

RICE UNIVERSITY

DATA TO KNOWLEDGE (D2K) LAB

DSCI 435: DATA SCIENCE PROJECTS

Investigating the Impact of Neighborhood Equity and Artificial Light on Children's Growth Trajectories: A Final Report

Authors

Zachre ANDREWS
Jacob COYLE
Caleb HUANG
Gail OUDEKERK
William PAN
Elijah SALES

Sponsor

Dr. Jennette MORENO
Baylor College of Medicine

D2K Fellow Mentor
Kevin MCCOY

Faculty Instructor

Dr. Arko BARMAN

Special Thanks

Dr. Salma MUSAAD
Baylor College of Medicine

July 30, 2024

Abstract

BACKGROUND While research has indicated patterns of seasonal trends in children's growth trajectories, their causes are not yet clear. In particular, the relationships between trends in children's body mass index (BMI) and their neighborhood-level physical and social factors are understudied.

OBJECTIVES To further investigate patterns and groupings in children's growth; to evaluate the relationship between seasonal growth trajectories (in BMI and obesity risk) and neighborhood-level environment measures (Child Opportunity Index (COI) and exposure to artificial light at night (ALAN)).

METHODS This analysis uses data from a longitudinal study of a cohort of elementary-school children, with twice-yearly measurements of weight and height from kindergarten to 5th grade. Child Opportunity Index and satellite light data are matched to each measurement point by each child's school's census tract. The analysis uses k-means time-series clustering to detect patterns in BMI over time. It also uses forecasting models (SARIMAX and Prophet) to test the roles of neighborhood-level factors and seasonality in individual growth trajectories, and employs mixed-effects modeling to assess those associations within the measured population more broadly.

RESULTS The k-means clustering method creates interpretable clusters, in line with previous group-based trajectory modeling methods (with approximately 5% deviation). Forecasting models perform relatively poorly on individual height and BMI trajectories, with neighborhood-level factors failing to improve model accuracy. Mixed-effects modeling results reveal that lower child opportunity and greater amounts of artificial light are weakly associated with higher BMI values as children age, and these neighborhood-level factors moderate seasonal patterns of growth.

CONCLUSIONS The k-means clustering results' similarity to the results of previous work speaks to the robustness of the BMI trajectory groups, at least within the study data [1]. Forecasting models may be unsuitable tools in revealing sample- and population-level patterns in these data. Mixed-effects model results show relationships between neighborhood-level factors and children's BMI, but the mechanisms of those relationships and of seasonal patterns bear further investigation.

Contents

1	Introduction	4
1.1	Domain Background	4
1.2	Project Background	5
1.3	Objectives	5
2	Data Sources	7
2.1	Student Data	7
2.2	Child Opportunity Index (COI) Data	7
2.3	Artificial Light at Night (ALAN) Data	8
3	Data Pre-Processing	8
3.1	Data Cleaning	8
3.2	Data Irregularities and Missing Data	10
3.3	Preparing for Validation	10
4	Data Exploration	12
5	Modeling	15
5.1	Unsupervised Model: K-Means Time Series Clustering	15
5.1.1	K-Means and K-Means Time Series	15
5.1.2	Hyperparameter Tuning	16
5.1.3	Results and Comparisons	17
5.1.4	Validation	18
5.1.5	Implications and Future Work	20
5.2	Supervised Model: Time Series Regression	20
5.2.1	Fitting Time Series Models	20
5.2.2	SARIMAX Model	21
5.2.3	Prophet Forecasting Model	22
5.2.4	Time Series Regression Results	22
5.3	Supervised Model: Linear Mixed-Effects Regression	25
5.3.1	Introduction to Mixed-Effects Models	26
5.3.2	Methodology	27

5.3.3	Model Validation and Stability Testing	28
5.3.4	Model Checking	30
5.3.5	Missing Data	31
5.3.6	Interpretation	32
6	Discussion	35
6.1	Conclusions	35
6.2	Project Limitations	36
6.3	Future Work	36
7	Appendix	38

1 Introduction

1.1 Domain Background

For more than three decades, childhood obesity has been a serious health concern in the United States [2]. In 2010, the prevalence of childhood obesity (defined as a child's Body Mass Index (BMI) falling at or above the 95th percentile) in the United States was approximately 17% [3]. By 2020, that figure had risen to 19.7%, impacting 14.7 million Americans under the age of 18 [4]. Childhood obesity is associated with a variety of comorbidities: these include cardiovascular conditions (such as hypertension, atherosclerosis, and vascular aging), metabolic diseases (such as type II diabetes and fatty liver disease), psychological and social states (depression, reduced cognitive function, and stigmatization), and others (such as cancer) [2, 5]. Moreover, these negative impacts can persist over time: obesity often follows individuals from childhood to adulthood, with the attendant increased risk for disease [5].

There are many complex factors that lead to childhood obesity. Some are pre- and immediately post-natal, including maternal diabetes, high infant birth weight, and lack of breastfeeding. Others are enduring characteristics (higher disease susceptibility among certain ethnic groups), or dependent on lifestyle or circumstance (diet, opportunities for physical exercise) [2]. Many of these factors are related to socioeconomic status. In fact, there has been an observed inverse relationship between childhood obesity and a caregiver's level of education, and the prevalence of childhood obesity in low- and middle-income families is almost twice that of high-income families in the U.S. [4].

One key set of factors associated with childhood obesity occur on the neighborhood level. A child's environment, especially in early childhood, has been shown to have a profound effect on long-term outcomes, including socioeconomic mobility and health [6–9]. One measure of physical and social environment is the Child Opportunity Index (COI), a Census-derived index that represents neighborhood-level factors that can foster healthy child development (including availability of early childhood education centers, walkability, and proximity to sources of healthy food) [10]. Children residing in areas with high COI have been found to have lower risk of obesity in childhood and adolescence and lower average BMI trajectories, independent of family prenatal and socioeconomic characteristics [11].

Another neighborhood-level factor is exposure to artificial light at night (ALAN). Artificial light has been theorized to disrupt children's circannual and circadian systems, leading to detrimental changes in sleep patterns, endocrinology, and metabolism that may contribute to obesity [12, 13]. Exposure to ALAN by area, duration, and intensity has been positively related to increases in children's BMI [14–16]. There is an additional seasonal aspect: a prospective study found children's later sleep midpoints during the summer (relative to the school year) predicted larger summer increases in BMI, through a process called the Circadian and Circannual Rhythms (CCR) Model [13, 17, 18]. Some researchers have hypothesized that children's summer behaviors may contribute to disrupted CCRs, and subsequent increases in BMI, through later bedtimes and thus greater exposure to artificial light at night [13].

Numerous studies have investigated how the social determinants of health, at the neighborhood level, relate to children's propensity for becoming overweight and obese [19–21]. Only a few studies have used Census-based proxy indices such as the COI metric or Social Vulnerability Index (SVI) [11, 22]. Similarly, although the effect has been investigated in adults, there is limited research evaluating the impact of exposure to ALAN on children's BMI, especially in the United States [14–16]. The current analysis seeks to synthesize these factors into a single, novel analysis, examining the relationship between neighborhood-level factors (COI and ALAN) and the well-studied growth trajectories of a group of elementary-school children.

1.2 Project Background

Beginning in 2015, researchers at the Baylor College of Medicine (BCM) studied growth patterns within a cohort of more than 7500 kindergarteners from Fort Bend Independent School District (ISD). The children were studied longitudinally across all 5 years of elementary school. Heights and weights were measured at the beginning and end of each school year, allowing Moreno et al. (2015) to examine patterns in children's BMI [23]. Contrary to expectations, it was observed that many children increase their body mass index during the summer at a faster rate than during the school year, contributing to their risk of becoming overweight and obese. Group-based trajectory modeling revealed that about 17% of children begin a trajectory toward an unhealthy weight status in elementary school, with summertime increases contributing substantially to these trends (Figure 1) [1].

These upward trends in BMI appear to be driven in part by a seasonal pattern in height gain that involves slower increases in height during summer, while weight gain remains consistent year-round (Figure 2). The causes of this seasonal variation in height and BMI are unknown, but may be related to neighborhood-level light exposure, as represented by ALAN measures [13].

The objectives of the current analysis are as follows:

1.3 Objectives

- Train unsupervised algorithms, e.g., time series clustering, to identify patterns and groupings in height and BMI gain over time for children.
- Build a model to evaluate impact of neighborhood child opportunity index on seasonal patterns in children's height and BMI gain and risk of developing overweight or obesity during elementary school.
- Examine the impact of exposure to light at night on seasonal patterns in children's height and BMI gain and risk of developing overweight or obesity during elementary school.

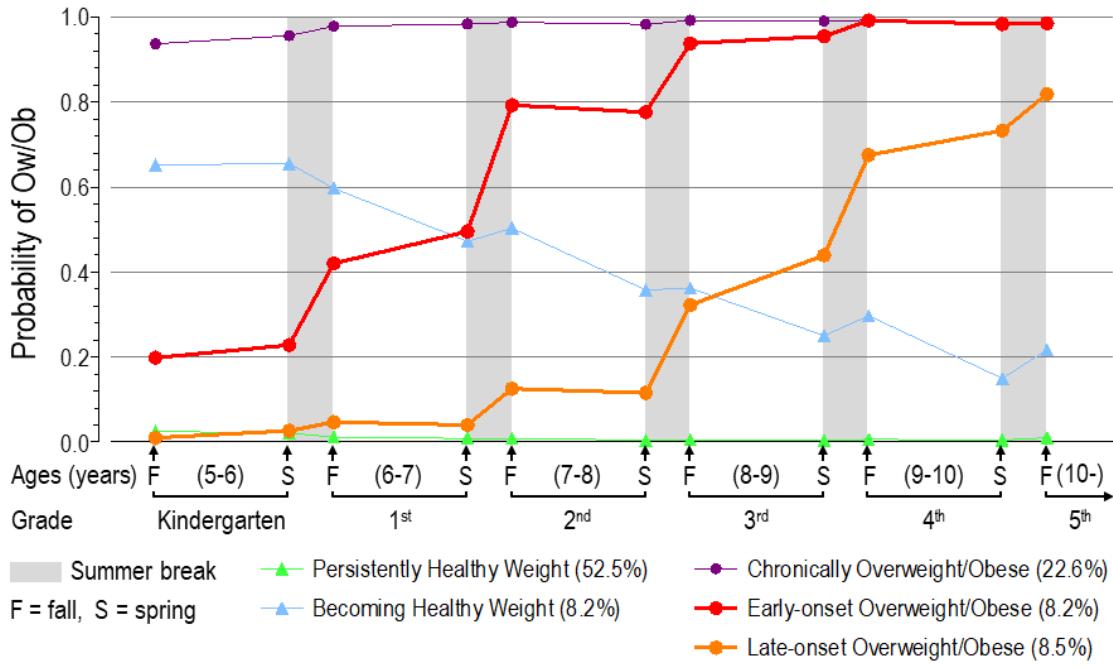


Figure 1: BMI trajectory groups' probability of being overweight/obese ($\text{BMI} \geq 85^{\text{th}} \text{ percentile}$) over the study period. Probability is calculated as the average BMI percentile of all children assigned to each group; observations were assigned to trajectory groups through the best-fitting semiparametric mixture, group-based trajectory model. [1, 23].

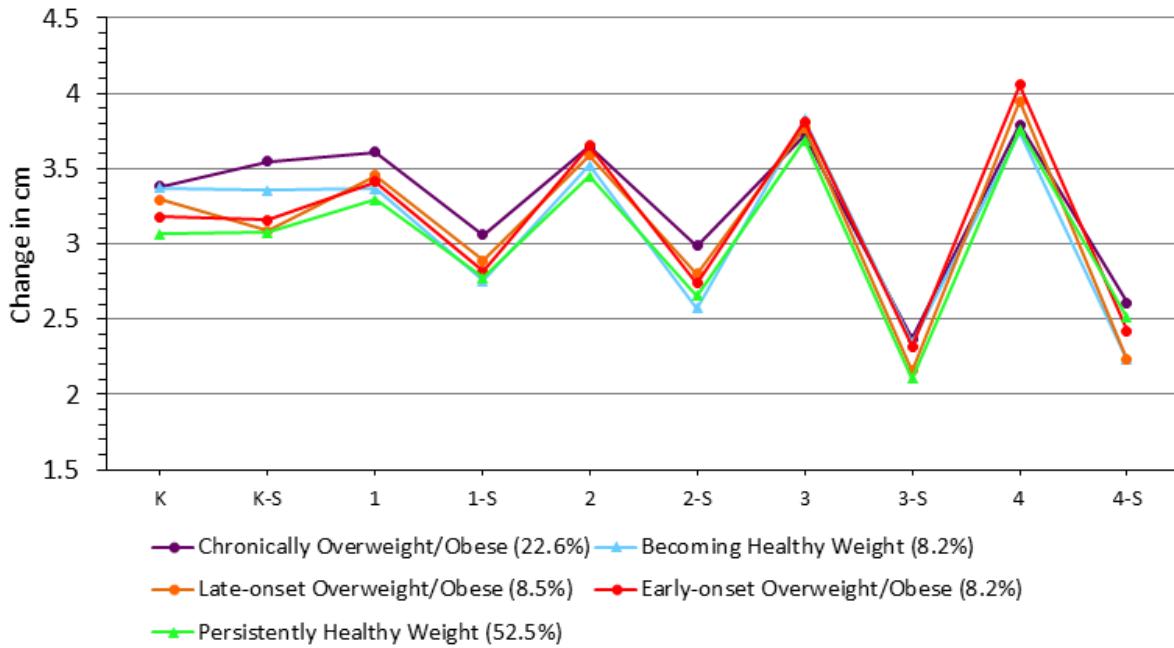


Figure 2: Seasonal variability in height gain across trajectory groups. Means adjusted for age (months), season, sex, and race/ethnicity (S=summer). [1].

2 Data Sources

2.1 Student Data

The data provided by the Baylor College of Medicine follow 7,648 kindergarteners from Fort Bend ISD from 2005 to 2010. Each participant is contained in a single row; there are 11 measurements for each variable (such as height and weight). These measurements were taken near the beginning and end of each school year, allowing researchers to examine patterns in children's BMI trajectories. Out of the 11 measurements, 6 were taken in the fall and 5 in the spring. Each child's school is recorded at each measurement point; we match the school label to census tract codes (in 2010) to tie in supplementary data sources. The neighborhood level factors in this project will also be summarized at the census tract level in order to match the spatial component of our student data.

Previous work has used these data to construct BMI trajectory groups in order to better understand children's growth patterns [1]. The group labels include "Persistently Healthy Weight," "Becoming Healthy Weight," "Chronically Overweight/Obese," "Early-onset Overweight/Obese," and "Late-onset Overweight/Obese" (see Figure 1). In our analysis, we compare these labels to those produced by our trajectory modeling.

2.2 Child Opportunity Index (COI) Data

The Child Opportunity Index (COI) is a Census-based metric which measures neighborhood resources and conditions that can impact a child's healthy development. For over 72,000 census tracts in the United States, 29 different indicators were measured. Each indicator is grouped into one of three domains: Education, Health and Environment, and Social and Economic [10]. These composite scores are standardized on a national level, with each census tract associated with a percentile rank. Previous research has split the numeric COI score into quintiles, ranging from "very low" ($< 20^{\text{th}}$ percentile) to "very high" ($\geq 80^{\text{th}}$ percentile) [11]. In Fort Bend ISD in the given time range, there were no census tracts in the "very low" category. The two-way frequency table of students in different BMI trajectory groups and COI categories is shown in Table 1.

The education domain measures a child's access to quality education. Some indicators include the number of early childhood education centers and their quality, teacher experience, high school graduation rate, third-grade reading proficiency, and more. The health and environment domain includes factors such as access to healthy food, access to parks and green spaces, and (lack of) exposure to extreme heat. The social and economic domain includes indicators such as median household income, poverty rate, employment rate, commute duration, and more [10]. Previous studies have concluded that living in neighborhoods with a higher COI score is associated with lower mean BMI trajectory and a lower risk of obesity throughout childhood[11].

