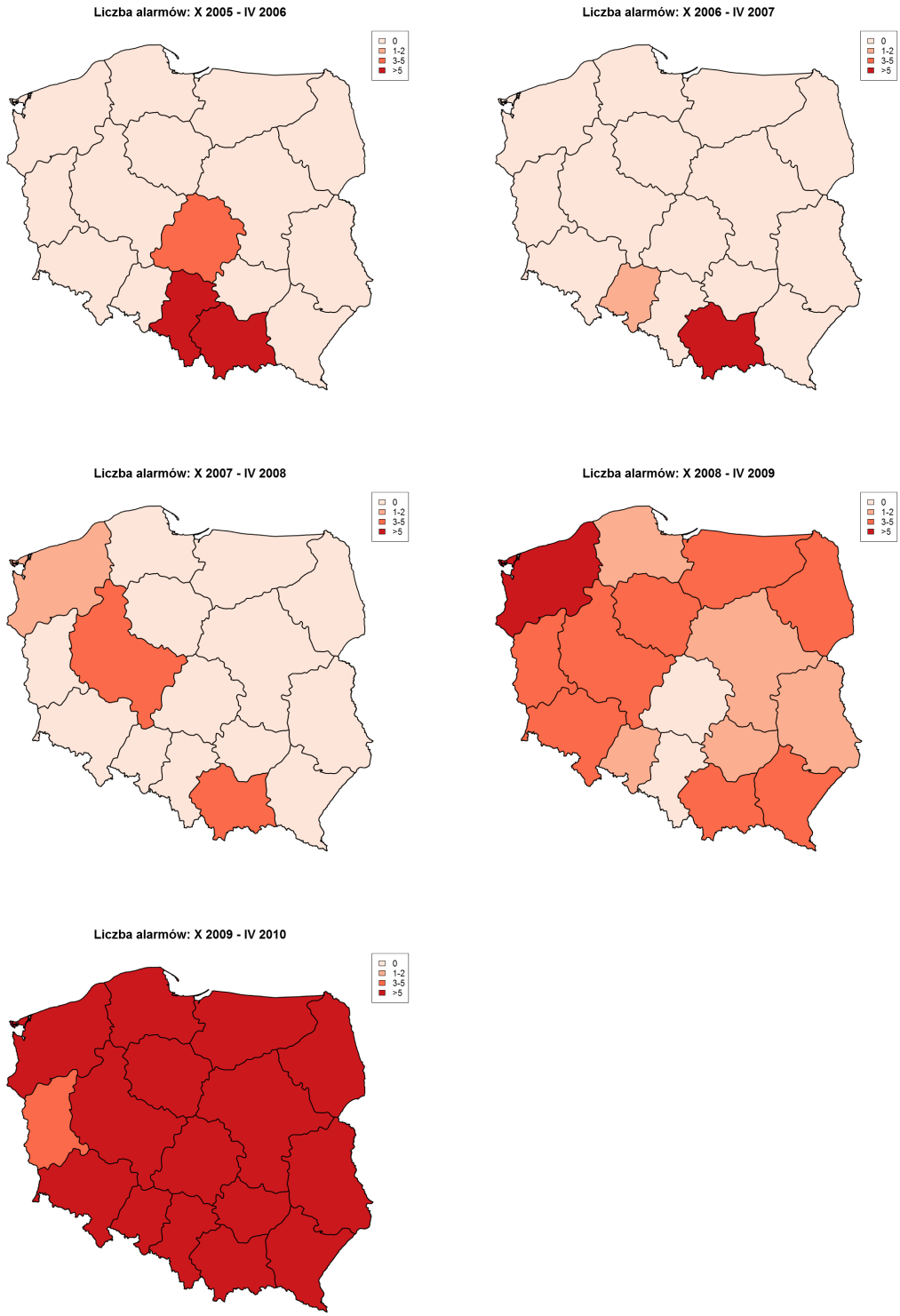
Rysunek 3.14: Analiza stanów alarmowych liczby zachorowań na grypę w Polsce przy wykorzystaniu metody RKI.

Wyniki otrzymane metodą RKI są identyczne jak w przypadku modelu Farringtona z poprzedniego podrozdziału. Mianowicie metoda zasygnalizowała jeden alarm w sezonie grypowym 2005-2006 roku związany z lekką zmianą struktury liczby zachorowań, 3 alarmy w sezonie 2008-2009 oraz całą serię alarmów w sezonie 2009-2010 związaną zarówno z przesunięciem struktury i liczby zachorowań na grypę.

Kolejnym etapem analizy było przeprowadzenie badania dla poszczególnych województw. W celu zwiększenia przejrzystości analizy zdecydowano się nie umieszczać w tym podrozdziale poszczególnych szeregów czasowych wraz z granicami alarmów i zaznaczonymi alarmami dla każdego z województw jak w przypadku analizy modelem Farringtona. Ograniczono się jedynie do przedstawienia map liczby odnotowanych alarmów w poszczególnych województwach - rys. 3.15. Odpowiednie wykresy znaleźć można w załączniku (B.1 - B.3, str. 69-70), podobnie jak kod w R do sporządzenia zarówno map jak i wspomnianych wykresów.

Z rys. 3.15 oraz z wykresów w załączniku wynika, że o ile wyniki w przypadku całego kraju nie różniły się, to wyniki na poziomie województw są zbliżone, ale nie identyczne.

W przypadku sezonu grypowego 2005-2006 alarmy odnotowano w województwach Małopolskim, Śląskim, Łódzkim. W przypadku województw Małopolskiego przyczyną wystąpienia alarmu była zmiana struktury liczby zachorowań w ciągu roku. Z perspektywy czasu wydaje się, że zmiana ta nie była znacząca, w związku z czym można uznać, że odnotowany alarm był fałszywy. W przypadku pozostałych województw oprócz zmiany wspomnianej struktury



Rysunek 3.15: Liczba odnotowanych alarmów w poszczególnych województwach w wybranych sezonach grypowych - metoda RKI.

zauważono także znaczący wzrost liczby zachorowań dla okresu jesiennego. W sezonie 2006-2007 podniesiono alarmy w województwach Małopolskim (na początku sezonu) i Opolskim (w momencie największej liczby zachorowań).

W sezonie 2007-2008 wykryto alarmy w województwach Małopolskim, Wielkopolskim i Zachodniopomorskim. Wszystkie wynikały nie z drastycznego wzrostu liczby zachorowań, ale raczej ze zmiany struktury zachorowań. W przypadku województw Małopolskiego i Zachodniopomorskiego podniesione alarmy wydają się być niepotrzebne i nieznaczące z perspektywy czasu. Z kolei w przypadku województwa Wielkopolskiego przesunięciu uległ okres największej zachorowalności.

Kolejne dwa okresy cechują się licznymi alarmami w niemal wszystkich województwach. W przypadku sezonu 2008-2009 większość alarmów wynikała ze zwiększonej liczby zachorowań. W mniejszym stopniu odpowiedzialne za alarmy były zmiany w strukturze liczby zachorowań. W przypadku sezonu 2009-2010 nałożyły się obydwa czynniki: po pierwsze cała struktura liczby zachorowań przesunęła się znacząco, tzn. okres największej zachorowalności przyszedł wcześniej niż zwykle, w wielu momentach odnotowano dużo większą zachorowalność niż wynikało to z granicy alarmu wyznaczonej metodą RKI.

