„Przeszłość Demograficzna Polski" 38, 2016, nr 2
DOI: 10.18276/pdp.2016.2.38-04

Radosław Poniat
Białystok

O możliwości wykorzystania regresji LOESS w analizie szeregów czasowych

Celem prezentowanego artykułu jest zwrócenie uwagi historyków na możliwość wykorzystania do analizy szeregów czasowych tzw. regresji LOESS[1]. Technika ta, pod pewnymi względami podobna do średniej ruchomej, posiada zalety, które powinny wzbudzić zainteresowanie badaczy oraz umożliwić pełniejsze i skuteczniejsze wykorzystanie dostępnych im zbiorów danych. Choć regresja LOESS jest metodą stosunkowo trudną pod względem obliczeniowym, to dzięki powszechnie dostępnym programom komputerowym sięgnięcie po nią nie powinno nastręczać historykom szczególnych trudności[2].

Spośród trzech sekcji artykułu, pierwsza stanowi wprowadzenie do zagadnienia analizy szeregów czasowych i zawiera krótkie omówienie wad i zalet technik w stosunku do regresji LOESS alternatywnych. W części drugiej zaprezentowano tytułową regresję. Z kolei w sekcji trzeciej przedstawiono procedurę jej wyliczania w programie R z pakietem ggplot2. Wszystkie podane w tekście przykłady

[1] Choć niektórzy autorzy tłumaczą LOESS na język polski jako regresję lokalnie ważoną, ważoną *regresję* lokalnie wielomianową lub nieparametryczną regresję lokalnie ważoną, to w literaturze najczęściej spotykane jest jednak wykorzystywanie nazwy anglojęzycznej bez prób jej przekładu.

[2] Obecność regresji LOESS w popularnych pakietach statystycznych będzie miała istotny wpływ na zawartość tego artykułu. Pominięte w nim zostaną zagadnienia związane z techniczną stroną wyliczania regresji, a uwaga skupiona zostanie na korzyściach płynących z jej zastosowania oraz interpretacji uzyskanych rezultatów.

pochodzą z dobrze znanej historykom pracy Jana Baszanowskiego poświęconej demografii nowożytnego Gdańska[3].

W artykule skupiono się na rocznych liczbach chrztów w latach 1601–1846. Baza danych zawierająca te wartości zmieszczona została na stronie internetowej czasopisma „Przeszłość Demograficzna Polski".

1

Zaprezentowany na wykresie 1a szereg czasowy opisujący roczne liczby chrztów w nowożytnym Gdańsku stanowić może podstawę do rozbudowanych analiz z zakresu historii społeczno-gospodarczej i demografii. Działanie takie można podjąć na kilka sposobów. Badacze o bardziej analitycznym zacięciu, zainteresowani przede wszystkim krótszymi odcinkami czasu i specyfiką poszczególnych lat, zwrócą przede wszystkim uwagę na punkty wyróżniające się w stosunku do swego otoczenia. Szczególna kulminacja liczby chrztów w roku 1646 (łącznie aż 2879) lub wyraźny spadek w 1790 (zaledwie 975) mogą stać się podstawą do rozważań na temat lat prosperity i kryzysów[4], wpływu klęsk elementarnych i wojen na stan populacji Gdańska[5]. Z kolei badaczy bardziej zainteresowanych zjawiskami długookresowymi, cyklami i trendami, na wykresie 1a zaciekawi kształt, w jaki układają się punkty oddające roczne liczby zgonów. Bez wielkiego trudu dostrzegą na nim kryzys liczby urodzeń w drugiej połowie XVIII wieku, odbicie następujące w pierwszej połowie następnego stulecia, względną stabilność charakteryzującą wiek XVII. Wielu badaczy, aby móc lepiej dostrzec opisywane tu prawidłowości, zdecydowałoby się na dodanie do wykresu centrowanej średniej ruchomej[6]. Jej zastosowanie pozwoliłoby na pełniejsze zilustrowanie generalnych prawidłowości, a zmniejszyło optyczne znaczenie obserwacji odstających od trendu.

