Welcome to the appendix of “Forcing of abrupt transitions of the last 300,000 yrs” 😊

Vanessa Skiba, Martin Trüssel, Christoph Spötl, Andrea Schröder-Ritzau, René Eichstädter, Birgit Plessen, Norbert Frank, Tobias Braun, Takahito Mitsui, Norbert Marwan, Niklas Boers & Jens Fohlmeister
Methods – Event detection

In order to identify abrupt transitions in the stalagmite records, we linearly interpolate the data to obtain a regular time scale (equidistant time steps). We then run a two neighbouring 300-year windows over the data (sketch on the right). For each time step, we compute the difference in mean values of the isotope data in each window. When the difference is above/below a certain threshold (one standard deviation (1σ) of the whole record of difference in neighbouring windows δ^{18}O) we mark the first time step crossing the threshold as corresponding to an abrupt transition.
Methods – Event detection

In the example plot below we show that procedure for one of the Swiss stalagmite composite records (Marine Isotope Stage 8). A higher isotope value in the first 300-yr window than in the subsequent 300-yr window (positive difference in means), marks transitions leading to warmer climate conditions (interstadials, pink dots in example plot below) and vice versa for transitions leading to colder climate conditions (blue dots in example plot below).

In Chinese stalagmites, the a lower δ¹⁸O corresponds to interstadials/transitions leading to warmer conditions (and higher δ¹⁸O to stadials), thus negative differences in means between the running windows here correspond to interstadal transitions (thus, would be pink dots).

Hit me up if you wanna know about what we think the Swiss stalagmite δ¹⁸O is a proxy for!
Methods – Zero-inflated Poisson regression model

Thus, a ZIP regression model consists of three parts:

1) Logistic regression
   1. A PMF $P(y_i=0)$ which is used to calculate the probability of observing a zero count.

2) Poisson regression
   2. A second PMF $P(y_i=k)$ which is used to calculate the probability of observing $k$ events, given that $k > 0$.
   3. A link function that is used to express the mean rate $\lambda$ as a function of the regression variables $X$.

© Sachin Date
https://timeseriesreasoning.com/
Methods – Zero-inflated Poisson regression model

Thus, a ZIP regression model consists of three parts:

1) Logistic regression
   1. A PMF $P(y_i=0)$ which is used to calculate the probability of observing a zero count.

2) Poisson regression
   2. A second PMF $P(y_i=k)$ which is used to calculate the probability of observing $k$ events, given that $k > 0$.
   3. A link function that is used to express the mean rate $\lambda$ as a function of the regression variables $\mathbf{X}$.

Probability distribution of the Logistic Model

\[
P(y'_i = 0) = \mu_i = \frac{e^{x_i \gamma}}{1 + e^{x_i \gamma}}
\]

\[
P(y'_i = 1) = (1 - \mu_i)
\]

Data set → 1. Transform $\mathbf{y}$ to a binary vector $\mathbf{y}'$

2. Train a Logistic regression model on ($\mathbf{y}'$, $\mathbf{X}$) to get vector of fitted probabilities $\boldsymbol{\mu}$

3. Set $\varphi$ to $\mu$

Maximum Likelihood Estimation (MLE)

"excess zeros" -> Logistic regression
Methods – Zero-inflated Poisson regression model

Thus, a ZIP regression model consists of three parts:

1. Logistic regression
   A PMF \( P(y_i=0) \) which is used to calculate the probability of observing a zero count.

2. Poisson regression
   A second PMF \( P(y_i=k) \) which is used to calculate the probability of observing \( k \) events, given that \( k > 0 \).

3. A link function that is used to express the mean rate \( \lambda \) as a function of the regression variables \( X \).

Maximum Likelihood Estimation (MLE)

“excess zeros” -> Logistic regression

Poisson distribution

©Sachin Date
https://timeseriesreasoning.com/
Methods – Zero-inflated Poisson regression model

Thus, a ZIP regression model consists of three parts:

1. **Logistic regression**
   1. A PMF $P(y_i=0)$ which is used to calculate the probability of observing a zero count.

