# Inne prace

## Predictive model for cocoa yield in Santander using Supervised Machine Learning

Gamboa, A.A., Cáceres, P.A., Lamos, H., Zárate, D.A. and Puentes, D.E., 2019, April. Predictive model for cocoa yield in Santander using Supervised Machine Learning. In 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA) (pp. 1-5). IEEE.

**Źródło zbioru danych:**

W tej pracy autorzy wykorzystują dane pozyskane z AGROSAVIA (Corporación Colombiana de Investigación Agropecuaria) z eksperymentalnej plantacji zlokalizowanej w centrum Rio Negro Santander w obszarze La Suiza. Dane dotyczą lat 2015, 2016, 2017. Istnieje 10 typów klonów najbardziej reprezentatywnych sklasyfikowanych w 5 regionach (SCC-19, SCC-52, SCC-6, SCC-52, SCC-6), oraz 5 uniwersalnych typów klonów (ICS-95, CNN-51, ETT-8, TSH565, ICS-1). W badaniach wzięto pod uwagę 3 czynniki: poziom nawiezienia, typy klonów kakaowca oraz ekspozycję (słońce, cień). W sumie dokonano 60 rodzajów obserwacji, gdzie każda obserwacja przypada na 15 roślin. Brakujące dane zastąpiono w drodze imputacji za pomocą średniej. Dodatkowo wykorzystano zmienne klimatyczne (dotyczące m.in. poziomu opadów) w celu uzupełnienia zbioru danych (skonsolidowany zbiór danych).

**Przygotowanie zbioru danych:**

Dane zostały zbadane pod względem korelacji oraz indeksu inflacji w celu uniknięcia problemów związanych z wieloliniowością. Na podstawie korelacji zmiennych dotyczących m.in. fotosyntezy, morfometrii, fizycznych i chemicznych cech gleby oraz zmiennych klimatycznych. (Wybór na podstawie istotności statystycznej). Wybrano zmienne:

Transpiracja, efektywne wykorzystanie wody, średnica pnia, poziom materii organicznej (OM), poziom sodu (Na), poziom fosforu (P), poziom magnezu (Mg), Radiacja, Temperatura, Wilgotność i opady skumulowane.

**Wykorzystane modele:**

* Uogólniony model liniowy (ang. Generalised Linear Model, GLM)
* Maszyna wektorów nośnych (ang. Support Vector Machines, SVM)

**Walidacja modeli:**

Do walidacji modelu wykorzystano technikę hold out. Dodatkowo został wykorzystany zbiór walidacyjny, tak by walidacja algorytmu odbywała się na zbiorze danych innym niż ten, który został użyty do estymacji jego parametrów. Użyto metryk RMSE (ang. root mean square error), przedziału ufności oraz p-wartości. Odpowiednio 1758,7421 RMSE dla SVM oraz 1700,0615 dla GSM.

## A Tool for Classification of Cacao Production in Colombia based on Multiple Classifier Systems

Plazas, J.E., López, I.D. and Corrales, J.C., 2017. A tool for classification of cacao production in colombia based on multiple classifier systems. In Computational Science and Its Applications–ICCSA 2017: 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part II 17 (pp. 60-69). Springer International Publishing.

Klony kakao są wiecznie zielonymi drzewami, które rosną do 10 metrów wysokości. Aby zapewnić prawidłowy wzrost i rozwój wymagane jest kilka czynników, takich jak wilgotny klimat o temperaturze od 20°C do 30°C, cień, ochrona przed wiatrem oraz porowata gleba bogata w azot, magnez i potas. Dodatkowo jej idealna wysokość uprawy wynosi 400 metrów n.p.m. Kolumbia jest krajem agrarnym a jej gospodarka w dużej mierze zależy od wydajności upraw. W tym sensie kako jest jedną z najważniejszych upraw; w niektórych rejonach stanowi główne źródło dochodu dla około 25000 rodzin. Głównymi producentami z udziałem około 75% całkowitej krajowej produkcji kakao są Santander, Huila, Norte de Santander. Np. w 2004 roku produkcja wyniosła 41703 tony. By rolnicy mogli liczyć na zyski (plony powyżej 1500 kg/ha) konieczne jest terminowe i odpowiednie stosowanie praktyk zapewniających opłacalność zarówno uprawy, jak i zbiorów. W związku z tym konieczne jest posiadanie modeli lub mechanizmów pozwalających na przewidywanie produkcji danej uprawy w krótkim okresie, zapobiegając negatywnym skutkom, które wypływają na działalność rolniczą w celu zrównoważenia produkcji roślinnej interesariuszy.

