Household Embeddings: Introducing continuous vector representations for car ownership on a household level

Ioanna Arkoudi[1]*, *Carlos Lima Azevedo*[1], *Francisco C. Pereira*[1]
[1]*DTU Technical University of Denmark*
* *corresponding author*, ioaar@dtu.dk

February 14, 2020

Abstract

This study presents a new method of representing travel-related categorical variables and observations using continuous vector representations, commonly known as embeddings. Specifically, we focus on generating embedding representations that aim to describe car ownership on a household level based on methods and approaches that are used in the field of Natural Language Processing (NLP) and Machine Learning (ML). We build on the previous work on traveling embeddings by Pereira [1] and extend it with a joint embedding space approach that allows us to leverage on the compositionality of the categorical vectors and introduce a method that uses embedding centroids to represent back individual observations, *i.e. household embeddings*. The efficiency of the centroid-based method is tested by comparing two binary logit models for car ownership: one that uses the embedding centroids encoding and one that uses the traditional dummy encoding. For our car-ownership modelling case, the results show that using embedding centroids encoding in utility specification performs better that direct categorical variables in out-of-sample prediction. We further demonstrate that the proposed method not only produces meaningful representations of the categorical space, but also allow us to define *prototypical* households for the behaviour at stake.

Keywords: categorical embeddings, machine learning, space representations, encoding methods, behavioral modeling, discrete choice

Contents

1 Introduction

The purpose of this study is to present a new method of representing categorical variables and observations with regards to a specific traveling behavior using continuous vector space representations, i.e. embeddings. The proposed method is based is on the recent work of Pereira on traveling embeddings [1] and uses methods and techniques borrowed from the field of Natural Language Processing (NLP) and vector space models of semantics.

Word embeddings are continuous-valued vector representations of words that manage to capture meaningful syntactic and semantic relationships between the encoded linguistic units based on their distributional properties [2]. They have become very popular in the recent years, especially after word2vec, a deep learning-based method for word embeddings generation introduced by Mikolov et al. [3]. The theoretical origins behind them are linked to the "distributional hypothesis" as suggested by Harris that words that occur in the same contexts tend to have similar meanings [4]. They have proven to be highly effective at a wide variety of downstream NLP tasks ranging from text classification and question answering to automatic machine translation [5][6].

Recent studies have demonstrated that such methods can also find applications outside the field of NLP. Several of these studies [7][8][9][10] focused on generating place/geospatial embedding representations in a similar manner to word2vec taking into account mobility patterns and spatial context information. The resulting representations offered new insights into the understanding of space semantics and place functionalities [9][10] and found applications in various tasks related to urban planning and policy design such as point-of-interest (POI) recommendation[8], and predicting users who will visit a given POI in a given future period [7]. Following the word to place analogy Yabe et. al [11] accurately translated place representations between different cities extending the existing NLP methods for automatic translation. Lastly, Pereira [1], presented a method of mapping discrete variables, that are typically used in travel demand modeling, into a latent embedding space. He further demonstrated that using the embedding encoding for categorical variables in a choice model outperforms traditional methods of data representations, such as dummy variables or PCA encoding in out-of-sample evaluation. These studies suggest a change of paradigm in representing information in the areas of geographic information, mobility and urban planning leveraging on data-driven methods and deep-learning techniques.

The current study specifically focuses on exploring the representation of discrete explanatory variables in discrete choice models as embedding vectors. More specifically, we focus on car ownership models at the household level. In contrast with Pereira's work in [1] where the vectors of the considered variables reside in separate spaces, we alter the training process and propose a joint space approach. This joint space approach comes with a number of advantages, namely:
(i) it allows for the participation of binary variables in the embedding space.
(ii) it allows the modeller to decide on the optimal number of embedding dimensions simultaneously for all the variables (instead of choosing the number of dimensions separately for each variable).
(iii) it allows to explore meaningful associations not only between categories within a variable but also across variables, and to identify meaningful clusters of categories that exhibit similar behavior with respect to the study variable.
(iv) it allows for creating new continuous representations on an observation level, i.e. household level, which are suitable for clustering tasks and similarity queries with respect to the target

behavior.