---

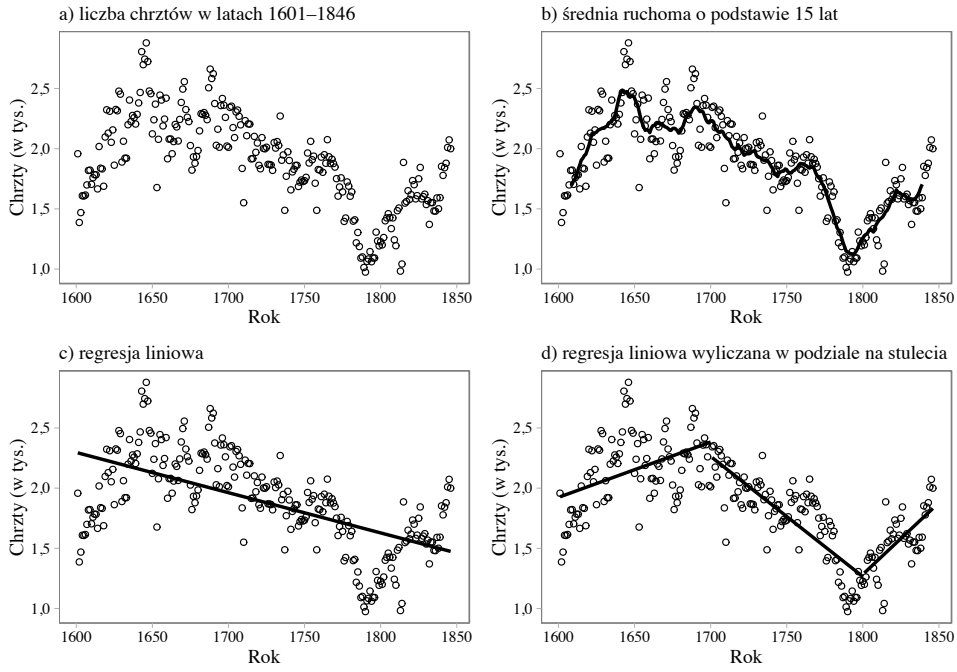
[3] Jan Baszanowski, *Przemiany demograficzne w Gdańsku w latach 1601–1846* (Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego, 1995), 348–354.

[4] Choć przy wyliczaniu kryzysów demograficznych kluczowe znaczenie ma przede wszystkim informacja na temat zgonów, istnieją metody badawcze biorące także pod uwagę wahania liczby chrztów. Z całą też pewnością znajomość natężenia urodzeń może mieć znaczenie przy ocenianiu wielkości analizowanej populacji. Cezary Kuklo, *Demografia Rzeczypospolitej przedrozbiorowej* (Warszawa: Wydawnictwo DiG, 2009), 249.

[5] W odniesieniu do Gdańska wskazać można na przykład pracy podejmującej takie zagadnienia – Edmund Kizik, red., *Dżuma, ospa, cholera. W trzechsetną rocznicę wielkiej epidemii w Gdańsku i na ziemiach Rzeczypospolitej w latach 1708–1711* (Gdańsk: Muzeum Historyczne Miasta Gdańska, 2012).

[6] Michał Kopczyński, *Podstawy statystyki. Podręcznik dla humanistów (*Warszawa: Oficyna Wydawnicza „Mówią Wieki", 2005), 188–190.

Wykres 1. Chrzty w Gdańsku w latach 1601–1648: średnia ruchoma i regresja liniowa



Źródło: Baszanowski, *Przemiany*, 348–354.

Widoczna na wykresie 1b czarna krzywa przedstawia 15-letnią średnią ruchomą gdańskich chrztów. Jej analiza potwierdza powyżej poczynione obserwacje. Okresy wzrostów i spadków są dzięki niej dobrze widoczne. W porównaniu z wykresem 1a dostrzec też można kolejne zjawiska, poprzednio ukryte przez znaczny rozrzut części punktów. Dotyczy to zwłaszcza wzrostów następujących w pierwszych dziesięcioleciach XVII wieku, poprzednio słabo dostrzegalnych. Można więc stwierdzić, że centrowana średnia ruchoma spełniła swe zadanie – trendy zostały uwypuklone, zaś znaczenie pojedynczych obserwacji uległo redukcji. Niestety, zastosowana tu technika, obok niewątpliwych zalet, obciążona jest także poważnymi wadami. Najważniejszą z nich jest zapewne, dobrze widoczny na wykresie 1b, brak oszacowania średnich na początku i końcu szeregu czasowego. Ponieważ wykorzystano tu średnią o podstawie 15 lat, aż w 14 punktach brak jest wystarczającej liczby obserwacji poprzedzających dany rok lub po nim następujących. W rezultacie, krzywa przedstawiająca średnią ruchomą ulega zauważalnemu skróceniu, a w prowadzonych wyłącznie na jej podstawie analizach niczego nie można powiedzieć na temat lat 1601–1607 oraz 1840–1846.

Oczywiście, strata taka zostałaby zminimalizowana dzięki zastąpieniu średniej o podstawie 15 lat nową miarą wyliczaną z krótszego okresu, na przykład 3–5 letniego, ale decyzja taka również pociągnie za sobą pewne koszty. Nowa średnia będzie bowiem znacznie bardziej podatna na obserwacje skrajne i znacznie gorzej opisze długookresowe trendy. Zresztą, należy też pamiętać, że wykorzystywany tu szereg czasowy jest stosunkowo rozbudowany. Zazwyczaj historycy nie mają do czynienia z aż tak bogatymi zbiorami danych, a ich obserwacje dotyczą znacznie krótszych okresów. W takiej sytuacji nawet wyłączenie 2–4 obserwacji z krańców rozkładu uznane być może za poważną stratę. Problem ten ulega zaostrzeniu w sytuacji braku danych w środku szeregu czasowego, które to zjawisko zostanie omówione nieco dalej.

