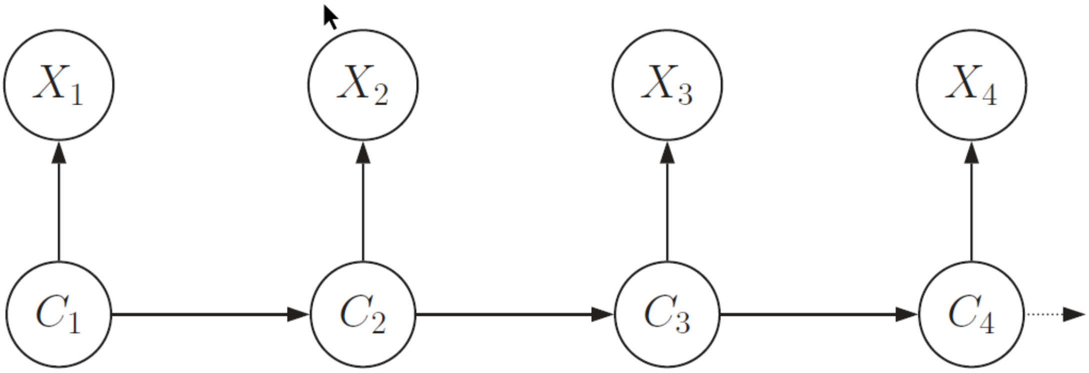
1. Introduction - Leo

Hong Kong is a city located in a subtropical area. The position intensely affects the rainfall level in Hong Kong, especially in summer. The availability of rainfall data from Hong Kong Observatory allows the formation of a Hidden Markov Model (HMM).

The main objectives of the project are to make predictions on the hidden state of the rainfall data by utilizing decoding problems and to find out the optimal sequence by using Algorithms, and also to make predictions on probability by utilizing the Algorithm in the Hidden Markov Model. The project takes the rainfall data in Hong Kong from 1947 to 2021.

The results of the prediction would show whether the Hidden Markov Model can be used to predict the probability of rainfall data availability.

2. Propose and fully specify a hidden markov model, or its extension - Kevin



The HMM is based on augmenting the Markov chain. A Markov chain is a model that tells us something about the probabilities of sequences of random variables, states, each of which can take on values from some set. These sets can be words, or tags, or symbols representing anything, like the weather. A Markov chain makes a very strong assumption that if we want to predict the future in the sequence, all that matters is the current state. The states before the current state have no impact on the future except via the current state. It's as if to predict tomorrow's weather you could examine today's weather but you weren't allowed to look at yesterday's weather.

We will use the Poisson Hidden Markov Model in the following parts. It is a mixture of two random processes, a Poisson process and a discrete Markov process, to represent counts based time series data. We will specify the Poisson mean to be (10,20) in the 2-states model and (10,20,30) in the 3-states model.

3. Simulate from this hidden markov model - Heidi

Two sets of simulated data with length 100 are created. The first set of simulated data is simulated from a 2-states non-stationary Poisson-HMM model, with Markov chain of , and Poisson parameter , with seed of 40112.

Simulated data set 1

7 17 13 23 15 18 21 14 18 18 21 16 17 14 19 17 22 17 20 20
20 27 23 11 16 16 25 22 26 14 22 15 25 16 26 22 21 16 21 12
15 25 18 19 17 20 17 24 30 17 16 11 9 12 14 22 15 3 4 14
15 26 14 5 13 9 9 8 8 12 20 19 18 19 17 12 21 7 29 22
18 19 15 19 22 24 11 20 21 26 21 23 18 20 24 20 16 20 19 18

Summary of simulated data set 1

Maximum	Minimum	First quartile	Median	Third quartile	Mean	standard deviation
30	3	14.75	18	21	17.62	5.55

Scatter plot & density plot of simulated data set 1

The second set of simulated data is simulated from a 3-states non-stationary Poisson-HMM model, with Markov chain of , and Poisson parameter , with seed of 40112.

Simulated data set 2

26 17 13 33 15 27 13 13 9 9 10 7 8 11 12 9 8 11 23 32

10 10 10 15 12 9 13 6 9 11 8 14 28 13 14 7 14 11 11 7

10 5 6 13 8 9 7 10 7 24 42 17 25 21 28 12 23 22 24 3

11 37 6 26 23 13 36 9 9 8 18 35 20 29 8 19 27 12 31 7

17 11 9 9 6 9 22 35 11 10 10 14 11 12 8 10 12 10 7 10

Summary of simulated data set 2

Maximum	Minimum	First quartile	Median	Third quartile	Mean	standard deviation
42	3	9	11	19.25	14.81	8.58

Scatter plot & density plot of simulated data set 2

4. **Perform parameter learning from this hidden Markov model (Both direct MLE or forward+backward+EM algorithms are fine) - Kitty**

In the following parts, we will use MLE and EM algorithms by Poisson-HMM to simulate 1 dataset.

