Effective Integration of Geotagged, Ancilliary Longitudinal Survey Datasets to Improve Adulthood Obesity Predictive Models

Saptashwa Mitra*, Yu Qiu*, Haley Moss†, Kaigang Li† and Sangmi Lee Pallickara*

*Department of Computer Science, †Department of Health and Exercise Science
*,†Colorado State University, Fort Collins, CO, USA
Email: *{sapmitra,yqiu,sangmi}@cs.colostate.edu,†{Haley.Elizabeth.Moss,kaigang.li}@colostate.edu

***Abstract*—Obesity is a critical health issue world-wide and has been identified as a leading cause of chronic diseases such as cardiovascular disease, type-2 diabetes, stroke and certain types of cancer. In 2014, 1.9 billion adults were overweight and 600 million were obese. In this study, to facilitate early detection of childhood obesity, we present our methodology for effective data integration that allows modelers to import new attributes from auxiliary datasets using geospatial proximity, alongside the associated data uncertainty for each data point that is caused by the data aggregation process while estimating that attribute. We have used the data uncertainty estimate as input to various machine learning algorithms, to improve on the obesity prediction. As a case study, we have integrated the National Longitudinal Survey of Youth 1997 dataset with the US Census 2000 dataset and the 2000 CDC Growth Charts dataset to augment our prediction model with behavioral and environmental features. Compared to models with only biometric attributes, our empirical experiments show accuracy improvements when we incrementally consider behavioral aspects (8.9 $\sim$ 10.2%), environmental aspects (12.1 $\sim$ 12.3%) and data uncertainty estimates (18.3 $\sim$ 25.6%).**

***Index Terms*—data integration, data science, machine learning, uncertainty, obesity**

I. Introduction

Childhood obesity has received significant attention as an urgent health challenge [1]. According to the National Health and Nutrition Examination Survey(NHANES), obesity prevalence in 2007-2008 was 33.8%; this is twice as large as the prevalence rates in 1976-1980 and a 50% rise from 1988-1994 [2]. Worldwide obesity has more than doubled since 1980 [3].

The psychological, physical and economic consequences of obesity have been well studied [4]. Obesity costs are estimated to be as high as $147 billion per year, or roughly 9% of the annual medical expenditure in the United States [3]. Because childhood obesity often continues through adolescence and adulthood, an increased number of adults will be at a risk of chronic diseases that result from obesity [5], [6].

Providing effective and accurate predictions to obesity [7]–[9] is key to identifying causes and developing proactive strategies to prevent it. Recently, there has been growing recognition of extensive factors such as the environment [10], [11], genetic predisposition [12], and human behavior and their complex interactions [13], [14] as influencers of obesity.

Modeling change of obesity levels requires tracking the same individuals repeatedly over a long period of time, often over decades. Although existing longitudinal studies provide extensive and accurate observation of the changes and trends, exploring new attributes for a model is prohibitively expensive and challenging if those attributes have not been tracked in all surveys.

Our system integrates external attributes with associated uncertainty estimates rooted in heterogeneity between datasets. We propose an approach that aggregates data, based on the underlying sample distribution, and estimate the uncertainty of importing an attribute. This uncertainty quantifies the risk of integrating a value for subsequent model generation. Providing a highly accurate predictive model with effective exploration of external features plays a vital role in developing personalized obesity prevention strategies.

A. Scientific Challenges

Flexible and accurate prediction of adolescent obesity using integrated datasets introduces a unique set of challenges:

1) ***Heterogeneity among the integrated datasets****:* Both the geospatial location of samples and frequency of data acquisitions may be different.

2) ***Accuracy and Interpretability****:* The analysis must be accurate in predicting future obesity. Furthermore, the models must be interpretable, allowing identification of the important attributes contributing change in obesity levels.

3) ***Scalability****:* The proposed approach must scale with increase in the number of datasets that are integrated and the number of features within these datasets.

B. Research Questions

Research questions that we explore in this study include the following:

1) ***How can we integrate attributes from different datasets? (RQ1)****:* This involves identifying matching attributes and manipulating data values to cope with geospatial and temporal misalignments.(§IV-B)

2) ***How can we measure and represent the data uncertainty of integrating an attribute?(RQ2)****:* The system should provide the measure of risk in integrating imported data attributes to the modelers.(§III)

3) ***How can we improve the accuracy of model with the quantified risk of integrated attributes?(RQ3)****:* The analysis should account for the uncertainty derived from the data integration process to provide a more realistic view of the imported attribute.(§IV-E)

C. Overview of Approach

Our methodology for predicting obesity with extensive external factors involves:

1) integrating attributes from external datasets via attribute matching and data preprocessing,
2) calculating and preserving the quality of integrated attributes, and
3) evaluating multiple machine learning models to assess the effect of integrating external attributes.

To import attributes from a dataset with finer grained geospatial coverage (e.g. block-level), we aggregate those values based on the geospatially matching area (e.g. zip-code or county) to generate an approximated value.

To quantify the quality of integrated attributes, we introduce the concept of data uncertainty during geospatial integration. Unlike previously discussed data uncertainty measures in integration [15], data uncertainty in geospatial integration is defined as the likelihood that the approximation of the integrated attribute does not represent the dataset accurately. In this study, we estimate the information loss that results from data aggregation for each data point to be imported. To exemplify the effectiveness of the uncertainty estimate,we perform different model fitting algorithms such as Artificial Neural Networks, Gradient Boosting, and Random Forest with and without uncertainty attributes to contrast accuracy.