Metodą nieobciążoną częścią wad średniej ruchomej jest regresja liniowa, traktowana powszechnie jako alternatywa w stosunku do niej[7]. Przykład jej zastosowania do analizy gdańskich chrztów zaprezentowano na wykresie 1c. Choć czytelnika należy już na wstępie ostrzec, że uzyskany tu rezultat jest bardzo daleki od satysfakcjonującego i niewłaściwie oddaje badane zjawisko, stanowi on dobry punkt wyjścia do dalszych rozważań. Widoczna na wykresie prosta opisuje w formie graficznej zależność między rokiem a liczbą chrztów. Relacja ta może być też oddana za pomocą równania:

$$\text{liczba chrztów w danym roku} = 7642 - 3{,}3 \times \text{rok}.$$

Dzięki takiej formule łatwo można stwierdzić, że w badanym okresie istniał trend polegający na spadku liczby chrztów przeciętnie o 3,3 rocznie oraz wyliczyć przewidywaną liczbę chrztów w danym punkcie szeregu czasowego. Rezultat będzie wprawdzie obciążony błędem, ale koszt taki często warto ponieść w imię poszukiwania generalnych prawidłowości. Tak więc, według modelu regresji w roku 1626 w Gdańsku powinno się odbyć 2276,2 chrztów ($7642 - 3{,}3 \times 1626$). Wartość ta jest bardzo zbliżona do liczby podanej w pracy Baszanowskiego, która wynosiła 2326 chrztów.

Czytelnik szybko jednak dostrzeże, iż uzyskany tu satysfakcjonujący wynik stanowi konsekwencję wybrania do analizy dość specyficznego roku. W przypadku wielu innych lat dysproporcje między wartościami przewidywanymi w regresji a zarejestrowanymi w źródłach są znaczne, a czasem wręcz alarmujące. Za przykład posłużyć może rok 1790 – rzeczywiste 975 chrztów odbiega wyraźnie od uzyskanych w regresji 1735. Istnienie takich rozbieżności stanowi konsekwencję założeń regresji liniowej[8]. W opisywanym tu przypadku jedna prosta nie

[7] Kopczyński, *Podstawy*, 193–197.

[8] Gwoli ścisłości należy tu jednak zastrzec, że model prostoliniowy jest tylko jednym z wielu modeli regresji. Równania zakładające inną formę zależności między chrztami a latami (np.

może oddać widocznych na wykresie trendów. Kolejne okresy wzrostu i spadku, czasowe stabilizacje i pojawiające się często znaczne wahania rocznej liczby chrztów w Gdańsku nie mogą być opisane jedną linią.

Znacznie lepsze rezultaty zastosowania regresji liniowej przedstawiono na wykresie 1d. W tym wypadku, zamiast jednego modelu opisującego cały badany okres, równania regresji wyliczone zostały oddzielnie dla każdego stulecia. Otrzymane w wyniku takiej procedury trzy linie znacznie lepiej opisują rozkład obserwacji. Dzięki nim wyraźnie dostrzec można chociażby postępujący spadek liczby chrztów w XVIII wieku albo ich wzrost w stuleciu kolejnym. Obok bardziej adekwatnego przedstawienia trendów, na omawianym wykresie dostrzec też można jeszcze jedną zaletę regresji liniowej. W przeciwieństwie do średniej ruchomej, wykorzystana tu technika pozwala na uzyskanie wyników w odniesieniu do wszystkich badanych lat. Linie regresji rozpoczynają się więc już w 1601 i sięgają aż do 1846 roku. Fakt, iż za pomocą równań regresji można wyliczać przewidywaną liczbę chrztów dla każdego z poddawanych analizie lat, będzie miał szczególnie istotne znaczenie w przypadku pojawienia się w wykorzystywanym zbiorze braków danych. O ile w klasyczny sposób otrzymywana średnia ruchoma z sytuacją taką sobie nie poradzi, równanie regresji wciąż pozwoli na oszacowanie przewidywanego poziomu badanego zjawiska. Trudno się więc dziwić, że metoda ta bardzo często wykorzystywana jest właśnie po to, aby choć w przybliżeniu określić wartości w miejscach, gdzie szereg czasowy uległ przerwaniu.