Strategie modelowania:

* Strategie wokół modelowania upraw pod kątem zarządzania w określonych warunkach klimatycznych.
* Metody stosowane do analizy danych statystycznych w pewnym okresie czasu (szeregi czasowe i analiza statystyczna).
* Algorytmy uczenia maszynowego (ML), techniki predykcyjne, które mają zdolność do klasyfikacji różnych wartości zasiewów i produkcji, na podstawie analizy zebranych danych historycznych. Techniki takie jak sztuczne sieci neuronowe (ANN), drzewa decyzyjne (DT), klastry k-means, K Nearest Neighbor (KNN), Support Vector Machine (SVM).
* Klasyfikacje dokonywane przez Multiple Classifier Systems (MCS) są często bardziej dokładne niż klasyfikacje dokonywane przez najlepszy pojedynczy klasyfikator, biorąc pod uwagę kombinację zestawu klasyfikatorów.
* AdaSVM i AdaNaive

**Źródło oraz opis danych:**

Dane zostały pozyskane z National Federation of Cacao Growers of Columbia (FEDECACAO). Organizacja ta udostępnia bezpłatną bazę danych dotyczących gospodarki kakaowej w Kolumbii, zawierającą trzy główne sekcje: Roczna produkcja ziarna kakaowego według departamentów (gdzie rejestry figurują od 2002 do 2014); powierzchnia, produkcja i roczna wydajność kakao (krajowe rejestry od 2005 do 2011) i miesięczna produkcja ziarna kakaowego (rejestry krajowe od 2006 do 2010). W celu określenia okresu i obszaru geograficznego, na którym uzyskano dane, wykorzystano informacje o rocznej produkcji ziarna kakaowego według departamentów.

Najważniejszym obszarem uprawy kakao w Kolumbii jest region Santander, który przyczynia się do około 50% rocznej produkcji krajowej. Region ten położony jest w środkowo-północnej części kraju, w regionie andyjskim; ma powierzchnię 30537 km^2.

Źródłami danych klimatologicznych są:

* Tutiempo.net, które zwiera dane zebrane na stacjach klimatycznych głównych lotnisk Kolumbii. Źródło to zawiera dane z dwóch regionów Santander: Barrancabermeja (1964-2016) i Bucaramanga (1975-2016);
* NCEP/NCAR Reanalysis firmy NOAA gromadzi dane satelitarne dotyczące wartości klimatycznych od 1948 roku.

**Wstępne przetwarzanie danych:**

Dane klimatologiczne posiadały 48 cech ogólnoświatowych jednak tylko część z tych atrybutów była dostępna dla regionu Santander: średnia dzienna temperatura (°C), maksymalna dzienna temperatura (°C), minimalna dzienna temperatura (°C), średnia dzienna relatywna wilgotność powietrza (%), całkowity dzienny opad atmosferyczny (mm/h).

Dzienna temperatura minimalna i maksymalna zostały odrzucone, ponieważ atrybuty te miały stałe wartości w całym zbiorze danych (czyli zostały uznane za wartości nietypowe). Negatywne wartości zostały zastąpione zerami.

Procentowy udział produkcji regionu Santander został wykorzystany do obliczenia miesięcznej produkcji kakao w tym regionie na podstawie danych o zasięgu krajowym.

W celu określenia który ze zbiorów danych jest najbardziej odpowiedni: dzienny, tygodniowy, dwutygodniowe, przeprowadzono testy eksperymentalne trzech algorytmów uczenia nadzorowanego: M5P, M5Rules oraz kNN. Przeprowadzono walidację krzyżową (ang. cross-validation), gdzie kryterium porównania stanowił współczynnik korelacji. Najlepszym okazał się zbiór o interwale dwutygodniowym. Ostatecznie zbiór danych zawiera około 1800 instancji oraz cztery atrybuty: miesiąc, średnia dzienna temperatura, średnia dzienna wilgotność względna i średni dzienny skumulowany opad w Santander.

Numeryczna klasa docelowa „produkcja kakao” została zdyskretyzowana. Zdefiniowano pięć zakresów wartości jako "bardzo niska produkcja" (0-876 ton), "niska produkcja" (876-1274 ton), "średnia produkcja" (1274-1673 ton). (1274-1673 ton), "wysoka produkcja" (1673-2071 ton) oraz "bardzo wysoka produkcja" (>2071 ton).