MLE in Poisson-HMMs

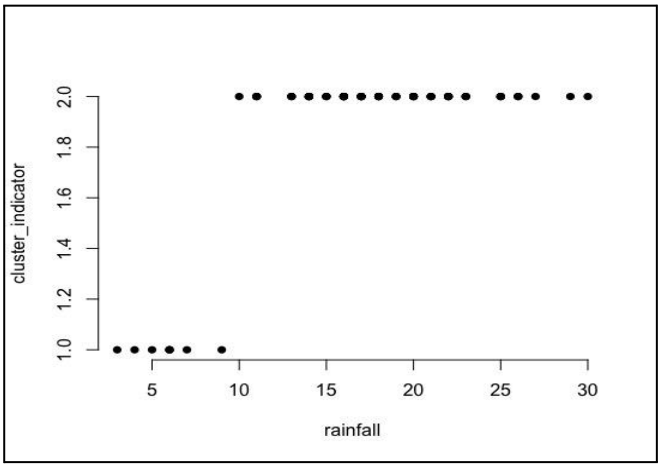
The estimates:

Parameter	**Two-state Poisson HMM**	**Three-state Poisson HMM**
Iteration	12	19
Log-likelihood	-226.890	-228.127
λ	(7.390, 18.958)	(8.546, 16.624, 21.236)
δ	(0.000, 1.000)	(0.0138, 0.9724, 0.01378)

Transition Matrix (Γ)	$\begin{bmatrix} 0.737 & 0.263 \\ 0.049 & 0.951 \end{bmatrix}$	$\begin{bmatrix} 0.678 & 0.000 & 0.322 \\ 0.130 & 0.000 & 0.870 \\ 0.000 & 1.000 & 0.000 \end{bmatrix}$

EM in Poisson-HMMs

Parameter	Two-state Poisson HMM
Iteration	50
Log-likelihood	-237.1185
δ	(0.114, 0.886)
State	State 1 ~ Poisson(6.263) State 2 ~ Poisson(18.580)



5. Check accuracy of the parameter learning - Kristy

Mean Percentage Error: (True Mean-Predict Mean)/True Mean

	Two-state Poisson HMM	Three-state Poisson HMM

RMSE	State 1: 11.431 State 2: 6.175	State 1: 10.458 State 2: 5.937 State 3: 7.173
Mean % error	State 1: 57% State 2: -10.4%	State 1: 50.2% State 2: 3.2% State 3: -23.7%

6. State and possibly interpret the difference among hidden states and possibly an application of this model - Kaysha

Hidden States

The HMM model assumes that the climate is composed of three states, either a dry state (D), corresponding to a low rainfall year, a wet state (W), corresponding to a moderate rainfall year, or a rainy state (R), corresponding to a high rainfall year. Each state has an independent rainfall distribution, assumed to be Poisson.

The parameter of interest, lambda, is the number of days in a year where rainfall amount exceeds 50mm:

State 1: Low Rainfall ~ Poisson(8.546)

State 2: Moderate Rainfall ~ Poisson(16.624)

State 3: High Rainfall ~ Poisson(21.236)

If the current year is in a dry state, the next year will most likely be a dry state (p = 0.678) or possibly transition towards the other extreme, a rainy state (0.322). However, there will be no probability that it goes to the wet state. On the other hand, if it is currently a wet state, the climate will not stay in the same state, but will either transition to a dry state (0.130) or rainy state (0.1870). Finally, if it is a rainy state, it will always be a wet state in the next year. Though this may not be the real world rainfall transition matrix, the simulated difference between hidden states is a concrete example of how the local and global climate can vary widely from year to year.