Choć przedstawione na wykresie 1d linie wydają się dość dobrze opisywać trendy zachodzące w nowożytnym Gdańsku, wciąż można wskazać na ich ograniczenia. Szczególnie łatwo są one dostrzegalne w przypadku XVII wieku, gdy regresja nie wychwytuje wzrostu liczby chrztów w jego początkowym okresie. Dostrzeżenie tego zjawiska wymagałoby wyliczania równań regresji dla jeszcze krótszych przedziałów: ćwierćwieczy lub dziesięcioleci. Strategia taka wiązałaby się jednak ze wzrostem liczby koniecznych obliczeń statystycznych, co wprawdzie w dobie powszechnie dostępnych programów komputerowych nie musi być szczególnie trudne, ale mimo wszystko wciąż nakładać będzie na badacza konieczność podjęcia dodatkowego wysiłku. Większym wyzwaniem może się tu jednak okazać problem kolejny, dotyczący wyboru właściwych przedziałów, dla których wyliczana będzie regresja oraz ich punktów startowych. Dobranie ich „na oko" skutkować może zarzutem o dostosowywanie rezultatów do z góry przyjętych założeń, zaś zdanie się na z pozoru „naturalne" i „neutralne" granice,

---

wielomian drugiego stopnia) mogłyby się tu okazać bardziej adekwatne. Należy też pamiętać, że w przypadku wielu innych zbiorów nawet zwyczajna regresja liniowa może zapewniać rezultaty satysfakcjonujące korzystającego z niej badacza.

czyli na przykład początek stulecia lub dekady, nie zawsze będzie właściwym rozwiązaniem. Łatwo to można dostrzec na wykresie 1d, gdzie charakterystyczny dla XIX wieku trend wzrostu liczby chrztów w rzeczywistości, wbrew temu co sugeruje linia regresji, wcale nie rozpoczął się w 1801 roku. W rzeczywistości jego punktu startowego należałoby szukać około czterech lat wcześniej. Próbą zaradzenia wspomnianym tu trudnościom jest właśnie regresja LOESS, która zostanie omówiona w dalszej części artykułu.

2

Przyswojenie regresji LOESS[9] nie powinno sprawić trudności czytelnikowi zaznajomionemu z regresją liniową i średnią ruchomą. Technika ta może być zresztą uznana za swoisty kompromis między opisanymi powyżej metodami lub wręcz ich amalgamat. Nieco upraszczając jej opis (i rezygnując z omówienia kryjącego się za nią aparatu matematycznego), regresja LOESS określona może być jako procedura polegająca na wyliczaniu oddzielnych równań regresji dla każdego punktu w szeregu czasowym[10]. Podobnie więc jak to miało miejsce na wykresie 1d, zamiast jednego równania regresji, uzyskanych jest ich tu wiele[11]. Liczba równań jest identyczna z liczbą obserwacji, czyli w omawianym tu przykładzie wynosi 246. Dla każdego roku pojawia się więc model jego dotyczący, co czyni regresję LOESS podobną do centrowanej średniej ruchomej, w której poszczególne lata otrzymują własne, lokalne średnie. Przy wyliczaniu LOESS największe znaczenie (wagę) mają obserwacje znajdujące się najbliżej danego punktu, zaś te bardziej oddalone w mniejszym stopniu wpływają na uzyskany wynik.

---

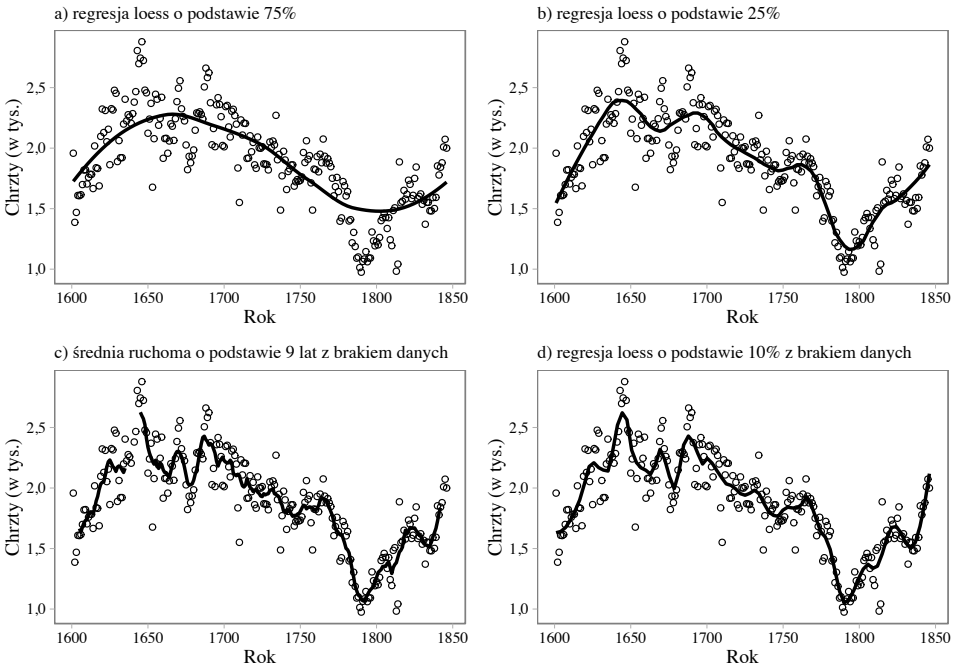
[9] Metoda zaproponowana została w roku 1979 przez Williama S. Clevelanda, który bazował na wcześniejszych koncepcjach Macauleya. Cleveland w swych kolejnych publikacjach poddawał ją dalszym modyfikacjom. Szczególne znaczenie ma tu jego artykuł z 1988 r., opracowany wspólnie z Susan J. Devlin. Cleveland odpowiada właściwie za stworzenie dwóch, blisko ze sobą powiązanych metod określanych mianem LOWESS (*Robust Locally Weighted Regression*) i LOESS (*Locally-Weighted Regression*). Ich związek jest jednak tak ścisły, że często traktowane są one jako jedna technika. William S. Cleveland, „Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association*, 74 (1979): 829–836, DOI:10.2307/2286407; William S. Cleveland, „Lowess: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression", *The American Statistician*, 35 (1981): 54; William S. Cleveland, Susan J. Devlin, „Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting", *Journal of the American Statistical Association*, 83 (1988): 596–610; William S. Cleveland., Susan J. Devlin, Eric Grosse, „Regression by Local Fitting: Methods, Properties, and Computational Algorithms", *Journal of Econometrics*, 37 (1988): 87–114.

