Census2Vec: Enhancing Socioeconomic Predictive Models with Geo-Embedded Data

Ravi Varma Kumar Bevara[1], Isabelle Wagenvoord[2], Farahnaz Hosseini[3], Himanshu Sharma[3], Vandana Nunna[3], and Ting Xiao[1,3]

[1] Department of Information Science, University of North Texas, Denton, TX, USA
[2] Department of Computer Science, Colorado College, Colorado Springs, CO, USA
[3] Department of Computer Science, University of North Texas, Denton, TX, USA

Abstract. Artificial intelligence's role in distilling insights from data has emerged as a pivotal solution to contemporary challenges within our data-centric society. Census data, while offering rich demographic and socioeconomic insights, is limited by its complex dimensionality posing obstacles to developing universally applicable models. This study introduces an approach to leverage census data through the creation of location embeddings across various domains. Utilizing the Optuna framework, this research tuned autoencoders to optimize the bottleneck layer size, producing compact, low-dimensional embeddings that encapsulate critical relationships. These embeddings are further enriched with Federal Information Processing Standard (FIPS) codes, maintaining geographic identifiers. The methodology's effectiveness is demonstrated through regression models trained on the American Community Survey, accurately predicting key indicators like median gross rent and per capita income with an 8-10% higher average accuracy compared to traditional PCA-based methods. These findings suggest a novel paradigm for overcoming the limitations of geographically specific huge dimensional data for synthesizing insights. In practical terms, such as in public policy, this approach could enable more precise targeting of socio-economic interventions based on nuanced community profiles. This innovative technique in representation learning shows considerable promise for enhancing machine learning applications across diverse sectors, including marketing, and real estate.

Keywords: Census Data, Location Embeddings, Autoencoders, Principal Component Analysis (PCA), FIPS Codes, Predictive Analytics

1 Introduction

Amidst the burgeoning landscape of data-driven decision-making, the advent of machine learning in analyzing socioeconomic trends has notably advanced the predictive capabilities of researchers and policymakers. However, a significant gap persists in effectively leveraging geo-embedded data within census information to capture the nuanced dynamics of socioeconomic phenomena. But, the complex, high-dimensional nature of this data, particularly its geographic specificity, poses significant challenges in developing models that are broadly applicable across various domains. This manuscript introduces Census2Vec, a novel methodology designed to bridge this gap by enhancing socioeconomic predictive models with rich, geographically contextualized embeddings. Our approach seeks to address the limitations of current methodologies that often overlook the spatial intricacies inherent in census data, leading to oversimplified predictions that may not accurately reflect complex regional disparities. By integrating geospatial embeddings, Census2Vec aims to provide a more nuanced and accurate representation of socioeconomic indicators, paving the way for more informed decision-making processes.

Specifically, it addresses the challenge of dimensionality and geographic constraints by employing tuned autoencoders to generate low-dimensional embeddings enriched with geographic identifiers, enhancing the predictive capabilities of socioeconomic models. The exponential increase in data generated by emerging technologies such as mobile devices, cloud computing, and the Internet of Things (IoT) necessitates robust analytical methods [1]. Deep learning, a subset of machine learning modeled after neural network architectures, has catalyzed numerous technological breakthroughs, expanding commercial possibilities and driving the impending technological revolution [2]. The growing dimensionality, which causes data sparsity and the “curse of dimensionality”, makes it more difficult to extract meaningful insights from the enormous amounts of data collected every day. This challenge can be mitigated using various dimensionality reduction techniques. Traditional methods like Singular Value Decomposition (SVD) and principal component analysis have been effective in reducing data dimensions in applications beyond prototype definition, but they require substantial computational resources when dealing with large datasets, including images, text, and videos [3],[4],[5]. In 2006, Hinton and Salakhutdinov introduced a deep learning approach for dimensionality reduction using neural networks capable of predicting their input, known as autoencoders. Once trained, autoencoders provide a non-linear dimensionality reduction superior to SVD-based techniques, proving crucial in model and representational theory [6].

This paper begins by outlining the current landscape of socioeconomic predictive modeling, highlighting the challenges faced by traditional methods, and positing Census2Vec as a pivotal solution. It further delineates the research problem and the objectives of leveraging geo-embedded data for enhanced predictive accuracy.

