# Guiding public health policy by using grocery transaction data to predict demand for unhealthy beverages *

Xing Han Lu**[1,2], Hiroshi Mamiya**[1]✉, Joseph Vybihal[2], Yu Ma[3], and David L. Buckeridge[1]

[1] Surveillance Lab, McGill Clinical and Health Informatics, Montreal, Canada
[2] School of Computer Science, McGill University, Montreal, Canada
[3] Desautels Faculty of Management, McGill University, Montreal, Canada
{xing.han.lu, hiroshi.mamiya}@mail.mcgill.ca
{joseph.vybihal, yu.ma, david.buckeridge}@mcgill.ca

**Abstract.** Sugar-Sweetened Beverages (SSB) are the primary source of artificially added sugar and cause many chronic diseases. Taxation of SSB has been proposed, but limited evidence exists to guide this public health policy. Grocery transaction data, with price, discounting and other product attributes, present an opportunity to evaluate the likely effects of taxation policy. Sales are non-linearly associated with price and are affected by the prices of multiple competing brands. We evaluated the predictive performance of Boosted Decision Tree Regression (B-DTR) and Deep Neural Networks (DNN) that account for the non-linearity and competition, and compared their performance to a benchmark regression, the Least Absolute Shrinkage and Selection Operator (LASSO). B-DTR and DNN showed a lower Mean Squared Error (MSE) of prediction in the sales of major SSB brands in comparison to LASSO, indicating a superior accuracy in predicting the effectiveness of SSB taxation. We have demonstrated how machine learning methods applied to large transactional data from grocery stores can provide evidence to guide public health policy.

**Keywords:** Public health informatics · Machine learning · Public health policy · Public health policy · Grocery transaction data · Taxation · Obesity · Sugar sweetened beverages · Public health nutrition.

## 1 Introduction

Unhealthy diet is the leading preventable cause of global death and disability, claiming 11 million lives and 241 million disability adjusted lost life years in 2012 [6]. Diet-related chronic diseases, such as obesity, cardiovascular diseases, cancers, and type-2 diabetes mellitus impose a considerable burden on society and

* This work was supported by the Public Health Agency of Canada

** The following authors contributed equally to this work individuals. Taxation has been proposed as a public health policy to discourage the purchasing of unhealthy foods [14], most notably Sugar Sweetened Beverages (SSB), which are the primary source of artificially added sugar with an established epidemiological association with obesity and major chronic diseases [5, 9]. SSB consists of beverages such as soda (carbonated soft drinks), fruits drinks, sports and energy drinks each containing many product brands (e.g. Coca-Cola and Pepsi in the category of soda). The expected effectiveness of taxation is determined by the magnitude of reduction in SSB purchasing likely to occur in response to an increase in the price of SSB. Formally, this key quantity is called the *price elasticity of demand* and quantified as the percent reduction in product purchased in response to a one percent increase in price.

Grocery transaction data can be used to predict SSB sales conditional on pricing, promotions and consumer demographic and economic attributes of the store neighborhood (e.g. income and family size). Because sales of a product are influenced by its features (*focal features*), but also by the features of competing products in the same store (*competing features*), the prediction of beverage purchasing must account for the influence of numerous competing brands. Due to correlations in price and promotion across many food products, feature selection is critical. Researchers previously performed ad-hoc dimensionality reduction, such as aggregating product sales and features into broader SSB categories or modeling only a small number of brands[1]. These approaches masks the complex patterns of competition among individual food products, emphasizing the importance of prediction at the level of individual food items or brands.

More importantly, associations between product features and sales are non-linear (i.e. deal-effect curve), and multiple product features can jointly affect sales through interactions due to competitive interference and synergistic effect of promotions [16]. While parametric estimators (e.g. linear regression) are traditionally used to model product demand, manual specification of non-linear functions and interactions is not feasible with dozens or hundreds of competing product features. In contrast, non-parametric algorithms, such as decision trees and artificial neural networks, naturally incorporate non-linear associations and interactions.

To date, SSB taxation is rarely implemented in developed nations, and the magnitude of consumer response to taxation on a large geographic scale (e.g. provincial and national scales) is for the most part unknown. Due to the paucity of real-world implementations, the main source of evidence about the likely effectiveness of SSB taxation is models that can predict the amount of the sales of SSB based on historical variation in price. We thus aim to provide computational approaches to evaluate the accuracy of non-parametric learning algorithms for predicting the quantity of SSB sales from scanner grocery transaction data.