[10] Zaznaczyć tu trzeba, że regresja LOESS jest też stosowana do danych o innym charakterze. Analiza szeregów czasowych stanowi więc tylko jeden z wielu obszarów, gdzie się z niej korzysta.

[11] Takie lokalne równanie nie musi mieć, a wręcz zazwyczaj nie ma, formy liniowej. W praktyce jednak kwestia ta nie ma istotnego znaczenia dla mniej statystycznie biegłych użytkowników regresji LOESS, a jej omówienie nie jest tu konieczne.

W zależności od parametrów regresji, pod uwagę brane być mogą albo wszystkie elementy szeregu czasowego (choć ze zmieniającymi się wagami), albo tylko ich część, położona w bezpośrednim sąsiedztwie wybranej obserwacji.

Wykres 2. Chrzty w Gdańsku w latach 1601–1648: regresja LOESS



Źródło: jak przy wykresie 1.

Wynik zastosowania takiej procedury zaprezentowano na wykresie 2a. Widoczna na nim krzywa ma kształt wyraźnie różny od regresji liniowej przedstawionej na wykresie 1c. W miejsce prostej i niezbyt prawdziwej opowieści o mającym dotykać nowożytny Gdańsk nieustannym zmniejszania się liczby chrztów, tutaj dostrzec można bardziej skomplikowaną narrację, w której obok okresów spadku pojawiają się też wzrosty. Właściwsze byłoby więc porównywanie wykresu 2a z 1d. Spośród tych dwóch regresja LOESS okazuje się przede wszystkim bardziej płynna, brak tu trzech oddzielnych linii wynikających z wyliczania trzech różnych modeli regresji liniowej, a w to miejsce pojawia się stały układ trendów. Uzyskanych dzięki zastosowaniu LOESS 246 wartości oszacowanych tworzy nieprzerwaną krzywą. Rezultat taki jest konsekwencją algorytmu, w którym dla każdego roku wyliczone zostało oddzielne równanie regresji, w którym

lata danemu punktowi bliskie miały większe znaczenie. Łączna liczba obserwacji branych pod uwagę przy tworzeniu poszczególnych modeli wynosiła 75% całego zbioru danych. Wybór takiego zakresu, będący domyślnym ustawieniem wykorzystywanego tu programu do analizy statystycznej, nie jest jednak jedynym możliwym. Na wykresie 2b zaprezentowano regresję LOESS, w której w każdym z równań pod uwagę brano 25% wszystkich obserwacji. Konsekwencją takiego ograniczenia wielkości wykorzystywanego zbioru jest większe pofałdowanie linii regresji. W znacznym stopniu kształt jej przypomina średnią ruchomą o podstawie 15 lat zaprezentowaną na wykresie 1b.

