William Choi-Kim
and
Juhi Bhatt

Cycles: Autoregressive Algorithms and Analyzing Monotonicity in Body Pattern Data Through Mobile Applications

Introduction

Often medical discrepancies and problems arise because of a complex network of factors that don't seem related. It's difficult for patients to identify what they should and should not disclose to their medical professional – and it's hard for a medical professional to determine potential root causes of many complex issues. Mobile applications are the ideal platform to bridge this gap – everyone carries a phone, and many already log their lives on social media and in personal journals.

We ourselves have seen this issue – William has struggled with intense migraines for almost a decade, and Juhi has long sought an answer to her struggle with irregular periods.

As children of the digital age, we **hypothesize** that mobile applications can be used to accurately determine relationships between periodic bodily patterns more effectively than commonly used medical techniques like in-office questionnaires. **Cycles** is our attempt at such a mobile application.

Research Question
Are the programs created for **Cycles** more effective at finding trends in and making predictions about bodily cycles than existing methods like **Kendall's Rank Correlation Coefficient**?

Project Objectives

From the association of caffeine and period cramps to Coca Cola and cancer, the engineering goal of **Cycles** is to use data analysis and machine learning to identify trends and enhance our understanding of the interplay between everyday phenomena and our bodily responses, enabling us to make more accurate and individualized conclusions.

We tackled this objective by:

- **Creating a novel algorithm to quantify positive and negative trends in data**
- **Applying a multivariate autoregressive model to predict bodily cycles**
- **Building an intuitive mobile application to distribute these tools to the end user**

References

Brannick, M. (n.d.). Regression with two independent variables. Regression with two independent variables by Michael Brannick. http://faculty.cas.usf.edu/mbrannick/regression/Part3/Reg2.html

Fehring, R. (2012). Menstrual Cycle Data. Randomized Comparison of Two Internet-Supported Methods of Natural Family Planning. https://parker.ad.siu.edu/Olive/mch12.pdf

Probert, C. J., Emmett, P. M., & Heaton, K. W. (1993). Intestinal transit time in the population calculated from self made observations of defecation. Journal of epidemiology and community health, 47(4), 331–333. https://doi.org/10.1136/jech.47.4.331

Uyan?k, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. In Procedia - Social and Behavioral Sciences (Vol. 106, pp. 234–240). Elsevier BV. https://doi.org/10.1016/j.sbspro.2013.12.027

HeinOnline. (2025, January 13). About - HeinOnline. https://heinonline.org/HOL/LandingPage?handle=hein.journals/ucinlr48&div=51&id=&page=

Kuo, C. (2024). Modern Time Series Forecasting Techniques for Predictive Analytics and Anomaly Detection: From Classical Foundations to Cutting-edge. (n.p.): (n.p.).

Engineering Methodology

Identifying Trends

Our first step was defining a new way to measure the "trendiness" of data. We wanted to create an expression that involves **both** correlation and monotonicity, so we kept that in mind while deriving our algorithm.
To identify trends, **Cycles** first min-max scales all N data points to between 0 and 1.
Then, let $(x_i, y_i) = \left(\frac{i}{N}, \frac{j}{N}\right)$. By transforming each data point to some $p_k = (x_i, y_j)$, where the i-th smallest x in the set of all x in the dataset becomes x_i and analogously for y, then any monotonically increasing function becomes $y = x$ and any monotonically decreasing function becomes $y = -x$.

Let $u = \frac{\sum_{k=1}^{N}(p_{k_y} - p_{k_x})^2}{N}$; $d = \frac{\sum_{k=1}^{N}(p_{k_y} - 1 + p_{k_x})^2}{N}$.

Let $M_u =$ the number of increasing coordinate pairs and $M_d =$ the number of decreasing coordinate pairs.

Then the resulting value of $\left|\frac{u-d}{u+d}\right| \times \left|\frac{M_u - M_d}{M_u + M_d}\right|$ is compared to a threshold of $0.2^2 = 0.04$.