Application: Crop Production

Farmers, scientists, and governments can participate in proper crop planning and cultivation before the growing season, leading to superior harvests. This is particularly important in developing countries where the limitations of technological capabilities is a fundamental problem due to the economic prominence of the agricultural industry in reference to the country's gross domestic output. By generating significant numbers of simulated rainfall data, a prediction model of crops is realized. Then the relevant stakeholders and policy makers can take risk assessments and deal with uncertainty regarding annual rainfall. Combining data amongst numerous weather stations within a country or region, the hidden markov model acts as a stochastic weather generator or mechanism whereby the influence of climate phenomena such as the El Niño–Southern Oscillation, one of the most important climate phenomena on Earth due to its ability to change the global atmospheric circulation, which in turn, influences temperature and precipitation across the globe, is simulated. Depending on the state it is in, namely neutral, El Niño, or La Niña, it may lead to drought or flooding, higher or lower daytime temperature, as well as less or more tropical cyclones in a region. In the present and in the future, rainfall data in tandem with various other meteorological data allow for the anticipation of extreme climate change and act as early warning signals to farmers and residents of natural disasters including floods and droughts. This forms the basis for disaster mitigation and policy strategy determination.

7. Viterbi Algorithm for global decoding

Viterbi algorithm

The process of discovering the sequence of hidden states, given the sequence
of observations, is known as decoding or inference. The Viterbi algorithm is
commonly used for decoding. It is also a dynamic algo that allows us to compute the most probable path. It takes the maximum over the previous path probabilities. And the package will help us recursively compute the probability of the most probable path.

In our example, we will consider the weather condition as a simple HMM:im

2 hidden states: dry or wet

3 possible observations(probabilities that year with rainfall >50mm):

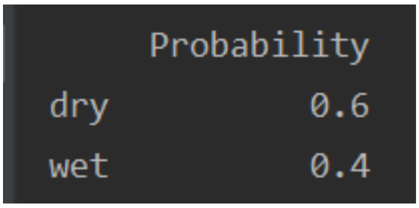
low(0-10 days), middle(11-15 days) or high(more than 15 days)

The model can be used to predict the weather's condition, at every time-step, from a given observation sequence. There are several paths through the hidden states (dry and wet) that lead to the given sequence, each with different paths.

Parameters:

Initial probability : A starting probability distribution over states P(initialState)

Our initial state vector:

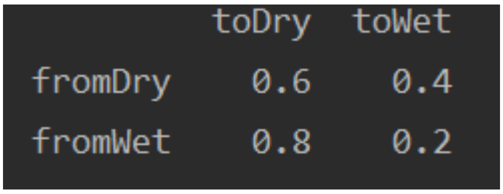


```
     Probability
dry          0.6
wet          0.4
```

Transition probability : A matrix with the probabilities from transitioning from one state to the next state, over time. P(nextState|currentState)

Transition probability matrix : Let's assume that if this year is dry, the probability that next year will be dry is 60%. So that leaves a 40% chance that next year will be wet.

Similarly, if this year is wet , there's a 20% chance that the next year will be wet and a 80% chance that the next year will become dry.

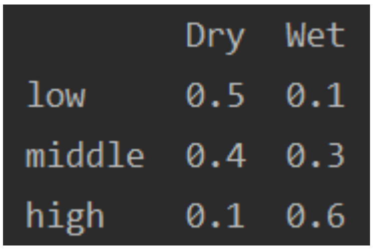


```
         toDry  toWet
fromDry    0.6    0.4
fromWet    0.8    0.2
```

Emission probability : A matrix with the probabilities of an observation (output) being generated from a state. P(Observation|currentState)

Let's say that we have assumed that the year is dry if there's a 50% chance that the number of rainfall days(>50MM) that year is between 0-10 days, a 40% chance that it is between 11-15 days, and a 10% chance that it is more than 15 days. Similarly, we have assumed that if the year is wet there's a 10%

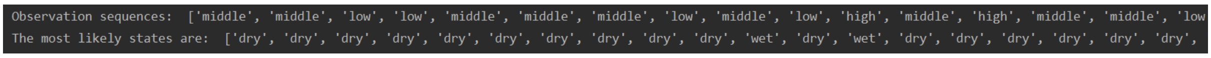
chance that the number of rainfall days(>50MM) that year is between 0-10 days, a 30% chance that it is between 11-15 days, and a 60% chance that it is more than 15 days.



```
        Dry  Wet
low     0.5  0.1
middle  0.4  0.3
high    0.1  0.6
```

Final probability: A final probability distribution over states

After running the Viterbi, we can calculate the viterbi path, which is the most likely sequence of states that generated the sequence given the full model



```
Observation sequences:  ['middle', 'middle', 'low', 'low', 'middle', 'middle', 'middle', 'low', 'middle', 'low', 'high', 'middle', 'high', 'middle', 'middle', 'low
The most likely states are:  ['dry', 'dry', 'dry', 'dry', 'dry', 'dry', 'dry', 'dry', 'dry', 'dry', 'wet', 'dry', 'wet', 'dry', 'dry', 'dry', 'dry', 'dry', 'dry',
```

8. **Apply your model to a dataset and do the followings (You may also do it on a simulated dataset, but less meaningful): (a) Predictions (b) Model Selection (c) Residual diagnostics / Outlier detection**

Predictions

Predictions can be performed based on the current dataset regarding future states as well as a probability distribution of the parameter of interest for a certain number of years later.