Obserwacja zaprezentowanych regresji LOESS pozwala na dostrzeżenie podstawowych zalet opisywanej metody. Tak jak miało to miejsce w przypadku regresji liniowej, analizie statystycznej poddane zostały wszystkie dostępne lata, bez straty obserwacji na początku i końcu szeregu[12], co miałoby miejsce w razie skorzystania ze średniej ruchomej. Zwłaszcza w przypadku stosunkowo krótkich szeregów czasowych uznać to trzeba za bardzo poważną zaletę omawianej techniki. Równocześnie stopień wygładzenia linii regresji, czyli jej dopasowania do rozkładu obserwacji pochodzących z poszczególnych lat, może być modyfikowany w bardzo szerokim zakresie i zapewnić rezultat charakteryzujący się podobną wrażliwością na trendy, jak to ma miejsce w przypadku średniej ruchomej. Co ważne, takie dostosowanie regresji do występujących w szeregu czasowym prawidłowości odbywa się w sposób ciągły i nieprzerwany, a otrzymany w wyniku takich działań rezultat nie wymaga wstępnych założeń na temat kształtu badanych zależności, ich punktów startowych i końcowych. Regresja LOESS niejako sama dopasowuje się do danych, a zadanie jej użytkownika sprowadza się przede wszystkim do określenia, jak bardzo szereg czasowy ma zostać wygładzony. Korzystający z metody historyk może więc w znacznym stopniu zdać się na procedurę obliczeniową i dopiero na jej podstawie dokonywać dalszych interpretacji zjawiska. Osoby korzystające z metod regresji opierających się na z góry przyjętych kształtach zależności między zmiennymi nie mają takiego luksusu i często odkrywają, że przyjęty przez nich model nieadekwatnie opisuje analizowane zjawisko. Z punktu widzenia wielu historyków zaletą regresji LOESS może być nawet to, co wśród części badaczy uchodzi za jej wadę – powiązanie metody z prezentacją graficzną. Podczas gdy większość technik regresji generuje przede wszystkim funkcję opisującą zależności między zmiennymi, a ich rezultaty są przedstawiane za pomocą tabel z odpowiednimi współczynnikami, do których wykresy stanowią tylko dodatek, często zresztą w publikacjach niezamieszczany, LOESS została opracowana i do dziś funkcjonuje jako metoda przedstawiania

[12] Należy jednak pamiętać, że oszacowania dotyczące skrajnych punktów szeregu czasowego obciążone są większym błędem niż dotyczące jego elementów centralnych.

danych w formie graficznej. Wynikiem zastosowania LOESS będzie więc zazwyczaj łatwa do zanalizowania linia na wykresie, nie zaś równanie, z którego interpretacją wielu historyków miałoby poważne trudności. Choć więc część badaczy może preferować techniki generujące jedną, choćby i skomplikowaną funkcję regresji poddającą się dalszej analizie statystycznej, wielu spośród ich kolegów preferować będzie prosty rezultat w postaci przejrzystego wykresu.

Wydaje się, że ostatnią z ważnych i wartych wymienienia zalet regresji LOESS jest jej względna odporność na braki danych w szeregu czasowym. Fakt ten można łatwo dostrzec dzięki porównaniu wykresów 2c oraz 2d. Na pierwszym z nich przedstawiono 9-letnią centrowaną średnią ruchomą. Co oczywiste, linia uzyskana dzięki jej wykorzystaniu jest nieco bardziej pofałdowana niż miało to miejsce na wykresie 1b, gdzie widoczna była średnia o podstawie 15 lat. W obydwu jednak przypadkach łatwe do dostrzeżenia są braki na krańcach szeregu czasowego. Na wykresie 2c brak jest więc oszacowań dla okresów 1601–1604 oraz 1843–1846. Czytelnik szybko dostrzeże jeszcze jedną przerwę w linii średniej ruchomej, obejmującą lata 1636–1644. Jej pojawienie się stanowi efekt celowego usunięcia z bazy danych informacji o liczbie chrztów w 1640 roku. Ten pojedynczy brak sprawił, iż średniej ruchomej nie można wyliczyć nie tylko dla tego roku, ale też czterech lat przed nim i po nim[13]. W konsekwencji, badacz korzystający z tej metody musi pogodzić się z faktem, że ze zbioru obejmującego informacje dotyczące 245 lat, trendy analizować może tylko w odniesieniu do 229 spośród nich. W prezentowanym przypadku oznacza to wprawdzie zniknięcie zaledwie 6% obserwacji, ale nawet taka strata jest warta uniknięcia. Należy też pamiętać, że większość historyków nie dysponuje tak długimi seriami danych i ewentualna utrata nawet kilku punktów może oznaczać całkiem poważną redukcję zbiorów wykorzystywanych w ich badaniach.

Zamieszczona na wykresie 2d regresja LOESS, choć wyliczona na podstawie tego samego szeregu czasowego z pojedynczym brakiem danych, w miejscu tym nie została przerwana. Bazując na informacjach opisujących liczbę chrztów w 245 latach, wyliczone tu zostały uśrednione wartości dotyczące 246 punktów w czasie. W odniesieniu do pominiętego w bazie roku 1640 równanie regresji przewidziało 2380 chrztów, co jest wartością całkiem bliską jej rzeczywistej liczbie wynoszącej 2285. W ten sposób nie tylko nie doszło do utraty informacji, ale też oszacowana została jedna dodatkowa wartość. Oczywiście, uzyskany wynik jest obciążony pewnym błędem[14]. Choć w podanym przypadku jest on stosunkowo mały, wyobrazić można sobie sytuacje, gdy rok z brakiem danych różniłby

[13] Istnieją wprawdzie techniki pozwalające na doszacowanie brakujących danych w średniej ruchomej, ale wiążą się one z odejściem od prostej, klasycznej formy tej metody.