## 2 Data

We obtained weekly transaction records of food products purchased from 44 stores sampled to be geographically representative of three large retail grocery chains in the province of Quebec, Canada between 2008 and 2013. The data were indexed by time (week), store identification code, product name, price, and three promotional activities: discounting, in-store display (placement of a product in a prominent location) and flyer advertising.

**Table 1.** Description of predictive features of SSB sales.

| Feature Description | Type |
|---|---|
| **Brand-level features** | |
| Chain code where product was sold | Categorical |
| Percent price discount (%) | Numerical |
| Prices in Canadian cents | Numerical |
| Display advertisement frequency | Numerical |
| Flyer advertisement frequency | Numerical |
| Brand name | Categorical |
| Store code where product was sold | Categorical |
| **Temporal features** | |
| Month of Sale | Categorical |
| Week of Sale | Categorical |
| **Store neighborhood features** | |
| Proportion of post-secondary certification | Numerical |
| Average family size | Numerical |
| Proportion of family with child | Numerical |
| Proportion of single parent family | Numerical |
| Median family income ($/family) | Numerical |
| Proportion of immigrants | Numerical |
| Number of dwellings (families) | Numerical |
| Total population (inhabitants) | Numerical |
| Dwelling density (families/km$^2$) | Numerical |
| **Target** | |
| Log of Weekly Sales of brand | Numerical |

There were 2,608 distinct SSB products defined by brand, flavor, and package type. As products in the same brand tend to exhibit similar pricing and promotional patterns, we aggregated the value of sales, pricing and promotion into a smaller set of 154 distinct SSB brands, such as Coca-Cola and Pepsi. Brand-level predictive features (i.e. price, discounting, display, and flyer advertisement) were calculated as the mean (price and discounting) and proportion promotion (display and flyer) across the products belonging to the brand.

Let $t := week$, $i := brand$, $j := store$. There were 1,509,280 weekly transaction records for the 154 SSB brands across all stores, with each record representing the brand-specific sales denoted as $Y_{ijt}$, which is the target variable and defined as the natural-log of the sales of brand $i$ in store $j$ at week $t$. The sales quantity was standardized to the U.S Food and Drug Administration serving size of 240 milliliters. Although the log transformation is relevant to parametric regression modeling [11], we applied this transformation in accordance with existing practice in demand modeling.

The vector of brand-level focal features is denoted as $X_{ijt}$ (Table 1, Brand-level features). We let $S_j$ be the categorical indicator of chain and store identification code and store neighborhood socio-economic and demographic features. We let $M_t$ and $W_t$ represent categorical features indicating the month and week for each record to account for temporal fluctuations in purchasing . As noted above, sales of a brand depend on the pricing and promotion of that brand (*focal brand features*) and on the features of popular competing brands (*competing brand features*). Because a few brands account for most of the market share in each SSB category (e.g. Coca Cola and Pepsi have nearly 70% of share in the soda category), their brand features have a strong influence on the sales of other brands. Thus, we extracted price and promotions of twenty brands with the highest market share among SSB that are denoted as $C_{kjt}$. The dimension of each feature vector was: ($X_{ijt}$, 245), ($C_{kjt}$, 80), ($S_j$, 9), ($M_t$, 12), and ($W_t$, 53).

We extracted the first five years (2008-2012) of the transaction data for training and validation. We randomly sampled 90% of these data as the training set for learning algorithm parameters, leaving the remaining 10% as the validation set for evaluating the prediction accuracy of the algorithms. The final year (2013) of data was reserved to estimate prediction accuracy, measured as Mean Squared Error (MSE). Data were managed using Numpy, Pandas and PostgreSQL.

## 3 Methods

We used two non-parametric methods: an ensemble of Decision Trees with Adaptive Boosting (B-DTR) and a fully-connected deep neural network (DNN). The baseline model was a regularized linear parametric model (LASSO, or Least Absolute Shrinkage Selection Operator). The DNN was implemented in Keras [3], and the other models were implemented in Scikit-Learn [13]. Normalization was done using standard mean shifting and variance scaling.