	Child Opportunity Index Category			
	Low	Moderate	High	Very High
Persistent Healthy Weight	0.37	0.46	0.55	0.59
Becoming Healthy Weight	0.14	0.06	0.08	0.09
Late-Onset Overweight/ Obese	0.05	0.11	0.08	0.09
Early-Onset Overweight/Obese	0.06	0.11	0.08	0.06
Chronically Overweight/ Obese	0.38	0.26	0.21	0.17

Table 1: Proportion of students (with complete case data) in each growth trajectory by COI quintile, excluding “Very Low” (not represented). Note growth trajectory labels are those assigned by Moreno et al. (2022) [1]

2.3 Artificial Light at Night (ALAN) Data

To measure children’s exposure to artificial light at night (ALAN), we leveraged data made available by the U.S. Air Force Defense Meteorological Satellite Program (DMSP), which collects day- and nighttime images of Earth’s surface [24]. The DMSP has been operating since 1972, with data starting from 1992 available online; therefore, the satellite has image data within the project date range (the years the children’s measurements were taken) [24]. The images are 30-arc-second grids, which are screened to remove data contaminated by sun- and moonlight, as well as cloud cover[24]. The satellite program measures lights from cities, towns, and other sites with persistent lighting [24].

We utilized ArcGIS to transform these images into quantitative data for analysis. ArcGIS is a family of software that employs a geodatabase, which contains spatial datasets and connects them using certain attributes, allowing us to convert images to “flat” numeric data. The resulting dataset contains every census tract within Fort Bend County, a variety of summary statistics regarding the average visible light band, and the date of collection. With ALAN data at the census level, we are able to use these data to understand the relationship between nighttime light and BMI gain among children in Fort Bend County.

3 Data Pre-Processing

3.1 Data Cleaning

We began by cleaning our main dataset, the students’ longitudinal growth trajectory data. Cleaning these data included selecting relevant columns, standardizing column names, and doing unit conversions. Because of extreme data irregularity for health-condition variables like asthma and type II diabetes, we excluded all health condition variables from our analysis.

The next data wrangling step involved calculating nationally standardized values for children’s mea-

surements. For this step, using CDC-provided methodology, we calculated z-scores (as well as percentiles, both being common in existing work on the subject) for children’s BMI, weight, and height [1, 23]. These values are based on the 2000 CDC growth charts for the U.S., with extended metrics for children with obesity [25]. Within the CDC documentation, there are also techniques to flag extreme values for age-specific height, weight, and BMI [26]. We reconstructed the latter functions in order to identify biologically implausible value (BIV) flags for observations that may be errors [26]. We used the most recent z-score cut points to construct these flags (see Table 2). Ultimately, based on the flag values, we excluded 45 additional observations.

		Cutoff for Extreme Z-Scores (several cut points changed in 2016)	
Description	Low	High	
Weight-for-age for children aged from 0 to < 240 months	< -5	> 8	
Height-for-age for children aged from 0 to < 240 months	< -5	> 4	
BMI-for-age for children aged 24 to < 240 months	< -4	> 8	

Table 2: Z-score cut points for childhood weight, height, and BMI, excerpted from CDC documentation [26].

For outliers that do not meet the CDC thresholds, we plan to evaluate the resilience of our methods and models to outliers and leverage points, and further discuss criteria for exclusion.

An additional data wrangling task was merging our data: we connected COI and ALAN data to individual students by the census tract code related to their school, since residential data are not available. The first task was collecting school addresses and their associated census tract codes. We then joined the tract codes into the data as a time-series variable, with 11 values per student, associated with each measurement point at which the child’s school was recorded. Next, we selected the relevant fields from our COI and ALAN data and merged them into our dataset, again as time series variables. With single-observation data, like COI, it was appropriate to match solely on census tract; with a dataset that is itself longitudinal, such as the satellite data, it became necessary to match on year as well. We also matched Moreno et al.’s (2022) trajectory group labels to the student data by student identification number [1].

We set standard assumptions for assigning each student to a single school when, for example, we wish to create a single spatial visualization rather than a series of visualizations over time. Children changing schools are, in the scope of our analysis, also changing census tracts. To account for that multiple-membership structure, we retrieved the school mode (the most-attended school among the 11 measurement points per student) [23]. If a student had two or more school modes, we selected the first mode, due to the greater relative importance of neighborhood factors in early childhood [7, 9]. However, in most cases, we employ longitudinal methods that do not require summary approaches, as these approaches introduce

information loss.

3.2 Data Irregularities and Missing Data

We then checked for duplicates, finding no duplicated student ID values. We checked for negative changes in height, ignoring missing values. We observed negative height changes in 1.1% of original cases. For cases with negative changes no greater than 1 inch in magnitude, we assumed a plausible slight measurement error and opted for inclusion, to maximize our usable data; for cases with negative changes greater than 1 inch in magnitude, we considered the error disruptive and excluded all records for that student. We excluded all observations with missing data for ethnicity; we excluded Indigenous American children from the sample due to low representation, as was done in previous work with these data [23]. This work with the ethnicity variable resulted in the removal of 117 observations.

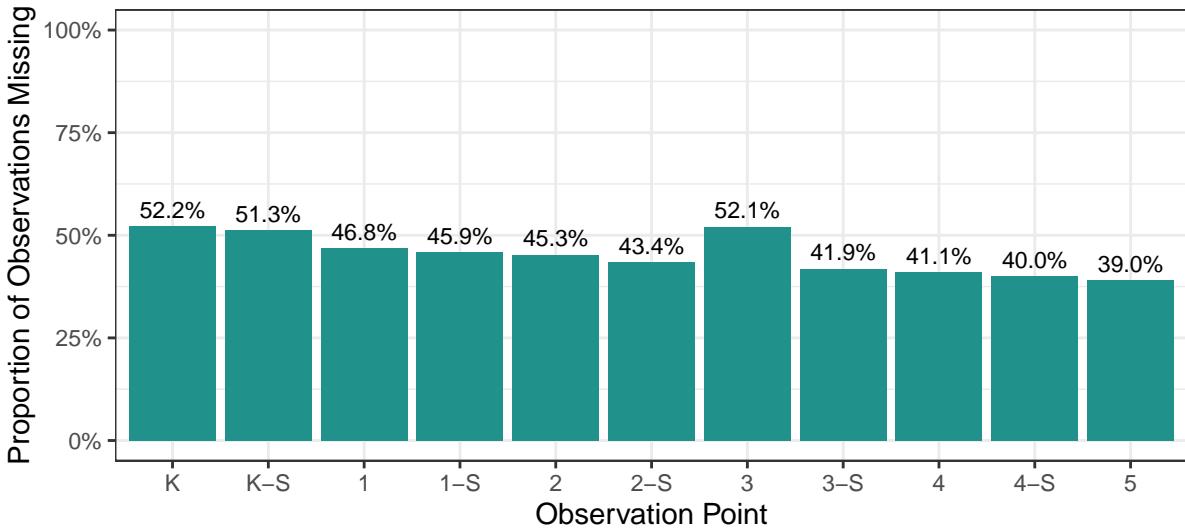
We must confront the issue of missing data more broadly, as well. The student data, after cleaning, contained only 1,642 complete records (of 7,759 total observations). Past work with these data have taken alternate approaches: some have used only complete records [23, 27], while others have leveraged the full data [1]. While our approaches may be resilient to some kinds of missingness (for example, one height value out of 11 missing), they may be vulnerable to others (all values for a child’s school code being missing, which leaves us unable to leverage our secondary datasets).

Patterns of missing values indicate that excluding missing observations may introduce complete-case bias (Figure 3). In the baseline data, significant differences between students with complete and with missing data were found: higher percentages of white, Hispanic, and Asian children have complete baseline data than Black children; more children with healthy weight status have complete baseline data than overweight and obese children; and a higher percentage of students in Title I schools have complete baseline data than those in other schools [23]. While exclusions of missing data are necessary for some methods, we retained as much data as possible, especially for the many observations which are missing some but not all elements of a time series.

3.3 Preparing for Validation

Before we begin working with the data, we must consider what data we will use to validate and test the efficacy and stability of our methods. While one simple standard procedure is splitting the data randomly into training and test sets, that is a fraught method in this context. Data with spatial or hierarchical aspects often feature autocorrelation based on internal dependence structures, where observations are more similar to those with whom they share group membership or a close spatial relationship than to those who are more “distant” [28, 29]. These dependence structures can lead to undesirable outcomes: these include the introduction of dependence structures into model residuals, which violates many models’ assumptions of independent samples, and predictive models overfitting to residual dependence structures through non-

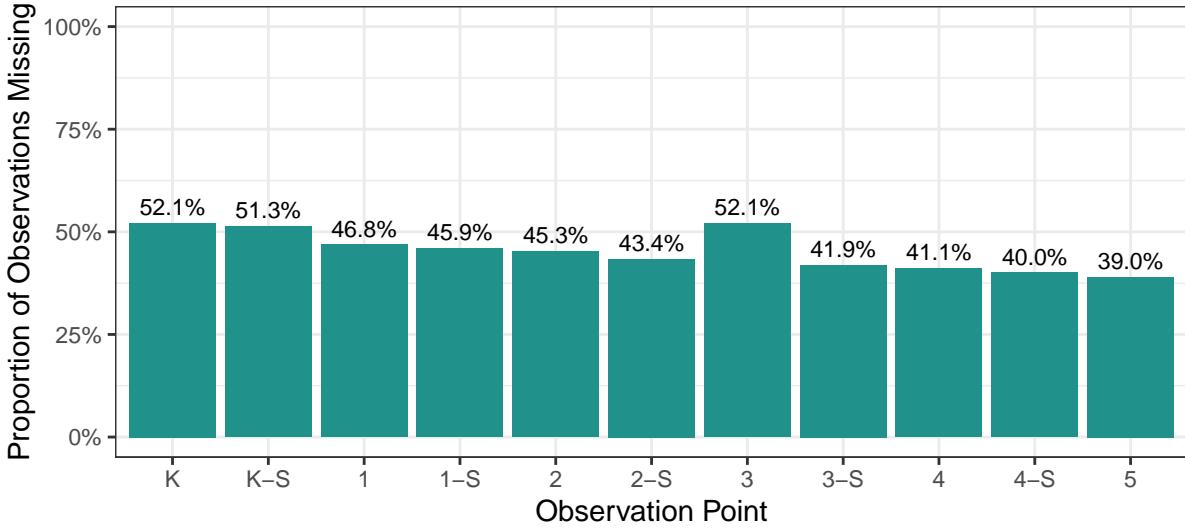
Missing data for height (in.) variable



49% of kids are missing at least 50% of observations for the height (in.) variable; 78.7% are missing at least one.

(a)

Missing data for weight (lbs) variable



49% of kids are missing at least 50% of observations for the weight (lbs) variable; 78.7% are missing at least one.

(b)

Figure 3: Missing data for student health attributes across measurement points

causal predictors [29, 30]. Because the student data feature internal dependence structures, a single random train-test split is clearly not appropriate, for reasons of data leakage between sets and for the above reasons.

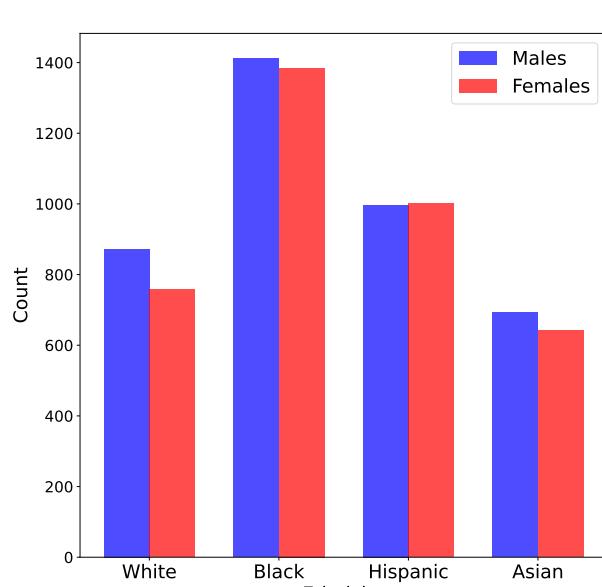
For purposes of model validation and encouraging stability, we instead turn to block cross-validation, where data are split deliberately (according to internal dependence structures) rather than randomly [29]. Strategic data blocking can resolve many of the issues above by accounting for data dependencies; however,

it can also introduce extrapolation through restricting the predictor space. In order to address concerns of extrapolation, it is recommended to use blocks as small as possible given the data structure, use the maximal amount of data for training, and have equal representation of predictors across validation folds [29]. In our case, the most logical blocks are the individual schools: they account for both intra-group dependence (as students are grouped together in a common school environment) and spatial dependence (as students are typically zoned to their closest school; this is not a perfect representation, but in the absence of student home addresses, it is the closest measure to which we have access).

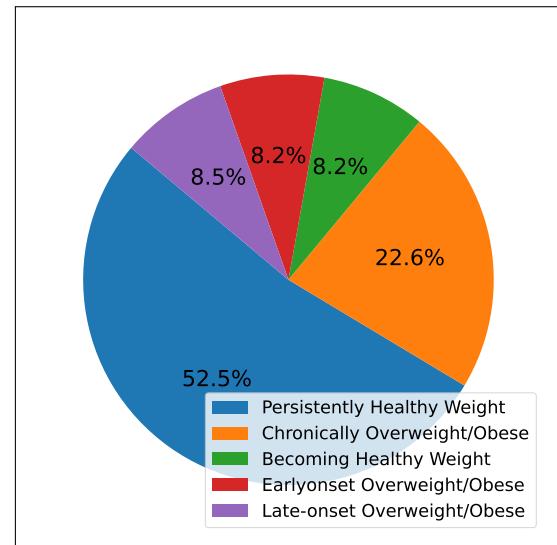
In accordance with the guidelines, we selected a k-fold cross-validation approach with block folds, where one school is held back as a validation set per iteration. The approach can also be conceptualized as a leave-one-out-cross-validation method (LOOCV), where the units are school-wide blocks rather than individual observations. This approach satisfies the minimizing-extrapolation requirements of using small blocks and maximal training data. It fails to satisfy the requirement of equal predictor representation across blocks and folds; the literature suggests manually distributing blocks with similar predictor values across folds as an alternative [29].

4 Data Exploration

In exploring the data, we have two main goals: descriptively analyze the data (including through visualizations) and gain a better understanding of the data for the purposes of predictive analysis.



(a) Ethnicity and gender



(b) BMI trajectory group membership (based on trajectory groups featured in Moreno et al. (2022)) [1]

Figure 4: Fort Bend ISD student data: distributions across sample

We first examined the demographic distribution of the students, as depicted in Figure 4a. Given the diverse range of ethnicity groups, we consider the inclusion of these groups, along with gender, as control variables in our models. Additionally, we have the distribution of the five trajectory groups constructed by Moreno et al. (2022), as illustrated in Figure 4b [1].

For many purposes, and for illustration in Figure 5, we utilized students' BMI z-scores. Recall that these z-scores are calculated not relative to our sample data, but to the CDC estimates for the national population [26]. Figure 5 illustrates the plotted BMI z-scores over time for selected students, providing a snapshot of children's developmental trends in the current sample.