Wydaje się, że metoda RKI dość dobrze sprawdza się w wykrywaniu stanów alarmowych grypy w Polsce. Wykrywa ona zarówno zmiany struktury jak i zwiększenie liczby zachorowań. Wyniki uzyskane przy pomocy tej metody odzwierciedlają charakter grypy jako choroby zakaźnej (rys. 3.15).

Metoda RKI dobrze odzwierciedla również wahania sezonowe. Widać to po kształcie granicy alarmu. Jest ona bardzo czuła na wszekie ich zaburzenia. Można to uznać zarówno za atut jak i za wadę. Wydaje się bowiem, że w przypadku danych dotyczących grypy w Polsce metoda ta czasami zbyt pochopnie sugerowała wystąpienie alarmu.

Innym mankamentem jest konieczność kompromisowego wyboru liczby obserwacji odniesienia. Szerzej problem ten został opisany w podrozdziale dotyczącym analizy przy użyciu modelu Farringtona (podrozdział 3.4.4).

Należy także przyznać, że dużą zaletą metody RKI jest jej prostota, a co za tym idzie szybkość wykonywania analiz przy pomocy pakietów statystycznych. Może to być ważne w przypadku przeprowadzania analiz na setkach zbiorów.

Podsumowując, metoda RKI mimo swych wad dość dobrze sprawdza się w wykrywaniu stanów alarmowych grypy w Polsce. Jej dużą czułość na małe zmiany w strukturze zachorowań można ograniczyć poprzez nałożenie bardziej restrykcyjnych warunków, o których była mowa w podrozdziale 3.5.1. Uznać jednak można, że metoda ta byłaby przydatna jako narzędzie wspomagające analizę podwyższonych stanów zachorowań.

3.6. Porównanie wyników

W niniejszym podrozdziale porównane zostaną po krótce wyniki wykrywania stanów alarmowych uzyskane różnymi metodami. Wnioski uzyskane w tym rozdziale bazują wyłącznie na wynikach dotyczących liczby zachorowań na grypę w Polsce i nie mogą zostać uogólnione bez dodatkowych analiz, które nie są przedmiotem tej pracy. Na wstępie przypomnieć należy, że z powodu mankamentów opisanych w podrozdziale 3.3.4, uznano, że algorytm CUSUM nie nadaje się do wykrywania stanów alarmowych zachorowań na grypę w Polsce. W związku z tym został on wyłączony z poniższego porównania.

Po pierwsze zaznaczyć należy, że wyniki uzyskane przy pomocy modelu Farringtona i metody RKI dały zbliżone, choć nie identyczne rezultaty. Widać to przede wszystkim na zagregowane dane dla całego kraju (rys. 3.9 i 3.14). Nie mniej jednak kształt granicy alar-

mu na uzyskanych wykresach oraz same alarmy sugerują pewne nieznaczne różnice między metodami.

Mapy sporządzone dla obu metod (rys. 3.13 i 3.15) wskazują na to, że metoda RKI wskazuje nieco częściej alarmy niż model Farringtona (np. dla województwa Małopolskiego w sezonie grypowym 2006-2007). Alarmy sygnalizowane przez metodę RKI a nie sygnalizowane przez model Farringtona występują przede wszystkim jesienią (np. woj. Małopolskie w sezonie 2007-2008 lub woj. Zachodniopomorskie w sezonie 2008-2009). Stąd wynika, że w przypadku danych o zachorowaniach na grypę w Polsce metoda RKI jest bardziej wrażliwa na lekkie zmiany w strukturze na początku okresu grypowego.

Przy ustalonych parametrach `b` i `w` zauważyć też można różnice w kształcie granicy alarmu na odpowiednich wykresach. W przypadku metody RKI wykres granicy alarmu jest bardziej „wygładzony", tzn. w sezonie grypowym różnice pomiędzy wartościami granic dla kolejnych momentów w czasie są mniejsze niż w przypadku modelu Farringtona. Szczególnie widoczne jest to w okresie największych zachorowań na grypę, czyli na początku roku kalendarzowego (np. woj. Kujawsko-Pomorskie, woj. Lubelskie, woj. Lubuskie, woj. Podkarpackie i inne). Nie jest to co prawda regułą dla wszystkich województw, jednakże w dużej większości przypadków zaznaczona własność została zaobserwowana. W związku z tym, wydaje się, że model Farringtona jest bardziej czuły na przesunięcia okresu maksymalnej liczby zachorowań.

Ważnym atutem metody RKI jest jej prostota i związany z nim czas wykonania odpowiednich funkcji w pakietach statystycznych. Przy wykorzystaniu funkcji `system.time` przeprowadzono krótką analizę czasu wykonywania odpowiednich funkcji. Jej wyniki jednoznacznie wskazują, że analiza metodą RKI wykonywana jest dużo szybciej. Przykładowo wykonanie funkcji `algo.farrington` dla danych dla Polski zajęło średnio 1,21 sekundy, zaś funkcji `algo.RKI` średnio 0,04 sekundy, czyli około 30 razy szybciej. Powyższe wyniki podano na podstawie 1000 symulacji. Dodatkowo zaznaczyć należy, że wykonanie całej analizy (wraz z wczytaniem pakietów i zbioru danych, wykonaniem funkcji i sporządzeniem odpowiednich wykresów wraz z zapisem na dysku) zajęło 25,10 sekund dla modelu Farringtona i 4,38 sekundy dla metody RKI (kody w załączniku). Podsumowując, model Farringtona jest dużo bardziej czasochłonny. Analizę przeprowadzono na komputerze z procesorem Intel Core i5-2430M 2.4 GHz, z pamięcią 4GB RAM, przy pomocy 64-bitowej wersji programu R.

Porównane dwie metody różnią się nieznacznie między sobą, a w przypadku badanego zjawiska dają zbliżone wyniki. Nie można wskazać metody lepszej, ponieważ wybór metody do analizy powinien zależeć od badanego zjawiska oraz celu analizy. Przykładowo, jeżeli celem jest wykrycie przesunięć się maksimów zachorowań na grypę w Polsce właściwszy byłby model Farringtona. W przypadku bardzo dużych zbiorów danych metoda RKI może być bardziej przydatna z powodu szybkości wykonania analiz przy jej pomocy.

3.7. Zestawienie uzyskanych wyników z rzeczywistością

W Polsce, co roku Główny Inspektorat Sanitarny (GIS) publikuje raport dotyczący stanu sanitarnego kraju. W niniejszym podrozdziale wyniki uzyskane w poprzednich podrozdziałach zostaną zestawione z rzeczywistymi obserwacjami pracowników GIS.