2 Background

Embeddings play a critical role in the field of machine learning, particularly in reducing the dimensionality of data in high-dimensional spaces while preserving its semantic significance [7]. They facilitate the grouping of similar inputs in a lower-dimensional space, thus enhancing the adaptability of machine learning models. Mathematically, embeddings can be seen as entities that capture the underlying structures in diverse data instances, aligning with the principles of representation theory. The unique applications of location data, due to its granular nature, span across various sectors from government organizations to businesses, aiding in establishing demographic patterns and making critical location-based decisions. Prior methodologies in the realm of embeddings have largely focused on image data, often confined to specific domains [3],[4],[5]. However, the comprehensive nature of census data, encompassing a wide range of features such as demographics, business, academic, and political data, necessitates a more versatile approach to embedding. This research introduces a novel technique for condensing the multifaceted characteristics of a location into a compact, low-dimensional space by creating location embeddings utilizing Federal Information Processing Systems (FIPS) codes. The use of autoencoders in this process addresses the challenges of redundancy and noise in the data, which are critical in making efficient and accurate business predictions while managing processing times and memory consumption [8],[9].

3 Related Work

Recent progress in socioeconomic predictive modeling has emphasized the significance of integrating spatial context into analytical frameworks. Although there have been breakthroughs, existing methods frequently fail to fully capture the intricate spatial dynamics that define socioeconomic events. Conventional techniques like principle component analysis (PCA) and linear regression models are fundamental but do not provide the detailed information required to include spatial variations. This section thoroughly analyzes the problems by referencing a variety of recent research to highlight the shortcomings of current approaches. By doing this, it prepares for the release of Census2Vec, emphasizing the need for creative methods that can fully utilize geo-embedded data.

In this section, the paper by Mikolov et.al [10] serves as a foundation for location embedding research. It introduced the concept of embeddings, which represent high-dimensional data in a lower-dimensional transformed space. Subsequently, numerous works in this field have contributed to our understanding of location embeddings. Wang et al. [11] proposed the ELSE model, which enriches location embeddings by incorporating both spatial and semantic information. They applied a multi-label model with Convolutional Neural Networks (CNNs) to train on map-tile images, extracting location embeddings to address practical business problems, such as recommending new service ports. Yin et al. [12] utilized the UTM coordinate system to create 2000-dimensional GPS semantic embeddings. These embeddings were obtained using a Multilayer Perceptron neural network model and included data from platforms like Twitter, Foursquare, and Flickr. Abonce et al. [13] developed a Siamese-like embedding model by training it on map-tile images and Google Street View images to capture location information effectively. Law and Neira [14] employed convolutional autoencoders and Principal Component Analysis (PCA) to generate location embeddings from street view images, even in the absence of explicit semantic information.

Recent studies have focused on integrating various features into a single model. For example, Yin et al. [12] designed a probability-generating model that jointly considered temporal cyclic effects, geographical influences, and semantic information. Xu et al. [15] integrated these features into a neural network framework for location recommendations. While [16] demonstrated the value of creating vector representations from unstructured text data for business analytics tasks, this research explores generating meaningful embeddings from structured census data to improve machine learning predictions. Typically, these solutions fused contextual factors using straightforward strategies like modeling them as weighting coefficients. In contrast to existing approaches, our work adopts a novel approach. We leverage nonlinear transformations to seamlessly integrate different features in an integrated manner. This approach not only smoothly combines various factors in a shared latent space but also qualifies as a generic method applicable across diverse scenarios. Additionally, in Spruyt's study [17], location embeddings were obtained using convolutional neural networks based on geographical location coordinates. These embeddings were effectively applied for venue mapping and transport classification, enhancing generalization capabilities by associating daytime and nighttime activities with specific areas such as industry zones, city centers, parks, and train stations. Furthermore, the paper by Jenkins et al. [18] stands out for its cross-modal embeddings, which incorporate data from various sources, including satellite images, human mobility patterns, point of interest locations, and spatial graphs. [19] By utilizing a RegionEncoder, this work significantly improved performance in downstream tasks related to urban environments and prediction tasks.

4 Methodology

The methodology underpinning this study begins with an intensive data preparation phase, concentrating on the U.S. census dataset to extract a low-dimensional representation for census block groups. The architecture of the Autoencoder model employed in our approach is illustrated in Figure 1. The initial step involves selecting columns with numerical data to ensure the integrity of the dataset, which is vital for the effective handling of missing values and guaranteeing a comprehensive dataset for meaningful analysis.