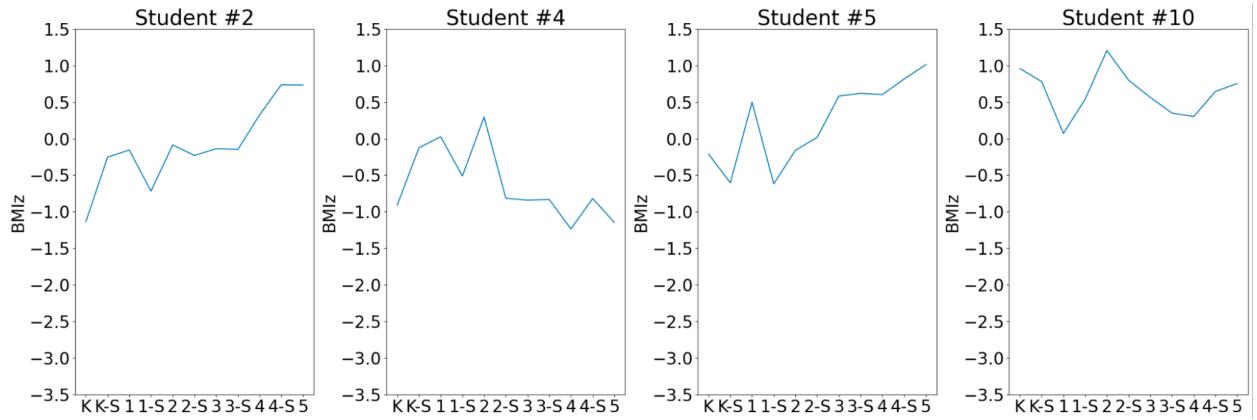
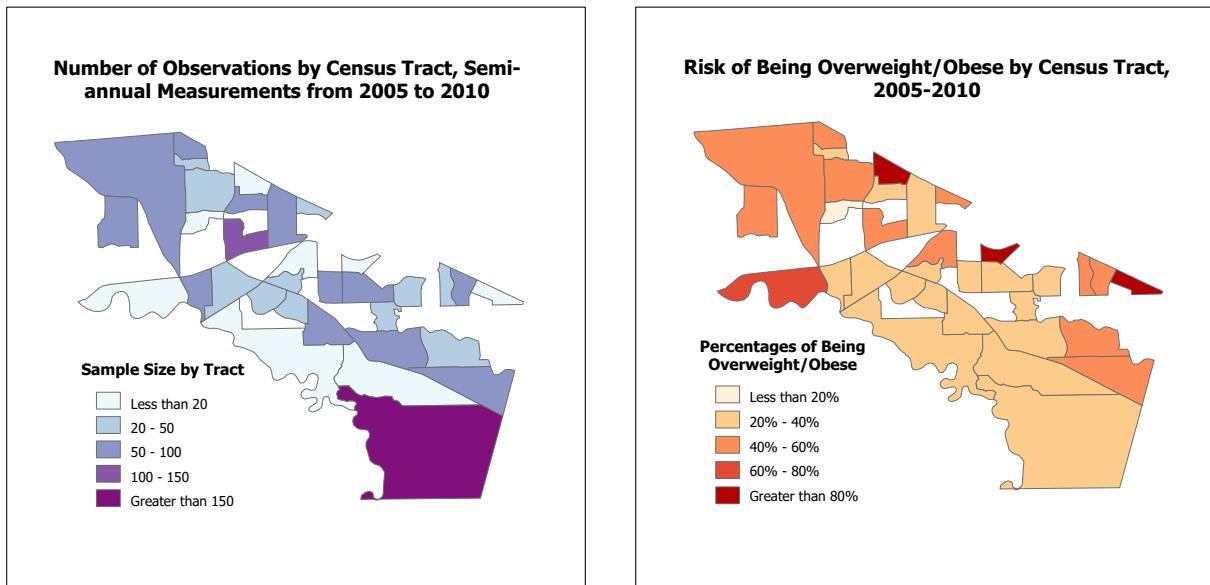


Figure 5: Standardized Body Mass Index (BMIz) trajectories of sample students

Visualizations such as a series of choropleth maps (where colors or hatching on different areas represent values on a cardinal or ordinal scale) can help diverse audiences gain a quick and intuitive sense of spatial data, such as the neighborhood patterns of this analysis and spatio-temporal trends in health factors like BMI [31].

Figure 6a maps the number of observations in the student data to each census tract. Note that we only use complete case data for these visualizations; the next step is recreating them with all available data. Evident through the figure, the majority of complete observations come from census tract 6745.06. Figure 6b visualizes the percentages of those overweight across the census tracts in Fort Bend County. When looking at the tracts, there appear to be three tracts that belong to the "greater than 80%" category.

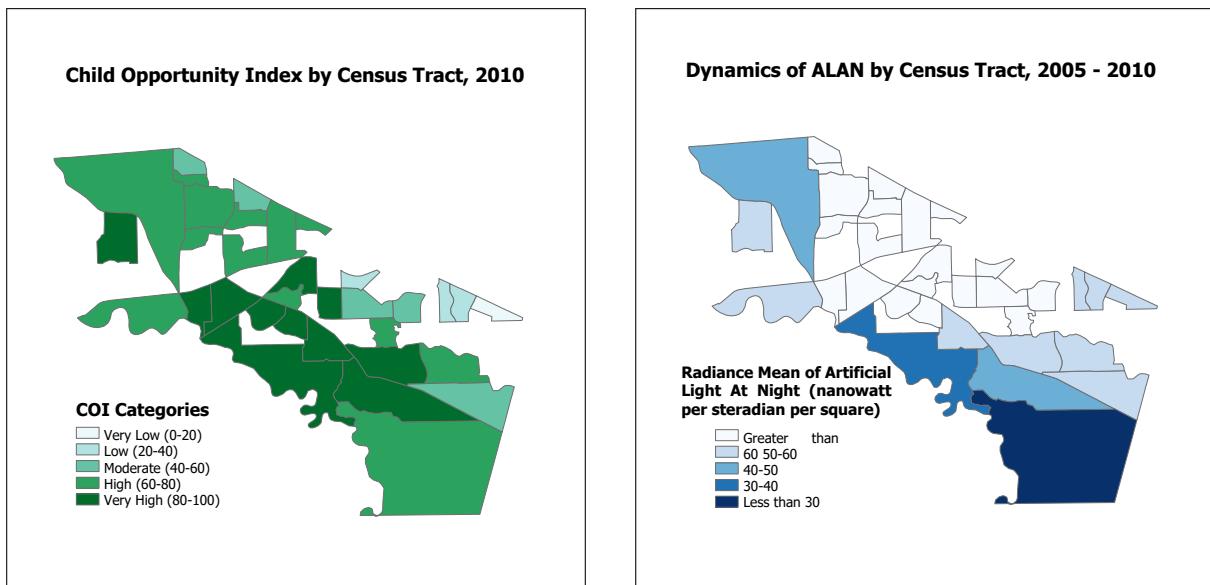
In Figure 7a, scores for COI are mapped to each census tract in Fort Bend County in 2010. There is a spread of values ranging from "less than 10" to "greater than 75" across the entire county. For COI and ALAN, these data are at the census level and therefore cover the entire county. Figure 7b maps radiance values for ALAN, and shows a clear regional difference. Northeast census tracts in Fort Bend County have higher exposure to ALAN whereas tracts to the south, west, and southwest are more variable.



(a) Observations per census tract across Fort Bend County

(b) Children's obesity risk across Fort Bend County

Figure 6: Spatial data analysis: exploring the landscape of student data (pictured: static values in fall 2010). Limited to census tracts with available school data.



(a) Childhood Opportunity Index values across census tracts, Fort Bend County (2010).

(b) Variability in average scores for artificial light at night across census tracts, Fort Bend County (2010).

Figure 7: COI and ALAN data visualized across census tracts in Fort Bend County.

5 Modeling

5.1 Unsupervised Model: K-Means Time Series Clustering

To better understand student growth trajectories and previous work with the dataset, our first approach is an unsupervised learning model that utilizes new methods to group students' health trajectories across the five-year span. Note that due to model limitations, analysis is limited to students with complete case data.

5.1.1 K-Means and K-Means Time Series

K-means is a prominent and widely-used partition-based clustering algorithm, common to both traditional clustering tasks and the specialized domain of time series clustering [32]. This algorithm unfolds through two primary stages. The first is an initialization stage, where the algorithm selects a set of representative cases to embody the characteristics of each cluster. The second is an update stage, where the example cases are iteratively refined until a predefined convergence criterion is met [32]. Formally, given a data set $X = \{x_1, \dots, x_n\}$ in a d -dimensional Euclidean space \mathbb{R}^d and c cluster centers $A = \{a_1, \dots, a_c\}$, k-means assigns each data point to a cluster using a matrix $z = [z_{ik}]_{n \times c}$, where z_{ik} is a binary variable (i.e. $z_{ik} \in \{0, 1\}$) indicating whether the data point x_i belongs to the k -th cluster, $k = 1, \dots, c$ [33]. The k-means objective function [33] is as follows:

$$J(z, A) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} \|x_i - a_k\|^2 \quad (1)$$

The algorithm iterates to minimize the objective function $J(z, A)$ by updating cluster centers and memberships using the following equations:

$$a_k = \frac{\sum_{i=1}^n z_{ik} x_i}{\sum_{i=1}^n z_{ik}} \text{ and} \quad (2)$$

$$z_{ik} = \begin{cases} 1 & \text{if } \|x_i - a_k\|^2 = \min_{1 \leq k \leq c} \|x_i - a_k\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $\|x_i - a_k\|$ is the Euclidean distance between data point x_i and cluster center a_k [33].

A time series x comprises a sequence of m observations, (x_1, x_2, \dots, x_m) , and we assume uniform length for all series within the dataset [32]. In the context of time series data represented by a set D containing n time series cases, the k-means clustering task entails dividing D into k distinct clusters of time series, denoted as $C = \{C_1, C_2, \dots, C_k\}$, where k is the predetermined number of clusters [32].

To align with prior research by Moreno et al. (2022) using a group-based trajectory model (GBTM), we adopted as our dependent variable whether a child was overweight, defined as a BMI \geq 85th percentile [1, 34]. For each student, we have a boolean (1 = BMI \geq 85th percentile, 0 = BMI $<$ 85th percentile) time series made up of 11 measurements.

5.1.2 Hyperparameter Tuning

The effectiveness of the k-means time series clustering algorithm can fluctuate based on several factors: centroid initialization, the predetermined number of clusters, and choice of distance metric. We optimized these parameters using unsupervised quality evaluation metrics, such as the silhouette score, which quantifies how well each data point in a cluster is separated from other clusters [35].

The k-means algorithm requires specifying the number of clusters, denoted as k , in advance. To determine the best-fitting model, we computed a range of models with increasing numbers of clusters. We assessed their fitness using several criteria: Bayesian Information Criterion (BIC) with $\Delta\text{BIC} > 10$ as a threshold [36]; clusters containing at least 5% of the sample [36]; and distinct and interpretable trajectories [36]. BIC quantifies the odds of a model's correctness and is typically a negative value, where a smaller absolute BIC suggests a better fit [37]. After evaluation, the optimal number of clusters for our dataset was found to be 5, aligning with the GBTM results (Table 3).

Number of Clusters	BIC	Minority Cluster Existence
2	-661.14	No
3	-598.04	No
4	-435.75	No
5	-380.85	No
6	-367.36	Yes

Table 3: BIC values for a general k-means model with different numbers of clusters (2 to 6). We also check for minority clusters ($\leq 5\%$ of the sample).

We also identified the optimal distance metric for our dataset. In k-means time series clustering, two widely used metrics are Euclidean distance and Dynamic Time Warping (DTW). Euclidean distance represents the L2 norm between series, while DTW accounts for time-axis distortions by aligning series for the best match [32]. Given our project's characteristics, which include sparse longitudinal data (only two time points per year) and a theoretical interest in age-specific rather than shape-specific effects, Euclidean distance tends to outperform (and be preferable to) DTW [38]. We took a set of seeds ranging from 0 to 99 and assessed the average model performance using Euclidean distance versus DTW. The mean silhouette score using Euclidean distance is 0.666, with a 95% confidence interval(CI) = [0.663, 0.668], which is higher than DTW (mean silhouette score is 0.627, with a 95% confidence interval(CI) = [0.624, 0.629]). This gap can be seen in Figure 8a; a high value for silhouette score is preferable.

Finally, we fine-tuned the model's initialization process. Instead of randomly selecting k observations from data, we adopted the k-means++ method, which enhances the distinctiveness of initial centroids for

quicker convergence [39]. We took a set of seeds ranging from 0 to 99 and assessed the average model performance using k-means++ versus the baseline (random k observations from data). From Figure 8b, the mean silhouette score of k-means++ is 0.666, with a 95% confidence interval(CI) = [0.663, 0.668], which is higher than the baseline (mean silhouette score is 0.653, with a 95% confidence interval(CI) = [0.647, 0.659]).

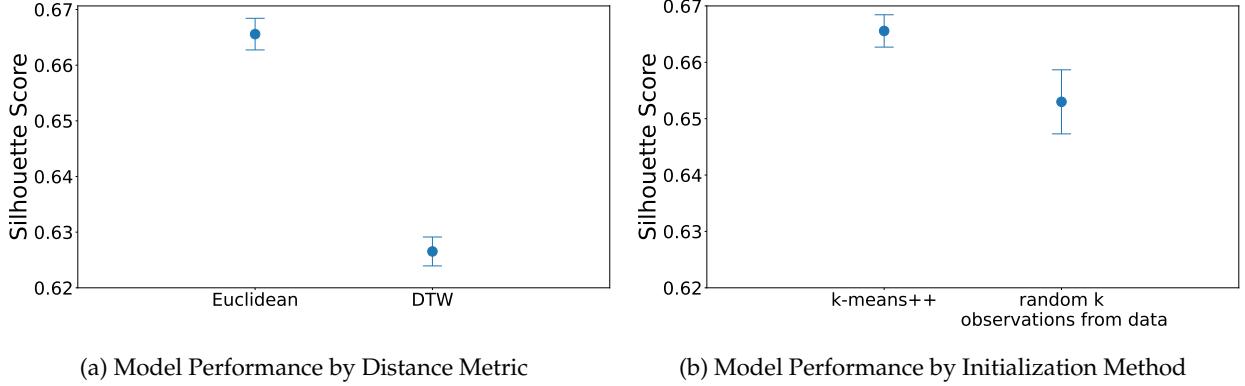


Figure 8: Hyperparameter Tuning: Model performance is evaluated on the mean silhouette score with 95% confidence intervals using seeds ranging from 0 to 99. Euclidean distance (mean silhouette = 0.666, [95% confidence interval (CI) = 0.663, 0.668]) is preferred over Dynamic Time Warping (DTW: mean silhouette = 0.627, [95% confidence interval (CI) = 0.624, 0.629]); k-means++ is preferred (mean silhouette = 0.666, [95% confidence interval (CI) = 0.663, 0.668]) over random selection (mean silhouette = 0.653, [95% confidence interval (CI) = 0.647, 0.659]).

5.1.3 Results and Comparisons

To enable reproducibility, we used the fixed seed (99) and the fine-tuned hyperparameters to fit the model. The k-means model identifies 5 trajectory groups, aligning closely with GBTM results from Moreno et al. (2022) [1]. The key transition points for early-onset and late-onset overweight/obesity groups—after kindergarten and after 2nd grade, respectively—are also consistent with previous findings [1]. The updated percentages of the cluster membership are shown in Table 4.

We assessed cluster membership in both k-means and GBTM, revealing an overall 5% variation in the labels. Specific differences are outlined in Table 5. We compared the clustering performance using the silhouette metric, and k-means had a score (0.679) slightly higher than GBTM (0.668). We hypothesize that individuals whose cluster membership diverged between the two models exhibit trajectories that deviate more substantially from the GBTM centers. This observation is made through a visual examination of the differences, and our subsequent validation process aims for quantitative measurement and confirmation. From both Figure 9 and Table 5, we saw that the clustering results of the two models diverge the most among the “Becoming Healthier”, “Early-onset Overweight/Obese”, and “Late-onset Overweight/Obese” groups. Accordingly, we are interested in not only comparing the overall clustering performance of the two

models but also looking into the specified three groups.

