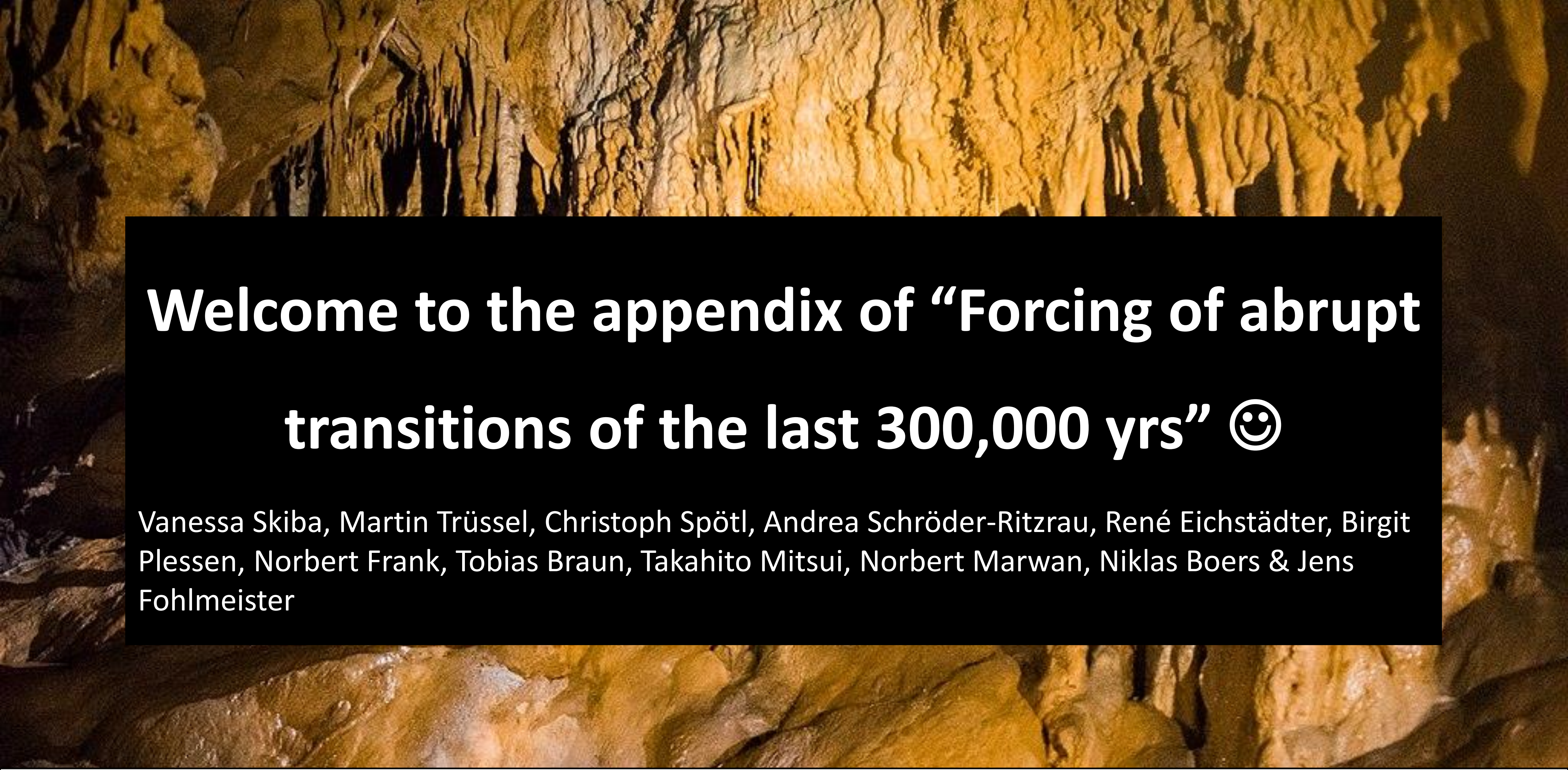
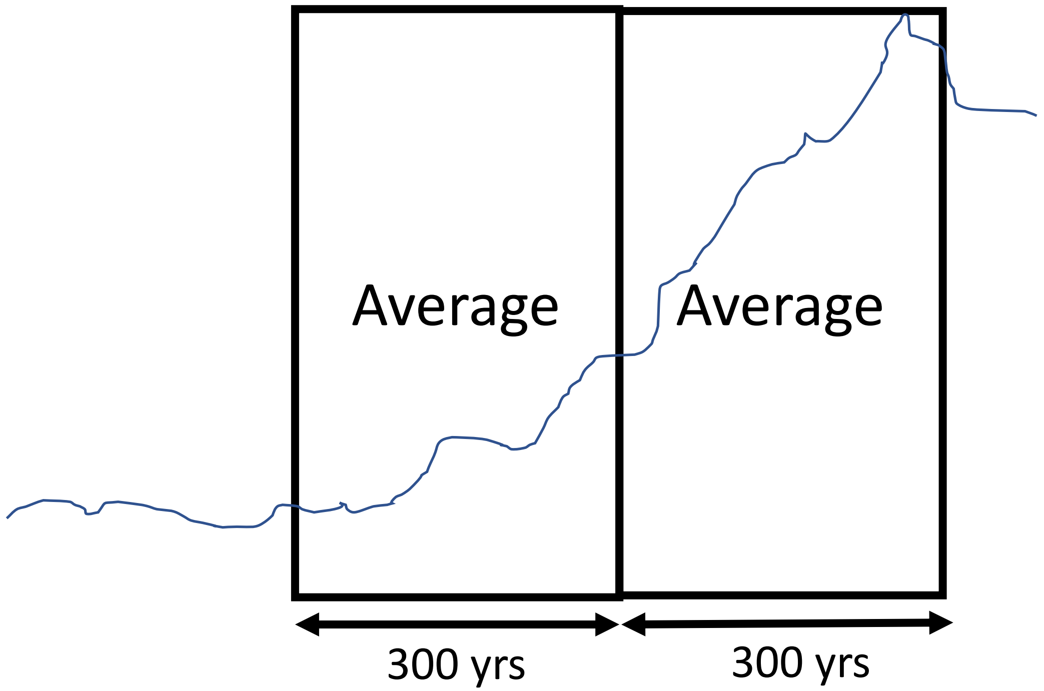