2. **Poisson regression**
   2. A second PMF $P(y_i=k)$ which is used to calculate the probability of observing $k$ events, given that $k > 0$.
   3. A link function that is used to express the mean rate $\lambda$ as a function of the regression variables $X$.

![Diagram of ZIP regression model](https://timeseriesreasoning.com/)

Where $\lambda_i = e^{x_i \beta}$

$\beta$ = Vector of regression coefficients

$x_i$ = Regression variables in the $i^{th}$ row

Maximum Likelihood Estimation (MLE)

Data set -> 1. Calculate the $\phi$ vector

2. Use this 2-part probability distribution to train the ZIP Model on the $(y, X)$ data set

3. Use the trained model to make predictions on the test data

© Sachin Date

https://timeseriesreasoning.com/
Methods – Zero-inflated Poisson regression model

Brushing up on Poisson regression

A Poisson Process with mean rate $\lambda=5$

$$P(y = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

Probability of seeing $k$ events in time $t$, given $\lambda$ events occurring per unit time

Probability of seeing zero events =

$$P(y = 0) = \frac{e^{-\lambda} \cdot \lambda^0}{0!} = e^{-\lambda} = e^{-5} = 0.0067$$
Methods – Training and testing procedure

Different (combinations) predictor variables

Model **training** on half of the data (predictor variables & count data)

Use fitted model to predict counts via remaining predictor variables (**test data**)

Estimate model **accuracy** using difference between predicted vs observed counts

Check for overfitting by comparing train and test accuracy

Vanessa Skiba, Potsdam Institute for Climate Impact Research, skiba@pik-potsdam.de
Methods – Training and testing procedure

Model training on half of the data (predictor variables & count data)

Use fitted model to predict counts via remaining predictor variables (test data)

Estimate model accuracy using difference between predicted vs observed counts

Model training involves hyperparameter optimisation: we include $n$th degree polynomials of the input variables and test whether the model accuracy improves for both test and training data (see next slides). If increasing the polynomial degree would improve prediction accuracy for the train data but not for the test data, while the variance of the test data prediction accuracy increases, we have overfitted the model (found relationships in the train data which are not a description of the data in general).

Check for overfitting by comparing train and test accuracy
Methods – Training and testing procedure

Model training on half of the data (predictor variables & count data) → Use fitted model to predict counts via remaining predictor variables (test data) → Estimate model accuracy using difference between predicted vs observed counts

Model training involves hyperparameter optimisation: we include $n^{th}$ degree polynomials of the input variables and test weather the model accuracy improves for both test and training data (see next slides). If increasing the polynomial degree would improve prediction accuracy for the train data but not for the test data, and the variance of the test data prediction accuracy increases, we have overfitted the model (found relationships in the train data which are not a description of the data in general).

By including $n^{th}$ degree polynomials we can model non-linear relationship between predictor variables and the dependent variable (the count data). However, the underlying model equation stays linear!

Check for overfitting by comparing train and test accuracy
Methods – Training and testing procedure

Model **training** on half of the data (predictor variables & count data)

Use fitted model to predict counts via remaining predictor variables (test data)

Estimate model **accuracy** using difference between predicted vs observed counts

By including $n^{\text{th}}$ degree polynomials we can model non-linear relationship between predictor variables and the dependent variable (the count data). However, the underlying model equation stays linear!

Check for overfitting by comparing train and test accuracy

Model **training** involves hyperparameter optimisation: we include $n^{\text{th}}$ degree polynomials of the input variables and test whether the model accuracy improves for both test and training data (see next slides). If increasing the polynomial degree would improve prediction accuracy for the train data but not for the test data, and the variance of the test data prediction accuracy increases, we have overfitted the model (found relationships in the train data which are not a description of the data in general).

See example of confusion matrix on the left. It consists of predicted counts (x-axis) vs observed counts (y-axis). From the confusion matrix we calculate the number of false predictions and calculate the fraction of false predictions from the total number of predictions.

Vanessa Skiba, Potsdam Institute for Climate Impact Research, skiba@pik-potsdam.de
Let’s have a chat! 😊

Now or:

skiba@pik-potsdam.de
References


Great online source
Sachin Date’s website: https://timeseriesreasoning.com/