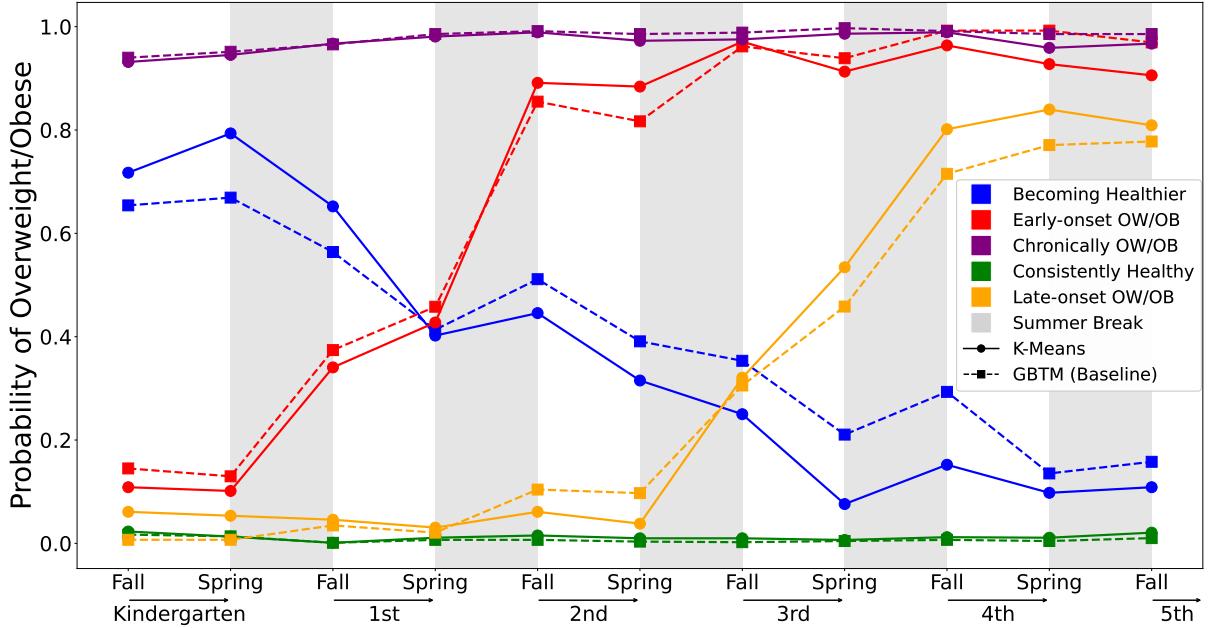


Figure 9: Probability of being overweight/obese among the five trajectory groups of k-means vs. GBTM over time. The trajectories of the two models diverge the most among the “Becoming Healthier”, “Early-onset Overweight/Obese”, and “Late-onset Overweight/Obese” groups.

Trajectory Group	k-means Membership	GBTM Membership
Becoming Healthier	5.6%	8.1%
Early-onset Overweight/Obese	8.4%	8.0%
Chronically Overweight/Obese	22.3%	21.3%
Persistently Healthy Weight	55.7%	53.8%
Late-onset Overweight/Obese	8.0%	8.8%

Table 4: Group membership for k-means vs. GBTM trajectory groups (complete cases only, n = 1,642).

5.1.4 Validation

Due to the absence of ground truth cluster membership for individual records, we devised a metric to assess the model’s overall accuracy. As our target variables are binary, it is challenging to represent cluster centroids directly. To address this, we chose to map cluster centroids back to their respective standardized BMI measurements using the same membership. Since the BMI z-scores are adjusted for age, this approach minimizes the distortion of the temporal information derived from our k-means model [40].

Next, we normalized the residuals from each cluster center to eliminate potential correlations with the

GBTM k-means	Becoming Healthier	Late-Onset OW/OB	Early-Onset OW/OB	Total
Consistently Healthy	15	16	0	31
Late-Onset OW/OB	6	0	9	15
Early-Onset OW/OB	9	12	0	21
Chronically OW/OB	11	0	5	16
Total	41	28	14	83

Table 5: The pivot table highlights the pairwise differences in cluster membership between k-means and GBTM. For example, it shows that 15 observations belong to the “Consistently Healthy” group by k-means but “Becoming Healthier” by GBTM. We will utilize the accuracy metric to evaluate the clustering performance in the validation section.

covariates [41]. Given that all measurements are standardized, we can assess the accuracy of each individual observation’s cluster membership using a threshold of 2 standard deviations from its centroid. This threshold is a commonly employed statistical criterion for identifying outliers [42]. Our validation metric is precisely defined as follows:

$$\text{Individual Accuracy} = \begin{cases} 1 & \text{if an observation's residual is within 2 standard deviations of its centroid} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Overall Accuracy} = \frac{\text{sum of individual accuracy}}{\text{sample size}}$$

In Table 6, we present a comprehensive comparison of clustering accuracy between the two models across both the entire dataset and individual groups. Consistent with the findings obtained from the silhouette metric analysis, k-means demonstrates a slightly higher overall accuracy (95.25%) compared to GBTM (95.07%).

Notably, in the subsets where the trajectory patterns diverge significantly between the two models, k-means exhibits superior accuracy in the “Becoming Healthier” group (94.57% vs. 93.23%) and the “Late-onset Overweight/Obese” group (94.66% vs. 93.75%). Interestingly, in the “Early-onset Overweight/Obese” group, k-means performs marginally lower than GBTM (93.48% vs. 94.66%).

A noteworthy observation is the distinct difference in membership percentages in the ‘Becoming Healthier’ group, with k-means at 5.6% and GBTM at 8.1%. This variation suggests that k-means has effectively improved clustering membership in the ‘Becoming Healthier’ group, contributing to its overall higher accuracy.

Method	Accuracy (%)			
	Overall	Becoming Healthier	Early-Onset OW/OB	Late-Onset OW/OB
K-Means	95.25	94.57	93.48	94.66
GBTM	95.07	93.23	94.66	93.75

Table 6: Accuracy comparison between k-means and GBTM for different groups.

5.1.5 Implications and Future Work

The application of k-means time series clustering to identify growth trajectory groups among elementary school children has significant potential impacts. By highlighting distinct developmental pathways related to childhood weight, this approach contributes to a deeper understanding of the complex interplay between growth patterns and environmental factors such as COI and ALAN. The application of a different grouping method emphasizes the robustness of the initial growth trajectories of Moreno et al. (2022) [1]. Such insights can inform targeted interventions, enabling more personalized strategies for addressing childhood obesity and its associated health risks.

In future, there are several potential avenues to explore. Further refinements in clustering algorithms, such as incorporating advanced distance metrics and exploring alternative parametric clustering techniques that are better suited to the distribution of our data, could enhance the precision of trajectory group identification. Validating the results with additional datasets would also lend them additional strength.

5.2 Supervised Model: Time Series Regression

5.2.1 Fitting Time Series Models

To better understand the trends over time in children’s health, we have constructed multiple supervised time-series models. We aim to predict height and BMI trajectories at the individual level. Our input variables are COI, ALAN, and health factors, while our specific output variables are height and BMI. The goal of this approach is to gain insight into the relationship between our neighborhood-level factors (childhood opportunity and artificial light) and their relationship to seasonality and the predicted trajectories of individual children’s growth. We employed both the SARIMAX model and Prophet Forecasting Model to better understand the relationship between these variables. The use of consistent random seeds across all models ensures that results are reliable and reproducible regardless of the time and device in which they are being trained [43]. As an initial investigation, we constructed time series models using a sample of 5 randomly-selected individuals from the training dataset. Values for COI for most individuals remain the same throughout the study period unless students changed schools. ALAN data may differ by year.

The resulting regression models are evaluated using various error and goodness-of-fit metrics, such as mean square error (MSE) and root mean square error (RMSE). Although the error metrics mentioned are

scale-dependent, their thresholds may be determined through discussions with the project mentors and sponsors. However, it may also be useful to utilize more scale-independent metrics for better interpretability, such as mean absolute percentage error (MAPE) [11, 44]. Assessing the performance of the SARIMAX model compared to that of the SARIMA model leads to an increased understanding of whether external factors such as COI and ALAN have significant impacts on children's growth [45]. More specifically, if the SARIMAX model's validation metrics are superior to those of the SARIMA model, these external neighborhood-level factors may have a correlation with changes in BMI as well as the probability of being overweight or obese [45]. Similarly, comparing Prophet model performance with and without additional regressors will inform researchers on the relationship of such factors to growth patterns.

5.2.2 SARIMAX Model

For our regression approach, one modeling option is a SARIMAX (Seasonal Autoregressive Integrated Moving Average Exogenous) model. This is an approach to modeling time series data that allows for the inclusion of both exogenous factors and seasonality [46]. Since much of our sponsor's recent work deals with the seasonality of growth trends, and we wish to include exogenous neighborhood-level variables, SARIMAX may be a good modeling option. Given that we only have twice-yearly growth measurements, seasonality is challenging to measure and to model, despite its theoretical relevance.

To investigate the effects of the neighborhood-level factors on children's growth trajectories, the SARIMA and SARIMAX models are implemented. The SARIMA model, like the SARIMAX model, belongs to the ARIMA family of models and may be described by the following equation:

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^Dy_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

where y_t is the dependent variable, $\varphi_p(B)$ represents the AR polynomial with order p , $\theta_q(B)$ represents the MA polynomial with order q , $\Phi_P(B^s)$ represents the seasonal AR polynomial with order p , and $\Theta_Q(B^s)$ represents the seasonal MA polynomial with order Q [46]. The SARIMAX model is described similarly as follows:

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^Dy_t = \beta_kx'_{k,t} + \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

where $x'_{k,t}$ represents the vector containing the k^{th} exogenous variable at time t and β^k represents the coefficient of the k^{th} exogenous variable [46].

Prior to running these models on the combined dataset, the data underwent certain necessary pre-processing steps. These included reformatting variables, re-sampling the data to represent a daily frequency, and interpolating values for rows containing missing data. Since the health data was collected only twice a year, re-sampling it to a daily frequency resulted in the majority of the dataset containing missing values, necessitating interpolation. The results of this method are included in Time Series Regression

Results section.

5.2.3 Prophet Forecasting Model

In addition to the SARIMAX modeling, another forecasting procedure we employed is the Prophet Forecasting Model, created by Meta [47]. Prophet is a versatile and accurate predictive tool that forecasts a desired variable with respect to time [47]. Prophet operates efficiently even if the dataset has numerous outliers and missing values [47]. In some studies, Prophet performed better than other widely accepted models like ARIMA and SARIMAX ([47]). While SARIMAX constructs a formula as a function of past values, Prophet operates by detecting significant changes in patterns [48]. It detects separate non-linear trends in the time series like yearly, weekly, and daily seasonality as well as holiday effects and then combines them together to obtain a forecast value ([48]). Prophet constructs a generative additive time series model that analyzes the growth (or trend), seasonality, and holidays within the time frame ([48]). The time series model can be decomposed as:

$$y(t) = g(t) + s(t) + e_t$$

where e_t encodes variation not included in the model ([49]). Given that Prophet is a generative additive model, the approach it follows for forecasting can be described as curve-fitting, initially with time as the only regressor [49]. Additional regressors can also be added to the model as well.

When applying Prophet to the current data, there were certain necessary pre-processing steps. These included altering indices, reconstructing dataframes, and interpolating the data linearly to prepare it for the Prophet model. We examined a random subset of four students and assessed their growth trajectories through the Prophet model. As with SARIMA and SARIMAX, rather than predict future values outside the scope of the study, we opted to train the model on the first four years and test on the final year of the study. The first Prophet model utilized only student health data. In the second model, we inserted the health data as well as COI and ALAN to understand if these additional regressors could better predict future values of children's height. The results of this method are included in Time Series Regression Results section.

5.2.4 Time Series Regression Results

In order to compare the Prophet and SARIMAX models, we visualized four students' trends in height and BMI, including their forecasted values. Figure 10 showcases the forecasted values for BMI using both time series regression methods without additional regressors. Blue denotes the Prophet forecast, while green denotes the SARIMA/SARIMAX forecast and yellow the test (observed) values. In Figure 10, the performance of each model appears to be inconsistent. In some instances, the SARIMA model follows the test values closely; in others, the Prophet model appears to perform better.

We conducted the same analysis with additional regressors such as COI and ALAN. Figure 11 displays these results. The SARIMAX model still appears to produce forecasts similar to the observed test data.

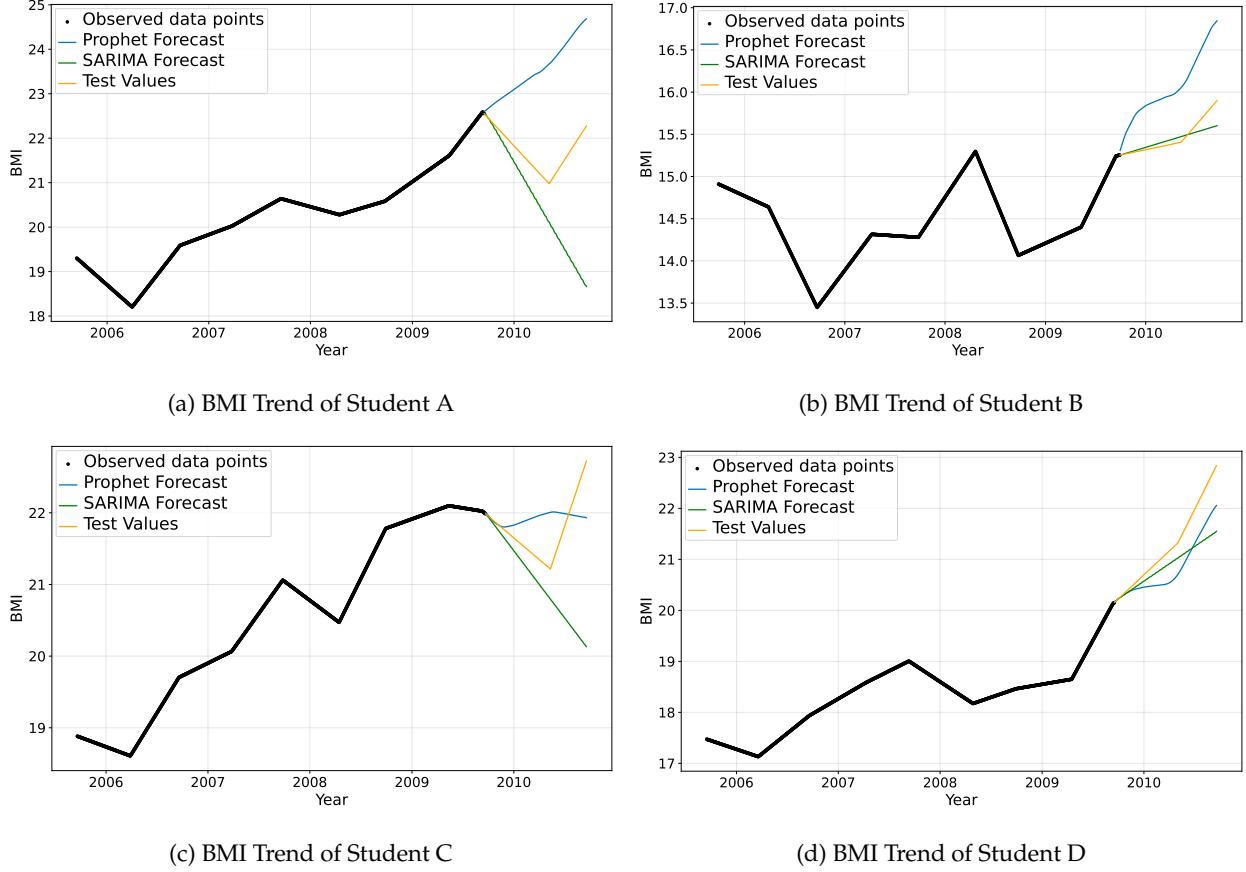


Figure 10: Forecasted values for BMI using Prophet and SARIMAX time-series regression models, without additional regressors (COI and ALAN).

However, the same cannot be said about the Prophet model. Based on the visual output, it appears that including the additional regressors seems to generally hurt the overall accuracy of the forecast. This finding is confirmed by the performance metrics, below.

In addition to measuring BMI, we also wanted to compare the performance of both modeling approaches when forecasting height. Figure 12 showcases the results of the Prophet and SARIMAX methods without additional regressors. Similarly to the models predicting BMI, both models appear to forecast future values fairly accurately, indicating that the models without additional regressors appear to predict future values with a reasonable level of success. The final four plots (Figure 13) showcase the Prophet and SARIMAX models forecasting height with the additional regressors. Unlike when predicting BMI, the Prophet model appears to forecast values more accurately than the SARIMAX, but it still appears to further from the test values than when modeling without additional regressors.