Welcome to the appendix of “Forcing of abrupt transitions of the last 300,000 yrs” ☺

Vanessa Skiba, Martin Trüssel, Christoph Spötl, Andrea Schröder-Ritzrau, 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 $\delta^{18}O$) we mark the first time step crossing the threshold as corresponding to an abrupt transition.

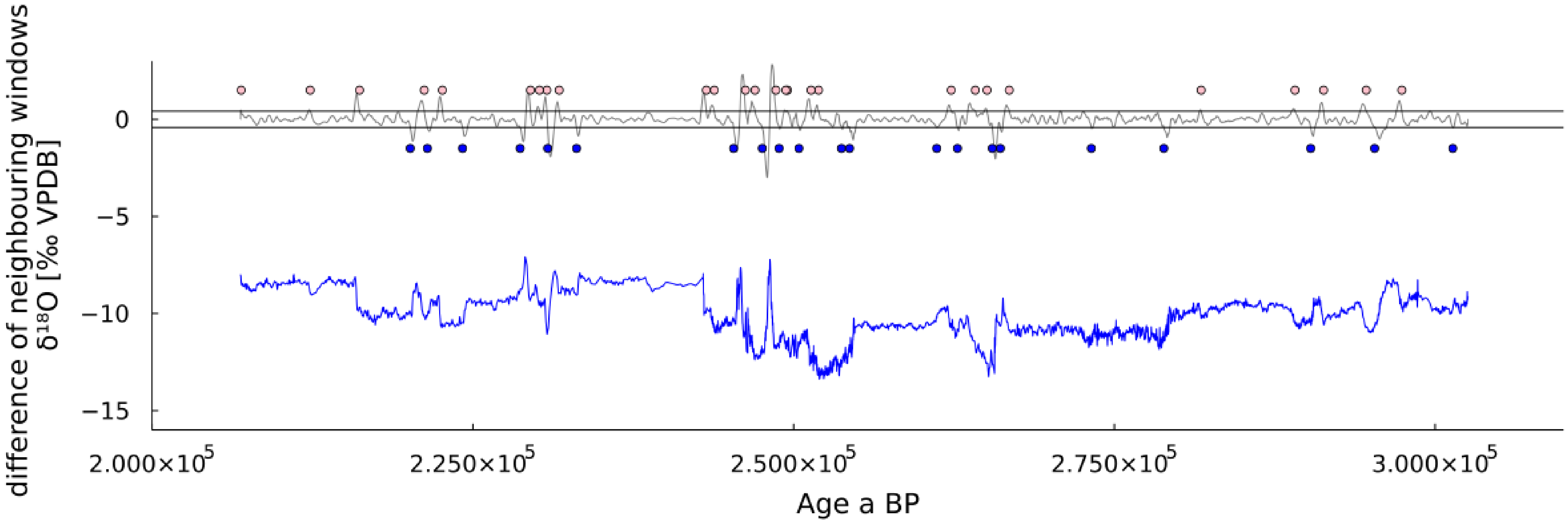


Methods – Event detection

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 $\delta^{18}O$ corresponds to interstadials/transitions leading to warmer conditions (and higher $\delta^{18}O$ to stadials), thus negative differences in means between the running windows here correspond to interstadial transitions (thus, would be pink dots).

Hit me up if you wanna know about what we think the Swiss stalagmite $\delta^{18}O$ is a proxy for!



Vanessa Skiba, Potsdam Institute for Climate Impact Research, skiba@pik-potsdam.de

Methods – Zero-inflated Poisson regression model

1) Logistic regression

2) Poisson regression

Thus, a **ZIP** regression model consists of three parts:

1. A PMF *P(y_i=0)* which is used to calculate the probability of observing a zero count.
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 λ as a function of the regression variables ***X*.**