4.1 Expolaroty Data Analysis (EDA)

Dataset Description Our dataset includes all U.S. census block groups related to Federal Superfund sites, categorized under the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) of 1980. It comprises 220,338 census block groups with 345 features representing a variety of location-specific characteristics, such as geographical, demographic, and socioeconomic data.

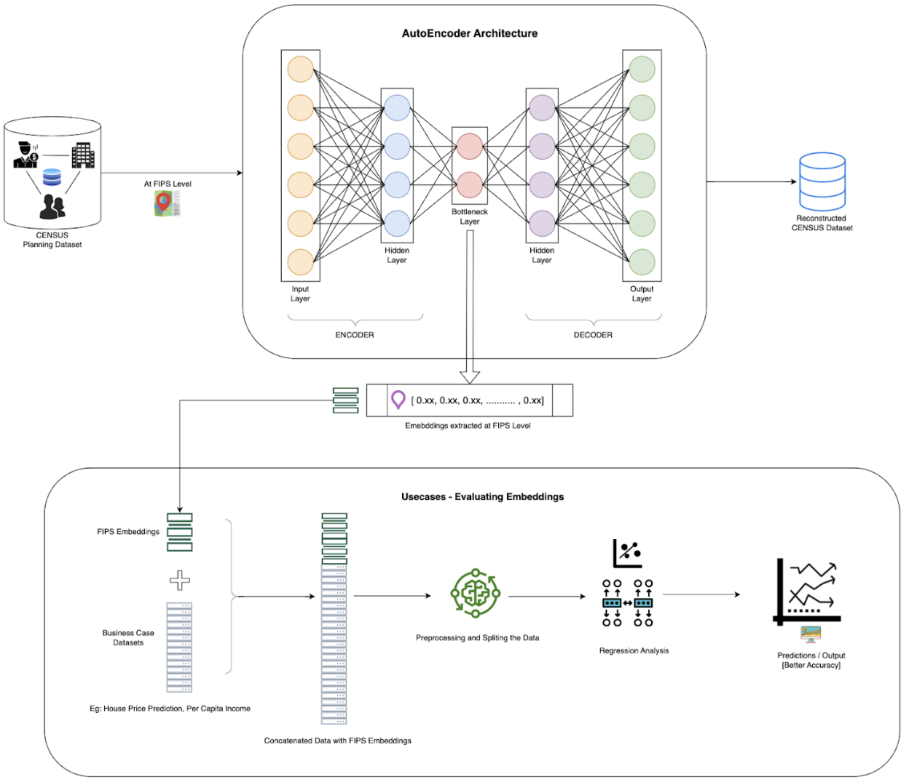

Fig. 1. Architecture of the Autoencoder model utilized for dimensionality reduction in census data analysis, showing the flow from input to compressed representation at the bottleneck layer, and reconstruction at the output.

Data Cleansing and Feature Selection The feature selection process was critical to optimize the census dataset by excluding non-essential features, such as county and state names, to focus on quantitative metrics. Features like "average household income" and "median rent", initially in U.S. Dollar string formats, were cleansed of special characters and transformed into numerical values.

Handling Zero Values Features with more than 10% zero values were excluded to prevent skewed distributions and potential biases in the embeddings. This refinement resulted in a dataset with 220,322 entries and 110 distinct numeric features.

Data Imputation and Normalization We addressed missing values indicated by "NaN" using the K-Nearest Neighbors Imputer (KNNImputer) method. This approach estimates missing values by utilizing the values of the nearest neighbors within the multidimensional feature space. The data was then normalized to a common scale, as is crucial for the subsequent machine learning processes.

4.2 Dimensionality Reduction using PCA with K-Fold Cross Validation

Principal Component Analysis (PCA) was employed as a prelude to Autoencoder-based dimensionality reduction to establish a baseline for comparison. PCA serves to transform the original high-dimensional feature space into a lower-dimensional space where the axes correspond to the directions of maximum variance. We methodically reduced the feature dimensions from 10 to 100, aiming to retain as much variance as possible while simplifying the dataset's complexity. To evaluate the adequacy of the reduced dimensions and prevent overfitting, we applied K-Fold Cross-Validation with eight splits. This rigorous statistical technique partitions the data into eight subsets, where each subset is used once as a validation set while the others form the training set. This cross-validation approach is depicted in Figure 2, where the performance of various regression models on PCA-reduced data is assessed. By averaging the performance across all folds, we obtain a more reliable estimate of the model's predictive power on unseen data.