In respect to (iv) we draw inspiration from similar NLP methods used for document representation originally introduced by Radev et. al [12], and propose a centroid-based method for representing households that exploits the compositional capabilities of the embedding vectors. Furthermore we hope that methods developed in this work set the premises for the future exploration and modeling of temporal behavioral dynamics using NLP borrowed methods for machine translation.

2 Data and Embeddings generation

For the purposes of this study we used the Danish National Travel Survey (TU) dataset covering the period from May 2006 to December 2019[1]. After excluding the observations for which some variables were not available we ended up with a total of 90,833 observations. The explanatory discrete variables considered for car ownership in household level are presented below (Table 1). Additionally we included the continuous variable 'HomeAdrDistNearestStation', i.e. the distance from the respondent's home address to the nearest public transport station in km. The discrete variables were used for learning the embedding vectors with respect to the target variable, i.e. car ownership, using PyTre (PYthon TRavel Embeddings) package[2]. 'HomeAdrDistNearestStation' was also used an input in the PyTre model so that its effects are taken into account when learning the embeddings for the categorical variables [1].

Discrete explanatory Variables	**Categories or Intervals**
NuclFamType (respondent's nuclear family type)	SingleM, SingleW, Couple, Single&Children, Couple&Children
HousehAccOwnOrRent (respondent's type of house ownership)	cooperative, owner, rent
HomeAdrNUTS (region of household address)	WZealand, CPH, EJutland, Funen, SJutland, NZealand, EZealand, NJutland, GrCPH, WJutland
RespPrimOcc (respondent's primary occupation)	Pension, SelfEmpl, EarlyPension, Employee, nonAgePension, Unempl, Student, HighSchool, PrimSchool,OtherOccupation
RespEduLevel (respondent's highest completed education)	1st-7th, 8th, 9th, 10th, MediumFurtherEdu, LongFurthEdu, Upper2ndCertif, HigherCertif, Vocational, OtherSchool
IncNuclFamily (household's total gross income in thousand DKK per year)- *discretized*	[0, 250],]250, 400],]400, 550],]550, 700],]700, 870]
Target variable: *Car Ownership* *on a household level*	No Car (0), Car (1)

Table 1: *The discrete variables used for generating the car ownership embeddings together with their corresponding categories (or intervals).*

Since the methods presented in this study are intended to be used in the future for exploring the temporal dynamics in regards to car ownership, we divided the data into 7 subsets by every

[1] for survey description please visit: https://www.cta.man.dtu.dk/english/national-travel-survey/

[2] https://github.com/camaraf/PyTre

2 years [3] in order to learn embedding representations separately for each time period. Each 2-year subset was further divided into a training set (80%), a development set (10%) and a test set (10%). The training sets were used for learning the embeddings for each period using the PyTre algorithm[4]. The embedding generation and evaluation process is briefly summarized in Figure 1. The experiment was repeated multiple ($\sim$ 200) times for each 2-years subset in order to find an optimal number of embedding dimensions as there is stochasticity in the training process. The embeddings selection criteria relied on adjusted R^2 performance in the development set of a binary logit model estimated on the training embeddings data. The test sets were reserved for later use, to estimate the out-of-sample performance of the embeddings encoding in a choice model and compare it with the dummy variables encoding method.

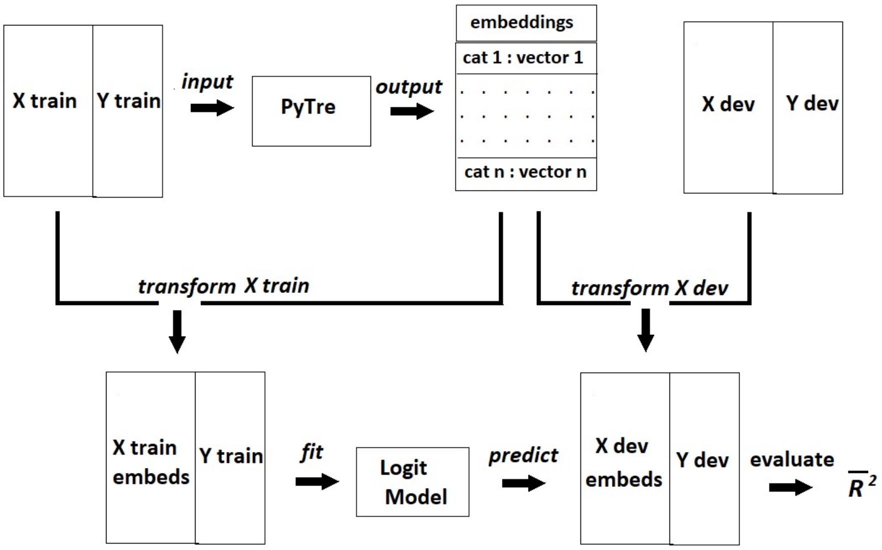

Figure 1: *Workflow diagram of the embeddings generation and evaluation process; X and Y as per Table 1*