Probability Mass Function of the Zero-Inflated Poisson regression model

When $y_i = 0$:

$$P(y_i = 0) = \phi_i + (1 - \phi_i) * e^{-\lambda_i}$$

ϕ_i= proportion of zeros for the i[th] row in the data set

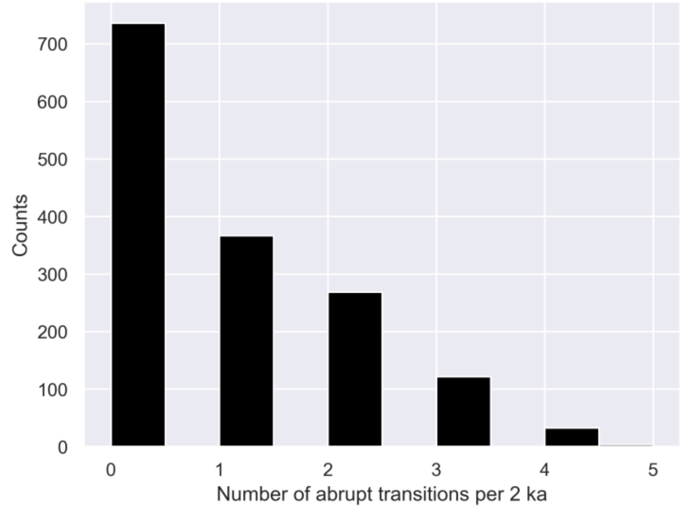
When $y_i = 1,2,3,\ldots$:

$$P(y_i = k) = (1 - \phi_i)\frac{e^{-\lambda_i} * \lambda_i^k}{k!}$$

$\boldsymbol{x_i}$=Regression variables in the i[th] row

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

β=Vector of regression coefficients



Methods – Zero-inflated Poisson regression model

1) Logistic regression

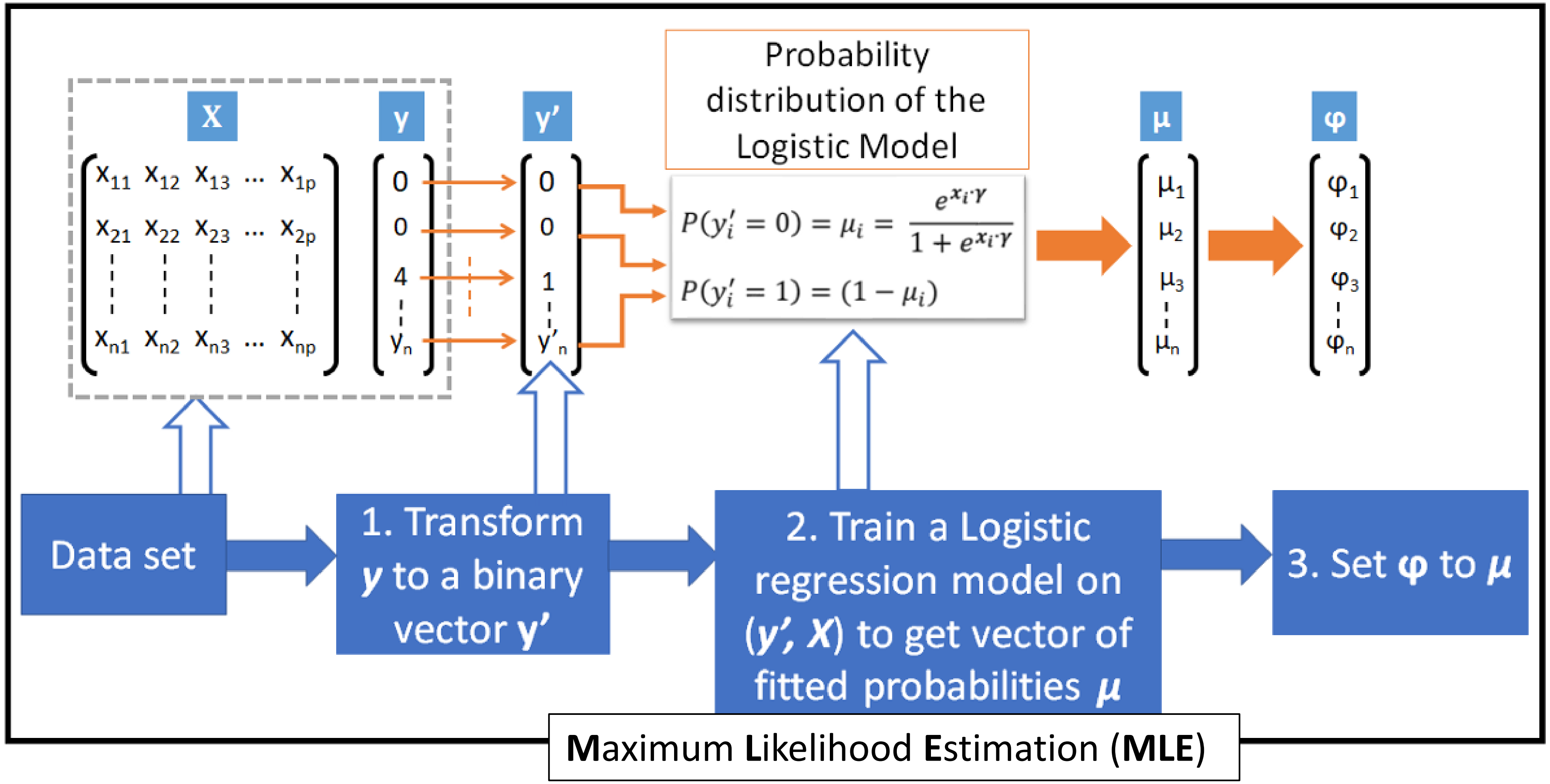
2) Poisson regression

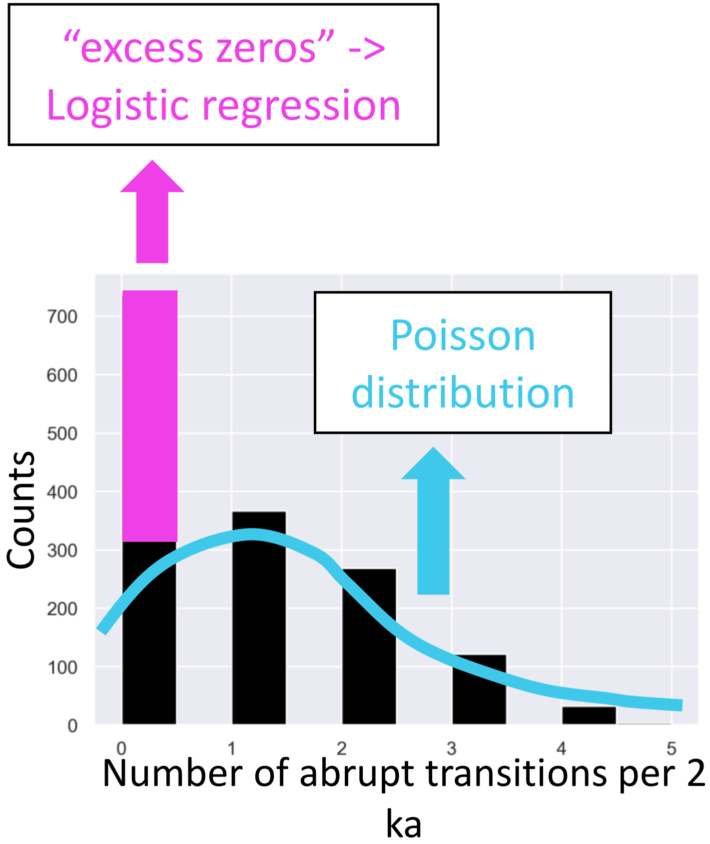
Thus, a **ZIP** regression model consists of three parts:

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

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 λ as a function of the regression variables ***X*.**