State prediction

In state prediction, $\Pr(C_{T+h} = i \mid X^{(T)} = x^{(T)})$ or probabilities that given the current rainfall data, the corresponding probabilities that the future underlying state falls in one of three, is computed for a range of values $h \in N$. For the purposes of this prediction we have used h = 50 but only show the first and last columns. Assuming a stationary model, 50 years or iterations later the state that the weather is in will increase for state 3. Assuming a nonstationary model, the probability will increase slightly for state 2. In both instances, the majority of the predictions fall in state 1 or 2, but the proportion that corresponds to state 3 is much more significant when fitting a stationary model. There are notable differences in probabilities depending on whether a stationary or nonstationary model is fitted, which should be carefully chosen when employing the model in practice.

Fit a stationary model:

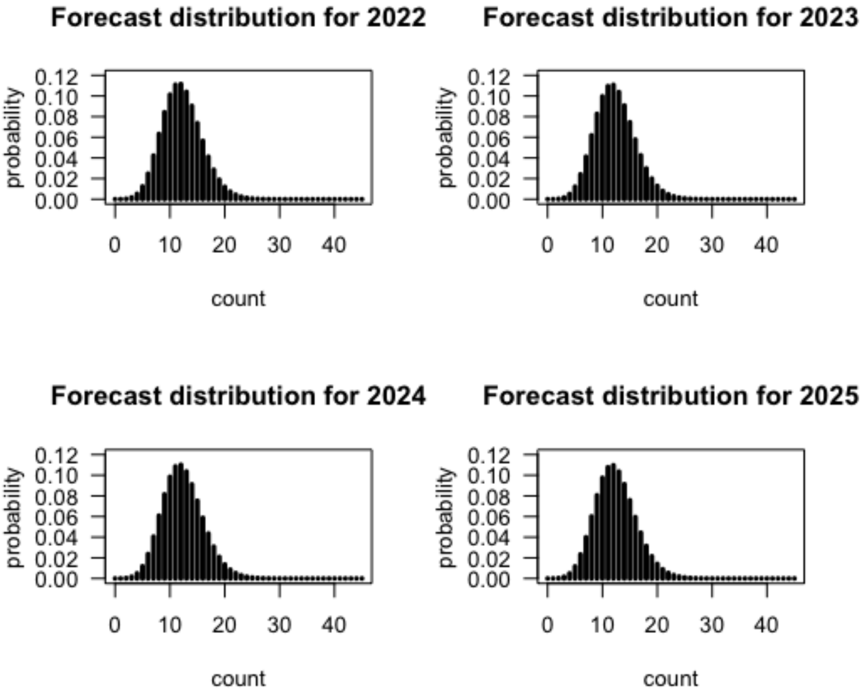
	h = 1	…	h = 50
State 1	0.4906045	…	0.4272820
State 2	0.4154271	…	0.3616145
State 3	0.0939684	…	0.2111036

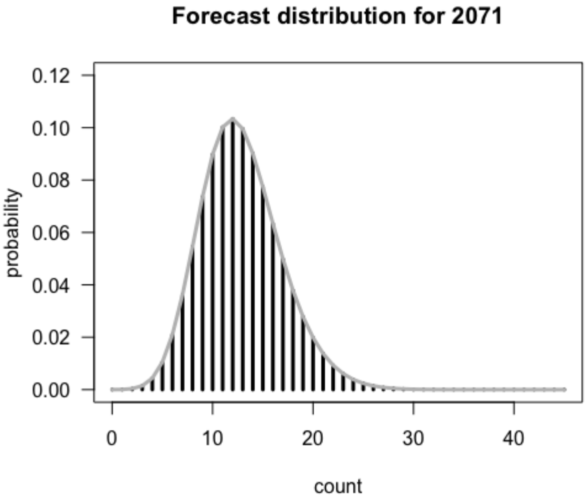
Fit a nonstationary model:

	h = 1	…	h = 50
State 1	0.8176067	…	0.7761904
State 2	0.1821655	…	0.2237968
State 3	0.0002277	…	0.0000127