3 Joint Embedding Space

After choosing the best performing embeddings for each (2 years) subset of the data, the t-Distributed Stochastic Neighbor Embedding (tSNE) projections of the categorical embedding vectors were plotted into a 2D-plane to obtain a simplified graphical representation of the multidimensional vectors[5]. In Figure 2 we present an example of tSNE embedding projections for the years 2010-2011. In order to demonstrate that the embedding vectors cluster together in a meaningful way in the space with respect to the target variable, i.e. car ownership, the mean probability for car onwership per category $\overline{P}_{cat}$ was computed and is included in the plot next to the name of each category. The colors of the scatterplot also indicate the value of $\overline{P}_{cat}$ according to the colorbar on the side. We can observe that $\overline{P}_{cat}$ seems to be associated with the geometry of the embedding vectors, as categories that have similar $\overline{P}_{cat}$ values tend to form larger or

[3] 2006-2007, 2008-2009, 2010-2011, 2012-2013, 2014-2015, 2016-2017, 2018-2019

[4] For a detailed description of the PyTre algorithm and the embeddings generation method please refer to [1]

[5] After experimenting on the number of embedding dimension in a range between 5 to 30, we found that using 25 dimensions yields better results for the task at hand.

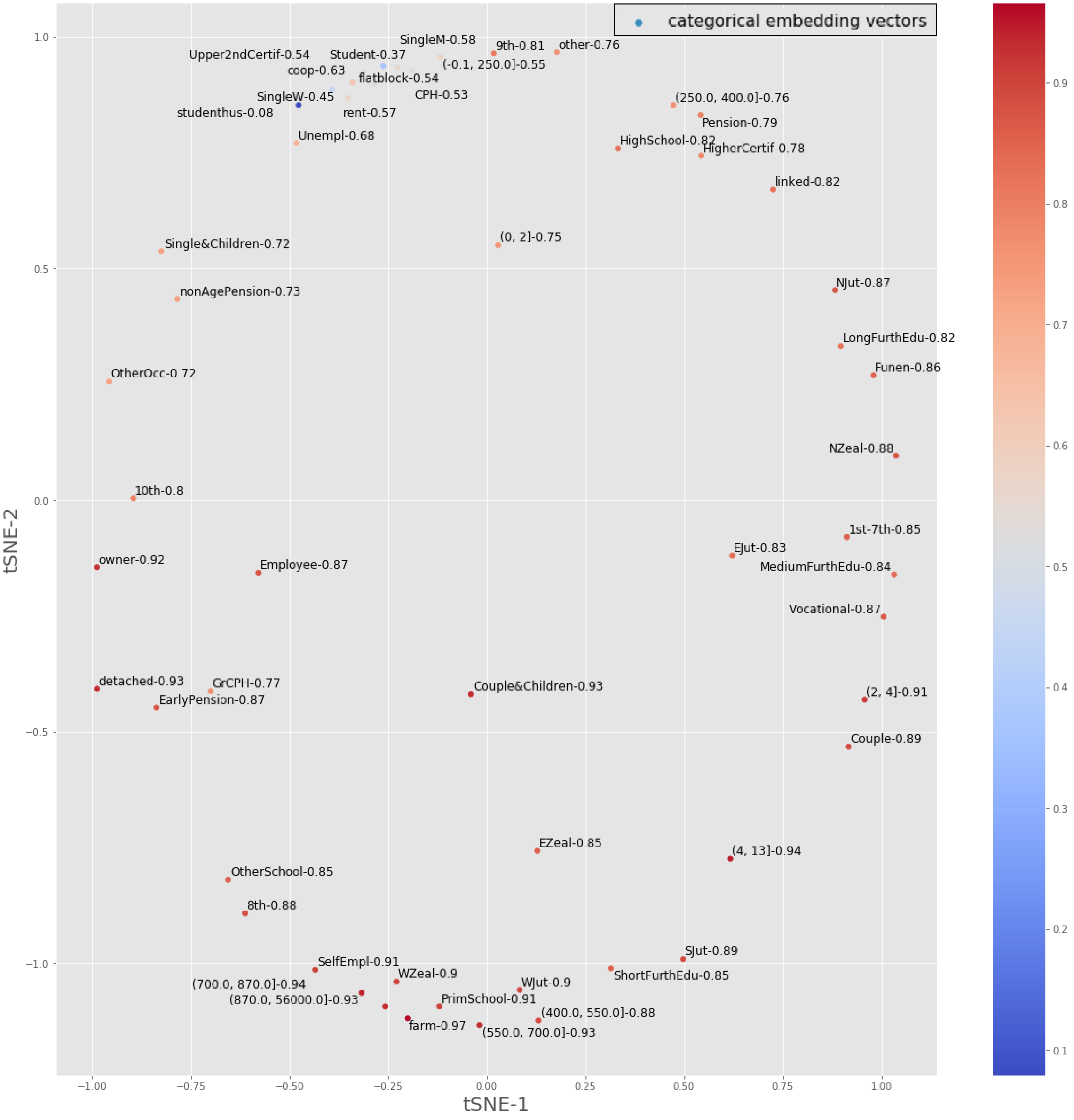

Figure 2: *2-D tSNE categorical embeddings projections for the years 2010-2011. The colorbar on the right indicates the value of* $\overline{P}_{cat}$ *(ranging from 0.08 to 0.97).*

[6] Please note that $\overline{P}_{cat}$ is used as an indicator that would allow us to make the embedding space more interpretable and to highlight a meaningful association and does not imply that the complex geometry of the high-dimensional embedding space can be reduced or be fully described by mere statistical properties of its elements inferred directly from the data.