Methods – Zero-inflated Poisson regression model

1) Logistic regression

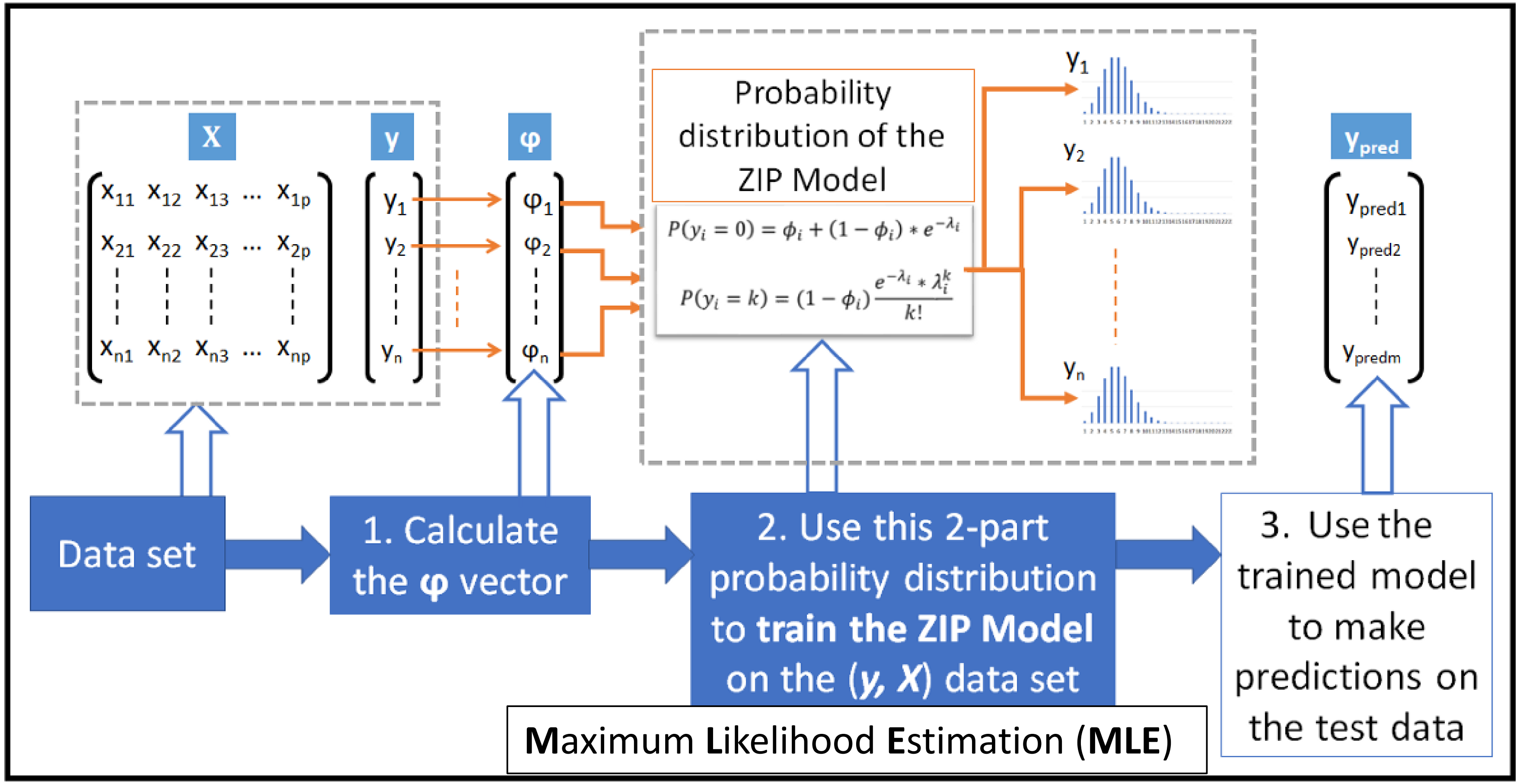
2) Poisson regression

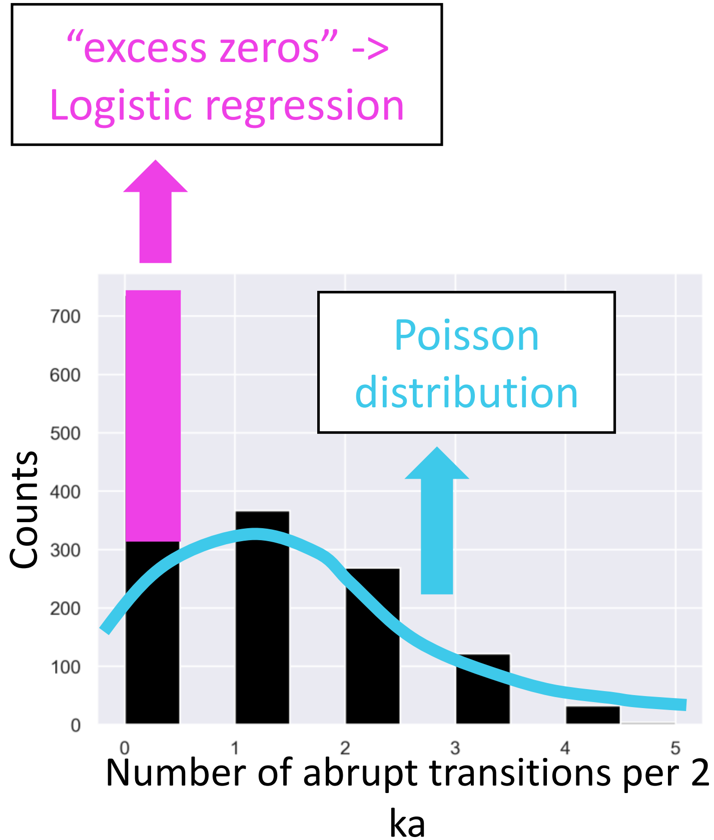
Thus, a **ZIP** regression model consists of three parts:

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

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 λ as a function of the regression variables ***X*.**