D. Paper Contributions

This study presents our methodology to improve the accuracy of predictive models and characterize external factors by analyzing and integrating longitudinal survey datasets with voluminous auxiliary dataset. Our specific contributions include:

1) We have designed a geospatial data integration framework to provide effective estimation of values with associated data uncertainty from the geospatial integration.
2) Support for interpretability of predictive analytics by identifying, ranking, and prioritizing core features that contribute to potential obesity.
3) Predicting personal BMI levels using artificial neural networks, gradient boosting and random forests with integrated attributes. The resulting model identifies characteristics of a particular factor that is likely to contribute to potential obesity or being overweight in the future.
4) Assimilation of data uncertainty during geospatial integration for predictive models.
5) Our approach avoids repeated I/O access to the raw datasets and generates intermediate integrated data for rapid subsequent model fitting.

E. Paper Organization

The rest of the paper is organized as follows. Section II outlines the longitudinal survey datasets used in this study, a short description of our distributed computing environment. Section III describes our uncertainty estimation method, followed by an overview of our obesity-level prediction model with data integration in Section IV. Section V provides a thorough evaluation of our methodology, followed by the related work in Section VI. Finally, conclusions and future research directions are described in Section VII.

II. Background

The Body Mass Index (BMI) is a measure of body fat, defined as the weight (in kilograms) divided by the square of the body-height (in meters). The range of the BMI is frequently used to determine a persons obesity level [16], [17]. A person's BMI could fluctuate over time due to several biometric, economic, environmental and familial factors. Building a predictive model for an individual's future BMI requires us to track these aspects over a large span of time, which can be challenging.

A. Data Selection

In order to predict future BMI for children over the US, the datasets we consider as candidates for integration need to satisfy a few conditions. First, the dataset should have a large geographic coverage, as that would feed a wide range of individuals from different regions/demographic to the model and reduce bias. Second, the datasets should have intersecting attributes that would facilitate their integration properly.

We have explored several longitudinal datasets as candidates, which contained information for surveyed individuals relating to their biometric, economic, geographical, familial aspects, to name a few. The candidate datasets were National *Longitudinal Study of Youth 97 (NLSY97) [18], NEXT Generation Health Study (NEXT) [19], National Longitudinal Study of Adolescent to Adult Health (Add Health) [20] and US Census data [21].*

The NLSY97 was one of the largest datasets that tracks 8,984 cohorts since 1997, originating from 6,819 unique households, selected via screening. It provides a relatively large geospatial coverage(338 US counties) with well defined geographic information for each of its participants. Therefore, due to both its large number of participants and its geospatial coverage, we have selected NLSY97 as the primary dataset.

The AddHealth data, despite having a good geographical coverage, deidentifies the state and county codes for each candidate. On the other hand, the NEXT dataset provides precise geographic locations of its candidates using a zipcode. However, its geospatial coverage was unsuitable for geospatial integration with other datasets with large coverage. On analysis, we found that in contrast to the 338 US counties that are covered by NLSY97's participants, NEXT covers only 52, with only $\sim 12.5\%$ of the NLSY97 data-points intersecting geospatially with NEXT.

Therefore, we have selected the Census dataset to be our auxiliary dataset to explore factors that are not tracked in the NLSY dataset. Since participants of these surveys are not the same, it is difficult to merge two records based on aspects such as biometric, behavioral, economic etc. One possible aspect on which they can be merged is using their geographic location, if the dataset being merged with the NLSY97 contains environmental information of the area an NLSY97 participant hails from. The Census dataset was such a dataset that had the proper geographic coverage (all of US) and had environmental, economic,familial and demographic information for each location with resolution up to block-level and could be summarised to provide relevant information regarding a cohort's area of origin.

B. Datasets

1) The National Longitudinal Study of Youth (NLSY97): The NLSY97 is funded by the Bureau of Labor Statistics and seeks to provide data on youth investments in education, training, labor force experiences, government program participation, health practices, household and geographical context variables, attitudes, expectations, family processes, and other factors, heavily influenced by the behavior of the labor market. The cohort of 8,984 adolescents has been surveyed 16 times, interviewed biennially starting from 1997 when respondents were aged 12-16, to the most recent one from round 16 which took place from 2013-2014. Data from round 16 includes 7,141 ($\sim 80\%$) of respondents included in the first sample. 4,599($\sim 51\%$) of respondents are male, and 4,385($\sim 49\%$) are female. For the first round of interviews, 1,771 respondents were 12 (born in 1984), 1,807 respondents were 13 (1983), 1,841 respondents were 14 (1982), 1,874 respondents were 15 (1981), 1,691 were 16 (1980).

2) The US Census Data 2000: The Census data(for year 2000) [21] is the summarization of US population over regions on factors such as employment, crime, health,demographic, consumer expenditures, housing etc. on a decennial basis by the United States Census Bureau. The Census data covers all 50 states of USA and comes at different summary levels, each providing summarisation over regions of different magnifications going down till block-group level. For the purpose of our research, we have used the summary file 3(nearly 100 GB in size), which provides magnification down to the block-groups in the US.