**Wykorzystane modele:**

Modele zostały wykonane i uruchomione w aplikacji WEKA. Modele:

* IBk. Implementuje k-NN. Testowane od 1 do 100. Ostatecznie została wybrana wartość k=1.
* J48. Implementacja drzewa decyzyjnego C4.5
* Naiwna sieć Bayesa (Naive Bayes)
* Drzewo losowe (Random Tree)
* Las losowy (Random Forest)
* Bagging. Bootstrap Aggregating
* AdaBoost. Adaptive Boosting
* Stacking
* Grading
* DECORATE. Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples. Tworzenie zróżnicowanych zespołów poprzez przeciwstawne przeklasyfikowywanie sztucznych przykładów szkoleniowych

Wygenerowanie modele zostały przetestowane **przy użyciu metody 10-krotnej walidacji krzyżowej**, która jest jednym z najbardziej wiarygodnych mechanizmów oceny dokładności modelu.

Jako metryki wykorzystano:

* RAE (Relative Absolute Error)
* Precision – precyzja
* Recall – Czułość
* Receiver Operating Characteristics Area (ROC Area)
* Czas treningu

Obraz zawierający stół

Opis wygenerowany automatycznie

Najlepszym modelem okazał się las losowy, uwzględniając błąd poniżej 10%.

* Wskaźnik poprawnie sklasyfikowanych instancji na poziomie 98,23% (1781 z 1813 instancji).
* RAE mniejszy niż 10% (jest większy niż średnia klasyfikacja o ponad 90%).
* Precyzja i Recall większe niż 0,980 (prawdopodobieństwo złej klasyfikacji jest mniejsze niż 2%).
* Obszar ROC równy 0,999 (model ma prawie idealny stosunek swoistości do czułości)
* Czas szkolenia poniżej 0,5 s (jeśli użytkownik będzie musiał ponownie wytrenować model, nie będzie odczuwalnego opóźnienia).

**Opis narzędzia:**

Odpowiedź systemu jest podana w postaci dyskretnych wartości odpowiadających poziomowi oczekiwanej produkcji kakao w Santander, szacowanej maksymalnej i minimalnej produkcji oraz procentowej precyzji klasyfikacji. Dodatkowo, użytkownikowi prezentowana jest pewna liczba użytecznych zaleceń dla rolników (zalecenia te zostały zaczerpnięte z [2], [18], i są przedstawione w tabeli 4), biorąc pod uwagę optymalizację ich logistyki w zależności od oczekiwanego poziomu produkcji.

Obraz zawierający stół

Opis wygenerowany automatycznie

## Forecasting Cocoa Yields for 2050

Kozicka, M., Tacconi, F., Horna, D. and Gotor, E., 2018. Forecasting cocoa yields for 2050.

**Dobre źródło informacji na temat czynników jakie wpływają na poziom produkcji ziaren kakaowych.**

Na przykład informacja o analizie wykonanej na podstawie czynników dotyczących obszarów Ghany, gdzie wykazano, że 60 % zmienności produkcji ziaren ma swoje źródło z połączenia całkowitych rocznych opadów z poprzedniego roku, całkowitych opadów w dwóch najsuchszych miesiącach i całkowitego czasu nasłonecznienia.

Obraz zawierający stół

Opis wygenerowany automatycznie

Międzynarodowy Model Analizy Polityki Towarów Rolnych i Handlu (IMPACT) jest przykładem przykładem strukturalnego modelu symulacyjnego, który pozwala w przyszłości analizować rynek kakao w skali globalnej. Model ten został opracowany w International Food Policy Research Institute (IFPRI), aby

rozważyć długoterminowe wyzwania stojące przed decydentami w zakresie redukcji głodu i ubóstwa zrównoważony. IMPACT jest głównym narzędziem ilościowym wykorzystywanym przez Global Futures & Strategic Foresight (GFSF), w której Bioversity International uczestniczy jako partner. Obecnie IMPACT to sieć powiązanych ze sobą modeli ekonomicznych, wodnych i modeli upraw (dla wybranych upraw). Jego rdzeń stanowi dynamiczny, wielorynkowy model ekonomiczny o częściowej równowadze, który symuluje krajowe i międzynarodowe rynki rolne w długoterminowym horyzoncie czasowym do 2050 roku. Wyróżnia 62 rynki towarowe i 320 Food Production Units (FPUs), które są subnarodowymi jednostkami geoprzestrzennymi, które są generowane przez przecięcie 159 krajów ze 154 zlewniami hydrologicznymi (Robinson i in., 2015).