4.3 Dimensionality Reduction using Autoencoders with Optuna

Building upon the insights gained from PCA, Autoencoders were trained to explore a more nuanced and sophisticated approach to dimensionality reduction. An Autoencoder is a type of unsupervised neural network that learns to compress (encode) the input data into a compact representation and then reconstruct (decode) it as closely as possible to the original input. This dual process forces the Autoencoder to capture and prioritize the most salient features in the bottleneck layer, which is the heart of the network characterized by its reduced number of neurons.

Table 1. Hyperparameter Ranges for Tuning with Optuna

Hyperparameter	Range
Number of Layers (*n_layers*)	1 – 10
Number of Neurons (*n_neurons*)	32 – 128
Activation Function (*activation*)	relu, tanh, linear
Learning Rate (*learning_rate*)	1e-5 – 1e-2
Batch Size (*batch_size*)	32, 64, 128
Bottleneck Dimension (*bottleneck_dim*)	2 – 32

Table 2. Tuned Results (Best Hyperparameters)

Hyperparameter	Value
Number of Layers (*n_layers*)	8
Number of Neurons (*n_neurons*)	128
Activation Function (*activation*)	tanh
Learning Rate (*learning_rate*)	0.00036
Batch Size (*batch_size*)	64
Bottleneck Dimension (*bottleneck_dim*)	21
Input Size	(110, 0)
Loss Function	MSE
Optimizer	Adam
Epochs	100

Optuna, an open-source hyperparameter optimization framework, was employed to systematically search for the optimal hyperparameters that would result in the most efficient encoding and decoding process. The framework's efficiency stems from its ability to perform trials of various hyperparameter combinations and evaluate them based on the reconstruction loss. The hyperparameter space explored by Optuna included the number of layers, number of neurons per layer, activation functions, learning rate, batch size, and bottleneck dimension, as detailed in Table 1. The final architecture of the Autoencoder, which provided the best balance between dimensionality reduction and data reconstruction fidelity, is summarized in Table 2. This Autoencoder was then trained and validated using the same K-Fold Cross-Validation method described earlier to ensure the robustness and generalizability of the learned embeddings.

4.4 Autoencoder Architecture

The architecture of our Autoencoder, illustrated in Figure 1, is meticulously designed to capture the intricate patterns inherent in high-dimensional census data. The network consists of an input layer that accepts the preprocessed census data, several hidden layers for feature transformation, a bottleneck layer for representing the compressed knowledge, and an output layer for reconstructing the input data. The hidden layers are composed of fully connected neurons with rectified linear unit (ReLU) activation functions to introduce non-linearity, enhancing the network's capacity to learn complex representations. The bottleneck layer, the core of the Autoencoder, acts as a constraint to force the network to distill the most essential information from the input data, resulting in a compact, low-dimensional representation of the original dataset. This process of encoding and subsequent decoding is optimized using backpropagation with an Adam optimizer, minimizing the reconstruction loss measured by the Mean Squared Error (MSE). The optimal architecture and hyperparameters, presented in Table 2, were determined through extensive hyperparameter tuning using the Optuna framework. The fine-tuning process balanced the model's complexity against its generalization ability, ensuring that the Autoencoder captures a robust and general representation of the census data features.

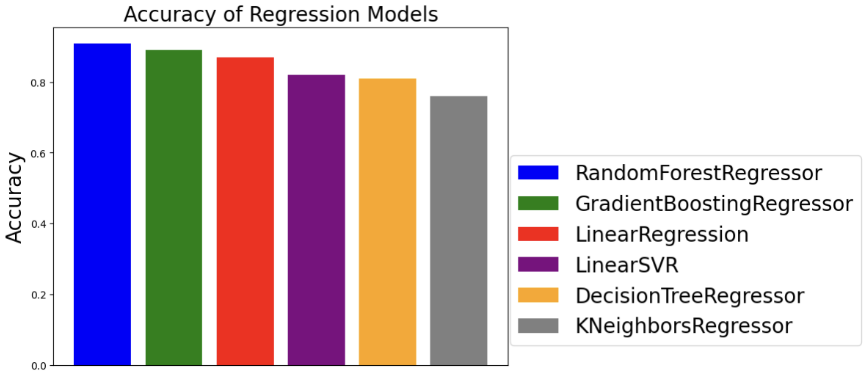

Fig. 2. Comparative accuracy of various regression models employed in the study, with the bars representing the K-Fold Cross-Validation scores for each model.