C. Distributed Computing Environment

We leverage the Apache Hadoop framework [22], [23] to provide scalable, fault-tolerant computing over a cluster of machines. Hadoop is used for writing Map-Reduce programs for processing large datasets stored over distributes file systems like HDFS,S3 etc. We use Hadoop Map-Reduce to get an aggregate value for different features, along with their standard deviations over each county over the US from the block-level attributes in the Census data. We could have used faster alternatives to Hadoop such as Apache Spark, capable of performing in-memory computations. But since the summarization and integration with Census dataset is a one-time process, Hadoop Map-Reduce should suffice.

III. Attribute based Uncertainty Estimation for Geospatial Data Integration

Importing geospatial attributes from an external dataset introduces uncertainty due to the mismatch of characteristics such as geospatial coverage and/or temporal range. The imported data is often estimated using interpolation or aggregation algorithms. Our study focuses on datasets with hierarchical geospatial demarcations such as political boundaries that are popularly used in longitudinal studies.

We propose a methodology, *attribute-based uncertainty estimation for data integration* ***(AUEDIN)***, that estimates uncertainty for each imported value to quantify the risk of assimilating imported attributes for subsequent computations such as a model generation.

Suppose dataset A imports a set of attributes from dataset B over a matching geospatial attribute, k, that is included in both datasets A and B. If records in the imported data(from dataset B) has higher geospatial precision, the records with the larger coverage(from dataset A) geospatially intersect with multiple records (from dataset B). In such a scenario, we calculate an aggregate of attribute values (for all records in B that intersect geospatially with a record in A) and use this as the imported value in the integrated data along with uncertainty of these imported records using the following method.

Let dataset A be a set of m geospatial units, A_0, A_1,..., A_{m-1} and their corresponding attributes. For a set of attributes of the dataset A, $S_A = attr_{A_1}, attr_{A_2}, attr_{A_3}, ...$, the value of an attribute $attr_{A_i}$, at the geospatial unit A_j, is denoted as $val(attr_{A_i}, A_j)$. Similarly, for dataset B with n geospatial units, B_0, B_1,..., B_{n-1}, each geospatial unit is mapped to a set of attributes values,$S_B = attr_{B_1}, attr_{B_2}, attr_{B_3}, ...$, and the value of an attribute $attr_{B_i}$ of the geospatial unit B_j is denoted as $val(attr_{B_i}, B_j)$. Let us assume that a geospatial unit A_i overlaps with a set of n' units in B, $\{B_0,B_1,...,B_{n'-1}\}$.

We estimate the imported value from B using weighted data aggregation based on the sample distribution. Assume that geospatial units in B, B_0, B_1,..., B_{n-1}, contains sample counts, C_0, C_1,...,C_{n-1} respectively. We calculate the weights, $\{W_0,W_1,...,W_{n'-1}\}$ for each geospatial unit in $\{B_0,B_1,...,B_{n'-1}\}$ as,

$$W_k = C_k/(\sum_{j=0}^{n'-1} C_j)$$

Once we calculate weights and normalize them, the aggregated attribute value of the record for attribute B_r, to be integrated with record A_i , denoted by $val(attr_{B_r}, \{B_0, B_1, ..., B_{n'-1}\})$, is calculated using the following formula:

$$val(attr_{B_r}, \{B_0, B_1, ..., B_{n'-1}\}) = \sum_{j=0}^{n'-1} (val(attr_{B_r}, B_j) \times W_j)$$

Each aggregated value is delivered with the associated uncertainty estimate, calculated using weighted standard deviation

as follows:

$$\sigma_{ri} = \sqrt{\frac{n' \times \sum_{j=0}^{n'-1}((val(attr_{B_r}, B_j) - \mu_{B_{ri}})^2 \times W_j)}{(n'-1) \times \sum_{k=0}^{n'-1} W_k}}$$

where $\mu_{B_{ri}} = val(attr_{B_r}, \{B_0, B_1, ..., B_{n'-1}\})$ and σ_{ri} denotes uncertainty associated with merging A_i with aggregate value of the attribute B_r from dataset B calculated as above.

In Section IV, we apply **AUEDIN** for integrating multiple datasets from longitudinal studies and models to improve the accuracy of the obesity-level prediction model.

IV. Generating the Obesity-Level Prediction Model with Data Integration

A. Preliminary Analysis

Our goal is to predict an individual's potential obesity down the line, given certain available information about him/her in the current day. As a preliminary analysis, we have considered the growth charts proposed by organizations such as the Centers for Disease Control and Prevention (CDC) [24] and World Health Organization (WHO) [25] to estimate BMI, relying only on individuals' biometric information- weight, height and age.

The growth chart is group of graphical measurements predicting the progression of weight and height among children from their birth till the age of 20 [24]. At any point in time, a child's age and the BMI is plotted to a point on the graph. The percentile curve which is the closest to the point is used to estimate the growth pattern for child over the next few years as shown in Fig.1

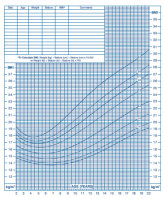

Fig. 1: **CDC growth chart of BMI progression with age for American boys aged 2-20 years.**

We used the CDC 2000 growth chart for BMI on participants of NLSY97 to see how effectively it predicts the BMI N years from day 1 of their interview, compared to the actual measurements (we have tested for N=1,2 and 3) recorded in NLSY97. Table I shows the results of that test.