Forecast probabilities

Another type of prediction that was performed on the dataset was probability distribution forecasting. Given current rainfall amount, the future amount h years later was forecasted, $\Pr(X_{T+h} = x \mid X^{(T)} = x^{(T)})$, according to a nonstationary model. Accordingly, the corresponding results are displayed for the next four years, as well as 50 years later. It is clear that the distribution is right skewed due to the assumption of Poisson distribution and takes a mean of approximately 12, approximating towards state 2, a moderate rainfall amount or wet state.





Model Selection

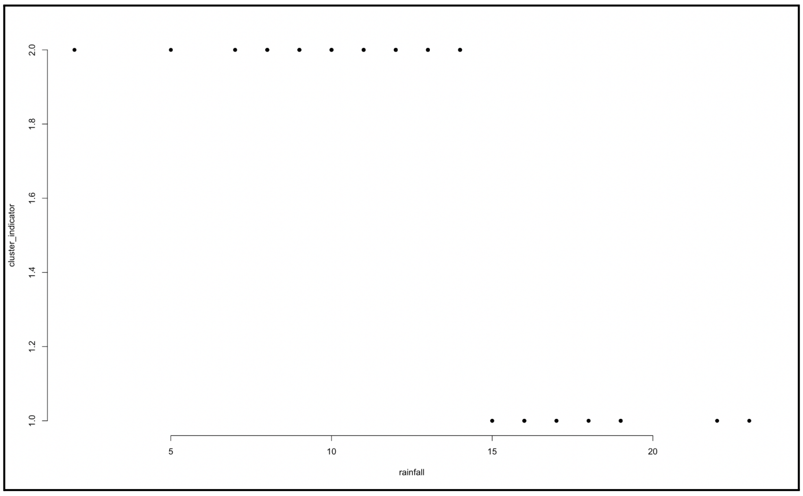
The estimates:

Parameter	Two-state Poisson HMM	Three-state Poisson HMM
Iteration	24	30
Log-likelihood	-207.265	-204.708
λ	(11.777, 15.091)	(2.037, 12.780, 7.102)
δ	(0.764, 0.236)	(0.0138, 0.9724, 0.01378)

Transition Matrix (Γ)	$\begin{bmatrix} 0.882 & 0.118 \\ 0.382 & 0.618 \end{bmatrix}$	$\begin{bmatrix} 0.000 & 1.000 & 0.000 \\ 0.000 & 0.986 & 0.014 \\ 1.000 & 0.000 & 0.000 \end{bmatrix}$

EM in Poisson-HMMs

Parameter	Two-state Poisson HMM
Iteration	50
Log-likelihood	-207.437
δ	(0.873, 0.127)
State	State 1 ~ Poisson(12.018) State 2 ~ Poisson(16.283)



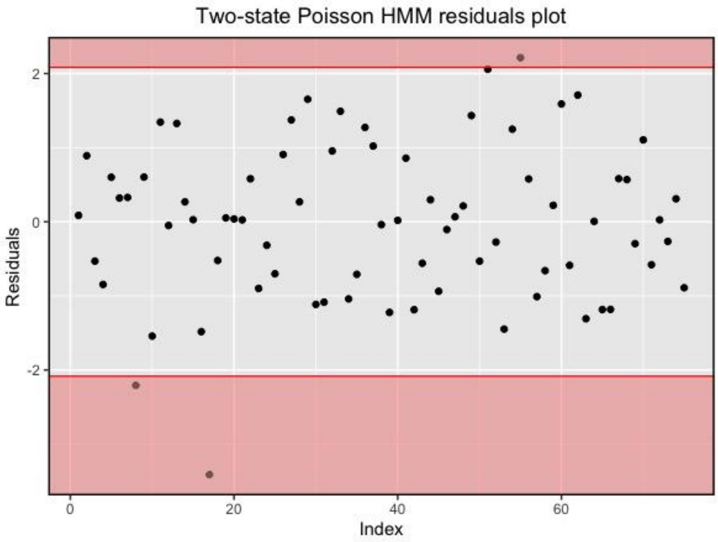
Residual diagnostics / Outlier detection