Zgodnie z uzyskanymi wynikami, w sezonach grypowych 2005-2006, 2006-2007 oraz 2007-2008 przyczynami zasygnalizowania alarmów były zmiany w strukturze liczby zachorowań na grypę (głównie w okresie jesiennym). W Raporty Głównego Inspektoratu Sanitarnego ([28], [29], [30]) nie wspomniano o jakiejkolwiek zmianie struktury zachorowań. Jest to potwierdzenie przypuszczeń z poprzednich rozdziałów, że alarmy te były fałszywe. Wyjaśnieniem zjawiska nieznacznej zmiany struktury liczby odnotowywanych zachorowań może być zagro-

żenie związane z grypą ptaków H5N1 (tzw. „ptasia grypa"). Co prawda z raportów ([28], [29], [30]) wynika, że w latach 2006-2008 nie odnotowano w Polsce zakażeń wirusem H5N1, jednakże informacje na temat wirusa mogły skłonić ludzi do wcześniejszego zgłaszania się do lekarza w razie grypopodobnych objawów. W przypadku sezonu grypowego 2008-2009 raporty Głównego Inspektoratu Sanitarnego nie pozwalają na wysnucie przyczyn zwiększonej w porównaniu do lat wcześniejszych liczby zachorowań.

We wcześniejszych podrozdziałach zauważono, że w całym kraju odnotowano liczne alarmy w sezonie grypowym 2009-2010. Związane one były z przesunięciem struktury liczby zachorowań, przede wszystkim okresu największej zachorowalności. Przyczyną tego zjawiska było niewątpliwie wirusa A(H1N1)v nazywanego początkowo terminem „świńska grypa". Pierwsze ogniska zachorowań wywołane tym właśnie szczepem zidentyfikowano w Meksyku w kwietniu 2009 roku [31]. Wirus bardzo szybko rozprzestrzeniał się na całym świecie, co skutkowało ogłoszeniem 11 czerwca 2009 roku pandemii grypy przez Światową Organizację Zdrowia (WHO). W Polsce pierwsze zachorowania odnotowano w maju 2009 r. Wystąpienie wirusa A(H1N1)v spowodowało „epidemiczny wzrost zachorowań na grypę z początkiem listopada to znaczy dwa miesiące wcześniej niż w latach ubiegłych, gdy epidemie były wywoływane przez szczepy sezonowe" [31]. GIS podaje także, że nowy szczep grypy dominował wśród wszystkich wirusów grypy na terenie Polski w czasie sezonu grypowego 2009-2010. W związku z zaprezentowanymi informacjami udostępnionymi przez GIS, uznać należy, że w tym przypadku zarówno model Farringtona jak i metoda RKI wzorowo zasygnalizowała niepokojące zmiany w zachorowalności na grypę.

Z powyższych rozważań wynika, że w wielu przypadkach model Farringtona i metoda RKI dobrze wykrywały stany alarmowe, zwłaszcza w sytuacjach najgroźniejszych. Potwierdza to, że metody te znalazłyby zastosowanie w wykrywaniu stanów alarmowych grypy w Polsce.

Podsumowanie

Celem niniejszej pracy było porównanie algorytmów CUSUM, RKI i Modelu Farringtona w zastosowaniach do wykrywania stanów alarmowych w zachorowaniach na grypę w Polsce. Wyniki badań przeprowadzonych za pomocą pakietu statystycznego R ukazały jak istotny jest odpowiedni wybór algorytmu w zastosowaniach praktycznych.

Podobieństwa wynikające z wyników działania algorytmu RKI i Modelu Farringtona sugerują, że przy wyborze algorytmu do monitoringu epidemiologicznego nie powinno się kierować teoretyczną wyższością bardziej skomplikowanego algorytmu. Mimo bardziej wyrafinowanych szczegółów w umiejętności wykrywania wahań sezonowych i mniejszej wrażliwości na zmianę w strukturze zachorowań, podczas implementacji algorytmów do danych rzeczywistych dla dużych baz danych bardziej stosowalny może być algorytm o szybszym czasie działania i mniej skomplikowanej implementacji.

Różnice wynikające z wyników zastosowania RKI i Modelu Farringtona a algorytmu CUSUM pokazały, że przy strukturze zachorowań w Polsce w latach 2000-2009 bardziej odpowiednie są dwa pierwsze z nich. Błędne sygnalizowanie dużej ilości fałszywych alarmów przez algorytm CUSUM pokazało, jak ważne przy implementacji tego algorytmu do danych rzeczywistych jest sprawdzenie, czy dane po zaproponowanej standaryzacji mają w przybliżeniu rozkład $N(0,1)$, gdyż w sytuacji takiej, jaka miała miejsce dla danych o zachorowaniach na grypę w Polsce, aproksymacyjny algorytm CUSUM błędnie odzwierciedlał sytuację. Wyniki otrzymane przy zastosowaniu algorytmu CUSUM pokazują również potrzebę nowej metody standaryzacji danych.

Wszystkie algorytmy wykazały sytuację alarmową w końcu 2009 roku, co sugeruje istotność epidemii, która miała miejsce w Polsce, jako skutek światowej pandemii wirusu A/H1N1. Algorytmy RKI i Model Farringtona, na podstawie dokonanej analizy danych, wcześnie wykryły nadchodzącą epidemię, w związku z czym mogłyby być wykorzystane w systemie monitoringu epidemiologicznego w Polsce. Algotym CUSUM powinien być zmodyfikowany do zastosowań praktycznych.

Dodatek A

Kody pakietu R użyte w pracy

Kod A.1: Stworzenie wykresu liczby zachororwań na grypę w Polsce w czasie.

```
CairoPNG("C:/Users/Kuba/Desktop/licencjat/dane/polska_liczba.png",
   width=1400, height=800)
par(bty="l", cex.axis=0.7, cex=2)
plot(dane[,17], type="h", xaxt="n", yaxt="n", ylab="Liczba␣zachorowań",
   xlab="Data", main="Liczba␣zachorowa␣na␣grypę␣w␣Polsce")
axis(side=2, at=seq(0,500000, by=50000), las=1)
axis(side=1, at=seq(0,499,by=48), labels=2000:2010,las=1)
dev.off()
```

Kod A.2: Tworzenie obiektu klasy disProg do analizy przy wykorzystaniu algorytmu CUSUM.

```
a<-read.table("C:/grypa1.csv", sep=",", header=TRUE)
pod <- a[,i+1]
n<-length(pod)
stan<-mat.or.vec(n, 1)
grypapod<-create.disProg(week=2:n, observed=pod, state=stan, start=c(2000,1),
 freq=48)
```

Kod A.3: Wykresy wyników działania algorytmu CUSUM dla poszczególnych województw stworzono przy wykorzystaniu pakietu Cairo.

```
daty=c(2005,2006,2007,2008,2009,2010)
nazwy<-read.table("C:/nazwy.csv", sep=",")
d=seq(1,259, by=48)
e=seq(0,30000, by=5000)
legend.cex=0.8
CairoPNG("C:/rysunek.png", width=229*8, height=324*8)
par(bty="l", cex.axis=0.8, mfrow=c(3,2), cex=2)
for (i in 1:6) {
  nazwa<-nazwy[,i]
  woj<-a[,i+1]
   n<-length(woj)
  stan<-mat.or.vec(n, 1)
  grypadisProg<-create.disProg(week=2:n, observed=woj, state=stan,
start=c(2000,1), freq=48)
  survglm<-algo.cusum(grypadisProg, control = list(range = (n-258):n,
 k =10, h =4, m="glm", trans="rossi"))
   plot(survglm, method="CUSUM", disease="Grypa", xaxis.years=F, yaxt="n",
 xaxt="n", startyear=(2005),
   ylim=c(0,17000), xlab="Data", ylab="",
 main=paste("Wojewodztwo:␣",nazwa,sep=""), legend=NULL)
```

```
  axis(side=1, at=d, labels=daty, las=0)
  axis(side=2, at=e, labels=e, las=1)
  legend("topright", lty=c(1,2,-1),pch=c(-1,-1,2),
        col=c("black","blue","red"), cex=legend.cex, bty="n",
        legend=c("liczba zachorowan", "granica alarmu", "wystapienie alarmu"))
    }
   dev.off()
```

Do stworzenia wykresów statystyki CUSUM i standaryzowanych danych posłużono się ręcznym wywołaniem funkcji `algo.cusum` (wraz z uwzględnieniem modyfikacji opisanych w rozdziale 3.3.1).