Generating Test Data

To generate test data for the trend-finding model, we referenced a study on the relationship between intestinal travel time and defecation frequency (DF), interdefecatory time interval (IDTI), and stool form scores (SFS) (Probert). The provided formula was:

$$intestinal\ transit\ time = 103 - 1.23(DF) - 4.69(SFS) + 0.638(IDTI)$$

with a correlation coefficient of 0.736.

By generating (DF, SFS, IDTI, intestinal transit time) quadruplets, then adding Gaussian noise derived from r = 0.736 to offset values from the perfect values given the expression, a dataset is generated that approximately fits the given expression and coefficient.

Predicting Future Incidences

Our next step was creating a model to **predict** future events. Most existing applications don't use unique models because they only track one type of cycle. To track many different types of cycles, we needed a model that was adaptable and could take into account many variables.

To make predictions, **Cycles** employs a **multivariate linear regression model** with a lag term.
In other words, the **Cycles** model takes time series data (a dataset of {time, value} pairs) and predicts not only the next time or value but both simultaneously.

Cycles uses two lag terms, so when it has insufficient data to build the lagged rows, it defaults to a standard linear regressive model.

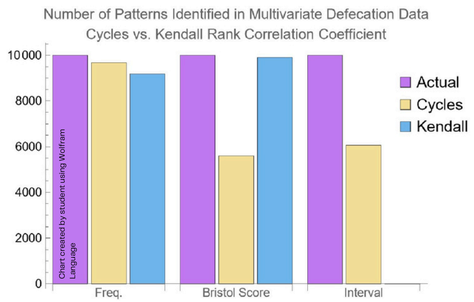
Sourcing Test Data

To test the prediction model, menstrual cycle data was sourced from an open public database. The database provided numerous attributes for each cycle, including Client IDs to group cycles by each subject. Discarding cycles of insufficient length (in our case, at least 4 cycles), the remaining 111 subjects then yielded cycle lengths and the attribute Total Menses Score.

By registering events in **Cycles**, we were able to run the predictive model on these subjects. By limiting the reference frame of data given, we collected predictions for cycles within the dataset, allowing us to make comparisons between predicted dates and menses scores and actual dates and menses scores.

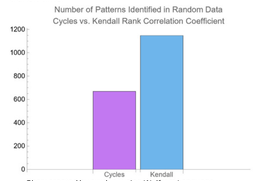
One goal when building **Cycles** was being able to make accurate predictions for irregular periods – typical applications and algorithms struggle with periodic cycles that don't conform to relatively stable distributions. The results of running our predictive model appear to indicate promising results past the first 8 menstrual cycles; that is, around three months of data is required before **Cycles** can make accurate, consistent predictions for those with irregular periods.

Results and Analysis

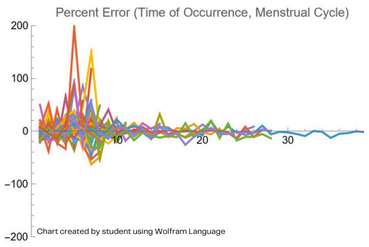


The data used in this test was generated from a multivariate linear distribution found in intestinal transit time. The three variables are the frequency of defecation over a week, the combined Bristol score of the last three defecations (a measure of the shape and hardness of a stool sample), and the time interval between the last two defecations.

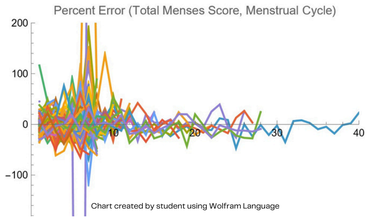
Cycles correctly identified more trends (from **10,000 trials**) when examining the relationships between frequency and intestinal transit time and interval and intestinal transit time, while the Kendall Rank Correlation Coefficient (tau) identified more trends (at a threshold of tau > 0.3) when examining the relationship between Bristol score and intestinal transit time.