Methods – Zero-inflated Poisson regression model

1) Logistic regression

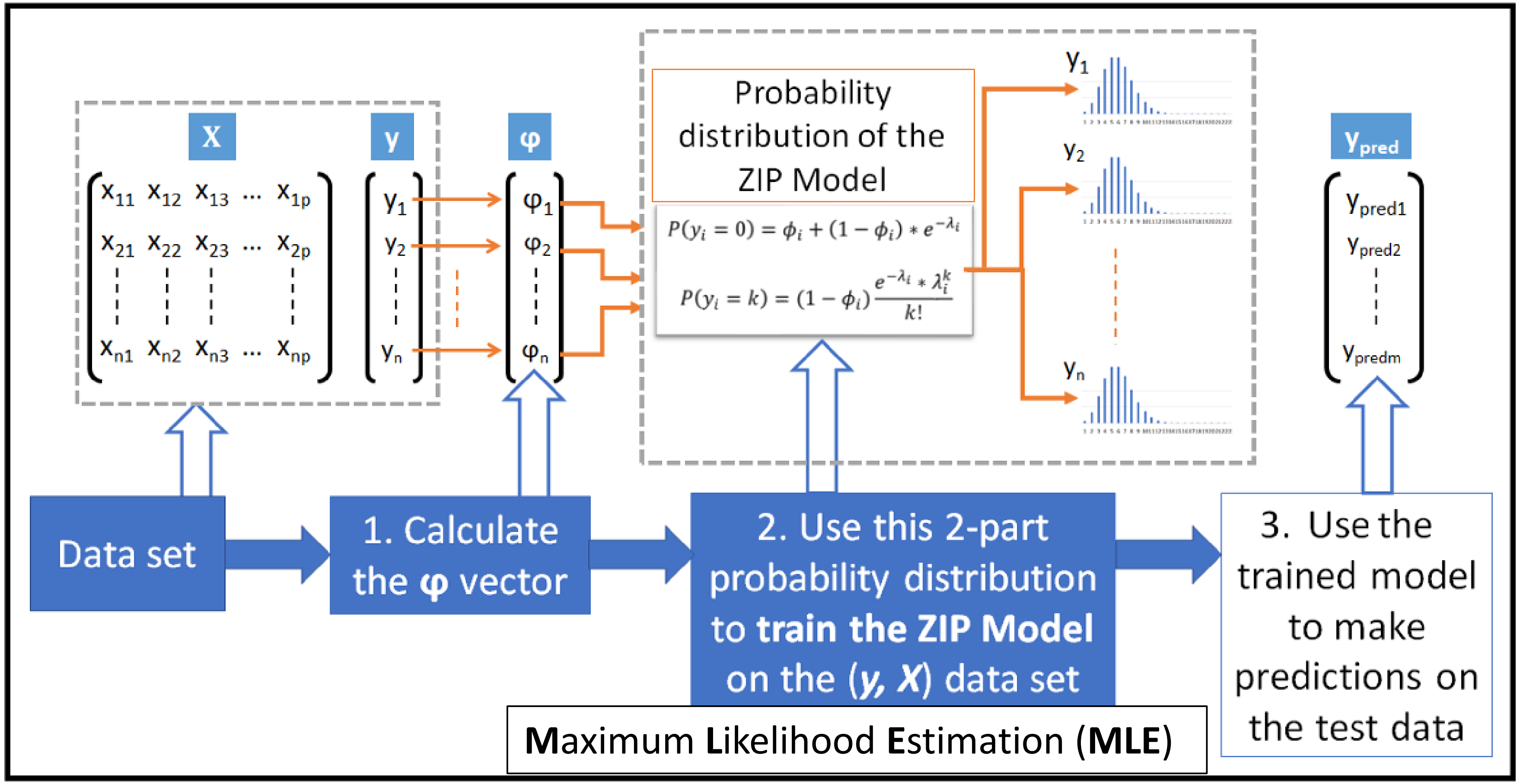
2) Poisson regression

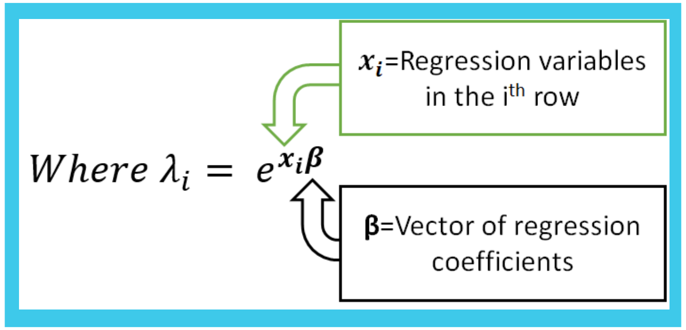
Thus, a **ZIP** regression model consists of three parts:

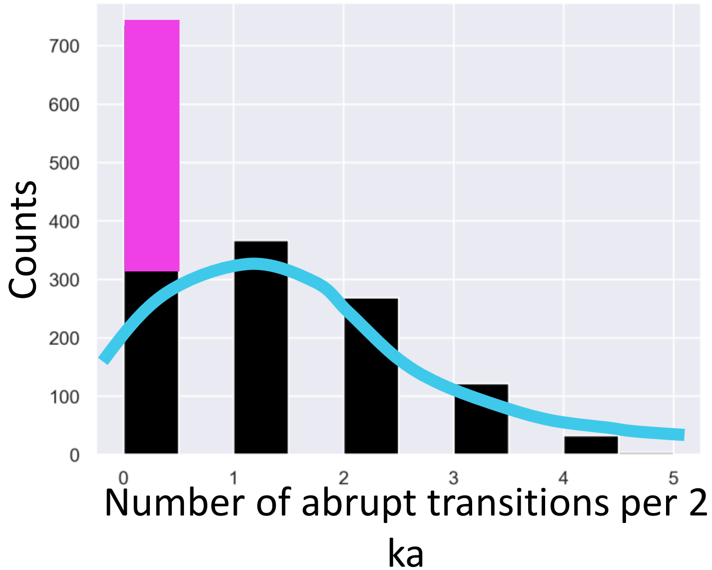
1. A PMF *P(y_i=0)* which is used to calculate the probability of observing a zero count.

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 λ as a function of the regression variables ***X***.