Kod A.4: Ręczne wywołanie algorytmu CUSUM z odpowiednimi modyfikacjami

```
CairoPNG("C:/cusum.png", width=229*8, height=324*8)
par(bty="l", cex.axis=0.8, mfrow=c(3,2), cex=2)
 for (i in 1:6) {
 nazwa<-nazwy[,i]
 pod <- a[,i+1]
 n<-length(pod)
  stan<-mat.or.vec(n, 1)
grypa2mazo<-create.disProg(week=2:n, observed=pod, state=stan,
start=c(2000,1), freq=48)
    control=list(range = (n-258):n, k =2.5, h =4,
 m="glm", trans="rossi")
    observed <- grypa2mazo$observed
    timePoint <- control$range[1]
    training <- 1:(timePoint - 1)
        t <- grypa2mazo$start[2] + training - 1
        x <- observed[training]
        p <- grypa2mazo$freq
        df <- data.frame(x = x, t = t)
        control$m.glm <- glm(x ~ 1 + cos(2 * pi/p * t) + sin(2 *pi/p * t) +
            sin(2*pi/p * 2*t) + cos (2*pi/p * 2*t)
            , family = poisson(), data = df)
        t.new <- grypa2mazo$start[2] + control$range - 1
        m <- predict(control$m.glm, newdata = data.frame(t = t.new),
            type = "response")
    x <- observed[control$range]
    standObs <-(x - 3 * m + 2 * sqrt(x * m))/(2 * sqrt(m))
    standard <- (x - m)/sqrt(m)
    cusum <- matrix(0, nrow = (length(control$range) + 1), ncol = 1)
    alarm <- matrix(data = 0, nrow = (length(control$range) +
        1), ncol = 1)
    for (t in 1:length(control$range)) {
     if ((t+23)%%48==0) {cusum[t]<-0}
     cusum[t + 1] <- max(0, cusum[t] + (standObs[t] - control$k))
        alarm[t + 1] <- cusum[t + 1] >= control$h
      }
    h <- control$h
    k <- control$k
    Ctm1 <- cusum[1:length(control$range)]
    upperbound <- 2 * h * m^(1/2) +
        2 * k * m^(1/2) - 2 * Ctm1 * m^(1/2) + 5 * m - 2 * (4 *
        m^2 + 2 * m^(3/2) * h + 2 * m^(3/2) * k - 2 * m^(3/2) *
        Ctm1)^(1/2)
    upperbound[is.na(upperbound)] <- 0
    upperbound[upperbound < 0] <- 0
    cusum <- cusum[-1]
    alarm <- alarm[-1]
```

```
    control$name <- paste("cusum:", control$trans)
    control$data <- paste(deparse(substitute(grypa2mazo)))
    control$m <- m
    result <- list(alarm = alarm, upperbound = upperbound, disProgObj = grypa2mazo,
        control = control, cusum = cusum)
    class(result) = "survRes"
plot(cusum, lty=1, pch='l', col='white', xaxt="n", ylim=c(0,1500),
 xlab="", ylab="", main=paste("Wojewodztwo: ",nazwa,sep=""))
abline(h=4,col="red",lty=1)
lines(standObs,col='black',lty=1)
lines(cusum, col='green3', lty=1)
tt <- seq(from = 1, to = 259, length = 259)
dtt <- cusum[tt]
polygon(x = c(1, tt, 259), y = c(0, dtt, 0), col = "#7FFF0022", border=NA)
axis(side=1, at=d, labels=daty, las=0)
legend("top", lty=c(1,1,1),pch=c(-1,-1,-1),
        col=c("green","black","red"), cex=legend.cex, bty="n",
        legend=c("CUSUM", "Standaryzowane dane", "Granica alarmowa"))
  }
  }
   dev.off()
```