4.5 Accuracy Measurement in Regression Models

Figure 2 presents the accuracy of various regression models assessed in this study, with the term 'accuracy' referring to the proportion of correct predictions in the validation sets during the K-Fold Cross-Validation process. The models included, such as Random Forest, Gradient Boosting, and K-Nearest Neighbors, were selected for their diverse approaches to regression, ranging from ensemble methods to instance-based learning, providing a comprehensive evaluation of the embedding's predictive quality. The validation process employed K-Fold Cross-Validation with eight folds, ensuring that each model was tested on all data points while also being trained on diverse subsets of the data. This rigorous validation method not only provides an unbiased estimate of the model's predictive performance but also helps in mitigating the risk of overfitting, as each fold serves as an independent check against the model's ability to generalize. The final accuracy metrics are the averaged results from all folds, offering a robust measure of performance and a comparative view of how each model leverages the embeddings produced by the Autoencoder to predict socioeconomic indicators such as median gross rent and per capita income. The superiority of the embeddings is evidenced by the consistent improvement in accuracy across all models when compared to their performance with the original high-dimensional data.

5 Results and Analysis

5.1 Embedding Performance Evaluation

To assess the quality of embeddings generated from both PCA and Autoencoder dimensionality reduction methods, we reserved one feature from the dataset as a target variable, excluded from the input features to the models. This approach enables a direct comparison of how well the compressed features from both embeddings predict the target variable. For PCA, we evaluated the performance of various regression models on this target feature to determine the effectiveness of the embeddings in predictive tasks. Figure 2 shows a comparative analysis of the accuracy achieved by different regression algorithms when applied to the validation set. The RandomForestRegressor emerged as the top-performing model, with GradientBoostingRegressor and LinearRegression also showing strong results. This suggests that ensemble methods are particularly effective in leveraging the reduced feature space for predictive analysis.

5.2 Comparative Analysis of PCA and Autoencoder Embeddings

We further examined the efficacy of embeddings produced by Autoencoders against those generated by traditional PCA. The R-squared scores, reflecting the proportion of variance captured by the models for median household income, are plotted against the number of dimensions retained in the models, as illustrated in Figure 3.

The Autoencoder embeddings displayed a remarkable ability to capture a significant amount of variance within just 10 dimensions, which is substantially more efficient than PCA, which required 60 dimensions to achieve a similar level of variance capture. As depicted in Figure 3, the embeddings consistently outperformed PCA in terms of R-squared scores across all tested dimensionalities. This underscores the Autoencoder's ability to distill more relevant information for the prediction task, particularly at higher dimensions.

5.3 Case Study: American Community Survey Data

Our analysis extended to a larger dataset from the American Community Survey (ACS), ensuring comprehensive coverage and enhancing the generalizability of