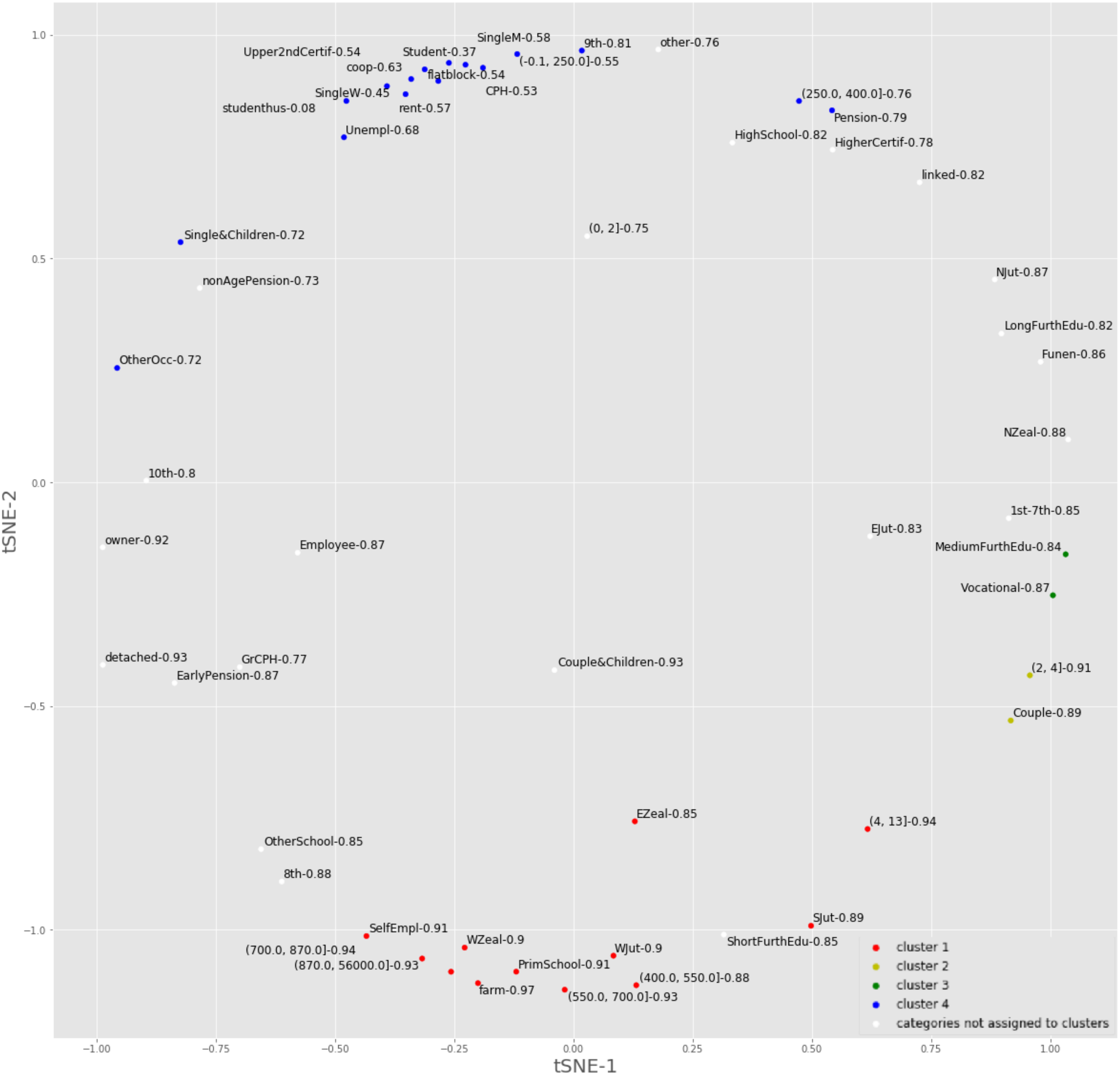

Figure 3: 2-D tSNE categorical embeddings projections for the years 2010-2011 after cosine similarity clustering.

The proximity between two categories in the space can be measured through a distance metric between their corresponding embedding vectors, such as cosine similarity. After applying cosine similarity clustering using a (tight) similarity threshold of 0.9, the results are plotted above in Figure 3. The categories belonging to the same cluster and highlighted with the same color, while the categories that were not assigned to any clusters are left white. Two major tightly connected clusters can be identified the blue and the red one. The first includes students, unemployed and pensioners, singles, living in the capital, renting a place to live and households of lower income. The second cluster corresponds to big families, self-employed people, living in rural areas/farms, households of greater income level and relatively lower educational level.

These clusters are intuitive regarding car ownership since, as one would expect, the first cluster is associated with lower and the second with higher probability for owning a car. However it should be emphasized that further work and exploration of the embedding space is required to uncover causality and deeper underlying patterns in the data before drawing any definitive conclusions about how these clusters are associated to the study behavior.

4 Household embeddings: The centroid-based method