Kod A.5: Analiza stanów alarmowych zachorowań na grypę z wykorzystaniem modelu Farringtona i pakietów pomocniczych do tworzenia wykresów i map

```
library(surveillance)
library(Cairo)
library(maptools)
library(RColorBrewer)
library(classInt)

dane<-read.table("C:/Users/Kuba/Desktop/licencjat/dane/proba/1tyg_kr.csv",
      header=TRUE, sep=";")
wojewodztwa=names(dane)
daty_fr<-read.table("C:/Users/Kuba/Desktop/licencjat/dane/daty.csv",
      header=FALSE, sep=";")
daty=daty_fr[,1]
liczba.map=5
liczba=16

# deklaracja macierzy do wypisywania liczby alarmow: w kolumnach
   poszczegolne wojewodztwa dla poszczegolnych lat
macierz.alarmow=1:(liczba.map*liczba)
macierz.alarmow=macierz.alarmow*0-1
dim(macierz.alarmow)=c(liczba.map,liczba)
e=seq(0,300000, by=5000)
d=seq(1,259, by=48)

#do wypisywania numerow kolejnych rysunkow
k=20
legend.cex=0.8

for (i in 1:liczba)  {
   wektor<-dane[,i]
   dp<-create.disProg(week=1:length(wektor),observed=wektor, state=wektor*0,
      freq=48)
   n<-length(dp$observed)
   control <- list(b=4,w=4,range=(n-258):n,reweight=TRUE, trend=F,
      verbose=FALSE,alpha=0.01,limit54=c(600,3))
   res <- algo.farrington(dp,control=control)
```

```
   sums=seq(-1,-1,length.out=liczba.map)
   for (n in 1:liczba.map) {
      okres=res$alarm[(33+48*(n-1)):(60+48*(n-1))]
      suma.okres=sum(okres)
      if (suma.okres==0) {suma.okres=runif(1,min=0,max=0.05)}
      macierz.alarmow[n,i]=suma.okres
   }
   tytul=paste("Województwo: ",wojewodztwa[i])
   if ((i%%6)==1)
      {rysunek=paste("C:/Users/Kuba/Desktop/licencjat/dane/farrington",k,".png",
      sep="")
      k=k+1
      if (liczba-i+1<6){
         if ((liczba-i+1)%%2==0)
            {CairoPNG(rysunek,width=229*8,height=324*8*(liczba-i+1)/6)
               par(bty="l",cex.axis=0.8, mfrow=c((liczba-i+1)/2,2), cex=2)}
               else {CairoPNG(rysunek,width=229*8,height=324*8*(liczba-i+2)/6)
               par(bty="l",cex.axis=0.8, mfrow=c((liczba-i+2)/2,2), cex=2)}
         } else {
               CairoPNG(rysunek,width=229*8,height=324*8)
               par(bty="l",cex.axis=0.8, mfrow=c(3,2), cex=2)}
         plot(res,disease="Salmonella",method="Farrington", xaxis.years=F,
               ylim=c(0,17000), xaxt="n", yaxt="n",
               legend.opts=NULL, ylab="", xlab="Data", main=tytul)
               axis(side=1, at=d, labels=daty, las=0)
         axis(side=2, at=e, labels=e, las=1)
         legend("topright", lty=c(1,2,-1),pch=c(-1,-1,2),
         col=c("black","blue","red"), cex=legend.cex, bty="n",
         legend=c("liczba zachorowań", "granica alarmu", "wystąpienie alarmu"))
      } else {
          plot(res,disease="Salmonella",method="Farrington", xaxis.years=F,
               ylim=c(0,17000), xaxt="n", yaxt="n",
               legend.opts=NULL, ylab="", xlab="Data", main=tytul)
          axis(side=1, at=d, labels=daty, las=0)
          axis(side=2, at=e, labels=e, las=1)
          legend("topright", lty=c(1,2,-1),pch=c(-1,-1,2),
          col=c("black","blue","red"), cex=legend.cex, bty="n",
          legend=c("liczba zachorowań", "granica alarmu", "wystąpienie alarmu"))
      }
   if (i%%6==0 || i==liczba) {
   dev.off()
        }
}

woj<-readShapePoly("C:/Users/Kuba/Desktop/licencjat/dane/POL_adm1.shp",
   proj4string=CRS("+proj=longlat+ellps=clrk80"))
CairoPNG("C:/Users/Kuba/Desktop/licencjat/dane/mapka20.png",
   width=229*10,height=324*10)
par(mfrow=c(3,2), mar=c(5, 4, 4, 2) + 0.1 - 1, cex=2)
przedzialy <- 4
podzial=c(0,1,3,6,30)
kolory <- brewer.pal(przedzialy, "Reds")
for (i in 1:liczba.map) {
    klasy=classIntervals(macierz.alarmow[i,],style="fixed", fixedBreaks=podzial,
    n=przedzialy)
    tytul=paste("Liczba alarmów: X ",2005+i-1," - IV ", 2005+i ,sep="")
    tabela.kolorow<-findColours(klasy, kolory)
    malowanie = as.vector(tabela.kolorow)
      [c(3,13,15,2,6,1,4,5,7,8,10,11,12,9,14,16)]
    plot(woj, col=malowanie, lwd=2)
```

```
    title(main=tytul)
    legend("topright", fill=kolory, cex=0.8, #bty="n",
         legend=c("0", "1-2", "3-5",">5"))
}
dev.off()

wektor<-dane[,17]
dp<-create.disProg(week=1:length(wektor),observed=wektor, state=wektor*0,
   freq=48)
n<-length(dp$observed)
control <- list(b=4,w=4,range=(n-258):n,reweight=TRUE, trend=F,
   verbose=FALSE,alpha=0.01,limit54=c(4000,3))
res <- algo.farrington(dp,control=control)
CairoPNG("C:/Users/Kuba/Desktop/licencjat/dane/far_polska.png",
   width=1200,height=1000)
par(bty="l", cex.axis=0.7, cex=2)
plot(res,disease="Salmonella",method="Farrington", xaxis.years=F,
   ylim=c(0,140000), xaxt="n", yaxt="n",
   legend.opts=NULL, ylab="", xlab="Data",
      main="Analiza␣stanów␣alarmowych␣grypy␣w␣Polsce")
axis(side=1, at=d, labels=daty, las=0)
axis(side=2, at=seq(0,300000,by=30000), labels=seq(0,300000,by=30000), las=1)
legend("topright", lty=c(1,2,-1),pch=c(-1,-1,2),
   col=c("black","blue","red"), cex=legend.cex, bty="n",
   legend=c("liczba␣zachorowań", "granica␣alarmu", "wystąpienie␣alarmu"))
dev.off()
```

Powyższy kod służy do wykonania wszystkich analiz i rysunków uwzględnionych w pracy, a odnoszących się do modelu Farringtona. W pierwszej kolejności wykonywane jest wczytanie wykorzystywanych pakietów oraz wczytanie danych. Następnie zadeklarowane wykorzystywane później w skrypcie zmienne. W pętli `for` wykonywane są analizy dla poszczególnych województw oraz odpowiednie rysunki. W między czasie tworzone są pliki na dysku w formacie PNG z wykonanmi rysunkami oraz zapisywane dane wykorzystane później przy tworzeniu mapek. Po największej pętli `for` wykonywane są mapki a następnie przeprowadzona jest analiza dla danych dla Polski. W powyższym kodzie wykorzystano funkcje z pakietów:

- `surveillance`- analiza stanów alarmowych [25],
- `Cairo` - wykonanie odpowiednich rysunków i wykresów [20],
- `maptools` - wczytanie danych geograficznych [22],
- `RColorBrewer` - znalezienie palety kolorów do map [23],
- `classInt` [21]- dokonanie podziału województw na odpowiednie klasy oraz dopasowanie do nich kolorów wykorzystanych później do narysowania map.

Kod A.6: Analiza stanów alarmowych zachorowań na grypę wykorzystaniem metody RKI.

```
library(surveillance)
library(Cairo)
library(maptools)
library(RColorBrewer)
library(classInt)

dane<-read.table("C:/Users/Kuba/Desktop/licencjat/dane/proba/1tyg_kr.csv",
   header=TRUE, sep=";")
wojewodztwa=names(dane)
daty_fr<-read.table("C:/Users/Kuba/Desktop/licencjat/dane/daty.csv",
```

```
header=FALSE, sep=";")
daty=daty_fr[,1]
liczba.map=5
liczba=16

# deklaracja macierzy do wypisywania liczby alarmow: w kolumnach poszczegolne
wojewodztwa dla poszczegolnych lat
macierz.alarmow=1:(liczba.map*liczba)
macierz.alarmow=macierz.alarmow*0-1
dim(macierz.alarmow)=c(liczba.map,liczba)
e=seq(0,300000, by=5000)
d=seq(1,259, by=48)
i=2

#do wypisywania numerow kolejnych rysunkow
k=20
legend.cex=0.8
limit=seq(0,0,length.out=259)

for (i in 1:liczba)  {
   wektor<-dane[,i]
   dp<-create.disProg(week=1:length(wektor),observed=wektor, state=wektor*0,
      freq=48)
   n<-length(dp$observed)
   control <- list(range=(n-258):n,b=4,w=4)
   res <- algo.rki(dp,control=control)

   for (z in 1:259) {
      limit[z] = res$disProgObj$observed[z+240]+res$disProgObj$observed[z+240-1]
         +res$disProgObj$observed[z+240-2]
      if (limit[z]<600) {res$alarm[z]=0} }
   sums=seq(-1,-1,length.out=liczba.map)
   for (n in 1:liczba.map) {
      okres=res$alarm[(33+48*(n-1)):(60+48*(n-1))]
      suma.okres=sum(okres)
      if (suma.okres==0) {suma.okres=runif(1,min=0,max=0.05)}
      macierz.alarmow[n,i]=suma.okres
    }
   tytul=paste("Województwo: ",wojewodztwa[i])
   if ((i%%6)==1)  {
      rysunek=paste("C:/Users/Kuba/Desktop/licencjat/dane/rki",k,".png",sep="")
      k=k+1
      if (liczba-i+1<6){
         if ((liczba-i+1)%%2==0)
            {CairoPNG(rysunek,width=229*8,height=324*8*(liczba-i+1)/6)
            par(bty="l",cex.axis=0.8, mfrow=c((liczba-i+1)/2,2), cex=2)}
         else {CairoPNG(rysunek,width=229*8,height=324*8*(liczba-i+2)/6)
            par(bty="l",cex.axis=0.8, mfrow=c((liczba-i+2)/2,2), cex=2)}
      } else {
      CairoPNG(rysunek,width=229*8,height=324*8)
      par(bty="l",cex.axis=0.8, mfrow=c(3,2), cex=2)}
   plot(res,disease="Salmonella",method="Farrington", xaxis.years=F,
      ylim=c(0,17000), xaxt="n", yaxt="n",
      legend.opts=NULL, ylab="", xlab="Data", main=tytul)
   axis(side=1, at=d, labels=daty, las=0)
   axis(side=2, at=e, labels=e, las=1)
   legend("topright", lty=c(1,2,-1),pch=c(-1,-1,2),
      col=c("black","blue","red"), cex=legend.cex, bty="n",
      legend=c("liczba zachorowań", "granica alarmu", "wystąpienie alarmu"))
    } else {
```