**Dobrym źródłem danych pogodowych może być** [**https://www.worldclim.org/**](https://www.worldclim.org/)**.**

**Wykorzystany model:**

ARIMA

**Metryki:**

* MAPE – Mean Absolute Percentage Error - Średni bezwzględny błąd procentowy
* MAE – Mean Absolute Error – Średni bezwzględny błąd
* CAGR – Comound Annual Growth Rate – Roczna składana stopa wzrtostu

****

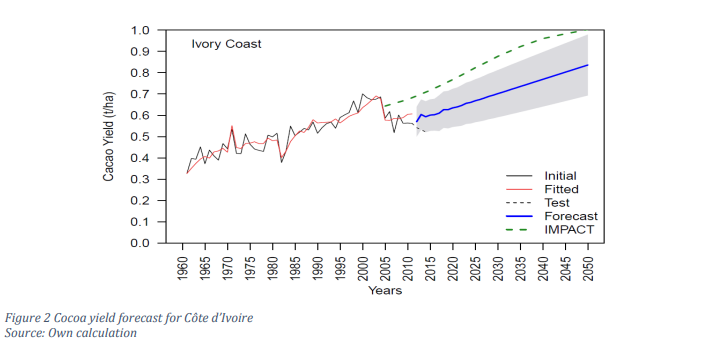
**Obraz zawierający tekst

Opis wygenerowany automatycznie**

**Porównanie odbywa się pomiędzy modelem ARIMA a opisanym wcześniej modelem IMPACT.** Przykład dla Côte d’Ivoire:

Obraz zawierający stół

Opis wygenerowany automatycznie



## Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia

Lamos-Díaz, H., Puentes-Garzón, D.E. and Zarate-Caicedo, D.A., 2020. Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia. Revista Facultad de Ingeniería, 29(54).

**Wykorzystane model:**

* Machine learning
  + Support Vector Machines (SVM)
  + Ensemble learning models (Random forests, Gradient boosting) – modele zespołowe
* Modele regresji
  + Least Absolute Shrinkage
  + Selection Operator (LASSO)

Rozpatrywanymi predyktorami były: **klimat warunki klimatyczne, odmiana kakao, poziom nawożenia i ekspozycja na słońce w eksperymentalnej uprawy zlokalizowanej w Rionegro, Santander**. Wyniki wykazały, że Gradient Boosting jest najlepszą najlepszą alternatywą predykcji z współczynnikiem determinacji (R2) = 68%, Mean Błąd bezwzględny (MAE) = 13,32, i Root Mean Square Error (RMSE) = 20,41.

Zmienność plonów jest wyjaśniana głównie przez promieniowanie na miesiąc przed zbiorem, sumą opadów w miesiącu zbiorów oraz temperaturą na miesiąc przed zbiorami. przed zbiorami. Podobnie, plony oceniane są na podstawie rodzaju ekspozycji na słońce. Stwierdzono, że promieniowanie na miesiąc przed zbiorami jest najbardziej wpływowym czynnikiem.

# Teoria – wykorzystanie uczenia maszynowego do analizy szeregów czasowych

**Bardzo przydatny link:** [**https://neptune.ai/blog/select-model-for-time-series-prediction-task**](https://neptune.ai/blog/select-model-for-time-series-prediction-task)

Modele do prognozowania szeregów czasowych można podzielić na 3 główne kategorie:

* Klasyczne modele regresji
* Modele nadzorowanego uczenia maszynowego
* Modele oparte na głębokim uczeniu

Dodatkowo dane szeregów czasowych można podzielić ze względu na liczność zmiennych docelowych.

|  |  |
| --- | --- |
| **Jednowariantowe modele szeregów czasowych** | **Wielowariantowe modele szeregów czasowych** |
| Użycie tylko jednej zmiennej | Używanie wielu zmiennych |
| Nie można korzystać z danych zewnętrznych | Można korzystać z danych zewnętrznych |
| Opiera się wyłącznie na związkach między przeszłością a teraźniejszością | Na podstawie związków między przeszłością i teraźniejszością oraz między zmiennymi |

**Dekompozycja szeregów czasowych** jest techniką pozwalającą na wyodrębnienie wielu rodzajów zmienności ze zbioru danych.