One of the most intuitive and intriguing properties of word embeddings is their compositionality, i.e. that they allow for performing meaningful algebraic operations between their vectors. [14]. A famous example is that of the male-female relationship such that the vector representation of "king" - "man" + "woman" results in a vector that is very close to that of "queen". A number of NLP studies have focused on exploiting the compositional capabilities of word embeddings to represent sentences or documents using the simple averages vectors (centroids) of the words they contain in order to model language data (e.g.[12][13]). In this section it will be shown that a similar method can be used effectively within our framework to obtain vector representations on an observation level, i.e. household embeddings. The proposed method is simple and easy to implement and can be summarized in the following formula. Let H be the set of households in the dataset and C be the set of all the categorical embedding vectors in the embedding space S, such that each $h \in H$ can described by $c \subseteq C$. Then the embedding vector of h in S, h_{vec}, can be computed by: $h_{vec} = \frac{\sum_{i=1}^{|c|} c_i}{|c|}$ (1)

We used (1) to obtain the vectors representations for all the households in our data. In order to visualize the results we used tSNE again[7] including the household embeddings and in Figure 4 we present an example of the 2D projections for the years 2010-2011. The household embeddings are plotted together with the categorical embeddings they compose of in order to exhibit their relative positions in the space.

[7] Note that tSNE algorithm can produce arbitrarily rotated projections every time we use it, which explains why the points in Figures 1 and 2 do not appear in the same positions as in Figure 3. However, the relative distances between the projected points remain the same.