[14] Wielkość takiego błędu może być zresztą z określonym prawdopodobieństwem oszacowana. Graficzne prezentacje regresji LOESS bardzo często, obok samej linii trendu, opisuje także

się tak znacznie od lat go okalających, że dokonane przy pomocy LOESS oszacowanie bardzo poważnie odbiegałoby od rzeczywistości, ale jest to ryzyko nie do uniknięcia. Można zresztą zakładać, że sytuacje takie będą stosunkowo rzadkie, a wystąpienie bardzo specyficznego roku będzie jednak zazwyczaj miało wpływ na lata po nim następujące, co już przez regresję LOESS zostanie dostrzeżone.

Wydaje się, że podane tu przykłady zastosowania regresji LOESS stanowią dobry i przekonujący argument na rzecz użyteczności tej techniki. W przeciwieństwie do średniej ruchomej nie wiąże się ona z utratą informacji dotyczącą części obserwacji oraz pozwala na przewidywanie wartości w przypadku braków danych. Z kolei w porównaniu z regresją liniową nie wymaga wstępnych założeń na temat kształtu badanych trendów. Jej niegdyś podstawowa wada – konieczność wykonania stosunkowo skomplikowanych obliczeń – straciła zaś na znaczeniu wobec rozwoju wyspecjalizowanych programów statystycznych. Przykład wyliczania regresji LOESS w jednej z takich aplikacji przedstawiony zostanie poniżej.

3

Wraz z postępującym upowszechnianiem się wykorzystania regresji LOESS użytkownicy rosnącej liczby programów statystycznych mogą ją odnaleźć wśród oferowanych przez takie aplikacje funkcji. Obok pakietów komercyjnych, takich jak SPSS, SAS czy Stata, zjawisko to dotyczy także programów freeware. Opisana poniżej procedura generowania regresji LOESS dokonywana jest właśnie w jednej z takich darmowych aplikacji: programie/języku R. Dokonany tu wybór po części odzwierciedla nawyki autora, ale znacznie ma tu też popularność i powszechna dostępność tego programu oraz bardzo duża elastyczność zamieszczonych w nim procedur obliczeniowych.

Choć podstawowym sposobem na wygenerowanie regresji LOESS w programie R jest polecenie loess(), podany tu przykład wykorzystywać będzie wersję tej komendy pochodzącą z pakietu do prezentacji graficznej ggplot2. Jest to jeden z najpopularniejszych dodatków do programu R i to właśnie za jego pomocą powstały wykresy zamieszczone w tym artykule[15]. W kontekście wyliczania regresji LOESS jego podstawową zaletą jest fakt, że rezultaty regresji od razu zostają umieszczone na wykresie, bez konieczności wykonywania dodatkowych działań

95% przedział ufności. Oczywiście, przedziały takie są szersze na krańcach szeregów czasowych oraz w okolicach braków danych.

[15] Opis podstawowych funkcji pakietu ggplot2 czytelnik odnaleźć może w artykule pochodzącym z jednego z wcześniejszych tomów *Przeszłości Demograficznej Polski*. Zawarty w nim też został opis wprowadzania danych do programu R – Radosław Poniat, „O wykorzystaniu wykresów pudełkowych do prezentacji danych demograficznych i o pożytku z użycia środowiska R z pakietem ggplot2", *Przeszłość Demograficzna Polski*, 34 (2014): 103–120.

i przekształceń. Podane poniżej polecenia pozwolą na wygenerowanie podstawowej wersji wykresu 2b. Potrzebną w tym celu bazę danych czytelnik odnajdzie na stronie internetowej czasopisma „Przeszłość Demograficzna Polski".

Pierwszym krokiem procedury obliczeniowej musi być zdefiniowanie zmiennych, które zostaną do niej wykorzystane. Dokonać tego można za pomocą polecenia:

```
ggplot(Gdansk, aes(Rok, Chrzty))
```

Termin ggplot rozpoczyna generowanie wykresu, zaś Gdansk to nazwa wykorzystywanej tu bazy danych. Po aes() podane zostały nazwy znajdujących się w tej bazie zmiennych. Zgodnie z obowiązującą w ggplot2 konwencją zapisu pierwsza z nich – Rok – oddana zostanie na osi X, druga zaś – Chrzty – na Y. Oczywiście, w razie wykorzystywania innej bazy lub zmiennych konieczne byłoby podanie w poleceniu innych nazw. Po zdefiniowaniu danych przystąpić można do generowania właściwej części wykresu. Zamieszczenie na nim punktów opisujących liczbę chrztów w kolejnych latach następuje dzięki komendzie geom_point(), z kolei regresję LOESS wprowadza odpowiednia forma polecenia geom_smooth(), w obrębie której zdefiniować należy preferowaną metodę wyznaczania trendu[16]. W przypadku sięgnięcia po LOESS oznacza to konieczność wpisania method='loess'. W podobny sposób przywołane też być mogą inne rodzaje regresji[17]. W razie niezdefiniowania metody program sam wybierze metodę wyznaczania trendu. W przypadku zbiorów liczących mniej niż 1000 obserwacji, będzie to LOESS. Pełny zapis komendy przybierze teraz formę:

```
ggplot(Gdansk,aes(Rok,Chrzty))+geom_point()+geom_smooth(method='loess')
```