After applying both the SARIMAX and Prophet models, we calculated metrics such as the mean absolute percentage error, mean square error, and root mean square error to determine their performance. These metrics were calculated for both height (Table 7) and BMI (Table 8). The Prophet model performed better

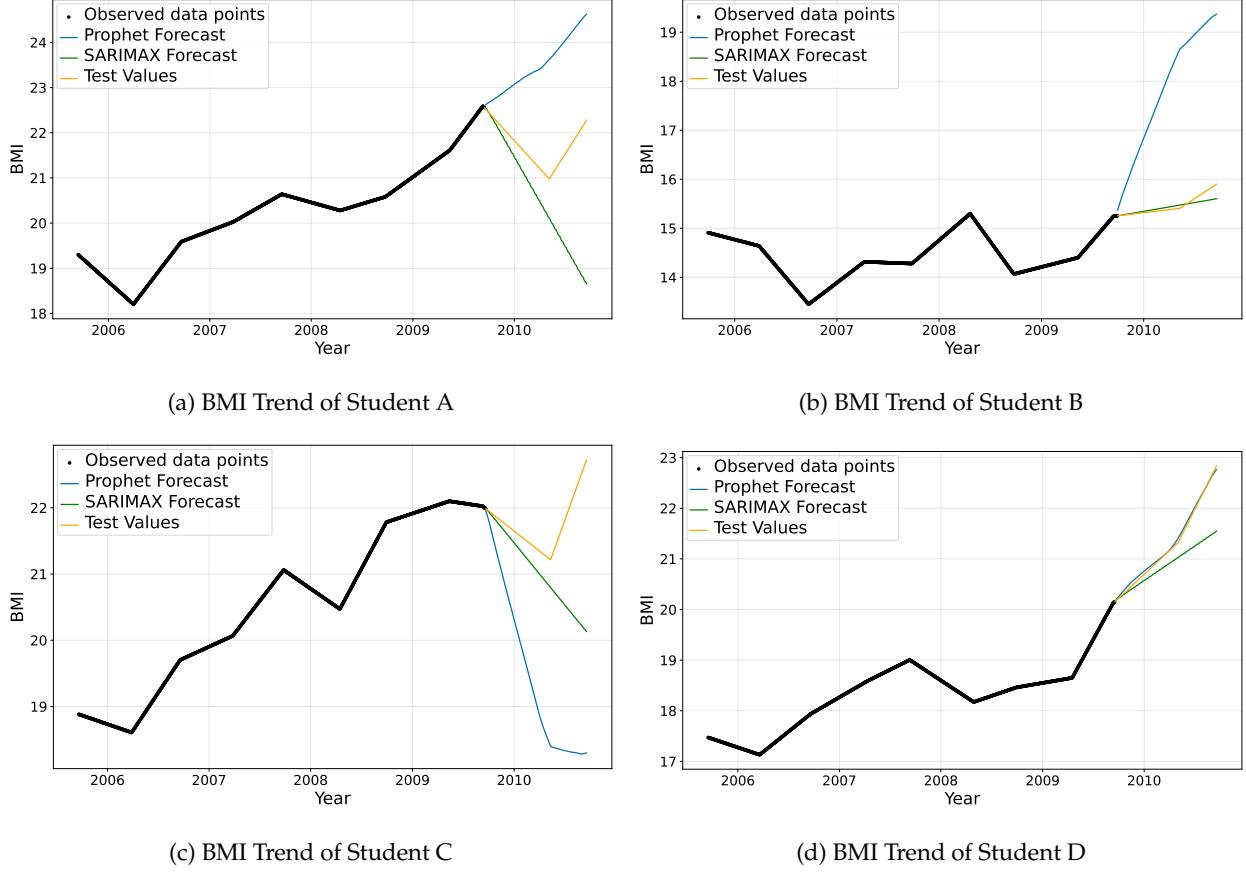


Figure 11: Forecasted values for BMI using Prophet and SARIMAX, with additional regressors.

	Mean MAPE	Mean MSE	Mean RMSE
Prophet	0.0070	0.3252	0.4467
Prophet with COI and ALAN	0.0244	3.3486	1.3967
SARIMA	0.007052	0.30762	0.483528
SARIMAX with ALAN	0.007052	0.30762	0.483528

Table 7: Model evaluation metrics for predicting height. Lower values indicate better performance.

	Mean MAPE	Mean MSE	Mean RMSE
Prophet	0.0345	0.9696	0.7785
Prophet with COI and ALAN	0.0945	6.6480	1.9043
SARIMAX	0.02397	0.770497	0.739724
SARIMAX with ALAN	0.02397	0.770497	0.739724

Table 8: Model evaluation metrics for predicting BMI.

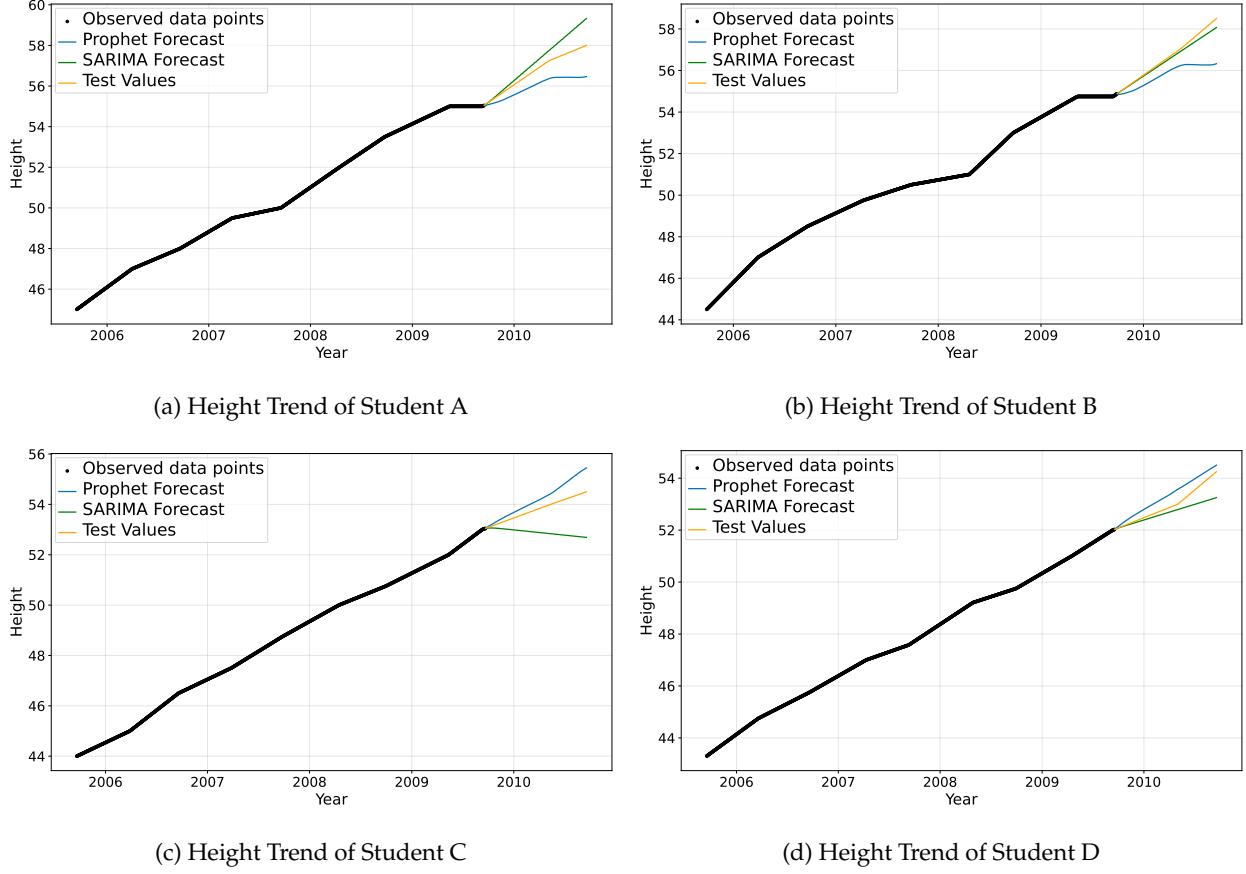


Figure 12: Forecasted values for height using Prophet and SARIMAX, without additional regressors.

when we did not include additional regressors such as COI and ALAN. For the SARIMA and SARIMAX models, model performance metrics were identical with and without the additional regressor; with this modeling approach to these trajectories, the exogenous ALAN variable had little to no impact on improving predictions.

When examining the results from both SARIMAX and Prophet modeling, the models that do not utilize COI and ALAN data often perform better or similar to those that do. Given the performance of both models, it's not immediately clear what, if any, relationship exists between these additional factors and childhood height and BMI trajectories. The results suggest that, at least within the context of the models and data used, the additional regressors do not seem to substantially enhance our understanding of these trends. A pivot to a different type of time-series regression may be in order.

5.3 Supervised Model: Linear Mixed-Effects Regression

Having gained insight into children's BMI trajectories through clustering them into groups and applying shape-based individual regressions, we shift to modeling the relationships between neighborhood-level factors (COI and ALAN) and students' BMI trajectories. With the final modeling analysis, we aim to better

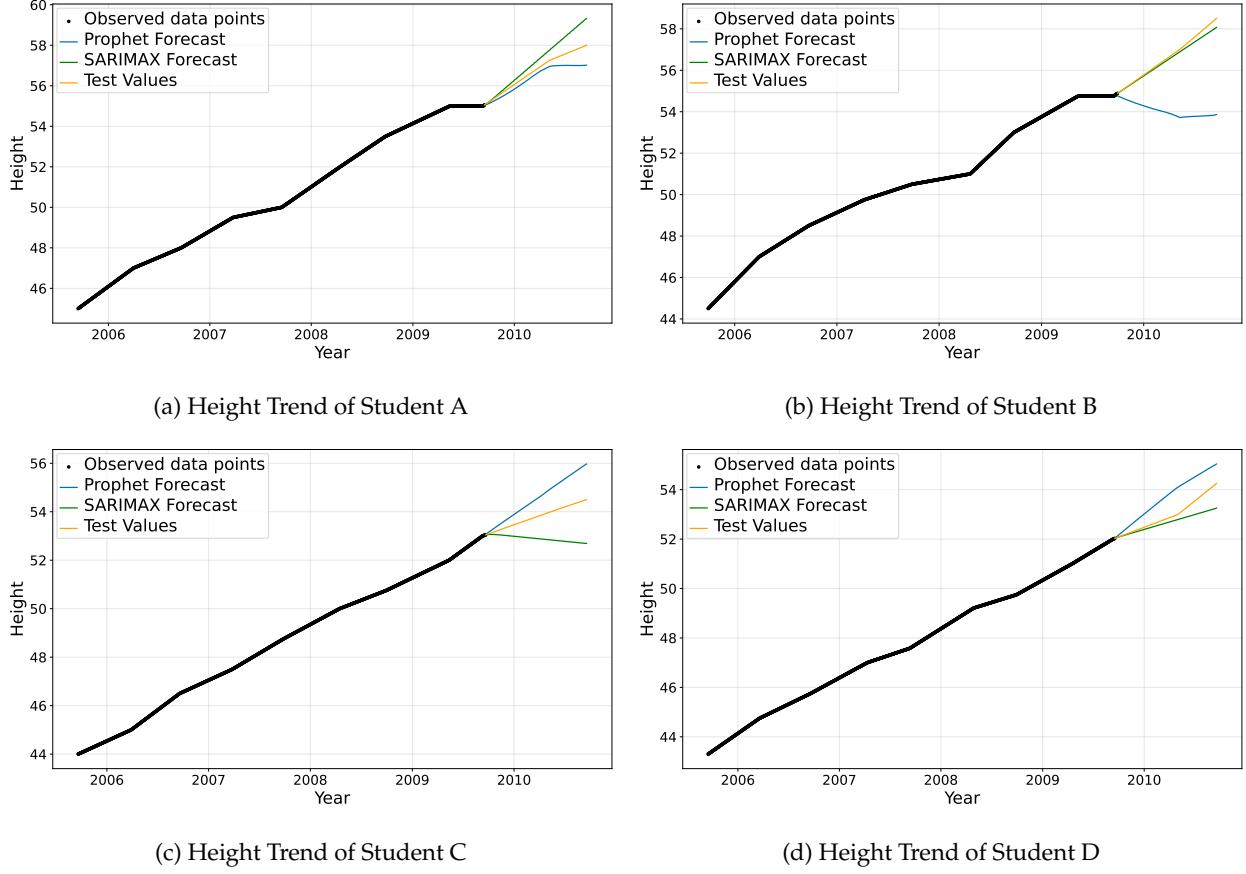


Figure 13: Forecasted values for height using Prophet and SARIMAX, with additional regressors.

understand the complex relationship between neighborhood opportunity and artificial light at night in children’s growth.

5.3.1 Introduction to Mixed-Effects Models

Mixed-effects models are a type of statistical model that incorporates two kinds of effects: fixed effects, parameters common to the population or to common groups or treatments; and random effects, which are modeled as random and specific to the individual units of observation [50]. They are typically used to analyze correlated data; these include data that are grouped by some characteristic, whether by time (longitudinal or repeated measures data) or by a structure or feature (multilevel or hierarchical data) [51]. Mixed-effects models are frequently seen in medical and social science research, especially when analyzing time-series data [52]. They are distinguished from other time-series analysis methods, such as repeated-measures ANOVA, by their capacity to handle missing data using maximum likelihood estimation, incorporating all available information from data with missing features [11, 53, 54]. Mixed-effects approaches have been used to model children’s BMI trajectories in a variety of recent studies [55–59]. Perhaps the most relevant are the work by Moreno et al. (2022) with mixed-effects models and the current data, and a recent

linear mixed-effects investigation of COI and BMI trajectories from infancy to adolescence by Aris et al. (2022) [1, 11].

5.3.2 Methodology

The project objectives include associating COI and ALAN to both children's growth trajectories (height and BMI gain) and to their risk of becoming overweight/obese over time. Aris et al. (2022) implement linear mixed-effects models to examine the relationship between COI (divided into quantile categories) and children's BMI in each life stage; they also utilized general mixed-effects models with a logit link (with all covariates held constant at mean values) to examine the relationship between COI and risk of obesity [11]. By implementing measures specific to our data, including ALAN categories, reproducing the broad strokes of the modeling approach of Aris et al. (2022) allows us to gain insight into the techniques they use, as well as draw meaningful and interpretable conclusions.

Covariates included in our mixed effects-models include the following fixed effects: COI or ALAN categories, children's socio-demographic attributes (such as age, race/ethnicity, and gender), season (measurements collected in the spring vs. the fall), and interaction terms (COI or ALAN interacted with season and age, in two-way and three-way interaction terms) [1]. Random effects in the linear mixed-effects model include children's intercept and linear slope for age, as well as for the nesting variable (common school) to allow for intra-group and spatial clustering [11]. To allow models to converge on the available data, we split our linear mixed-effects models into one model with COI included in the predictors, and one with ALAN, which serves to reduce our number of predictors and increase our degrees of freedom to within a functional range.

Both previous approaches recommend the inclusion of additional terms for age to capture non-linearity in BMI trajectories over time: Moreno via age^2 , and Aris via natural spline terms for age (with knots at 0.05, 1, 4.5, and 10 years, their location and number determined by optimal Bayesian Information Criterion (BIC)) [1, 11]. Due to the shorter study period in the current data (5 years), we test for the inclusion of an age^2 term only. We also "centered" the age term (and by extension, the inputs into the age^2 term) by subtracting 5 years from all values; in the past few decades, 5 years old has been the minimum age for a child to start kindergarten in most U.S. states, including Texas [60]. The shift in our age variable allows our models' intercept values to be more meaningful (for most students, placing the "zero" age within the range for which we have data) without changing the structure of the model or any continuous predictors' regression coefficients.