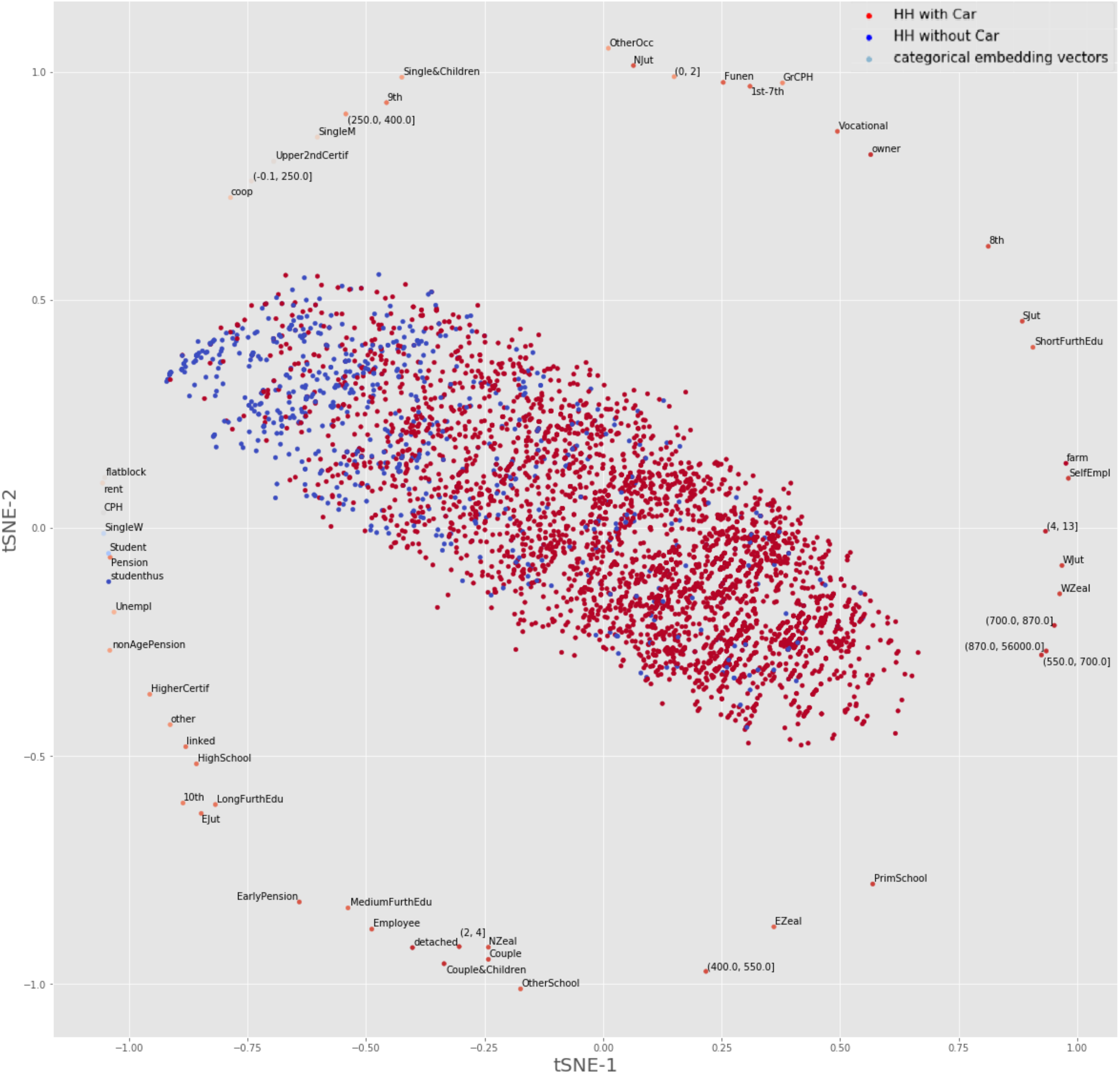

Figure 4: *2-D tSNE projections of the household embeddings and the categorical embeddings for the years 2010-2011.*

We can observe that the household embeddings are distributed in the space in a meaningful way, as they tend to cluster towards a direction based on their target labels (car ownership - no car ownership). The embedding space S can be thus, roughly divided in 3 segments or areas: (i) an area of low probability for car ownership (upper-left), (ii) an area of high probability for car ownership (lower-right), and (iii) an area with more balanced probabilities in the middle that contains observations with overlapping labels. It is also worth-noticing that the direction along which the data is spread, is aligned with the positions of the 2 major clusters of categorical vectors described in the previous section that are associated to lower and higher probability for car ownership respectively.