W szeregach czasowych można wyróżnić 3 najważniejsze składniki:

* **Sezonowość –** powtarzający się ruch, który jest obecny w zmiennej szeregu czasowego. Na przykład temperatura w danym miejscu będzie wyższa w miesiącach letnich i niższa w miesiącach zimowych. Można obliczyć średnie miesięczne temperatury i użyć tej sezonowości jako podstawy do prognozowania przyszłych wartości.
* **Trend -** może być długoterminową tendencją wzrostową lub spadkową. W szeregu czasowym temperatury, trend może być obecny z powodu globalnego ocieplenia. Na przykład, oprócz sezonowości lato/zima, można zauważyć niewielki wzrost średnich temperatur w czasie.
* **Szum -** jest częścią zmienności w szeregu czasowym, która nie może być wyjaśniona ani przez sezonowość, ani przez trend. Budując modele, łączymy różne składniki w formułę matematyczną. Dwiema częściami takiego wzoru mogą być sezonowość i trend. Model, który łączy oba te elementy, nigdy nie będzie idealnie reprezentował wartości temperatury: zawsze pozostanie błąd. Jest on reprezentowany przez współczynnik szumu.

**Autokorelacja**

Autokorelacja to korelacja pomiędzy obecną wartością szeregu czasowego z wartościami przeszłymi. Jeśli tak jest, można wykorzystać wartości obecne do lepszego przewidywania wartości przyszłych.

Autokorelacja może być pozytywna lub negatywna:

* Dodatnia oznacza, że wysoka wartość teraz prawdopodobnie przyniesie wysoką wartość w przyszłości i odwrotnie. Np. na giełdzie jeśli wszyscy kupują akcje to cena idzie w górę. Kiedy cena wzrasta, ludzie myślą, że jest to dobra akcja do kupienia i kupują ją również, tym samym napędzając cenę jeszcze wyżej. Jeżeli jednak cena spada, to wszyscy obawiają się krachu, sprzedają swoje akcje i cena staje się niższa.
* Ujemna autokorelacja jest przeciwieństwem: wysoka wartość dzisiaj implikuje niską wartość jutro, a niska wartość dzisiaj implikuje wysoką wartość jutro. Powszechnym przykładem jest populacja królików w środowisku naturalnym. Jeśli latem jednego roku jest dużo dzikich królików, zjedzą one wszystkie dostępne zasoby naturalne. Podczas zimy nie będzie już nic do jedzenia, więc wiele z nich umrze.

W wykryciu autokorelacji w zbiorze danych mogą pomóc dwa znane wykresy: wykres ACF i PACF. Wykres ACF (autocorrelation function) pozwala na zdiagnozowanie czy autokorelacja występuje.

PACF jest alternatywą dla ACF. Zamiast podawać autokorelacje, podaje autokorelację częściową, oznaczającą, że z każdym krokiem wstecz w przeszłości, tylko dodatkowe autokorelacje są wymienione. Różni się tym, że zmienność może być tu wyjaśniona przez wiele punktów w czasie. Na przykład jeśli wartość dzisiejsza jest taka sama jak wartość wczorajsza, ale także taka sama jak przedwczorajsza, ACF pokazałby dwa wysoce skorelowane kroki., natomiast PACF pokazałby tylko dzień wczorajszy i usunąłby dzień przedwczorajszy.

**Stacjonarność**

Stacjonarny szereg czasowy to taki który nie posiada trendu. Niektóre modele szeregów czasowych nie są w stanie poradzić sobie z trendami (więcej na ten temat później). Można wykryć niestacjonarność za pomocą testu Dickey-Fullera oraz można usunąć niestacjonarność za pomocą różnicowania.

**Test Dickeya-Fullera**

Jest to statystyczny test hipotez, który pozwala wykryć niestacjonarność. Np. można przyjąć 5 % przedział ufności (p-value = 0.05) dla hipotezy, że wykres ADF posiada trend, oraz przyjąć alternatywną hipotezę, że badany szereg czasowy jest stacjonarny.