TABLE I: **RMSE for BMI predictions(for year 1997) using CDC Growth Chart**

	1 Yr	2 Yr	3 Yr
Male	3.52	3.38	3.40
Female	3.72	4.25	3.94

The prediction using the CDC growth chart demonstrates a sizable error range that can impact the detection of obesity. If the BMI is between 25 and 29.9, the individual is considered overweight, BMI of 30 or over is considered obese. We believe this happens because a child's growth velocity can jump from one percentile curve to another during his/her developing years due to a number of external factors such as stress, family problems, genetic and chronic diseases. Our goal is to come up with a prediction model that can make better predictions of BMI by incorporating the growth chart data along with external factors from other available datasets.

B. Integrating External Factors

The external factors that we want in our input data along with the biometric information, are behavioral, economic, familial and environmental. The NLSY97 survey has a good variety of questions relating to various behavioral, familial and biometric aspects of an individual's life but lacks a good collection of environmental and economic attributes. The US Census data consists of a set of records relating to the country's demographic distribution,age, income, family structures down to the block-group level. Using this dataset, we can form summary of areas of the country on factors such as median income, household size, percentage of population based on ages, family structures and so on. Thus in order to incorporate the missing aspects in NLSY97 data, we have considered the summarised US Census data for integration.

Integrating Datasets Based on Geospatial Proximity: Between the NLSY97 and the Census 2000 datasets, the only common attributes suitable for a merge is the geospatial attribute. This, however, introduces a few challenges to the aggregation process. First, the maximum resolution of the Census datasets is at a block-group level, while NLSY97 is county level. Second, individual participant's data is not available from Census, only block-level aggregates are available. Therefore, the best way to combine the two datasets is to use the county information of each individual. However, since we are trying to combine an individual's record on the NLSY side with an aggregate of county-level information from the Census dataset, this will introduce some level of uncertainty in integrated Census information. As explained later, we have attempted to estimate this uncertainty and use it in our machine learning model to enhance its prediction accuracy.

Fig.2 gives a pictorial description of the mapping between the data-points of both the datasets. The circular regions colored red, green and blue represent three counties A,B and C. Now, on the NLSY97 side, there can be multiple participants who come from the same county whereas, with the Census aggregate, there would be a single record per county. Thus, geospatial integration between the NLSY97 and Census data would result in a many-to-one mapping.

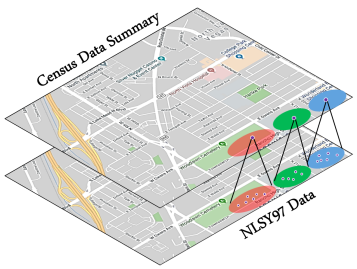

Fig. 2: **Strategy for merging Census 2000 & NLSY97 data**

C. Data Pre-Processing and Feature Selection

The NLSY97 dataset was collected by conducting a series of interviews with each participant over a period of twenty years. The responses are a mixture of quantitative and non-quantitative values. Some non-quantitative attributes (e.g. race or gender of the person) may have a set of choices but have no numerical significance. We re-constructed non-quantitative responses as a set of propositions with associated boolean responses. This re-construction of non-quantitative questions increased the number of attributes: The NLSY97 data has over 67,000 attributes in total. This input vector was large enough to suffer from the *curse of dimensionality* [26].

We reduced the dimensionality of the input vector using the Lasso algorithm [27] and performed Gradient Boosting [28] and Random Forest [29] to explore important features. We avoided PCA to simplify interpretability of model.

D. Estimating uncertainty

In our case, the Census dataset provides higher geospatial precision(block-group level) than NLSY97 dataset (county level). Attribute integration from the Census dataset involves data aggregation to match the lower geospatial precision in NLSY97.

We have used our proposed approach, AUEDIN, to aggregate values and estimate uncertainty. The population information included in the Census dataset is used to capture the distribution of samples. If an imported attribute is selected (see the section IV-3), the model input vector contains the uncertainty estimate for that attribute. Uncertainty estimates for eliminated attributes are not included in the model.

Overview of Model Building: The input data to our predictive model went through the following stages. Initially, we started out with the NLSY97 data and used simply those features available in that dataset to make our predictions. In order to improve on this initial accuracy, we introduced a new feature to our training dataset- the prediction of BMI after n years, using the CDC growth chart given the stats of the individual at the current date. In the next phase, we combined the input dataset with the county-wise summary from the Census dataset and used that to train our model. Finally, along with this data, we used a vector of uncertainty estimates for each selected feature and used it to train our model.

E. Uncertainty Aware Modelling

Since we only have an estimate of the error for integrated values from Census data, we tried out two approaches. First, we have used the integrated data, with potential errors as it is and applied machine learning models on it. In case the data has relatively low noise, the prediction algorithm might tune itself to the noise, thus becoming resistant. We have applied two commonly used machine learning models: *feed-forward neural networks*, to attempt deep learning and *gradient boosting*. We tried out gradient boosting here because, even if the data-noise affects a single model, combining multiple models as in ensemble methods, might help mitigate its effect.

In our second approach, we have incorporated the uncertainty estimates into our machine learning algorithm. Here, we would be working with two input vectors - one for the actual integrated dataset(d_{int}) and the other representing the error in measurement of each of the feature value in that integrated dataset (e_{int}). Evidently, both d_{int} and e_{int} must have the same dimension.