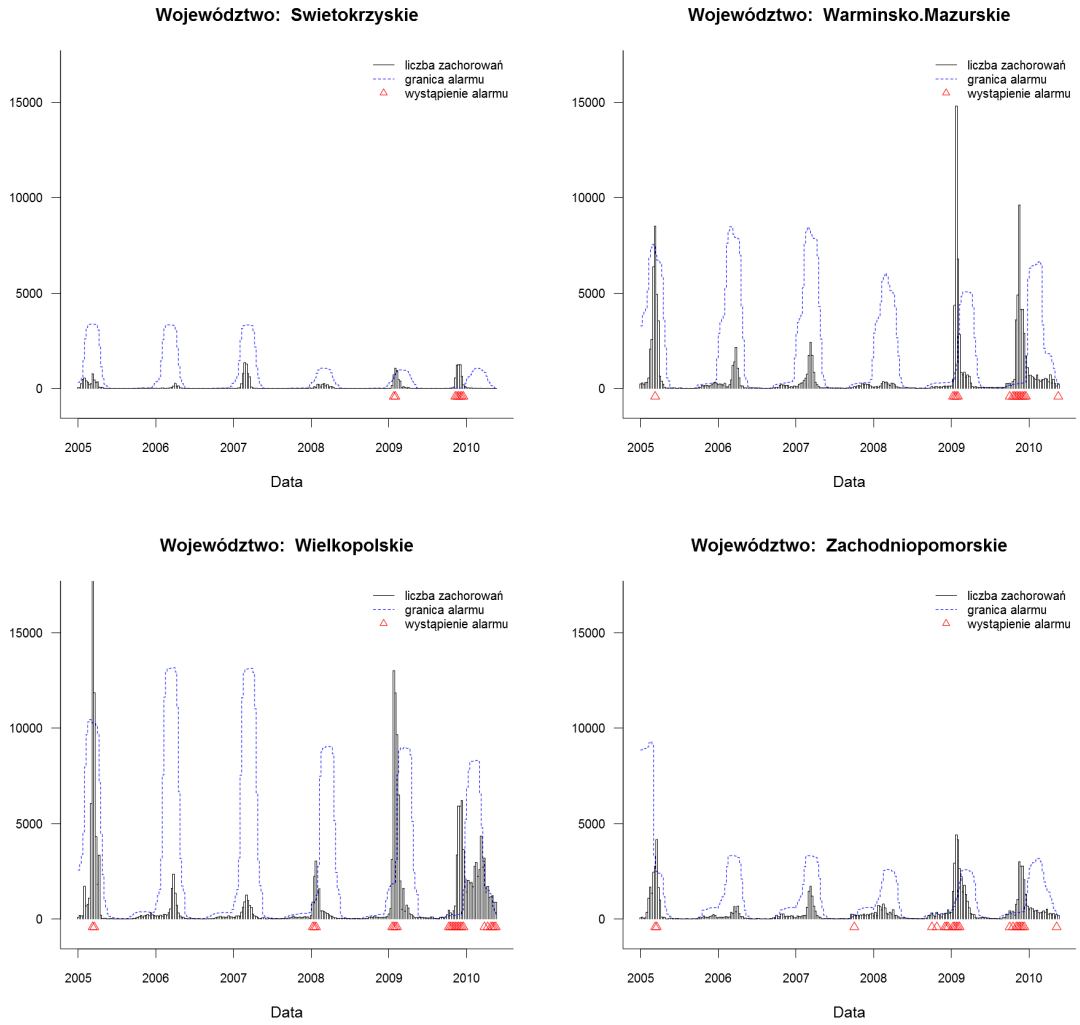
```
      plot(res,disease="Salmonella",method="Farrington", xaxis.years=F,
         ylim=c(0,17000), xaxt="n", yaxt="n",
         legend.opts=NULL, ylab="", xlab="Data", main=tytul)
      axis(side=1, at=d, labels=daty, las=0)
      axis(side=2, at=e, labels=e, las=1)
      legend("topright", lty=c(1,2,-1),pch=c(-1,-1,2),
         col=c("black","blue","red"), cex=legend.cex, bty="n",
       legend=c("liczba␣zachorowań", "granica␣alarmu", "wystąpienie␣alarmu"))
      }
   if (i%%6==0 || i==liczba) {
          dev.off()}
}


#rysowanie mapek
woj<-readShapePoly("C:/Users/Kuba/Desktop/licencjat/dane/POL_adm1.shp",
   proj4string=CRS("+proj=longlat+ellps=clrk80"))


CairoPNG("C:/Users/Kuba/Desktop/licencjat/dane/mapka_rki.png",
   width=229*10,height=324*10)
par(mfrow=c(3,2), mar=c(5, 4, 4, 2) + 0.1 - 1, cex=2)
przedzialy <- 4
podzial=c(0,1,3,6,30)
kolory <- brewer.pal(przedzialy, "Reds")
for (i in 1:liczba.map) {
   klasy=classIntervals(macierz.alarmow[i,],style="fixed", fixedBreaks=podzial,
   n=przedzialy)
   tytul=paste("Liczba␣alarmów:␣X␣",2005+i-1,"␣-␣IV␣", 2005+i ,sep="")
   tabela.kolorow<-findColours(klasy, kolory)
   malowanie = as.vector(tabela.kolorow)
      [c(3,13,15,2,6,1,4,5,7,8,10,11,12,9,14,16)]
   plot(woj, col=malowanie, lwd=2)
   title(main=tytul)
   legend("topright", fill=kolory, cex=0.8, legend=c("0", "1-2", "3-5",">5"))
}
dev.off()

wektor<-dane[,17]
dp<-create.disProg(week=1:length(wektor),observed=wektor, state=wektor*0,
   freq=48)
n<-length(dp$observed)
control <- list(range=(n-258):n,b=4,w=4)
res <- algo.rki(dp,control=control)
for (z in 1:259) {
   limit[z]= res$disProgObj$observed[z+240]+res$disProgObj$observed[z+240-1]+
   res$disProgObj$observed[z+240-2]
   if (limit[z]<4000) {res$alarm[z]=0} }

CairoPNG("C:/Users/Kuba/Desktop/licencjat/dane/rki_polska.png",
   width=1200,height=1000)
par(bty="l", cex.axis=0.7, cex=2)
plot(res,disease="Salmonella",method="Farrington", xaxis.years=F,
   ylim=c(0,140000), xaxt="n", yaxt="n",
   legend.opts=NULL, ylab="", xlab="Data",
      main="Analiza␣stanów␣alarmowych␣grypy␣w␣Polsce")
axis(side=1, at=d, labels=daty, las=0)
axis(side=2, at=seq(0,300000,by=30000),
      labels=seq(0,300000,by=30000), las=1)
legend("topright", lty=c(1,2,-1),pch=c(-1,-1,2),
```

```
      col=c("black","blue","red"), cex=legend.cex, bty="n",
      legend=c("liczba␣zachorowań", "granica␣alarmu", "wystąpienie␣alarmu"))
dev.off()
```