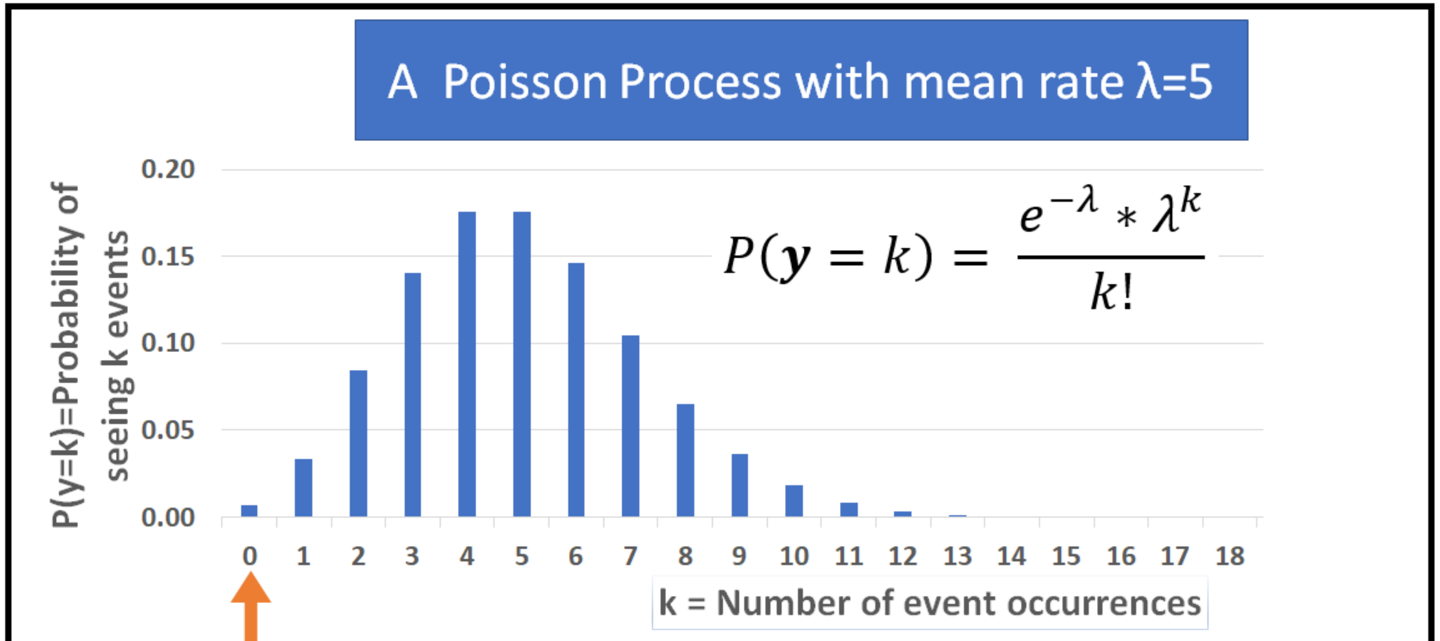






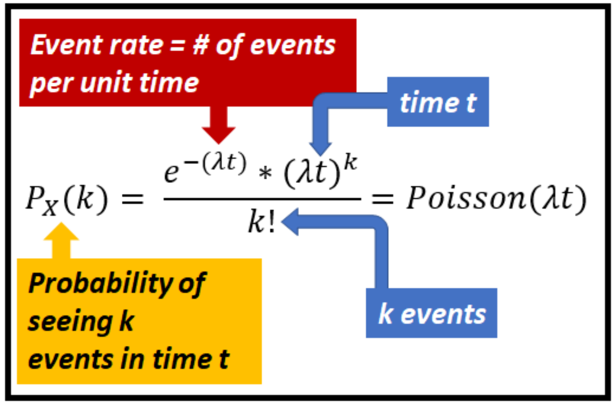
Methods – Zero-inflated Poisson regression model

Brushing up on Poisson regression



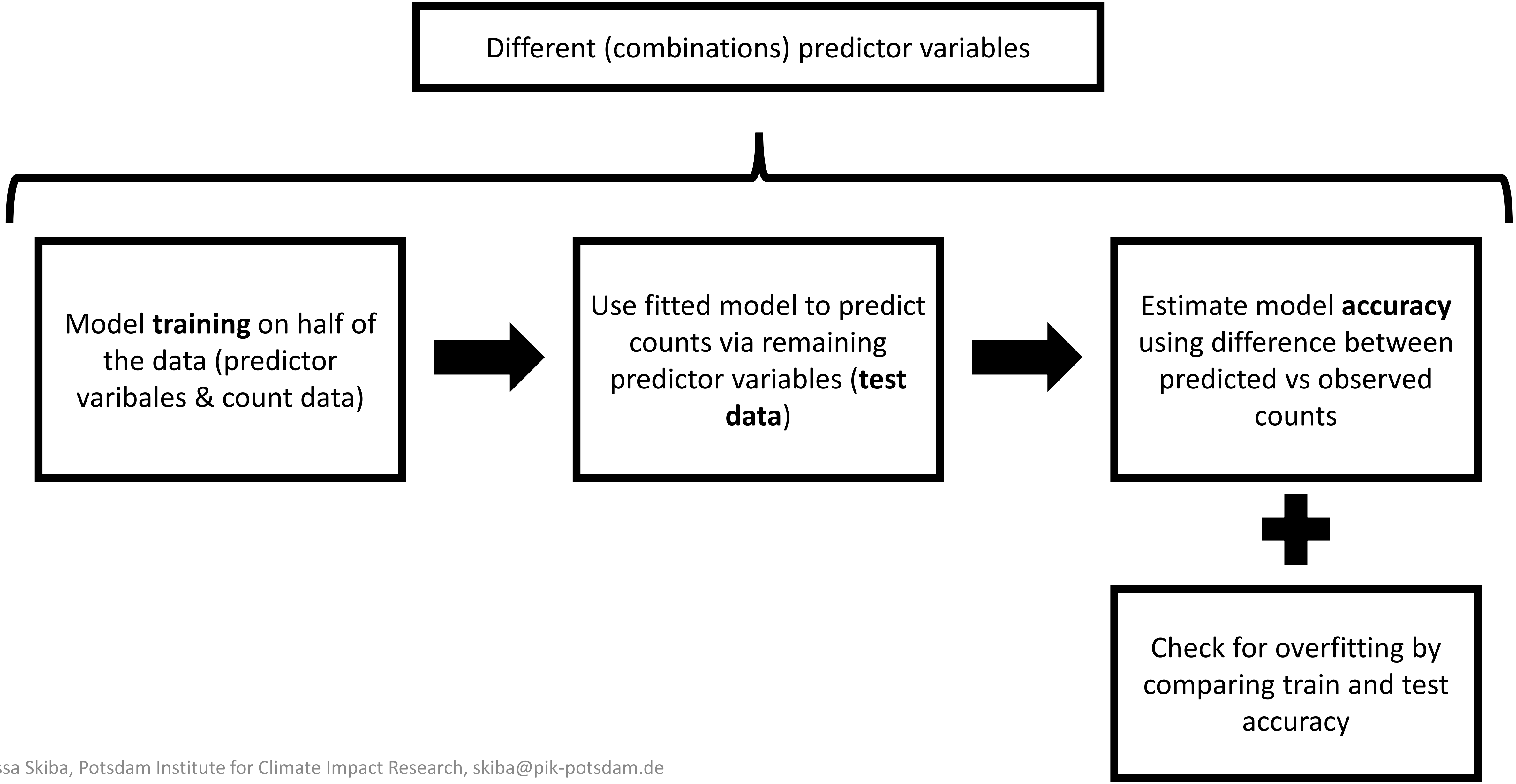
$$Probability\ of\ seeing\ zero\ events =$$