We assume that the features from the NLSY97 dataset would all have zero measurement errors, i.e. the values of variables acquired from interview of the participants are all assumed to be correct. Only the values of features derived and integrated from Census must have possible error associated with them. So, we can create an uncertainty matrix with the same dimension as the input vector, where all columns corresponding to NLSY97 features would have 0 values and only columns associated to Census attributes would have corresponding weighted standard deviation for that measurement.

For the purpose of modeling with uncertainty, we have used two different techniques to test which performs better. In the first approach, we have performed feature selection on d_{int} using Lasso Regression and then used the N(We have tried N=12,15,20) most important features, joined with the matrix representing their corresponding measurement error as an input vector to a neural network. In our second approach, we have performed feature selection as before, but this time instead of clubbing two vectors into one input vector, we have used Maximum-Likelihood Principal Component Regression(MLPCR) [30], which takes the input vector and a vector of measurement standard deviations as its input.

The reason we have used MLPCR instead of PCR(Principal Component Regression) is that although PCR is effective for quantitative analysis of multicomponent mixtures, it relies on the Singular Value Decomposition(SVD) to obtain a reliable estimation of a p-dimensional subspace. When the measurement errors in the input are all *iid normal*, the p-dimensional hyperplane determined by SVD will be an optimal model for the data in a maximum likelihood sense. However, in case the measurement errors are not independent with uniform variance, the p-dimensional estimation will not be optimal. MLPCR provides two advantages. First, it allows inclusion of measurement uncertainties, assuming the errors are uncorrelated, into the calibration process. Second, it provides a maximum likelihood estimate of the PCA [31] model, which is generally superior to that obtained through SVD.

V. Evaluation

We have used separate models for predicting BMI for males and females. This is because two people with the same BMI can have very different body compositions. This is especially true when comparing males and females because women typically have a higher percentage of body fat than men. We have noticed an increase in accuracy with models trained on gender-segregated data.

A. Experimental Setup

1) Data Pre-processing: We will be working with our datasets at 3 different stages. The first stage is simply the information available from the NLSY97 data, clubbed with the predictions from the CDC Growth Chart from year 2000. This was a simple data integration step where we appended a new column representing Growth Chart's BMI predictions 3 years from the current interview date.

2) Data Integration: In the second stage, we have integrated the data from the previous stage with the Census data. The calculations regarding summarization and uncertainty estimation and integration with the NLSY97 data were carried out in consecutive Map-Reduce tasks which were executed on Hadoop version 2.7.3 with the OpenJDK JVM, version 1.8.0 92. The Map-Reduce operations for this study were performed on a cluster of 20 HP Z420 servers (8- core Xeon E5-2560V2, 32 GB RAM, 1 TB disk).

3) Training and Testing of Predictive Models: In the final stage, we have used the Python's Scikit Learn library [32] for most of our machine learning processes. Since the data at this point is relatively small in size(nearly 200 Mb), the training and testing phases are carried out on a single machine whose configuration is the same as that of each cluster node.

B. Scalability Evaluation

Due to the large size of the Census dataset and the NLSY dataset, by adopting Hadoop framework, we are able to profit from its scalability feature. For experimental purposes, we gradually adjusted the number of nodes in the Hadoop cluster, starting from 2 nodes, till 20. The result of the experiment is shown in Fig.3. Fig.3, shows that with increasing cluster size,

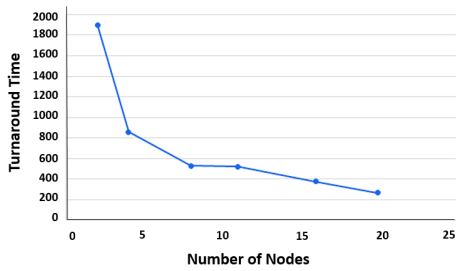

Fig. 3: **Turnaround time with increasing cluster size**

the execution time for the job decreased, meaning addition of machines to the cluster does improve the performance.

C. Experimentation and Accuracy Evaluation

1) Experiment 1 (Effect of data size): In the first step of our experimentation, we wanted to check whether using greater number of data-points does actually improve the prediction accuracy, as that would mean that the training data is properly distributed. We have first taken the NLSY97 data, appended with the Growth Chart predictions and truncated and fed it to two different machine learning models(deep learning model and an ensemble model) in an incremental fashion using Scikit Learn's `neural_network` package and `ensemble`

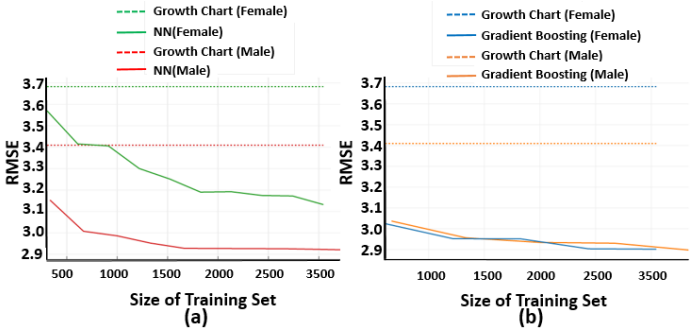

Fig. 4: **Prediction RMSE with change in training data size for (a) Neural Network and (b) Gradient Boosting Models**

package(for Gradient Boosting). The data-set is first randomized and 10% of it is kept aside for testing purpose, so that training data remains constant and then the remaining 90% data points are fed as training data to a machine learning model incrementally to see if larger training data size helps improve accuracy. The RMSE we have reported in the Fig.4a is actually an average RMSE over 50 separate trials of training and testing the data.