LASSO regression identifies a sparse set of features through shrinkage via $L_1$ regularization [1] [15] and was previously used for demand forecasting in high-dimensional feature space [12], even though explicit specification of non-linear features (e.g. spline) becomes unrealistic when modeling the sales of a large number of brands. We selected the regularization parameter $\lambda$ by iterating over a range of values and selecting the one with lowest average mean squared error (MSE) through three-fold Cross Validation.

Decision Tree Regression (DTR) is a rule-based learning algorithm that identifies a binary segmentation of predictive features, where the cut-point for each feature represents a decision boundary that minimizes the prediction loss (e.g sum of squared errors) for a target vector $Y_{ijt}$. The partitioning ends when pre-specified criteria, such as a maximum number of branches or a minimum number of observations at each terminal node, are met. We used Drucker's improved Adaptive Boosting [4] meta-estimator to form an ensemble of 100 weak learners. The weight of each learner was determined by a linear loss. Each learner was a Decision Tree with varying depths, set to a maximum depth of 30 nodes. The value of each node was determined by the partition that best minimized the MSE.

The Deep Neural Network (DNN) model with the best results had four fully connected layers. Adam optimization was used to enable convergence with large data and noisy gradients [10]. The optimum values of exponential decay rates and fuzzy factors were selected based on training stability and the ability to converge. The network weight parameters were initialized using Normalized Initialization [7]. We trained the model using mini-batches of 128 samples to leverage the richness of the data and to provide inherent regularization [2], while maintaining a stable training process. We chose the activation function to be a Rectified Linear Unit (ReLU) due to its biological properties and strong experimental results on high-dimension datasets [8], due in part to its non-linearity, which allows the DNN to learn complex relationships between features.

The DNN had an input layer dimension of 389, and fed a 400-dimension vector to the first hidden layer. The first hidden layer output a 100-dimension vector to the next layer with a $L_1$ regularization and ReLU activation. The last hidden layer output was a 25-dimension vector to the output layer. The final layer outputs a single numerical value corresponding to the predicted log of sales, using a linear activation function to take into account negative target values (brands with extremely low sales has negative log values).

**Table 2.** MSE of most popular brands of SSB

| | PEPSI | COCA COLA | SEVEN UP | CRUSH |
|---|---|---|---|---|
| B-DTR | **0.17** | **0.16** | 0.22 | 0.28 |
| DNN | 0.19 | 0.23 | **0.21** | **0.23** |
| LASSO | 0.51 | 0.44 | 0.46 | 0.35 |

## 4 Results

The Mean Squared Error (MSE) for the prediction of all SSB brands in the 2013 transaction data was 0.67, 0.72, and 0.91 for DNN, B-DTR, and LASSO, respectively. At the individual brand level, DNN, B-DTR, and LASSO showed best predictive performance for 80, 31, and 21 brands present in the test data, respectively. Prediction error of four most popular SSB brands driving overall sales of SSB is presented in Table 2. The DNN and B-DTR had comparable prediction accuracy for these brands, while LASSO showed the lowest accuracy except for the Nestle brand.

Using the most accurate predictive algorithm (DNN), we generated predictions of the percent reduction in SSB sales due to increases in beverage prices in reference to SSB sales with the observed price for a random sample of four stores in the 2013 test data (Figure 1). We present store-specific predicted effectiveness of taxation (i.e. price elasticity), since consumer demographic characteristics (e.g. income) around each store result in a varying level of price sensitivity, thus allowing public health researchers to identify neighborhoods where the taxation policy is least or most effective in reducing the sales of SSB. As an example, the store coded as 35973 (dotted line with the sharpest decrease of percent sales) exhibits the highest sensitivity to the increase of SSB pricing, implying the presence of consumers who are most likely to be discouraged to consume SSB upon taxation around this store.

## 5 Discussion

The superior prediction accuracy demonstrated by B-DTR and DNN over LASSO is likely due to their ability to model non-linear relationships and interactions across predictive features of the 154 brands. The finding indicates that traditional linear demand models such as LASSO may be a suboptimal approach in predicting the sale of SSB in competitive retail environment due to its linear constraint. Although it is theoretically possible to manually specify appropriate non-linear functional forms guided by model-fit criteria (e.g. Akaike's Information Criterion) in LASSO, this approach is not feasible as the number of competing brands grows large.