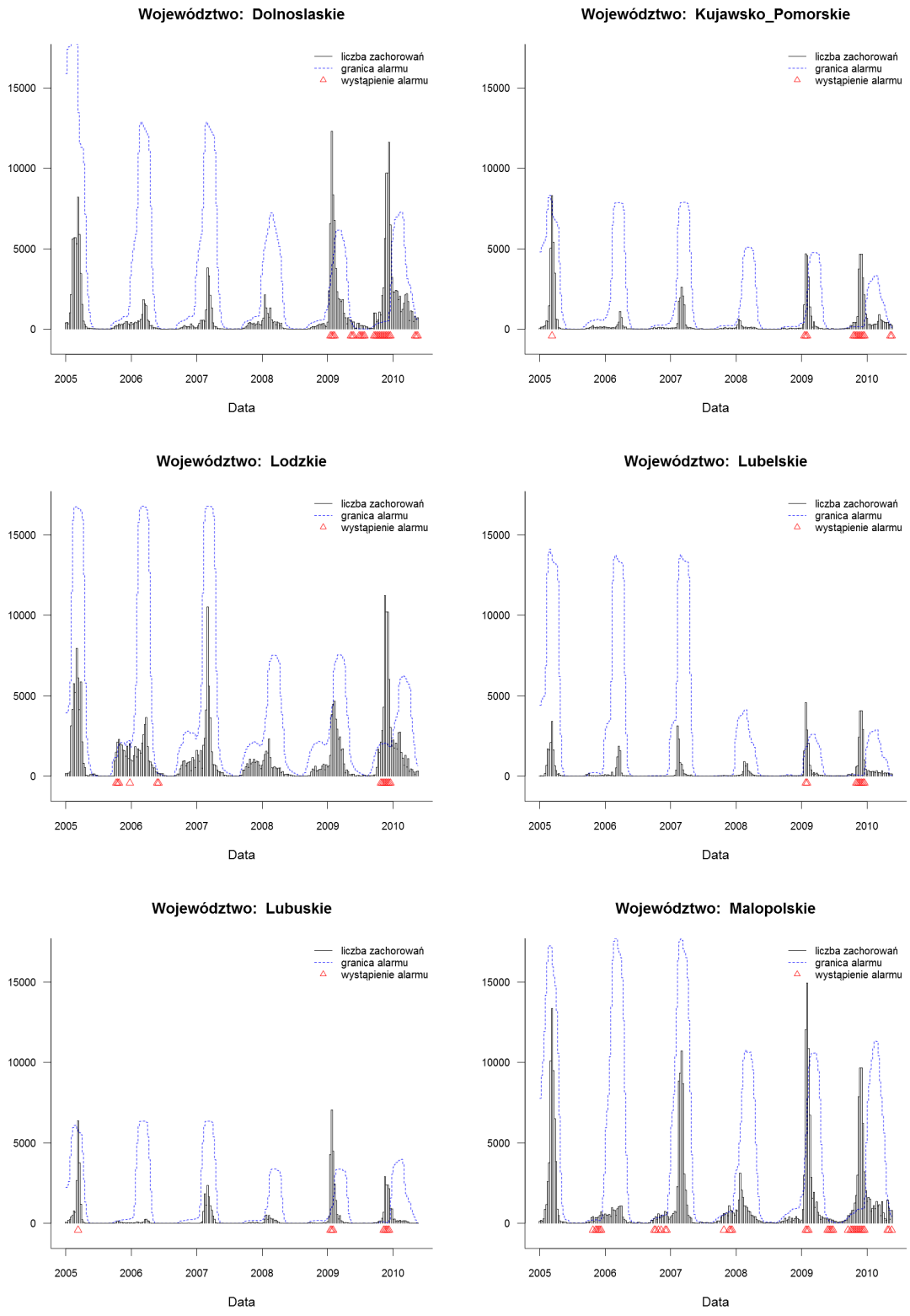
Powyższy kod wykorzystany został do przeprowadzenia wszystkich analiz i ilustracji. Jest on bardzo podobny do kodu wykorzystanego w przypadku modelu Farringtona. Poszczególne czynności wykonywane są w tej samej kolejności. Dodatkowo uwzględniono warunek konieczny wystąpienia alarmu, o którym wspomniano w podrozdziale 3.5.

Dodatek B

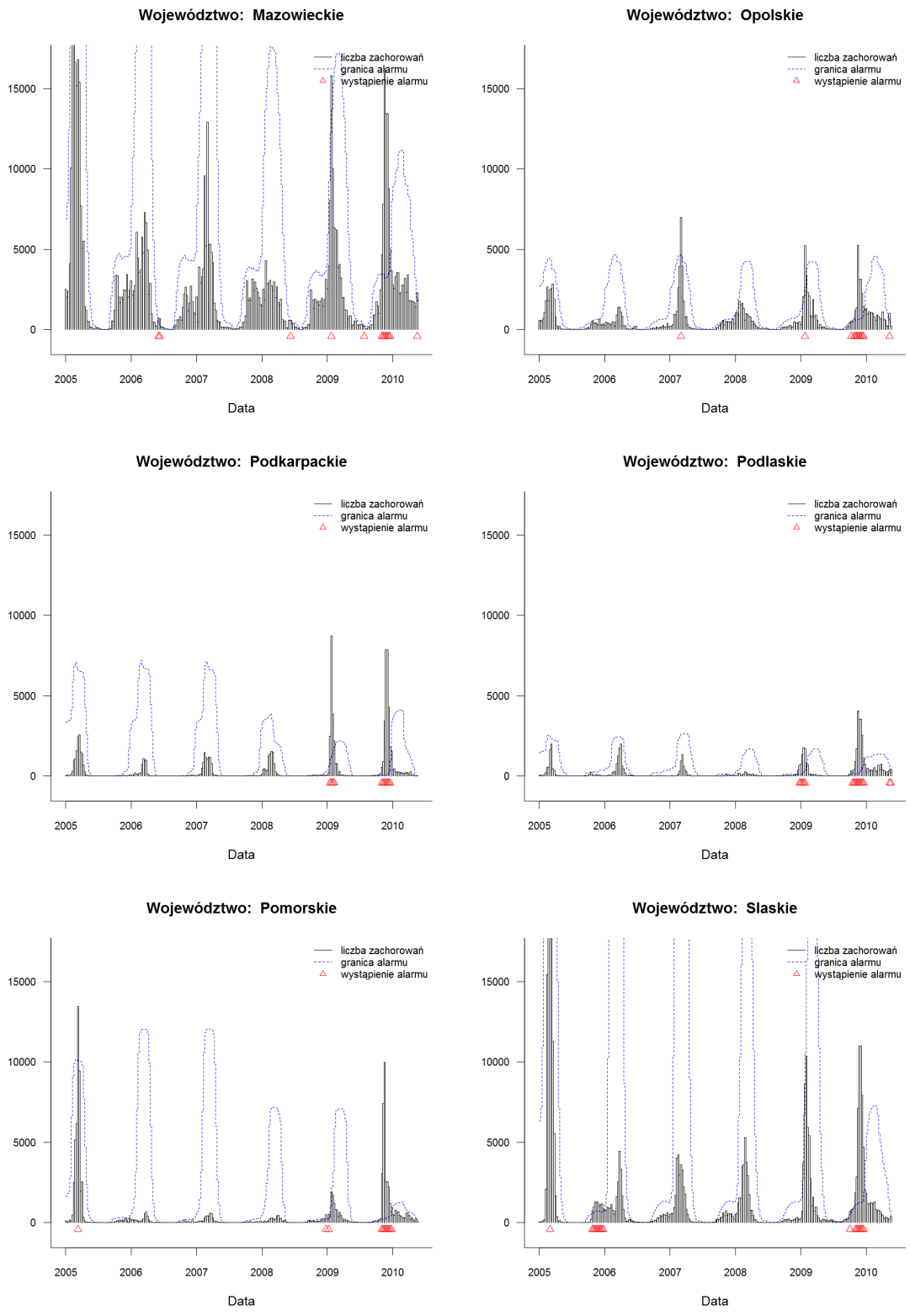
Rysunki nieumieszczone w głównej części pracy