Below is the equation for a "full" linear mixed regression model, which includes all of our potential predictors. *NBR* is a stand-in for our neighborhood-level variables, COI and ALAN. Note that in accordance with the methodology of Aris et al. (2022), we utilized COI categories whose percentile cutoffs are determined by overall childhood opportunity index distributions. [11, 61]. The categorical variable necessitates additional terms and interaction terms within the formula representation, which we have omitted here for

space and clarity. ALAN is a continuous variable, measured as the radiance mean of the geographic area (in nanowatts per steradian per square). Baseline demographic categories are $sex = male$ and $ethnic = Black$, the modes observed in the data.

$$BMI_i \sim N(\mu, \sigma^2)$$

$$\begin{aligned} \mu = & \alpha_{j[i],k[i]} + \beta_1(NBR) + \beta_2(c(season)_{spring}) + \beta_3 j[i],k[i](age) + \beta_4(age^2) + \\ & \beta_5(c(season)_{spring} \times NBR) + \beta_6(NBR \times age) + \beta_7(c(season)_{spring} \times age) + \\ & \beta_8(c(season)_{spring} \times NBR \times age) \end{aligned}$$

$$\begin{pmatrix} \alpha_j \\ \beta_{3j} \end{pmatrix} \sim N \left(\begin{pmatrix} \gamma_0^\alpha + \gamma_1^\alpha(c(sex)_{female}) + \gamma_2^\alpha(c(ethnic)_{white}) + \gamma_3^\alpha(c(ethnic)_{Hispanic}) + \gamma_4^\alpha(c(ethnic)_{Asian}) \\ \mu_{\beta_{3j}} \end{pmatrix}, \begin{pmatrix} \sigma_{\alpha_j}^2 & \rho_{\alpha_j \beta_{3j}} \\ \rho_{\beta_{3j} \alpha_j} & \sigma_{\beta_{3j}}^2 \end{pmatrix} \right), \text{ for student } j = 1, \dots, J$$

$$\begin{pmatrix} \alpha_k \\ \beta_{3k} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_{\alpha_k} \\ \mu_{\beta_{3k}} \end{pmatrix}, \begin{pmatrix} \sigma_{\alpha_k}^2 & \rho_{\alpha_k \beta_{3k}} \\ \rho_{\beta_{3k} \alpha_k} & \sigma_{\beta_{3k}}^2 \end{pmatrix} \right), \text{ for school } k = 1, \dots, K$$

Within the equation, one can see the random intercepts for student and school in the subscript for the intercept α , and the random slopes for age (based on student and school) in the subscript for slope coefficient β_3 . Most fixed-effect predictors have the standard slope or intercept coefficients (similar to a linear regression model), including the neighborhood-level variable, season, age^2 , and interactions. Student variables that stay constant over time (here, sex and ethnicity) contribute to each student-specific intercept term.

5.3.3 Model Validation and Stability Testing

For purposes of model validation, we implement our block cross-validation method. Of course, on unseen schools and students, the model can rely only on the fixed coefficients and observation-level error, not on our school- and student-specific random effects. In addition, we have some data leakage, since students who switch schools (around 20% of students in the sample change schools at least once) may appear in both the training and test data in a cross-validation iteration; however, students change schools via highly complex patterns and mechanisms, so it would be prohibitively difficult to devise a cross-validation method (aside from CV on the student level, which is computationally burdensome) that omits any such student-level leakage. Fang demonstrated that for linear mixed effects models, the marginal Akaike Information Criterion (mAIC, or typically the AIC) is asymptotically equivalent to leave-one-cluster-out cross-validation [62].

For model selection and validation, we implement both our block cross-validation method (where we assess predictive performance with the commonly-used root-mean-square error (RMSE) and mean absolute

error (MAE)) and the AIC and BIC metrics on models trained on all available data. The models under consideration are the full model (see equation above), a model without an age² term, a model without random slopes for age, and a model without an age² term or random slopes for age. The interaction terms (between season, age, and our neighborhood-level variables) are of theoretical interest, and thus included by default.

	RMSE	MAE	AIC	BIC
Full COI Model	2.9352	2.0217	147437.2	147682.2
COI Model (no age ² term)	2.9394	2.0226	148245.5	148481.7
COI Model (no random slopes for age)	2.9990	2.1403	172787.7	172997.7
COI Model (no age ² term or random slopes for age)	3.0031	2.1517	173459.4	173660.6

Table 9: Model evaluation metrics for linear mixed-effects models, with COI as a predictor. RMSE and MAE calculated via block cross-validation (leave-one-block-out, on the school level); AIC and BIC calculated on models trained on all available data. Lower values indicate better performance, with lowest values above in bold.

	RMSE	MAE	AIC	BIC
Full ALAN Model	2.9530	2.0357	147440.1	147615.1
ALAN Model (no age ² term)	2.9445	2.0263	148249.0	148415.2
ALAN Model (no random slopes for age)	3.0151	2.1509	173195.4	173335.4
ALAN Model (no age ² term or random slopes for age)	3.0091	2.1553	173780.9	173912.2

Table 10: Model evaluation metrics for linear mixed-effects models, with ALAN as a predictor.

For the COI models (see Table 9), we see that the full COI model has the best performance in all cases. For the ALAN models (see Table 10), the model without a nonlinear term for age performed best in the cross-validation procedure, but the full model performed best when trained on all data. This may indicate that the age² term is leading to overfitting on training data for this model; because of that concerns and the preferability of parsimonious models, we select the second ALAN model (with no age² term) for our modeling procedure.

Model stability is another important consideration. Models are stable when selected predictors are not sensitive to small changes in the data [63]. Bootstrap resampling, where a dataset is repeatedly sampled with replacement to generate many “simulated” datasets, is a commonly-proposed method for evaluating model stability [63]. Bootstrap-based confidence intervals for model features may be compared to “naive” confidence intervals [64]. We implement a bootstrap resampling procedure to estimate our coefficient confidence intervals and a simulation-based sampling procedure that captures model uncertainty (an alternative to bootstrap) to estimate our prediction intervals.

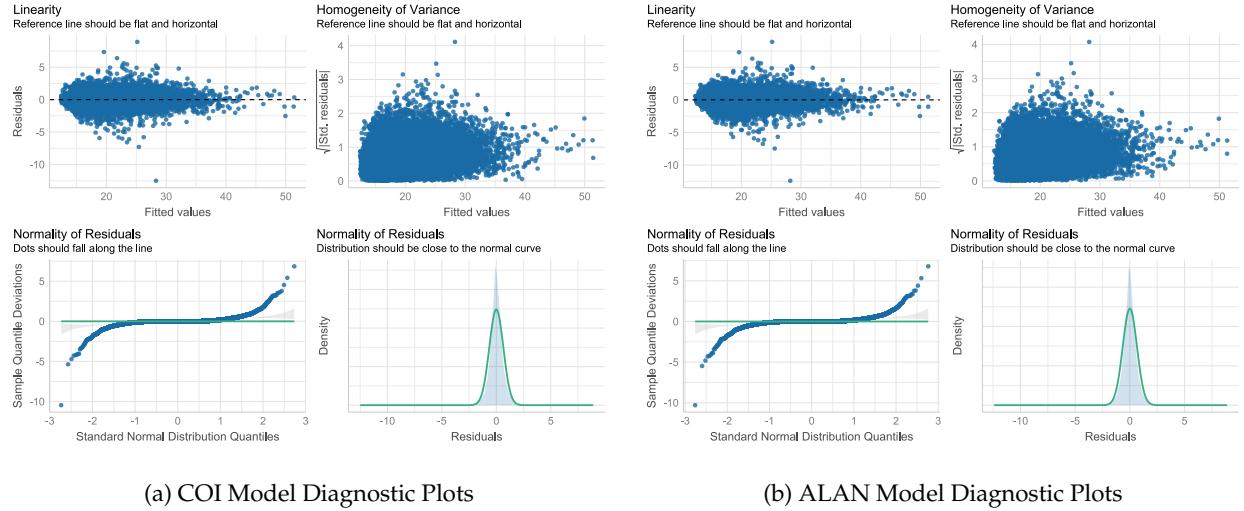


Figure 14: Assumptions Checks for Linear Mixed Effects Models

5.3.4 Model Checking

It is also essential to check model assumptions for mixed-effects models: for example, for linear models, these include validity of the model (correctly assigned fixed and random effects, appropriateness of structure), a linear relationship between the predictor and response, homogeneity of residuals, exogeneity (independence between covariates and random effects), data missing completely at random, and certain distributional assumptions (the residuals having a normal distribution with constant variance σ^2 and random effects having a multivariate normal distribution with constant covariance matrix) [65, 66]. Mixed-effects models are generally robust to violations of distributional assumptions and missing values, with violated assumptions usually producing unbiased estimates with increased variance [65]. However, we must still investigate model assumptions to determine the extent of these impacts on the precision of our estimates, as well as ensure the model is appropriate.

The diagnostic plots, seen in Figure 14, for the COI and ALAN models show very similar results with regard to checking the assumptions for linearity, homogeneity of variance, and the normality of residuals. With regards to determining the linearity between the predictor and response, the COI and ALAN models seem to meet this assumption since the residuals are mostly scattered randomly around a cluster with the exception of relatively few fitted values. This may indicate that the model generally follows a linear relationship despite the existence of potential outliers [67]. Outlier detection was performed according to CDC thresholds, prior to fitting the linear mixed effects models [26]. The assumption of homogeneity of variance appears to be violated since the respective plots show a “fan” shape in the points. This indicates increasing variance within the models’ predictors, which is generally expected in longitudinal data [68]. Lastly, the assumption of normality of residuals appears to be met since the probability density plot of residuals shows that the residuals are approximately normally distributed [69]. However, the quantile plots show otherwise. Although most points fall along the center of the reference line, many points at the left

and right ends of the respective plots do not. This may further point to the presence of outliers remaining in the data [70].

5.3.5 Missing Data

When implementing a full mixed-effects model, we must revisit the issue of missing data. There are several types of missingness: missing completely at random (MCAR), where the failure to observe the data does not depend on any data features; missing at random (MAR), where given the observed data, missingness does not depend on the unobserved data; and missing not at random, where failure to observe the data depends on the unobserved data [53, 54, 71]. When working with frequentist inference, MCAR mechanisms are typically ignorable (in that one can proceed only with observed data, without constructing a parametric model for missing data); in likelihood and Bayesian methods, MCAR and MAR are ignorable. MNAR mechanisms are almost always non-ignorable [54]. There are also theoretical distinctions in longitudinal analysis between monotone missingness patterns (also called drop-out, where as variables increase in proportion of missingness, subjects missing for one variable are a subset of the subjects missing for the next; in other words, an observational unit, once missing, does not reappear) and non-monotone missingness [72, 73].

In the current data, we have non-monotone missingness for all longitudinal data. The question of missingness type is more complex. It is not possible to determine which pattern of missingness is occurring based only on the sample data [71]. It is possible to think of several mechanisms by which the student data could be MNAR; for instance, if a student's data is missing (meaning that their school data is also missing) for certain observations, that could be because of some school-specific characteristic like the structure of the daily schedule or minimal encouragement of regular attendance. The associations between high childhood BMI and negative health outcomes are well-documented [2, 5]. It is not an unreasonable supposition that children with extreme BMI values might be more likely to stay home sick from school and miss a measurement opportunity. However, it is also possible that these mechanisms are not relevant, and the data are truly MAR.

The issue of addressing missing data in longitudinal contexts is not a new one, and there exist a large variety of proposed methods [54, 74]. These methods include maximum likelihood (a model-based approach), multiple imputation (creating multiple complete datasets by substituting values for missing data), and fully Bayesian methods (specifying prior distributions for parameters and for missing data), and weighted estimating equations [71]. The question of modeling non-ignorable, non-monotone missing data without introducing bias is particularly complex; likelihood-based approaches, for example, require the joint distribution of the data and the missing data mechanism to be specified, through a selection, shared-parameter, or pattern-mixture model [54].

Standard longitudinal data methods, including mixed-effects modeling, limit complete-case bias by incorporating observational units with some missing data [54]. In particular, these models use likelihood-

based estimation, with a within-subject error correlation structure [53]. Mixed-effects models produced unbiased estimates under the MAR framework, without the need for a parametric missing-data model structure, and are appropriate in many contexts [53]. Due to time and computational constraints, we utilize the MAR framework for the current data. We recommend that future modeling work with these data include sensitivity analysis comparing likelihood-based MAR methods to likelihood methods robust to data MNAR, as methods that fail to model non-ignorable missing data mechanisms inevitably introduce bias [53, 54].

5.3.6 Interpretation

With mixed-effects models, it is possible to directly interpret the effect size through confidence intervals and values of the estimated parameters β , though (as with many models), one must keep variable context and transformations in mind [50]. Estimated effect sizes may also be compared to one another.

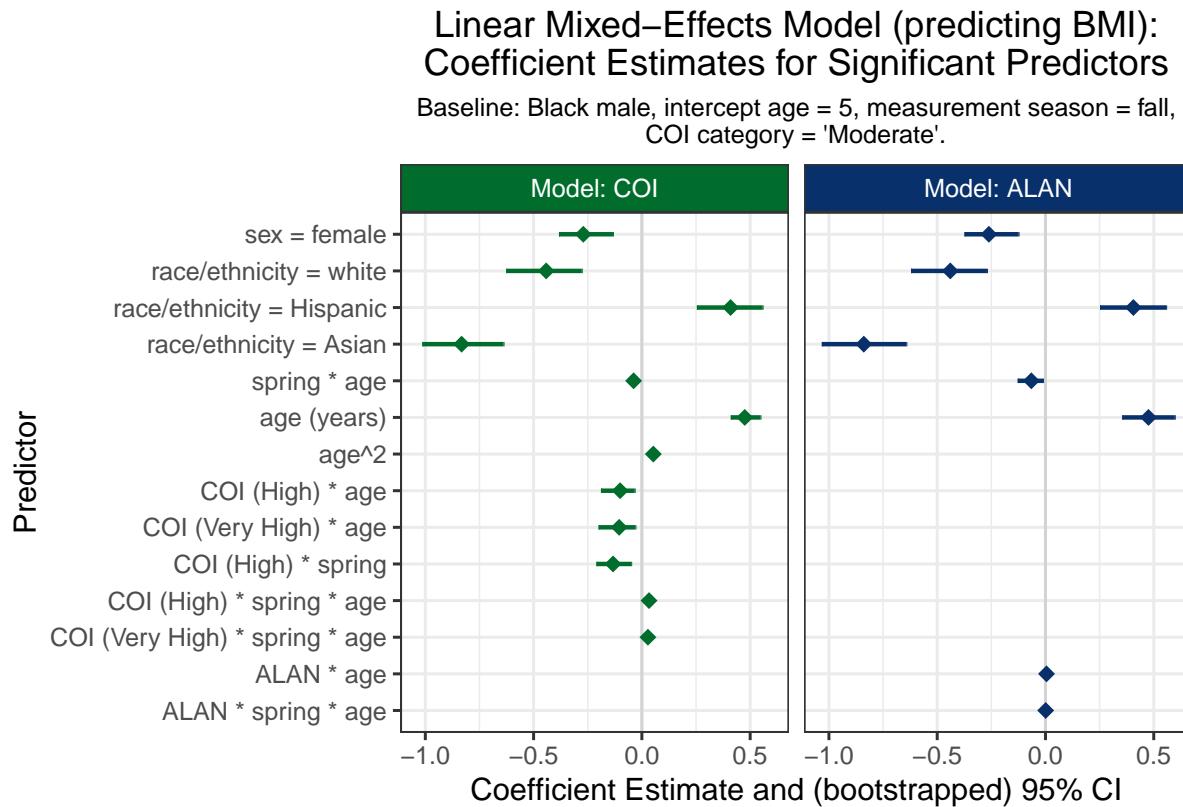


Figure 15: Linear mixed effects models' coefficient estimates and bootstrapped 95% confidence intervals. Only includes significant predictors, according to 95% CIs.