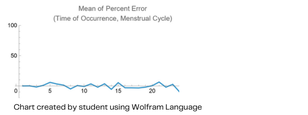
Cycles also found less trends in random generated data (from **10,000 trials**) than existing methods, indicating that it is less prone to identifying false trends or nonexistent patterns.

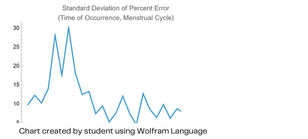


Given the date and time of **111** subjects' menstrual cycles with length ranging up to 40 cycles, the percent errors of the predictions made by **Cycles** is graphed above. Data was sourced from a public dataset collected from users of an app created by an external party. **Cycles** was fed partial datasets, and its predictions were compared to the actual values given by the full datasets. There are a few outliers between 0 and 10 data points, but prediction errors stabilize to acceptable values with further data.

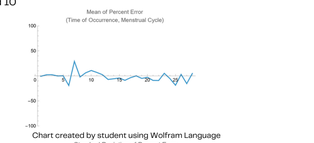


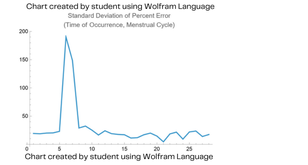
Graphed above are the same **111** cycles as previously used, but the predictions and data are in regard to the Total Menses Score (a quantification of menstrual blood loss) of each cycle rather than the time of occurrence. There is greater error overall, but the percent error does indeed stabilize as more data is collected. Once again, we see a few odd outliers between 0 and 10 data points, which we conjecture is a result of irregular menstrual cycles present in the data.





The mean error remains quite close to 0 at all quantities of data points, while we see that peak in standard deviation we attribute to irregular cycles. The standard deviation does stabilize around between 5 to 10 percent after that peak.





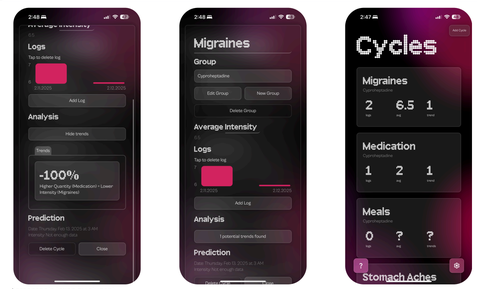
The mean error for predicting the Menses Score is more variable, as well as the standard deviation. This indicates that while **Cycles** is adept at predicting time in time series data, predicting values of bodily cycles poses a steeper challenge. Still, the standard deviation does stabilize after around 10 data points.

Mobile Application

Cycles is a set of tools and algorithms for an end user, so we designed a mobile application for this purpose.

The main features we focused on were:

- **Data Privacy**
  - All calculations, predictions, and trend finding is done **on-device**. This presented challenges with making our models and algorithms as lightweight as possible.
- **Intuitive Presentation**
  - With the app, we wanted to take very esoteric, low-level results from our algorithms and present them in an intuitive, comprehensible manner.
  - To do this, we used charts and design language common in apps, especially health trackers – bar charts, log books, and more.
  - We also implemented a tutorial to reduce barrier of entry.
- **Low Cost of Operation**
  - Because all computation is done on-device, the only costly architecture in our application is initially delivering the app to users. With a headless static website, this becomes much easier.



Screenshots taken by student

Conclusions

Based on our tests with data on **intestinal travel time** and **menstrual cycles**, we conclude that **Cycles** is **better at identifying positive and negative trends in data** than existing methods (Kendall Rank Correlation Coefficient) and **accurate in its predictive abilities for cyclical bodily patterns, even when handling irregular data**.

Future Work

To expand on our work in this project, we'd like to **improve our statistical models**, create **better systems for communicating results to medical professionals**, and provide **pre-trained models for specific bodily cycles**.

We're interested in if a **multiregressive quadratic or cubic model** might be more accurate than our linear one.
We'd also like to implement models pre-trained with data for **menstrual cycles**, **bowel movements**, and more.

Acknowledgements

Thank you to Dr. Patricia Chapela for being our wonderful school sponsor, for helping us with documentation, and for always being our sounding board.