As we can see from the Fig.4a, the prediction RMSE for both models for male and female BMI prediction goes down as the size of the training dataset increases. The dotted lines are for when only the Growth Chart curves are used to make a prediction for males and females. Similar results have been observed when we have used an ensemble method(Gradient Boosting) as can be seen in Fig.4b.

2) Experiment 2 (Effect of Data Integration without Uncertainty Estimates): In the next step of our experimentation, we evaluate the effect of integrating new features from the Census dataset. Our expectation is that some of the features integrated from the new dataset could turn out to be important and thus enhance our model's predictive capability.

Fig.5 shows the change in RMSE of BMI predictions with different models and input data. It is to be noted that the three bars(orange, green and red) in Fig.5 represent the results from doing a feature selection using Lasso Regression [27] and then training the top N features(we have tried with N=12,15,20) with an artificial neural network. Feature selection by Gradient Boosting by taking the most important 15 features and then applying artificial neural network gave similar results, but features selected by Lasso Regression were better. The RMSE values are the result of a K(=10) Fold validation on the input dataset.

As evident from the Figure5, inclusion of Census information does increase the prediction accuracy. The prediction RMSE for men comes down to ∼2.99 and that for women comes down to ∼3.23 in the case of Lasso Regression followed by training on neural networks. Also, from table III, we can see that the prediction RMSE with integrated features using Gradient Boosting also decreases to ∼2.9 for male and to ∼3.06 for female.

Another aspect of this experimentation is to check whether the features integrated from the Census dataset are actually important enough to contribute to the BMI prediction. Table

II shows a list of top 10 features that were selected using Lasso Regression in one of our experiments. We can see a mixture of attributes from both NLSY97 and the Census data summary make the list as the top influencers in BMI prediction. Also, some of the behavioral and biometric factors that were chosen as the top features intuitively make sense.

TABLE II: **Results from Feature Selection on Integrated Data**

	Selected Features
NLSY97	Predicted BMI Weight Of Participant(lbs) Is Participant Limited by Missing/Deformed Body Part Participant is Unhappy,Sad or Depressed Is Participant Limited By Sensory Problems Is Participant Limited By Mental Conditions
Census	Percent Families With Children Above 18Y/O Percentage of Seniors in Area Percentage of Married People in Area Average Household Size

D. Effect of Data Integration with Uncertainty Estimates

The final phase of our experimentation involves using the uncertainty information we had generated from the Map-Reduce task into our learning models to see if we get any improvement in prediction accuracy. In order to do so, we tried two approaches.The first approach is similar to what we did previously, i.e. we tried feature selection and then applied neural network. The difference is, here, along with the selected features, we appended the column representing the uncertainty value of that feature. For feature selection, we have found that using Gradient Boosting and taking out the top 15, along with their uncertainties gave us the best result. The uncertainty column is added only if it is for a feature relating to the geographically higher resolution dataset(Census), because the measurement uncertainty for NLSY97 is considered to be 0. In Fig.5, the red bar represents the output from this experiment and it shows that there is considerable improvement in terms of RMSE using this method both for male and female.

TABLE III: **Prediction RMSE for Gradient Boosting Models on Different Data**

	RMSE(Male)	**RMSE(Female)**
NLSY+Growth Chart	3.0387	3.1999
NLSY+Growth Chart+Census	2.889	3.0599

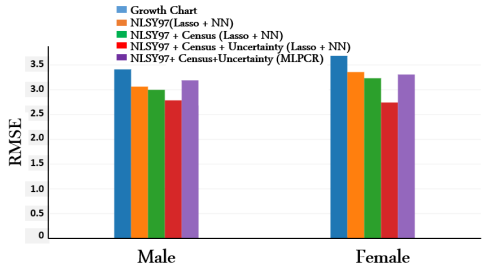

Fig. 5: **Comparison in RMSE for different datasets**

In the second approach, we tried out the MLPCR model, which allows us to include information on measurement uncertainties in its learning process. However, as we can see from the Fig.5(last bar in the clustered bar chart), we do not get an improved RMSE. A possible explanation for this could be that MLPCR, which relies on Maximum Likelihood Principal Component Analysis(MLPCA) requires that the measurements have uncorrelated errors. Since the errors/weighted standard deviations are the same for each measurement for each county, the condition for uncorrelated error condition for each measurement is not satisfied.

To summarize, we see that, while the RMSE of prediction using Growth Chart was 3.4 and 3.68 for male and female respectively, our integrated training data was able to reduce RMSE to 2.99 for male($\sim$8.9% improvement) and 3.23 for female($\sim$12.3% improvement) and further incorporation of uncertainty measures took the RMSE to 2.78 for male($\sim$18.3% improvement) and 2.74 for female($\sim$25.6% improvement).

VI. Related Work

Several research have dealt with both computing uncertainty for integrated data and childhood obesity prediction. The calculation of a point estimate for the value of an object of interest from a set of claims has been covered in various research. [33] [34] dealt with not only finding out the point value of an object of interest from a given set of claims but also with quantifying a confidence interval for the estimate based on the calculated reliability score of each of the sources, along with the number claiming sources.