Figure 15 features all predictors that are significant given bootstrapped 95% confidence intervals (to see all predictors, reference Figure 18 in Appendix A). The demographic factors such as sex and race/ethnicity are strong and significant: girls have lower predicted BMIs than boys, while Asian and non-Hispanic

white children have lower predicted BMIs than Black children, and Hispanic children have higher. We see, overall, greater predicted seasonality (a difference between fall and spring measurements) as a child gets older, seen in the $\text{spring} * \text{age}$ coefficient. Recall that fall measurements measure BMI after summer break, while spring measurements (roughly) measure BMI after the school year. The age term (and age^2 in the COI model) show a trend towards higher BMI as a child progresses through elementary school.

More relevant to our project are the coefficient estimates for predictors involving COI and ALAN. First addressing the COI model, we see that relative to the baseline ‘Moderate’ COI category, the ‘High’ and ‘Very High’ categories see progressively lower predicted BMIs as children age. For students in the ‘High’ category, there is more pronounced seasonality (a lower predicted BMI value in the spring relative to the baseline fall measurement) at the intercept age of 5 years. However, as children in the ‘High’ and ‘Very High’ categories age, we see that seasonality become less pronounced, as their predicted BMI in spring (relative to fall) increases. For the ALAN model, all significant predictors involving ALAN are quite weak; the coefficient estimates for the $\text{ALAN} * \text{age}$ and $\text{ALAN} * \text{spring} * \text{age}$ terms are weakly positive. We have higher predicted BMI over time for children in brighter areas (exposed to greater amounts of artificial light), and more pronounced seasonality over time (lower values for spring relative to fall) for children in less-bright areas.

Visualizing predictions can aid in model interpretation. Aris et al. (2022) predict population average BMI and obesity risk by COI category, then plot the trajectories and prediction intervals, with all covariates held at mean values [11]. We create a hypothetical unseen student in an unseen school, who has all covariate values at the sample means and modes (in this case, he is a Black boy with the mean age of students at each measurement point). We predict this child’s BMI values for each COI and ALAN category, then plot the corresponding trajectories and prediction intervals (see Figures 16, 17).

Within the relatively limited time-span of the study, we do not see statistically significant differences in trajectories: the prediction intervals overlap for both the COI and ALAN models. However, the predicted trajectories can help us interpret the model results (as seen in Figure 15). For both models, we see that the COI and ALAN categories are relatively undifferentiated at the beginning of the study period, with greater spread as the children progress through elementary school. In both cases, the seasonality (differences between spring and fall measurements, visible in the “jaggedness” of the predicted trajectory) is moderate at the beginning and grows more pronounced as the children age: for the COI model, the ‘Very Low / Low’ and ‘Moderate’ categories show greater predicted seasonality in BMI than the ‘High’ and ‘Very High’ categories, while for the ALAN model, the children in darker areas have stronger predicted seasonality than children in brighter areas.

Predicted BMI Trajectories by COI Category

Covariate values held constant at sample means, modes.

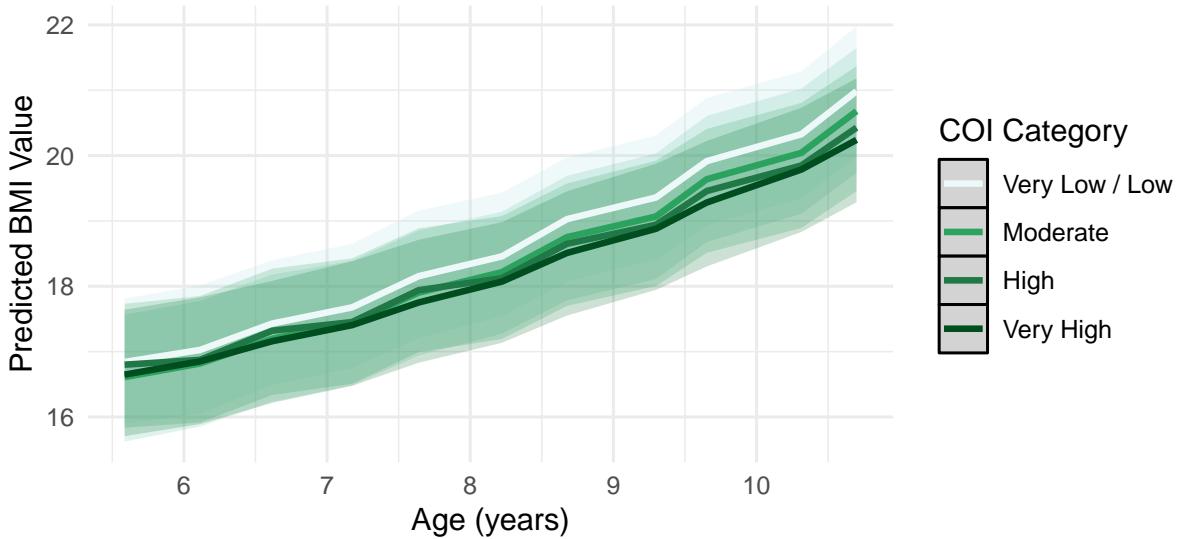


Figure 16: Predicted BMI trajectories by COI category for a hypothetical new student.

Predicted BMI Trajectories by ALAN Value

Covariate values held constant at sample means, modes.

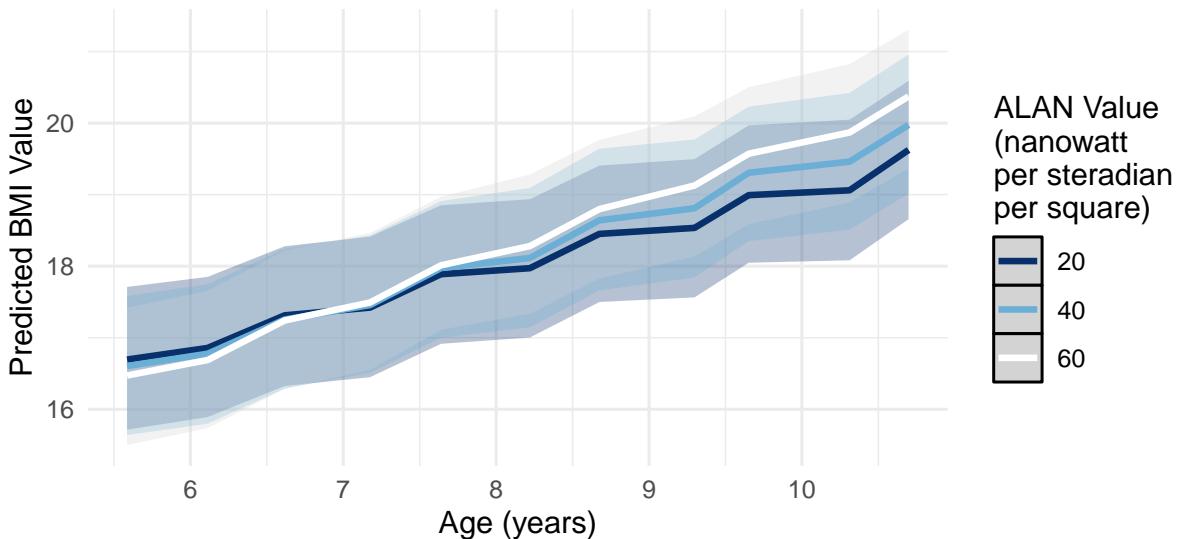


Figure 17: Predicted BMI trajectories by selected ALAN values for a hypothetical new student.

6 Discussion

6.1 Conclusions

Past research in children's growth trajectories has suggested significant trends towards an unhealthy weight status among elementary-school children, most notably during the summertime [17]. However, the exact factors that contribute to these trends have not been confirmed. Understanding the relationships between neighborhood-level child opportunity as well as exposure to artificial light at night on children's growth trajectories is crucial to determining the underlying causes of these trends [11, 18]. When examining the results of the k-means model, the clustering results were very similar to trajectory groups identified by Moreno et al. (2022) [1]. The application of an alternate grouping method speaks to the robustness of the initial trajectory groups.

The output of the initial time-series regressions examined the influence of COI and ALAN on childhood obesity and height rates. In both implementations of the time-series modeling, models with additional regressors had similar or worse performance than those without. Based solely on these results, it may appear that COI and ALAN are not important in the development of children's growth patterns, but previous studies contradict that conclusion [11]. Time-series modeling such as SARIMAX and PFM may simply be unsuitable to the data at hand.

We applied mixed-effects modeling to more precisely estimate the relationships between our key predictors (COI and ALAN) and our response (children's BMI). The modeling revealed that low neighborhood child opportunity and exposure to lower amounts of artificial light are associated with higher BMI values as children age. The results for COI are consistent with previous research on the topic [11]. Moreover, these neighborhood-level factors seem to moderate the seasonality of children's growth as they age: children in less-advantaged neighborhoods and darker areas have stronger seasonality over time. The mechanisms of these seasonal patterns are not apparent in the available data, but bear further investigation. Clearly, efforts to promote healthy childhood growth patterns should take neighborhood opportunity and artificial light exposure into account.

The current analysis could potentially serve as a foundation for policy development to better explore the role of neighborhood-level factors in childhood obesity. Ultimately, the findings of this project may allow for more efficient resource allocation to children who are most vulnerable to unhealthy weight gain. Utilizing census tracts of Fort Bend ISD schools can help to identify these vulnerable communities within a more localized area. Improved understanding of how neighborhood equity impacts children's growth trajectories may lead to an effort to eliminate risk factors. Children not only in the Fort Bend area but also worldwide could therefore be positively impacted by beneficial changes to their environment and lifestyle.

6.2 Project Limitations

The data used in this project presented a series of limitations in the generalization of our project. Given that there existed only a single COI value for our study period, we were not able to capture changes in neighborhood-level child opportunity across the five-year period. Our COI values were measured in 2010, which falls at the end of our time frame for our student data. The temporal elements of our data prevent the validation of the temporal precedence assumptions and limits our ability to conclude any causal relationships.

In addition, the limited satellite imagery of artificial light at night created large radii for ALAN values, where the measured spatial categories were much larger than desired neighborhood regions. Given the large measurement radius, the ALAN values were fairly homogeneous, leading to similar trends for the entirety of Fort Bend. Analyzing neighborhood-level factors at the census tract level limits our ability to consider the nuances of smaller regions where variation of lighting and resources differ. Another limitation for the ALAN data is the use of satellite measurements (of outdoor light) as a proxy for children's actual light exposure. One analysis of Dutch children found, on average, no association between satellite-measured light and children's bedroom lux levels [75]. Children may spend time in the evening and night outside their bedrooms (in common spaces or playing outside, for example); however, a better measurement of children's light exposure would be some kind of personal device taking real-time measurements, including exposure to screens and other indoor light.

Finally, our proportions of missing data suggests there are trends in groups and time frames that potentially indicate sections of our data are missing not at random. This can create a bias where there is under-representation of students in some neighborhoods, particularly in lower-resourced areas. The data presents imbalanced proportions of different neighborhoods and groups, limiting our generalization to neighborhoods outside of our population sample.

6.3 Future Work

Moving forward, it would be valuable to compare these findings to other school districts across the United States. Although Fort Bend is one of the most diverse districts in regards to population in the United States, geographical influences play a major role in a child's development, so conducting similar studies in other parts of the U.S would be beneficial in understanding childhood obesity. Incorporating data from earlier stages of life could also provide clarity on the role of early-childhood environment in childhood BMI.

In addition to early childhood, future studies might track these children's BMI and obesity trends into their teenage years and early adulthood. The resulting analysis could provide a more extensive understanding of how childhood upbringings influence long-term health outcomes. Additionally, supplemental data regarding air quality, access to parks and recreational areas, food deserts, etc. could shed light on the

mechanisms underlying the studied relationships and additional factors that influence childhood obesity. Ultimately, the findings from these studies could guide policymakers and legislators to construct interventions that prioritize children's growth, health, and long-term well-being.

7 Appendix

Appendix A: Linear Mixed-Effects Models' Coefficient Estimates

Linear Mixed-Effects Model (predicting BMI): Coefficient Estimates for All Predictors

Baseline: Black male, intercept age = 5, measurement season = fall,
COI category = 'Moderate'.

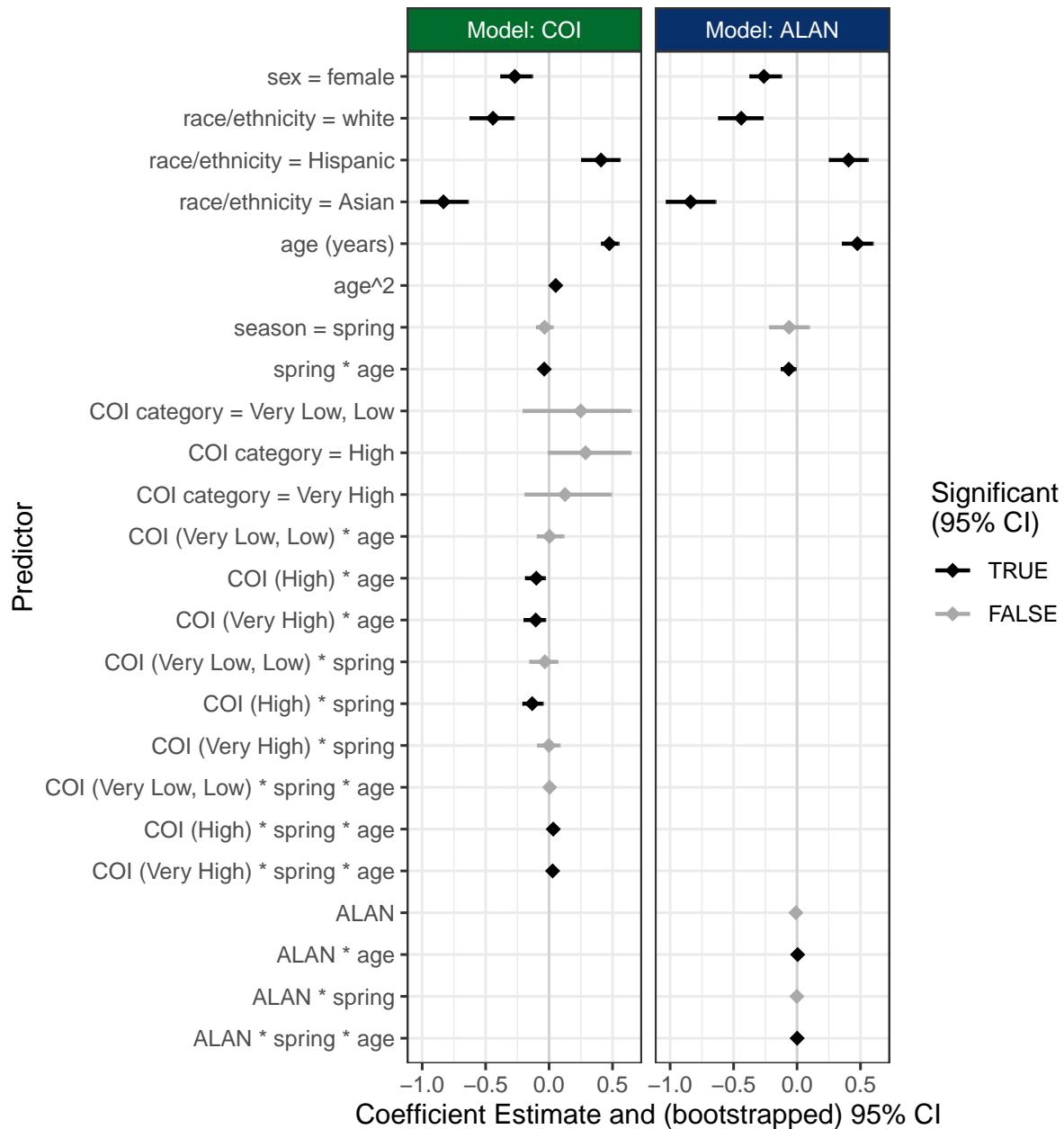


Figure 18: Linear mixed-effects models' coefficient estimates and bootstrapped 95% confidence intervals for all predictors. Note: Estimates and intervals in black are significant (95% CIs), while estimates and intervals in gray are not.