$$P(y=0) = \frac{e^{-\lambda} * \lambda^0}{0!} = e^{-\lambda} = e^{-5} = 0.0067$$



Probability of seeing ***k*** events in time **t**, given **λ** events occurring per unit time

Methods – Training and testing procedure



Vanessa Skiba, Potsdam Institute for Climate Impact Research, skiba@pik-potsdam.de

Methods – Training and testing procedure

Different (combinations) predictor variables

Model **training** on half of the data (predictor varibales & 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

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, 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).

Vanessa Skiba, Potsdam Institute for Climate Impact Research, skiba@pik-potsdam.de

Methods – Training and testing procedure

Different (combinations) predictor variables

Model **training** on half of the data (predictor varibales & 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

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 impr prediction accuracy for the train data but not for the test da the variance of the test data prediction accuracy increases, overfitted the model (found relationships in the train data w not a description of the data in general).

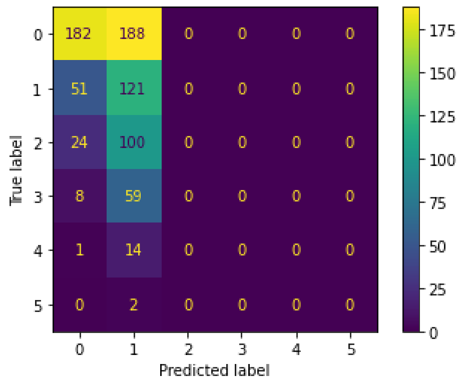
By including n^{th} degree polynomials we can model non-linear relationship between predictor variables and the dependent variable (the count data). However, the underlying model equation stays linear!

Vanessa Skiba, Potsdam Institute for Climate Impact Research, skiba@pik-potsdam.de

Methods – Training and testing procedure

Different (combinations) predictor var

Prediction **accuracy** = 1 - **Misclassification rate**
Misclassification rate is based on **confusion matrix**



See example of confusion matrix on the left. It consists of 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

Model **training** on half of the data (predictor varibales & 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 impr prediction accuracy for the train data but not for the test da the variance of the test data prediction accuracy increases, overfitted the model (found relationships in the train data w 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

Vanessa Skiba, Potsdam Institute for Climate Impact Research, skiba@pik-potsdam.de

Let's have a chat! ☺

Now or:

skiba@pik-potsdam.de

References

Bereiter, B., Eggleston, S., Schmitt, J., Nehrbass-Ahles, C., Stocker, T. F., Fischer, H., ... & Chappellaz, J. (2015). Revision of the EPICA Dome C CO_2 record from 800 to 600 kyr before present. *Geophysical Research Letters*, *42*(2), 542-549.

Cheng, H., Edwards, R. L., Sinha, A., Spötl, C., Yi, L., Chen, S., ... & Zhang, H. (2016). The Asian monsoon over the past 640,000 years and ice age terminations. *Nature*, *534*(7609), 640-646.

Grant, K. M., Rohling, E. J., Ramsey, C. B., Cheng, H., Edwards, R. L., Florindo, F., ... & Williams, F. (2014). Sea-level variability over five glacial cycles. *Nature communications*, *5*(1), 1-9.

Jouzel, J., Masson-Delmotte, V., Cattani, O., Dreyfus, G., Falourd, S., Hoffmann, G., ... & Wolff, E. W. (2007). Orbital and millennial Antarctic climate variability over the past 800,000 years. *science*, *317*(5839), 793-796.

Laskar, J., Robutel, P., Joutel, F., Gastineau, M., Correia, A. C. M., & Levrard, B. (2004). A long-term numerical solution for the insolation quantities of the Earth. *Astronomy & Astrophysics*, *428*(1), 261-285.

Lisiecki, L. E., & Raymo, M. E. (2005). A Pliocene-Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography*, *20*(1).

Loulergue, L., Schilt, A., Spahni, R., Masson-Delmotte, V., Blunier, T., Lemieux, B., ... & Chappellaz, J. (2008). Orbital and millennial-scale features of atmospheric CH_4 over the past 800,000 years. *Nature*, *453*(7193), 383-386.

Spratt, R. M., & Lisiecki, L. E. (2016). A Late Pleistocene sea level stack. *Climate of the Past*, *12*(4), 1079-1092.

Willeit, M., Ganopolski, A., Calov, R., & Brovkin, V. (2019). Mid-Pleistocene transition in glacial cycles explained by declining CO2 and regolith removal. *Science Advances*, *5*(4), eaav7337.

Great online source

Sachin Date’s website: https://timeseriesreasoning.com/