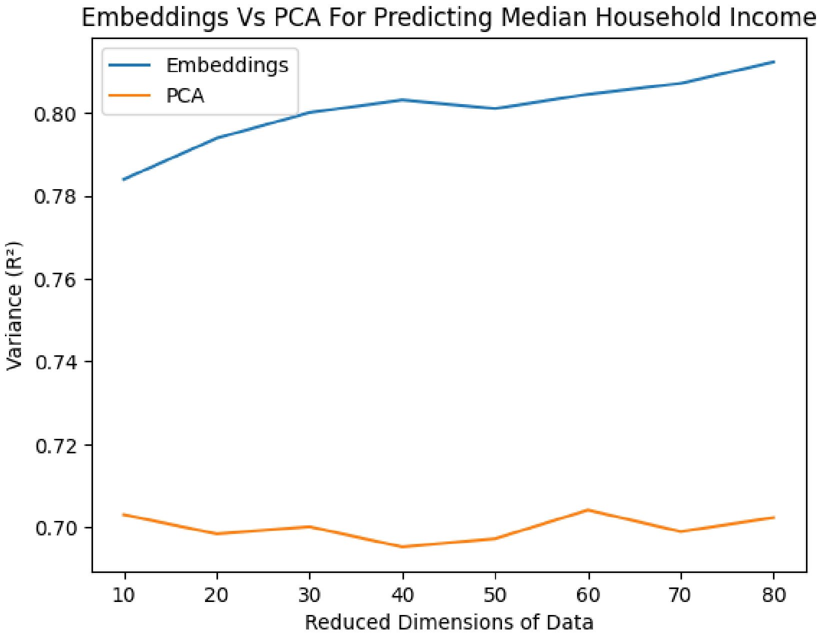

Fig. 3. Comparison of R-squared scores for PCA and Autoencoder embeddings across different dimensions.

our findings. The ACS dataset provides a wide array of social, economic, demographic, and housing characteristics. The five-year estimates from the ACS, representing data collected over a period, offer enhanced statistical reliability for less populated areas and smaller population subgroups. The data was retrieved at the block group level using the ACS 5-Year Data API, aligning with the granularity of our embeddings. Using the ACS data, we conducted an expanded evaluation, incorporating additional FIPS block group data. We considered attributes such as 'Per capita income', 'Median household income', 'Median gross rent', and 'Aggregate earnings', which are closely aligned with the demographic data represented by the embeddings. The 'download()' and 'censusgeo()' functions facilitated efficient retrieval of these attributes, with a looping mechanism implemented to traverse through each state's data.

We constructed four distinct datasets, each with one of the aforementioned attributes designated as the target variable. A RandomForest regression model was utilized to calculate the R-squared and variance scores for each target, yielding scores that highlight the predictive strength of the embeddings.

In Figures 4, 5, and 6, the variance results for models using Autoencoder embeddings for different socioeconomic indicators are presented. The merged datasets containing both PCA and Autoencoder embeddings with ACS data demonstrated that the Autoencoder embeddings significantly enhance model

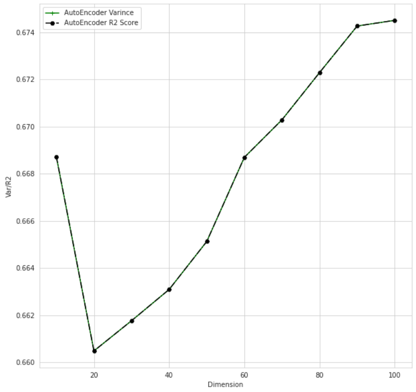

Fig. 4. Variance results for regression models using Autoencoder embeddings for Per Capita Income.

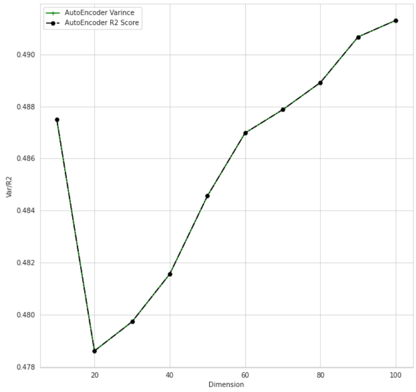

Fig. 5. Variance results for regression models using Autoencoder embeddings for Aggregate Earnings.

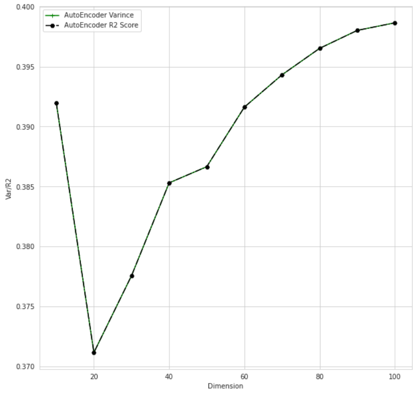

Fig. 6. Variance results for regression models using Autoencoder embeddings for Median Gross Rent.