Future work includes in-depth investigation of store-level difference in the estimated effectiveness of taxation, or price elasticity. Identification of store-level features (e.g. promotion and the number of competing items) and neighborhood features driving differential store-level elasticity is a critical public health interest, since the analysis allows the characterization of communities that are less likely to benefit from taxation and consequently in need of community-specific interventions addressing local obstacles of healthy eating.

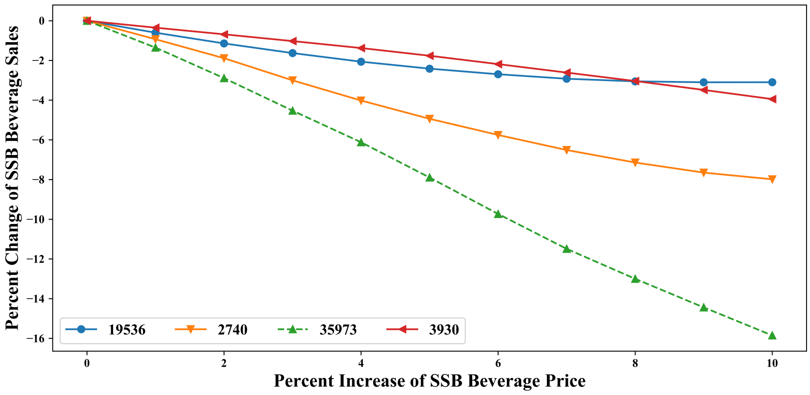

**Fig. 1.** Predicted percent reduction of SSB sales by DNN at various price levels simulating taxation, four randomly sampled stores from the 2013 test data

Analytical strategies for learning food demand from high-dimensional data were lacking to date. From a public health perspective, unique aspects of our study include the evaluation of the effectiveness of health policy using a large amount of transactional data, which were not available to public health researchers until recently.

## References

1. Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M.: Demand estimation with machine learning and model combination. Working Paper 20955, National Bureau of Economic Research (February 2015). https://doi.org/10.3386/w20955
2. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: Neural networks: Tricks of the trade, pp. 437–478. Springer (2012)
3. Chollet, F., et al.: Keras (2015)
4. Drucker, H.: Improving regressors using boosting techniques. In: ICML. vol. 97, pp. 107–115 (1997)
5. Escobar, M.A.C., Veerman, J.L., Tollman, S.M., Bertram, M.Y., Hofman, K.J.: Evidence that a tax on sugar sweetened beverages reduces the obesity rate: a meta-analysis. BMC public health **13**(1), 1072 (2013)
6. Forouzanfar, M.H., Afshin, A., Alexander, L.T., Anderson, H.R., Bhutta, Z.A., Biryukov, S., Brauer, M., Burnett, R., Cercy, K., Charlson, F.J., et al.: Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the global burden of disease study 2015. The Lancet **388**(10053), 1659–1724 (2016)
7. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 249–256. PMLR, Chia Laguna Resort, Sardinia, Italy (13–15 May 2010)
8. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 315–323 (2011)
9. Hu, F.B.: Resolved: there is sufficient scientific evidence that decreasing sugar-sweetened beverage consumption will reduce the prevalence of obesity and obesity-related diseases. Obesity reviews **14**(8), 606–619 (2013)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Leeflang, P., Bijmolt, T., Pauwels, K., Wieringa, J.: Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making. International Series in Quantitative Marketing, Springer (2015)
12. Ma, S., Fildes, R.: A retail store sku promotions optimization model for category multi-period profit maximization. European Journal of Operational Research **260**(2), 680 – 692 (2017). https://doi.org/10.1016/j.ejor.2016.12.032
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)

14. Thow, A.M., Downs, S., Jan, S.: A systematic review of the effectiveness of food taxes and subsidies to improve diets: understanding the recent evidence. Nutrition reviews **72**(9), 551–565 (2014)
15. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
16. Van Heerde, H.J., Leeflang, P.S., Wittink, D.R.: Semiparametric analysis to estimate the deal effect curve. Journal of Marketing Research **38**(2), 197–215 (2001)