(non-stationary)	Two-state Poisson HMM	Three-state Poisson HMM
RSS	80.390	72.082

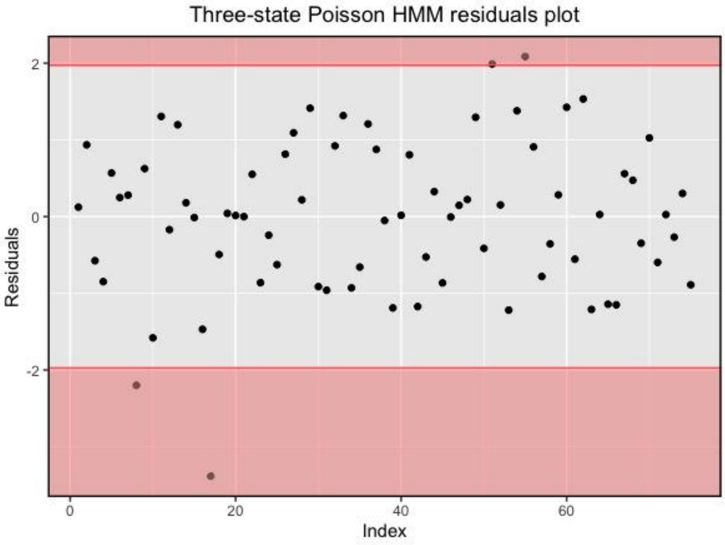
Mean of residuals (u)	-0.002559944	0.003220353
Standard deviations of residuals (s)	1.086	0.974
No. of outliers	3	4

By definition, residual > 2s is the outlier. We add an outlier boundary to the residuals plot for better detection. The data point that lies in the red region is the outlier.

Two-state Poisson HMM residuals plot



Three-state Poisson HMM residuals plot



Comparing the above residuals plot, we can observe that three-state has more outliers than two-state. Although the former has fewer outliers (i.e.contains more inappropriate information when fitting the Poisson-HMM model), it has a larger RSS (80.39>72.082). In short, the two-state model fits the data poorer than the 3-state in regard to RSS.

9. Conclusion - Leo

A stochastic process that satisfies probabilistic properties, assuming that there are properties of unobservable events, and that the properties of future events depend on the properties of present and past events, is called a Hidden Markov Model (HMM). In the project, the Poisson Hidden Markov Model which includes a Poisson process and a discrete Markov process is used to represent counts based time series data and to simulate the data set with length 100. Theoretically, the 2-state model would be a better approach, however, in practice, the 3-state model would be better due to less outliers and smaller RSS. As for the decoding problem on HMM using the Viterbi algorithm, it can be used to predict the weather condition. The results of probability rainfall data accessibility can be available. For potentiality, the results of probability rainfall accessibility in agriculture are to help provide information about extreme climate change and early warning to the communities about drought proofing. In addition, this information can also help in disaster mitigation as the basis for determining the policy to be taken.

(14 Pages excluding the cover page)

10. Reference - Leo

Climate prediction center - warm episodes. (n.d.). Retrieved December 4, 2022, from https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/impacts/warm_impacts.shtml

Comprehensive R Archive Network (CRAN). (n.d.). *Package depmixs4*. CRAN. Retrieved December 4, 2022, from https://cran.r-project.org/web/packages/depmixS4/

The EM algorithm for Poisson data - universiteit leiden. (n.d.). Retrieved December 4, 2022, from https://www.math.leidenuniv.nl/~vangaans/BASB2014/schmidthieber-EMalgorithm.pdf

The EM-algorithm for Poisson data - universiteitleiden.nl. (n.d.). Retrieved December 4, 2022, from https://www.universiteitleiden.nl/binaries/content/assets/science/mi/scripties/bsc-vanoosten.pdf

Ramadhan, R., & Devianto, D. (n.d.). *A hidden markov model for forecasting rainfall data availability at the weather station in West Sumatra*. Science and Technology Indonesia. Retrieved December 4, 2022, from https://sciencetechindonesia.com/index.php/jsti/article/view/223

https://web.stanford.edu/~jurafsky/slp3/A.pdf

ROLAND., Z. U. C. C. H. I. N. I. W. A. L. T. E. R. M. A. C. D. O. N. A. L. D. I. A. I. N. L. L. A. N. G. R. O. C. K. (2021). *Hidden markov models for time series: An introduction using R, second edition*. ROUTLEDGE.

Speech and Language Processing (3rd ed. draft) Dan Jurafsky and James H. Martin. Speech and Language Processing. (n.d.). Retrieved December 5, 2022, from https://web.stanford.edu/~jurafsky/slp3/