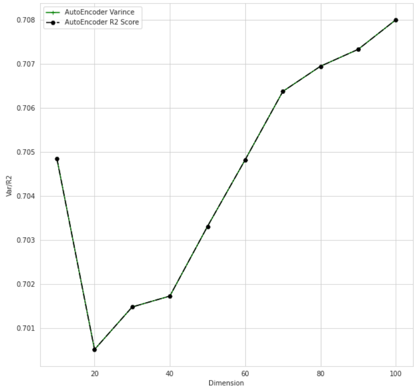

Fig. 7. Variance results for regression models using Autoencoder embeddings for Median Household Income.

performance across various socioeconomic predictions, validating the effectiveness of our dimensionality reduction approach. Figure 7 shows the variance results for models using Autoencoder embeddings for Median Household Income, further evidencing the superior performance of Autoencoder embeddings over PCA in predictive tasks. These results substantiate the effectiveness of PCA and Autoencoder embeddings and underscore the Autoencoder's efficiency in creating compact yet informative representations, particularly when applied to complex datasets like the ACS FIPS block group data.

6 Discussion

Our research primarily concentrated on four pivotal business cases: “Per Capita Income”, “Median Household Income”, “Median Gross Rent”, and “Aggregate Earnings”. These cases were meticulously chosen to embody the demographic characteristics foundational to our embedding creation process. The transformation of 110 original features into compact embeddings yielded dual advantages. First, it streamlined the dataset for regression models, enhancing manageability without compromising predictive efficiency. Second, it evidenced the versatility of embeddings in encapsulating trends across diverse demographic data types. Parallel to our approach, a study delineated in [12] leveraged user-generated content from social platforms to create embeddings that characterize distinct locales, such as addresses, photos, and phone numbers. These embeddings facilitated K-NN searches and the simplification of duplicate removal in location data graphs. While this method focuses on encoding information from decentralized sources, our strategy, drawing from comprehensive demographic data, showcases a different yet equally effective technique in capturing the multidimensional essence of locations. This comprehensive approach endows our embeddings with wide-ranging applicability.

In a related vein, the work presented in [15] integrates semantic data, temporal-spatial context, and sequential place connections to forge location embeddings. These embeddings are instrumental in predicting user check-ins by identifying close matches in a multidimensional space. While Venue2Vec excels in user location recommendations and tracking, we advocate for the use of place embeddings, given their efficacy in clustering models and their ability to identify groups with similar characteristics. This capability is invaluable for governmental and business entities in assessing and analyzing risks ahead of implementing high-impact policies. The versatility and potential applications of place embeddings, as illuminated in our study, are vast, spanning various domains. This discussion underscores the critical role of embeddings in data-driven decision-making, highlighting their capability to discern intricate patterns in demographic data, their adaptability across various fields, and their broader, far-reaching implications.

7 Future Research Directions and Advancements

The Census2Vec architecture has great potential for future study in the field of socioeconomic analysis. One such approach involves modifying Census2Vec for various geographical settings. Applying the technique to fast urbanizing cities in underdeveloped nations might reveal patterns of social segregation and infrastructure inequality. This involves overcoming obstacles including acquiring accurate census data at a detailed level and considering the changing nature of informal settlements. Integrating Census2Vec with agricultural and land-use data can help reveal the socioeconomic factors and impacts of rural change, such as out-migration and evolving livelihood options. Additionally, integrating Census2Vec embeddings with high-resolution satellite imagery might facilitate the real-time monitoring of socioeconomic changes by analyzing variables such as variations in building density and infrastructure growth. Integrating graph neural networks (GNNs) into Census2Vec might improve embedding quality by capturing intricate interconnections across census tracts, considering aspects like physical proximity, infrastructural networks, or social ties. Attention-based embedding processes might assist in identifying the most important census factors that influence socioeconomic distinction in certain circumstances.

Comparative studies in many socioeconomic settings are necessary to fully assess the effectiveness of Census2Vec. An experiment might be designed to compare Census2Vec models trained on datasets from high-income nations vs low-income countries or urban regions versus rural regions. The independent factors consist of socioeconomic background, data granularity, and embedding methodologies, whereas the dependent variables comprise prediction accuracy and the interpretability of created embeddings. Including a time aspect in Census2Vec would enhance predictive demographic modeling. Time-series census data, combined with real-time auxiliary data sources, might be used to predict Census2Vec embeddings over time using recurrent neural networks (RNNs) or their variations (LSTMs, GRUs). This would enable the examination of changing socioeconomic patterns and maybe guide preemptive governmental actions.