Other related work on interpolating values of unknown locations based on values from related surrounding regions have been covered in works such as [35]. The focus of this work has been to use the inverse-distance weighting interpolator (IDW), with cross-validation as a method of predicting the unknown value of a parameter. This was followed by a subsequent jackknife resampling was then used to reduce bias of the predictions and estimate their uncertainty.

Current works that predict obesity are mainly focused on the following three aspects: 1) *Connecting obesity with external environment factors*. For instance, [10] applied discrete-time hazard models on SECCYD dataset, showing associations between early life poverty and adolescent obesity. [11] examined the relationship between family environment and adolescent obesity [11]. 2) *Connecting obesity with genetic predisposition*. [11] showed that overweight adolescents in the study tended to have parents with problematic weight history. [12] concluded that models based on traditional predictors, such as family history of obesity, outperform those based solely on genetic predictors. 3) *Connecting obesity with human behavior and their lifestyle*. [13] conducted a systematic review to examine the relationship between obesity and obesity-related behaviors such as disinhibited eating, sedentary activity, and lower physical activity. [14] tried to explore the evidence for food addiction and its role in the rise of obesity in youth.

Efforts have explored the use of queries to launch analytic tasks [36], [37]; these queries performed targeted analytics and are not designed to support general analytic operations. Budgaga et al have explored the use of parameter space sampling and the use of ensemble methods to construct models over spatiotemporal phenomena [38]. Unlike our effort, this approach does not entail data integration. The Synopsis

[39] system constructs sketches of spatiotemporal data that can be subsequently used to construct synthetic datasets for particular portions of the feature space. This is synergistic with our methodology and can be used to support interpolation operations. Our methodology can also interoperate with spatiotemporal data storage systems [40], [41]. Efforts to integrate data from diverse data sources have explored use of metadata as the basis for integration [42]. However, metadata focused efforts do not perform uncertainty quantification.

VII. Conclusions and Future Work

We presented our methodology to predict obesity in adolescents using a longitudinal survey data,augmented with attributes integrated from ancillary datasets, that will allow doctors and parents to make targeted recommendations to help children make better choices to mitigate future health risks.

1) Research Question 1(RQ1): Integration of ancillary datasets must account for the spatiotemporal characteristics of auxiliary data to align them with the primary dataset. In cases where the spatial resolution is finer-grained than the longitudinal datasets, aggregations can ensure representatives. To integrate attributes from US Census to NLSY97 dataset, we have calculated weighted means of attribute values based on the distribution of the population.

2) Research Question 2(RQ2): Uncertainty introduced by data aggregation must be accounted for while computing representative estimates for larger spatial scopes. This uncertainty calculation must be scaled with the spatial or temporal range. Our estimates are amenable to calculation using MapReduce framework as demonstrated in our benchmarks.

3) Research Question 3(RQ3): Assimilating features from related ancillary datasets and accounting for uncertainty estimates improves model accuracy, as seen using models such as Neural Network, Gradient Boosting and Random Forests. In contrast to models using only biometric attributes, our methodology improved model accuracy by 8.9-10.2% on including behavioral attributes; by $\sim$12% on also considering environmental attributes, and by 18.3-25.6% on also considering uncertainty estimates.

As part of our future work we plan to explore integrating datasets that do not share geospatial aspects. We will also explore extending our uncertainty measurements to cope with attributes that are not geospatially related.

VIII. Acknowledgment

This research has been supported by funding from the US Department of Homeland Security (D15PC00279), the US National Science Foundations Advanced Cyberinfrastructure (ACI-1553685), and Colorado School of Public Health at Colorado State University (2016 Seed Grant).

References

[1] C. L. Ogden, S. Z. Yanovski, M. D. Carroll, and K. M. Flegal, "The epidemiology of obesity," *Gastroenterology*, vol. 132, no. 6, pp. 2087–2102, 2007.

[2] N. C. for Health Statistics (US *et al.*, "Health, united states, 2004: With chartbook on trends in the health of americans," 2004.

[3] E. A. Finkelstein, J. G. Trogdon, J. W. Cohen, and W. Dietz, "Annual medical spending attributable to obesity: payer-and service-specific estimates," *Health affairs*, vol. 28, no. 5, pp. w822–w831, 2009.

[4] J. A. Gazmararian, D. Frisvold, K. Zhang, and J. P. Koplan, "Obesity is associated with an increase in pharmaceutical expenses among university employees," *Journal of obesity*, vol. 2015, 2015.

[5] N. K. Güngör, "Overweight and obesity in children and adolescents," *Journal of clinical research in pediatric endocrinology*, vol. 6, no. 3, p. 129, 2014.

[6] S. Manna and A. M. Jewkes, "Understanding early childhood obesity risks: An empirical study using fuzzy signatures," in *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1333–1339.

[7] C. Riedel and R. e. a. von Kries, "Overweight in adolescence can be predicted at age 6 years: a cart analysis in german cohorts," *PloS one*, vol. 9, no. 3, p. e93581, 2014.

[8] K. E. Bevelander, K. Kaipainen, R. Swain, S. Dohle, J. C. Bongard, P. D. Hines, and B. Wansink, "Crowdsourcing novel childhood predictors of adult obesity," *PloS one*, vol. 9, no. 2, p. e87756, 2014.