References

- [1] Jennette P Moreno et al. "Seasonality of Children's Height and Weight and Their Contribution to Accelerated Summer Weight Gain". In: *Frontiers in Physiology* 13 (2022). ISSN: 1664-042X. URL: <https://www.frontiersin.org/articles/10.3389/fphys.2022.793999>.
- [2] Matthias Barton. "Childhood obesity: a life-long health risk". In: *Acta Pharmacologica Sinica* 33 (2 2012), pp. 189–193. ISSN: 1745-7254.
- [3] Centers for Disease Control and Prevention. *Defining Child BMI Categories*. 2022. URL: <https://www.cdc.gov/obesity/basics/childhood-defining.html> (visited on 10/10/2023).
- [4] Centers for Disease Control and Prevention. *Childhood Obesity Facts*. 2022. URL: <https://www.cdc.gov/obesity/data/childhood.html> (visited on 10/10/2023).
- [5] Vivica A Kraak, Catharyn T Liverman, and Jeffrey P Koplan. "Preventing childhood obesity: health in the balance". In: (2005). ISSN: 0309133408.
- [6] Eva Sellström and Sven Bremberg. "The significance of neighbourhood context to child and adolescent health and well-being: a systematic review of multilevel studies". In: *Scandinavian journal of public health* 34.5 (2006), pp. 544–554.
- [7] George C Galster. "The mechanism (s) of neighbourhood effects: Theory, evidence, and policy implications". In: Springer, 2011, pp. 23–56.
- [8] Hayley Christian et al. "The influence of the neighborhood physical environment on early child health and development: A review and call for research". In: *Health & place* 33 (2015), pp. 25–36.
- [9] Anita Minh et al. "A review of neighborhood effects and early child development: How, where, and for whom, do neighborhoods matter?" In: *Health & Place* 46 (2017), pp. 155–174. ISSN: 1353-8292. DOI: <https://doi.org/10.1016/j.healthplace.2017.04.012>. URL: <https://www.sciencedirect.com/science/article/pii/S1353829216303525>.
- [10] Dolores Acevedo-Garcia et al. "Racial And Ethnic Inequities In Children's Neighborhoods: Evidence From The New Child Opportunity Index 2.0". In: *Health Affairs* 39 (10 2020), pp. 1693–1701. ISSN: 0278-2715. DOI: [10.1377/hlthaff.2020.00735](https://doi.org/10.1377/hlthaff.2020.00735).
- [11] "Associations of Neighborhood Opportunity and Social Vulnerability with Trajectories of Childhood Body Mass Index and Obesity among US Children". In: *JAMA Network Open* 5 (12 Dec. 2022), E2247957. ISSN: 25743805. DOI: [10.1001/jamanetworkopen.2022.47957](https://doi.org/10.1001/jamanetworkopen.2022.47957).
- [12] Ka Yan Lai et al. *Exposure to light at night (LAN) and risk of obesity: A systematic review and meta-analysis of observational studies*. Aug. 2020. DOI: [10.1016/j.envres.2020.109637](https://doi.org/10.1016/j.envres.2020.109637).
- [13] Jennette P. Moreno et al. "Potential circadian and circannual rhythm contributions to the obesity epidemic in elementary school age children". In: *International Journal of Behavioral Nutrition and Physical Activity* 16 (1 Mar. 2019). ISSN: 14795868. DOI: [10.1186/s12966-019-0784-7](https://doi.org/10.1186/s12966-019-0784-7).

- [14] Cassandra L Pattinson et al. "Environmental light exposure is associated with increased body mass in children". In: *PLoS One* 11.1 (2016), e0143578.
- [15] Li-Zi Lin et al. "Outdoor light at night, overweight, and obesity in school-aged children and adolescents". In: *Environmental Pollution* 305 (2022), p. 119306.
- [16] Jiajia Dang et al. "Artificial light-at-night exposure and overweight and obesity across GDP levels among Chinese children and adolescents". In: *Nutrients* 15.4 (2023), p. 939.
- [17] Jennette P. Moreno et al. "Physiological mechanisms underlying children's circannual growth patterns and their contributions to the obesity epidemic in elementary school age children". In: *Obesity Reviews* 21 (3 Mar. 2020). ISSN: 1467789X. DOI: 10.1111/obr.12973.
- [18] Jennette P. Moreno et al. "Later sleep timing predicts accelerated summer weight gain among elementary school children: a prospective observational study". In: *International Journal of Behavioral Nutrition and Physical Activity* 18 (1 Dec. 2021). ISSN: 14795868. DOI: 10.1186/s12966-021-01165-0.
- [19] Tse-Chuan Yang and Scott J South. "Neighborhood effects on body mass: temporal and spatial dimensions". In: *Social science & medicine* 217 (2018), pp. 45–54.
- [20] Maria A Gartstein et al. "Featured article: community crime exposure and risk for obesity in preschool children: moderation by the hypothalamic–pituitary–adrenal-axis response". In: *Journal of pediatric psychology* 43.4 (2018), pp. 353–365.
- [21] Shimels Hussien Mohammed et al. "Neighbourhood socioeconomic status and overweight/obesity: a systematic review and meta-analysis of epidemiological studies". In: *Bmj Open* 9.11 (2019), e028238.
- [22] Jennifer L Gay et al. "Can the social vulnerability index be used for more than emergency preparedness? An examination using youth physical fitness data". In: *Journal of Physical Activity and Health* 13.2 (2016), pp. 121–130.
- [23] Jennette P. Moreno et al. "Seasonal variability in weight change during elementary school". In: *Obesity* 23 (2 Feb. 2015), pp. 422–428. ISSN: 1930739X. DOI: 10.1002/oby.20977.
- [24] Christopher D Elvidge et al. "Overview of DMSP nighttime lights and future possibilities". In: *2009 Joint Urban Remote Sensing Event*. IEEE. 2009, pp. 1–5.
- [25] Centers for Disease Control and Prevention. *cdcanthro: CDC ANTHROpometry values*. 2022. URL: <https://github.com/CDC-DNPAO/CDCAnthro> (visited on 10/11/2023).
- [26] Centers for Disease Control and Prevention. *Growth Chart Training: The SAS Program for CDC Growth Charts that Includes the Extended BMI Calculations*. 2023. URL: <https://www.cdc.gov/nccdphp/dnpano/growthcharts/resources/sas.htm> (visited on 10/10/2023).
- [27] Jennette P. Moreno, Craig A. Johnston, and Deborah Woehler. "Changes in Weight Over the School Year and Summer Vacation: Results of a 5-Year Longitudinal Study". In: *Journal of School Health* 83 (2013), pp. 473–477. ISSN: 00224391. DOI: 10.1111/josh.12054.

- [28] Jose J Salazar et al. "Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy". In: *Journal of Petroleum Science and Engineering* 209 (2022), p. 109885.
- [29] David R Roberts et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure". In: *Ecography*. 40 (2017), pp. 913–929. ISSN: 0906-7590. DOI: 10.1111/ecog.02881.
- [30] Jennifer Miller, Janet Franklin, and Richard Aspinall. "Incorporating spatial dependence in predictive vegetation models". In: *ecological modelling* 202.3-4 (2007), pp. 225–242.
- [31] Jochen Schiewe. "Empirical studies on the visual perception of spatial patterns in choropleth maps". In: *KN-Journal of Cartography and Geographic Information* 69 (3 2019), pp. 217–228. ISSN: 2524-4957.
- [32] Christopher Holder, Matthew Middlehurst, and Anthony Bagnall. *A review and evaluation of elastic distance functions for time series clustering*. 2023. DOI: 10.1007/s10115-023-01952-0.
- [33] Kristina P. Sinaga and Miin Shen Yang. "Unsupervised K-means clustering algorithm". In: *IEEE Access* 8 (2020). ISSN: 21693536. DOI: 10.1109/ACCESS.2020.2988796.
- [34] Tzu-An Chen et al. "Obesity status trajectory groups among elementary school children". In: *BMC Public Health* 16 (1 2016), p. 526. ISSN: 1471-2458. DOI: 10.1186/s12889-016-3159-x. URL: <https://doi.org/10.1186/s12889-016-3159-x>.
- [35] R. Lletí et al. "Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes". In: *Analytica Chimica Acta* 515 (1 2004). ISSN: 00032670. DOI: 10.1016/j.aca.2003.12.020.
- [36] Li Ju Lin et al. "The trajectory and the related physical and social determinants of body mass index in elementary school children: Results from the child and adolescent behaviors in long-term evolution study". In: *Journal of Obesity* 2014 (2014). ISSN: 20900716. DOI: 10.1155/2014/728762.
- [37] Gideon Schwarz. "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6 (2 2007). DOI: 10.1214/aos/1176344136.
- [38] Chotirat Ann Ratanamahatana and Eamonn Keogh. "Three Myths about Dynamic Time Warping Data Mining". In: Society for Industrial and Applied Mathematics, Apr. 2005, pp. 506–510. ISBN: 978-0-89871-593-4. DOI: 10.1137/1.9781611972757.50. URL: <https://pubs.siam.org/doi/10.1137/1.9781611972757.50>.
- [39] David Arthur and Sergei Vassilvitskii. "K-means++: The advantages of careful seeding". In: vol. 07-09-January-2007. 2007.
- [40] David Monasor-Ortolá et al. "Degree of accuracy of the BMI Z-score to determine excess fat mass using DXA in children and adolescents". In: *International Journal of Environmental Research and Public Health* 18 (22 2021). ISSN: 16604601. DOI: 10.3390/ijerph182212114.

- [41] Oliver Pain, Frank Dudbridge, and Angelica Ronald. "Are your covariates under control? How normalization can re-introduce covariate effects". In: *European Journal of Human Genetics* 26 (8 2018). ISSN: 14765438. DOI: 10.1038/s41431-018-0159-6.
- [42] Juanying Xie et al. "The Unsupervised Feature Selection Algorithms Based on Standard Deviation and Cosine Similarity for Genomic Data Analysis". In: *Frontiers in Genetics* 12 (2021). ISSN: 16648021. DOI: 10.3389/fgene.2021.684100.
- [43] Pranava Madhyastha and Rishabh Jain. "On model stability as a function of random seed". In: *arXiv preprint arXiv:1909.10447* (2019).
- [44] Ali Javed, Byung Suk Lee, and Donna M Rizzo. "A benchmark study on time series clustering". In: *Machine Learning with Applications* 1 (2020), p. 100001.
- [45] Md Junayed Hossain. "Comparative Analysis of ARIMA, SARIMAX, and Random Forest Models for Forecasting Future GDP in Relation to Unemployment Rate". In: (2023).
- [46] S I Vagropoulos et al. "Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting". In: 2016, pp. 1–6. DOI: 10.1109/ENERGYCON.2016.7514029.
- [47] McCalla S Shen J Valagolam D. "Prophet forecasting model: a machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O₃, NO₂, SO₂, CO) in Seoul, South Korea". In: *PeerJ* vol. 8 e9961 (2020).
- [48] Sujata et al Dash. "Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics". In: *Pattern recognition letters* 151 (2021).
- [49] Menculini L et al. "Comparing Prophet and Deep Learning to ARIMA in Forecasting Wholesale Food Prices". In: *Forecasting* 3 (2021).
- [50] Andrzej Gałecki and Tomasz Burzykowski. *Linear Mixed-Effects Models Using R A Step-by-Step Approach / by Andrzej Gałecki, Tomasz Burzykowski*. eng. 1st ed. 2013. Springer Texts in Statistics. New York, NY: Springer New York, 2013. ISBN: 1-299-33543-8.
- [51] José C Pinheiro and Douglas M Bates. "Linear mixed-effects models: basic concepts and examples". In: *Mixed-effects models in S and S-Plus* (2000), pp. 3–56.
- [52] Martijn PF Berger and Frans ES Tan. "Robust designs for linear mixed effects models". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 53.4 (2004), pp. 569–581.
- [53] Craig H Mallinckrodt et al. "Assessing and interpreting treatment effects in longitudinal clinical trials with missing data". In: *Biological psychiatry* 53 (8 2003), pp. 754–760. ISSN: 0006-3223.
- [54] Joseph G Ibrahim and Geert Molenberghs. "Missing data methods in longitudinal studies: a review". In: *Test* 18 (1 2009), pp. 1–43. ISSN: 1133-0686.
- [55] Paola Chivers et al. "Parental and early childhood influences on adolescent obesity: a longitudinal study". In: *Early Child Development and Care* 182 (8 2012), pp. 1071–1087. ISSN: 0300-4430.

- [56] Eva D'Hondt et al. "A longitudinal study of gross motor coordination and weight status in children". In: *Obesity* 22 (6 2014), pp. 1505–1511. ISSN: 1930-7381.
- [57] Aline Andres et al. "Longitudinal body composition of children born to mothers with normal weight, overweight, and obesity". In: *Obesity* 23 (6 2015), pp. 1252–1258. ISSN: 1930-7381.
- [58] Novi Hidayat Pusponegoro, Khairil Anwar Notodiputro, and Bagus Sartono. "Linear mixed model for analyzing longitudinal data: A simulation study of children growth differences". In: *Procedia computer science* 116 (2017), pp. 284–291. ISSN: 1877-0509.
- [59] Melyssa Roy et al. "Bedtime, body mass index and obesity risk in preschool-aged children". In: *Pediatric obesity* 15 (9 2020), e12650. ISSN: 2047-6302.
- [60] Gitanjali Saluja, Catherine Scott-Little, and Richard M Clifford. "Readiness for School: A Survey of State Policies and Definitions." In: (2000).
- [61] *Child Opportunity Index 2.0 Technical Documentation*. diversitydatakids.org/researchlibrary/research-brief/how-we-built-it.. 2020.
- [62] Yixin Fang. "Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models". In: *Journal of data science* 9.1 (2011), pp. 15–21.
- [63] Patrick Royston and Willi Sauerbrei. "Bootstrap assessment of the stability of multivariable models". In: *The Stata Journal* 9.4 (2009), pp. 547–570.
- [64] Patrick Royston and Willi Sauerbrei. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. John Wiley & Sons, 2008.
- [65] Holger Schielzeth et al. "Robustness of linear mixed-effects models to violations of distributional assumptions". In: *Methods in ecology and evolution* 11.9 (2020), pp. 1141–1152.
- [66] Roel Bosker and Tom AB Snijders. "Multilevel analysis: An introduction to basic and advanced multilevel modeling". In: *Multilevel Analysis* (2011), pp. 1–368.
- [67] *Residuals vs. Fits Plot*. <https://online.stat.psu.edu/stat462/node/117/>. 2018.
- [68] *Homoscedasticity*. <https://sscc.wisc.edu/sscc/pubs/RegDiag-R/homoscedasticity.html>. 2021.
- [69] *Normal Probability Plot of Residuals*. <https://online.stat.psu.edu/stat462/node/122/>. 2018.
- [70] *Understanding QQ Plots*. <https://library.virginia.edu/data/articles/understanding-q-q-plots>. 2023.
- [71] Joseph G Ibrahim, Haitao Chu, and Ming-Hui Chen. "Missing data in clinical studies: issues and methods". In: *Journal of clinical oncology* 30 (26 2012), p. 3297.

- [72] Joseph G Ibrahim, Ming-Hui Chen, and Stuart R Lipsitz. "Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable". In: *Biometrika* 88 (2 2001), pp. 551–564. ISSN: 1464-3510.
- [73] Joseph G Ibrahim et al. "Missing-data methods for generalized linear models: A comparative review". In: *Journal of the American Statistical Association* 100.469 (2005), pp. 332–346.
- [74] Joseph L Schafer and Recai M Yucel. "Computational strategies for multivariate linear mixed-effects models with missing values". In: *Journal of computational and Graphical Statistics* 11 (2 2002), pp. 437–457. ISSN: 1061-8600.
- [75] Anke Huss et al. "Shedding some light in the dark—a comparison of personal measurements with satellite-based estimates of exposure to light at night among children in the Netherlands". In: *Environmental health perspectives* 127.6 (2019), p. 067001.