8 Conclusion

Through the incorporation of geo-embedded data utilizing FIPS codes as unique identifiers, this work proposes Census2Vec, a novel approach that substantially enhances the accuracy and granularity of socioeconomic projections. Census2Vec provides a new perspective for analyzing and comprehending the multifaceted geographical dynamics of socioeconomic indicators by overcoming the limitations imposed by traditional prediction models. Also, this study explored different strategies for creating these embeddings, including an unsupervised deep autoencoder model to analyze publicly accessible Census data together with FIPS codes at state and county levels. This method has uncovered a plethora of information on location embeddings, demonstrating their strong prediction accuracy in comparison with adept contemporary approaches. Although the results demonstrate potential, this research acknowledges specific limitations. The study focused primarily on U.S. Census data, which could not entirely represent the multifaceted nature of different geographical areas or demographic differences. This analytical approach serves a purpose yet might require adaptations to accommodate datasets from various countries or areas, taking into consideration cultural and socio-economic differences.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable suggestions. The National Science Foundation grant 2244259 helped to support this work in part. The views, opinions, and/or findings expressed in this material are those of the authors and should not be interpreted as representing the official views of the National Science Foundation, or the U.S. Government.

References

1. Humby, C.; Palmer, M. (2006, November 3). Data is the New Oil. https://ana.blogs.com/maestros/2006/11/dataisthenew.html (accessed December 9, 2019).
2. Hinton, G. E., Osindero, S., Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.
3. M. Vikram, R. Pavan, N. D. Dineshbhai, and B. Mohan.Performance evaluation of dimensionality reduction techniques on high dimensional data,2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)
4. Q. Fournier and D. Aloise. Empirical comparison between autoen coders and traditional dimensionality reduction methods. In 2019 IEEE Second International Conference on Artificial Intelli gence and Knowledge Engineering (AIKE), pages 211–214, 2019. doi: 10.1109/AIKE.2019.00044.
5. Joshi, S. K., Machchhar, S. (2014, December). An evolution and evaluation of dimensionality reduction techniques|A comparative study. In 2014 IEEE International Conference on Computational Intelligence and Computing Research (pp. 1-5). IEEE.
6. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. ”Reducing the dimensionality of data with neural networks.” science 313.5786 (2006): 504-507.
7. Adachi, M. (2012). Embeddings and immersions. American Mathematical Soc..
8. Kraus, J., Tobiska, T., Bubla, V. (2009, June). Loooseless encodings and compression algorithms applied on power quality datasets. In CIRED 2009-20th International Conference and Exhibition on Electricity Distribution-Part 1 (pp. 1-4). IET.
9. Schizas, I. D., Aduroja, A. (2015). A distributed framework for dimensionality reduction and denoising. IEEE Transactions on Signal Processing, 63(23), 6379-6394.
10. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
11. Wang, Y., Wang, C., Ling, Y., Yokoyama, K., Wu, H. T. and Fang, Y. (2020, December). Leveraging an Efficient and Semantic Location Embedding to Seek New Ports of Bike Share Services. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 1273-1282). IEEE.
12. Y. Yin, Z. Liu, Y. Zhang, S. Wang, R. R. Shah, and R. Zimmermann. Gps2vec: Towards generating worldwide gps embeddings. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 416–419, 2019.
13. O. S. Abonce, M. Zhou, and A. Calway. You are here: Geolocation by embedding maps and images. 2019.
14. S. Law and M. Neira. An unsupervised approach to geographical knowledge discovery using street level and street network im ages. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, pages 56– 65, 2019.
15. Xu, S., Cao, J., Legg, P., Liu, B., Li, S. (2019). Venue2vec: An efficient embedding model for fine-grained user location prediction in geo-social networks. IEEE Systems Journal, 14(2), 1740-1751.

16. Gerling, C. (2023). Company2Vec–German Company Embeddings based on Corporate Websites. arXiv preprint arXiv:2307.09332.
17. V. Spruyt. Loc2vec: Learning location embeddings with triplet loss networks. Sentiance web article: https://www. sentiance. com/2018/05/03/venue-mapping, 2018.
18. P. Jenkins, A. Farag, S. Wang, and Z. Li. Unsupervised represen tation learning of spatial data via multimodal embedding. In Proceedings of the 28th ACM International Conference on Infor mation and Knowledge Management, pages 1993–2002, 2019.
19. Jolliffe, I. T., Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.