Rysunek B.1: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu metody RKI, cz.1.



Rysunek B.2: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu metody RKI, cz.2.



Rysunek B.3: Analiza stanów alarmowych liczby zachorowań na grypę w województwach przy wykorzystaniu metody RKI, cz.3.

Bibliografia

[1] B. Barratta, R. Atkinson, H. R. Anderson, S. Beeversa, F. Kellya, I. Mudwaya, P. Wilkinson. *Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme.*

[2] B. Y. Choi, H. Kim, U. Y. Go, JH. Jeong, J. W. Lee 2010, *Comparison of various statistical methods for detectng disease outbreaks* Springer-Verlag, Published online.

[3] W. D. Ewan, K. W. Kemp *Sampling inspection of continuous processes with no autocorrelation between successive results*, Biometrica 47, 363-380, 1960.

[4] J. J. Faraway, *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Boca, rRaton, London, New York, Chapman & Hall CRC, 2006.

[5] C. P. Farrington, N. J. Andrews, A. D. Bealy, M. A. Catchpole, *A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease.* Journal of the Royal Statistical Society, Series A, 547-563, 1996.

[6] B. A. Griffin, A. K. Jain, J. Davies-Cole, C. Glymph, G. Lum, S. C. Washington, M. A. Stoto *Early detection of influenza outbreaks using the DC Department of Health's syndromic surveillance system.*

[7] A. Hashimoto, Y. Murakami, K. Taniguchi, M. Nagai *Detection of epidemics in their early stage through infectious disease surveillance*, International Journal of Epidemiology 905-910, 2000.

[8] M. Hohle, *Surveillance: An R package for the monitoring of infectious diseases*, 2010.

[9] M. Hohle, *Poisson regression charts for the monitoring of surveillance time series. Technical report, Department of Statistics, University of Munich.*

[10] L. Hutwagner, E.Maloney, N. Bean, L. Slutsker, S. Martin *Using laboratory-based surveillance data for prevention: an algorithm for detecting salmonella outbreaks*, 1997.

[11] C. D. Montgomery *Introduction to Statistical Quality Control*, 1991.

[12] E. S. Page *Continuous inspection schemes*, Biometrika 41, 100–115, 1954.

[13] P. Rolfhamre *Outbreak Detection of Communicable Diseases – Design, Analysis and Evaluation of Three Models for Statistically Detecting Outbreaks in Epidemiological Data of Communicable Diseases*

[14] G. Rossi, L. Lampugnani, M. Marchi, *An approximate CUSUM procedure for surveillance of health events.*

[15] R. E. Serfling *Methods for current statistical analysis of excess pneumonia-influenza deaths*, Public Health Reports,494–506, 1963.

[16] M. A. Stoto, J. Matheson, V. B. Foster *CUSUM Techniques to Identify 'Flu' Outbreaks*

[17] H. E. Tillett, I. L. Spencer *Influenza Surveillance in England and Wales Using Routine Statistics. Development of Cusum Graphs to Compare 12 Previous Winters and to Monitor the 1980-81 Winter*, The Journal of Hygiene, Vol. 88, 83-94, 1982.

[18] S. Unkel, C. P. Farrington, P. H. Garthwaite, C. Robertson, N. Andrews *A review of statistical methods for the prospective detection of infectious disease outbreaks.*

[19] *SAS/INSIGHT 9.1 User's Guide, Volumes 1 and 2* SAS Institute Inc, Cary, NC, 2004.

[20] *http://cran.r-project.org/web/packages/Cairo/Cairo.pdf*

[21] *http://cran.r-project.org/web/packages/classInt/classInt.pdf*

[22] *http://cran.r-project.org/web/packages/maptools/maptools.pdf*

[23] *http://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf*

[24] *http://cran.r-project.org/web/packages/spc*

[25] *http://cran.r-project.org/web/packages/surveillance*

[26] *Development and evaluation of outbreak detection algorithms for communicable diseases*, County of Los Angeles — Department of Health Services, Special Studies Report, 1999.

[27] *Meldunki epidemiologiczne. Zachorowania i podejrzenia zachorowań na grypę w Polsce*, Narodowy Instytut Zdrowia Publicznego - Państwowy Zakład Higieny, *http://www.pzh.gov.pl/oldpage/epimeld/grypa/index.htm*

[28] *Stan sanitarny kraju za 2006*, Główny Inspektorat Sanitarny, Warszawa 2007, opublikowano online

[29] *Stan sanitarny kraju za 2007*, Główny Inspektorat Sanitarny, Warszawa 2008, opublikowano online

[30] *Stan sanitarny kraju za 2008*, Główny Inspektorat Sanitarny, Warszawa 2009, opublikowano online

[31] *Stan sanitarny kraju za 2009*, Państwowa Inspekcja Sanitarna, opublikowano online

[32] *Raport Głównego Inspektora Sanitarnego, stan sanitarny kraju w roku 2010*, opublikowano online

[33] Statistics of Sweden *http://www.scb.se/statistik/be0101/Be0101tab8samdrag.asp*, 2003.

[34] Wrocławski Zlot Użytkowników R - strona internetowa, *http://www.biecek.pl/WZUR*

[35] *Zagrożenia okresowe występujące w Polsce* Wydział Analiz i Prognoz Biura Monitorowania i Analizy Zagrożeń Rządowego Centrum Bezpieczeństwa, 2010.