[9] X. Wen, K. Kleinman, M. W. Gillman, S. L. Rifas-Shiman, and E. M. Taveras, "Childhood body mass index trajectories: modeling, characterizing, pairwise correlations and socio-demographic predictors of trajectory characteristics," *BMC medical research methodology*, vol. 12, no. 1, p. 38, 2012.

[10] H. Lee, M. Andrew, A. Gebremariam, J. C. Lumeng, and J. M. Lee, "Longitudinal associations between poverty and obesity from birth through adolescence," *American journal of public health*, vol. 104, no. 5, pp. e70–e76, 2014.

[11] L. M. Hooper, J. J. Burnham, R. Richey, J. DeCoster, M. Shelton, and J. C. Higginbotham, "The fit families pilot study: preliminary findings on how parental health and other family system factors relate to and predict adolescent obesity and depressive symptoms," *Journal of Family Therapy*, vol. 36, no. 3, pp. 308–336, 2014.

[12] R. J. Loos and A. C. J. Janssens, "Predicting polygenic obesity using genetic information," *Cell Metabolism*, vol. 25, no. 3, pp. 535–543, 2017.

[13] J. Liang, B. Matheson, W. Kaye, and K. Boutelle, "Neurocognitive correlates of obesity and obesity-related behaviors in children and adolescents," *International journal of obesity (2005)*, vol. 38, no. 4, p. 494, 2014.

[14] H. L. Yardley, J. Smith, C. Mingione, and L. J. Merlo, "The role of addictive behaviors in childhood obesity," *Current Addiction Reports*, vol. 1, no. 2, pp. 96–101, 2014.

[15] H. Dong and Yu, "Data integration with uncertainty," *Proceedings of the 33rd international conference on Very large data bases*, pp. 687–698, 2007.

[16] "Defining adult overweight and obesity." [Online]. Available: https://www.cdc.gov/obesity/adult/defining.html

[17] "Defining childhood obesity." [Online]. Available: https://www.cdc.gov/obesity/childhood/defining.html

[18] "The national longitudinal study of youth 97 (NLSY97)." [Online]. Available: https://www.bls.gov/nls/nlsy97.htm

[19] "Next generation health study (NEXT)." [Online]. Available: https://www.nichd.nih.gov/about/org/diphr/hbb/research/Pages/next.aspx

[20] "National longitudinal study of adolescent to adult health (add health)." [Online]. Available: http://www.cpc.unc.edu/projects/addhealth

[21] "Census 2000 data for the united states." [Online]. Available: https://www.census.gov/census2000/states/us.html

[22] T. White, *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.

[23] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar *et al.*, "Apache hadoop yarn: Yet another resource negotiator," in *Proceedings of the 4th annual Symposium on Cloud Computing*. ACM, 2013, p. 5.

[24] "Cdc growth charts." [Online]. Available: https://www.cdc.gov/growthcharts/index.htm

[25] "WHO growth charts." [Online]. Available: https://www.cdc.gov/growthcharts/who_charts.htm

[26] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*. Springer, 2011, pp. 257–258.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[28] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[29] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[30] P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, "Maximum likelihood multivariate calibration," *Analytical chemistry*, vol. 69, no. 13, pp. 2299–2311, 1997.
[31] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
[32] Scikit-learn. [Online]. Available: http://scikit-learn.org/stable/
[33] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng, and A. Zhang, "Towards confidence in the truth: A bootstrapping based truth discovery approach," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.
[34] H. Xiao, J. Gao, Z. Wang, S. Wang, L. Su, and H. Liu, "A truth discovery approach with theoretical guarantee," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.
[35] M. Tomczak, "Spatial interpolation and its uncertainty using automated anisotropic inverse distance weighting (idw) - cross-validation/jackknife approach," *Journal of Geographic Information and Decision Analysis*, vol. 2, no. 2, pp. 18–30, 1998.
[36] M. Malensek, S. Pallickara, and S. Pallickara, "Analytic queries over geospatial time-series data using distributed hash tables," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 6, pp. 1408–1422, 2016.
[37] ——, "Fast, ad hoc query evaluations over multidimensional geospatial datasets," *IEEE Transactions on Cloud Computing*, vol. 5, no. 1, pp. 28–42, 2017.
[38] W. Budgaga, M. Malensek, S. Pallickara, N. Harvey, F. J. Breidt, and S. Pallickara, "Predictive analytics using statistical, learning, and ensemble methods to support real-time exploration of discrete event simulations," *Future Generation Computer Systems*, vol. 56, pp. 360–374, 2016.
[39] T. Buddhika, M. Malensek, S. L. Pallickara, and S. Pallickara, "Synopsis: A distributed sketch over voluminous spatiotemporal observational streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 11, pp. 2552–2566, 2017.
[40] M. Malensek, S. L. Pallickara, and S. Pallickara, "Galileo: A framework for distributed storage of high-throughput data streams," in *Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on*. IEEE, 2011, pp. 17–24.
[41] M. Malensek, S. Pallickara, and S. Pallickara, "Polygon-based query evaluation over geospatial data using distributed hash tables," in *Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing*. IEEE Computer Society, 2013, pp. 219–226.
[42] S. L. Pallickara, S. Pallickara, M. Zupanski, and S. Sullivan, "Efficient metadata generation to enable interactive data discovery over large-scale scientific data collections," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*. IEEE, 2010, pp. 573–580.