**Modele jedno- i wieloetapowe (one- and multi-step)**

Niektóre modele świetnie sprawdzają się w przewidywaniu następnego kroku dla szeregu czasowego ale nie mają zdolności do przewidywania wielu kroków jednocześnie. **Te modele są modelami jednoetapowymi**. Można z nimi tworzyć modele wieloetapowe przez okienkowanie swoich przewidywań, ale istnieje ryzyko: kiedy używa się przewidywanych wartości do robienia prognoz, błędy mogą się szybo sumować i stać się bardzo dużymi.

**Modele wieloetapowe** to modele, które mają wewnętrzną zdolność do przewidywania wielu etapów jednocześnie. Są one zazwyczaj lepszym wyborem dla prognoz długoterminowych, a czasami również dla prognoz jednoetapowych. Kluczowe jest, aby zdecydować o liczbie kroków, które chce się przewidzieć przed rozpoczęciem modeli.

**Rodzaje modeli szeregów czasowych**

* **Klasyczne modele**

Są one silnie oparte na zmienności czasowej wewnątrz szeregu czasowego i działają dobrze z jednowymiarowymi szeregami czasowymi. Istnieją pewne zaawansowane opcje, aby dodać do modeli również zmienne zewnętrzne. Modele te są generalnie stosowane tylko do szeregów czasowych i nie są przydatne w innych typach uczenia maszynowego.

* **Modele nadzorowane (supervised models)**

Modele nadzorowane mogą być używane dla serii czasowych, o ile ma się sposób na wyodrębnienie sezonowości i umieszczenie jej w zmiennej. Przykłady obejmują stworzenie zmiennej dla roku, miesiąca lub dnia tygodnia, itp. Są one następnie używane jako zmienne X w twoim nadzorowanym modelu, a "y" jest rzeczywistą wartością serii czasowej. Można również włączyć opóźnione wersje y (przeszłą wartość y) do danych X, aby dodać efekty autokorelacji.

* **Uczenie głębokie i najnowsze modele**

Rosnąca popularność głębokiego uczenia w ostatnich latach otworzyła nowe drzwi również dla prognozowania, ponieważ wynaleziono specyficzne architektury głębokiego uczenia, które bardzo dobrze sprawdzają się na danych sekwencyjnych.

Chmura obliczeniowa i popularyzacja AI jako usługi również dostarczyły wielu nowych wynalazków w tej dziedzinie. Facebook, Amazon i inne duże firmy technologiczne otwierają swoje produkty prognostyczne lub udostępniają je na swoich platformach chmurowych. Dostępność tych nowych modeli "czarnych skrzynek" daje praktykom prognozowania nowe narzędzia do wypróbowania i przetestowania, a czasem nawet pokonania poprzednich modeli.

**Klasyczne modele prognozowania szeregów czasowych**

**Dokładniejszy opis ARIMA oraz SARIMA:** [**https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide**](https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide)

**Rodzina algorytmów ARIMA**

* **Autoregresja (AR)**

Model regresji, który wyjaśnia przyszłą wartość zmiennej za pomocą jej przeszłych (opóźnionych) wartości. Rząd modelu AR jest oznaczany jako p i reprezentuje liczbę wartości opóźnionych, które należy uwzględnić w modelu. Najprostszym modelem jest model AR(1): używa on tylko wartości z poprzedniego kroku czasowego do przewidywania bieżącej wartości. Maksymalna liczba wartości, które można użyć, to całkowita długość szeregu czasowego.

* **Moving average (MA) – średnia krocząca**

Wartościami przeszłymi w odróżnieniu od AR nie są wartości zmiennej a błędy predykcji w poprzednich krokach czasowych. Jest to dobry sposób na wykrycie sezonowości bez konieczności, skąd ona pochodzi.

* **Autoregressive moving average (ARMA) – autoregresyjna średnia ruchoma**

Wykorzystuje zarówno wartości jak i błędy predykcji wartości z przeszłości. ARMA może mieć różne wartości opóźnienia procesów AR i MA, np. ARMA(3, 1) ma rząd AR o z wartościach opóźnionych i używa 1 wartości opóźnionej dla MA.

* **Autoregressive integrated moving average (ARIMA) – autoregresyjna zintegrowana średnia ruchoma**

Model ARMA wymaga stacjonarności szeregu czasowego. Zawsze można zastosować różnicowanie ręcznie, jednak model ARIMA dodaje automatyczne różnicowanie do modelu ARMA. Posiada dodatkowy parametr, który można ustawić jako ilość razy, które szereg czasowy musi być zróżnicowany, np. ARIMA(3,1,1).