Dalszych modyfikacji regresji dokonać można za pomocą kolejnych poleceń dodawanych do komendy geom_smooth(). Kluczowe znaczenie będzie tu zwłaszcza miała zmiana zakresu danych branych pod uwagę przy wyliczaniu regresji w każdym punkcie szeregu czasowego. W miejsce domyślnego ustawienia 75%, które nie wymaga żadnych dalszych zapisów, polecenie span= pozwala na ustawienie innych wartości[18], co przełoży się na zmianę stopnia wygładzenia szeregu czasowego. Usunięcia domyślnie wyświetlanego przedziału ufności dokonuje komenda se=F, zaś kolor linii regresji modyfikuje się za pomocą zapisu

---

[16] Polecenie geom_smooth() wygenerować może kilka form regresji. Za każdym razem odbywa się to poprzez przywołanie funkcji pochodzących z bazowej wersji programu R. W przypadku regresji LOESS jest to, wspomniane już wcześniej, polecenie loess().

[17] Na przykład regresja liniowa zaprezentowana na wykresie 1c wymagała zapisu method='lm'.

[18] Określać je należy za pomocą ułamków dziesiętnych zapisywanych w notacji anglosaskiej, czyli z kropką w roli znaku dziesiętnego. Na przykład ułamek 0.5 oznaczać będzie zakres równy 50%.

colour='...'. W cudzysłów wpisać należy po angielsku nazwę koloru, który ma zastąpić domyślny niebieski. W przypadku regresji widocznej na wykresie 2b komenda przybrała więc formę:

```
ggplot(Gdansk,aes(Rok,Chrzty))+
geom_point()+
geom_smooth(method='loess',se=F,colour='black',span=0.25)
```

Jej dalsze modyfikacje, sprowadzające się głównie do przekształcania wyglądu wykresu pod względem estetycznym, pozostawić należy czytelnikowi i jego własnym preferencjom lub wymogom wydawniczym tekstu, w którym miałaby się znaleźć przygotowywana grafika.

Bibliografia

Baszanowski, Jan. *Przemiany demograficzne w Gdańsku w latach 1601–1846*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego, 1995.

Cleveland, William S. „Robust Locally Weighted Regression and Smoothing Scatterplots". *Journal of the American Statistical Association*, 74 (1979): 829–836. DOI:10.2307/2286407.

Cleveland, William S. „LOWESS: A program for smoothing scatterplots by robust locally weighted regression". *The American Statistician*, 35 (1981): 54. DOI:10.2307/2683591.

Cleveland, William S., Susan J. Devlin. „Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting". *Journal of the American Statistical Association*, 83 (1988): 596–610. DOI:10.2307/2289282.

Cleveland, William S., Susan J. Devlin, Eric Grosse. „Regression by local fitting: Methods, Properties, and Computational Algorithms". *Journal of Econometrics*, 37 (1988): 87–114.

Kizik, Edmund, red. *Dżuma, ospa, cholera. W trzechsetną rocznicę wielkiej epidemii w Gdańsku i na ziemiach Rzeczypospolitej w latach 1708–1711.* Gdańsk: Muzeum Historyczne Miasta Gdańska, 2012.

Kopczyński, Michał. *Podstawy statystyki. Podręcznik dla humanistów.* Warszawa: Oficyna Wydawnicza „Mówią Wieki", 2005.

Kuklo, Cezary. *Demografia Rzeczypospolitej przedrozbiorowej*. Warszawa: Wydawnictwo DiG, 2009.

Poniat, Radosław. „O wykorzystaniu wykresów pudełkowych do prezentacji danych demograficznych i o pożytku z użycia środowiska R z pakietem ggplot2". *Przeszłość Demograficzna Polski* (2014) nr 34: 103–120.

Streszczenie

Artykuł poświęcono metodzie statystycznej znanej jako regresja LOESS i możliwości jej zastosowania w analizie szeregów czasowych. Zalety tej metody omówiono w porównaniu z technikami alternatywnymi: centrowaną średnią ruchomą i regresją liniową wykorzystywaną do wyliczania trendów w czasie. Końcowa część artykułu zawiera instrukcję wyliczania regresji LOESS w programie R z pakietem ggplot2.

Słowa kluczowe: regresja LOESS, szeregi czasowe, ggplot2

On the Possibility of Using the LOESS Regression in the Analysis of Time Series

Summary

The article presents the statistical method known as the LOESS Regression and a possibility of its application in the analysis of time series. The advantages of the method have been compared to the alternative techniques: the central moving average and the linear regression to calculate the trends in time. The final part of the article contains the instruction of how to calculate the LOESS Regression in the R program with the package ggplot2.

Keywords: LOESS regression, time series, ggplot2