* **Seasonal autoregressive integrated moving-average (SARIMA) – sezonowa autoregresyjna zintegrowana średnia ruchoma**

Dodaje sezonowość to modelu ARIMA. Notacja to SARIMA(p,d,q)(P,D,Q)m, gdzie m to liczba obserwacji w ciągu roku, (p,d,q) to współczynniki niesezonowe, natomiast (P,D,Q) to współczynniki sezonowe. Każdy z komponentów otrzymuje parametr sezonowy.

* **Seasonal autoregressive integrated moving-average with exogenous regressors (SARIMAX) - Sezonowa autoregresyjna zintegrowana średnia ruchoma z regresorami egzogenicznymi (SARIMAX)**

Przegrupowuje on AR, MA, różnicowanie i efekty sezonowe. Na dodatek dodaje X: zmienne zewnętrzne.

**Vector autoregression (VAR) – Autoregresja wektorowa oraz jej pochodne VARMA i VARMAX**

Zamiast przewidywania jednej zmiennej zależnej, oferuje przewidywanie wielu serii czasowych w tym samym czasie. Może to być szczególnie przydatne, gdy istnieją silne związki między różnymi seriami czasowymi. Model VARMA jest wieloczynnikowym odpowiednikiem modelu ARMA. VARMA jest dla ARMA tym, czym VAR jest dla AR: dodaje do modelu składnik średniej ruchomej. VARMAX pozwala na wprowadzenie zmiennych egzogenicznych. Bardziej zaawansowane wersje, takie jak sezonowy VARMAX (SVARMAX) istnieją, ale stają się tak złożone i specyficzne, że trudno będzie znaleźć implementacje, które robią to łatwo i skutecznie.

**Wygładzanie**

Szeregi czasowe mogą zawierać szum, który obniża jakość stawianych prognoz. By uwypuklić główne składowe, można zastosować proces wygładzania.

* **Simple moving average – prosta średnia ruchoma**

Zastąpienie wartości bieżącej przez średnią z wartości bieżącej i kilku wartości przeszłych.

* **Simple exponential smoothing (SES) – proste wygładzanie wykładnicze**

Zamiast zwykłej średniej, bierze ważoną średnią z wartości przeszłych, gdzie im dalsza wartość tym mniej się liczy.

* **Double exponential smoothing (DES) - Podwójne wygładzanie wykładnicze**

Gdy w danych z szeregu czasowego występują trendy, należy unikać stosowania prostego wygładzania wykładniczego: nie działa ono dobrze w tym przypadku, ponieważ model nie może prawidłowo dokonać rozróżnienia między zmiennością a trendem. Można jednak użyć podwójnego wygładzania wykładniczego.

W DES mamy do czynienia z rekurencyjnym zastosowaniem filtra wykładniczego. Pozwala to na usunięcie problemów z trendem.

* **Holt Winter’s exponential smoothing (HWES) - Wykładnicze wygładzanie Holta Wintera**

Potrójne wygładzanie wykładnicze gdy w danych szeregu czasowego znajdują się trzy ważne składowe, np. trend oraz sezonowość tygodniowa oraz miesięczna.

**Zaawansowane modele**

* **Generalized Autoregressive Conditional Heteroskedasticity (GARCH)** – głównie do danych giełdowych/finansowych
* **TBATS**

T - Trigonometric seasonality

B - Box-Cox transformation

A - ARMA errors

T - Trend

S - Seasonal components

Model został stworzony w 2011 roku jako rozwiązanie do prognozowania szeregów czasowych z wieloma okresami sezonowymi. Jako że jest stosunkowo nowy i stosunkowo zaawansowany, jest mniej rozpowszechniony i nie tak często wykorzystywany jak modele z rodziny ARIMA.

## Supervised machine learning – nadzorowane uczenie maszynowe w prognozowaniu szeregów czasowych

Przykłady:

* Regresja liniowa
* Lasy losowe
* XGBoost
* SVM
* Płytkie sieci neuronowe
* maszyny wektorów nośnych (SVM – support vector machine)
* Naiwny klasyfikator Bayesa

## Deep learning – głębokie uczenie maszynowe w prognozowaniu szeregów czasowych

Przykłady:

* LSTM (Long Short-Term Memory)
* Prophet – model typu black-box dostarczony przez Facebook
* DeepAR - model typu black-box dostarczony przez Amazon