In order to test the effectiveness of the new centroid representations we used them as input in a logit choice model and compared their performance to the traditional dummy variables encoding method. Both the in-sample and out-of-sample performance of the resulting models will be presented using the training and the (reserved) test sets respectively. Since the centroid-encoding models and the dummy encoding models have different degrees of freedom, respectively 26 versus 47, McFadden's adjusted R-squared will be used as an evaluation metric that accounts for the number of predictors included in a model. The results are presented below in Figure 5 (for all the 2-year subsets).

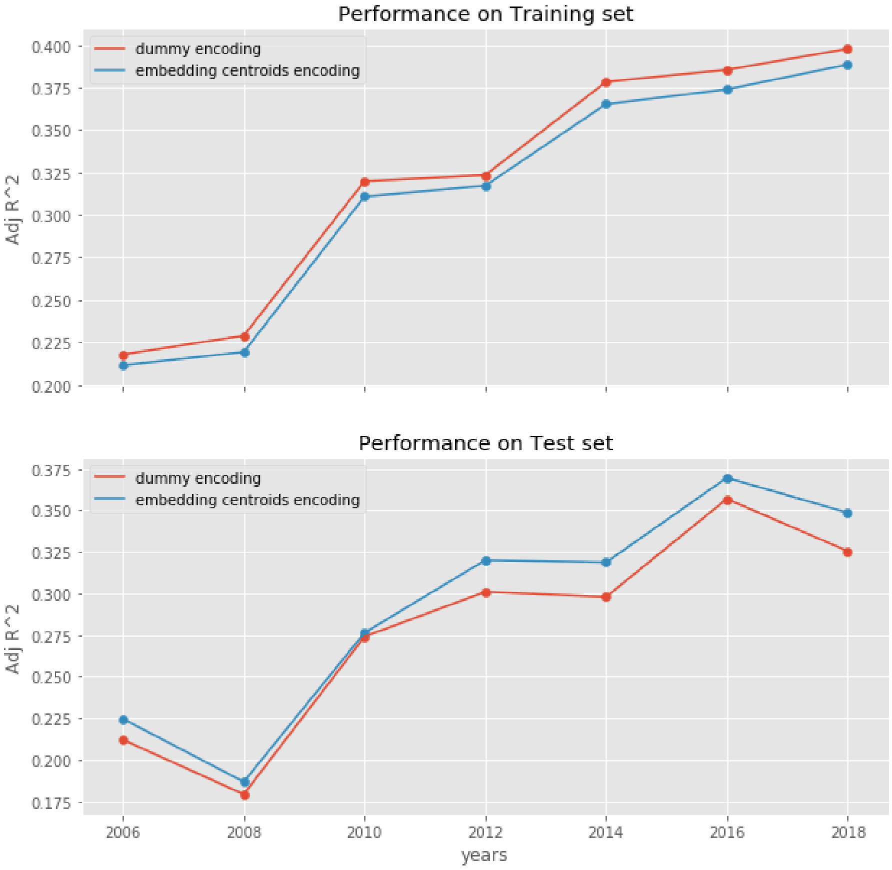

Figure 5: *Training and Test set results for embedding centroids-encoding vs the dummy encoding models (for all the 2-year subsets).*

We can observe that while the dummy encoding model performs better than the proposed model in the training set, the embeddings centroid-encoding model yields better results in the test set for all the 2-year subsets considered[8]. This proves the compositionality of the categorical embeddings and the effectiveness of the proposed centroid-based method for obtaining meaningful embedding representations on a household level.

5 Prototypical Households

In this section we will demonstrate how the embeddings centroid-based method introduced in the previous section can allow us to define theoretical households that are representative or

[8] We also note that additionally to adjusted R^2 Akaikes information criterion (AIC) was calculated to compare the 2 models which produced similar results.

prototypical to the classes of the study variable - in our case car ownership *vs* no car ownership.

The prototypical households vectors are defined as the centroid embedding vectors composed of *categorical vectors that exhibit an homogeneous and pronounced behavior with respect to the study variable*. These prototypical households are theoretical, and thus, they may be observed or not observed in the data. In order to obtain such representations we follow a 2-step process described below.

Step 1: finding a set of households U *that are composed by categories that exhibit homogeneous/uniform behavior towards car ownership.*
Let V be the set of the n categorical explanatory variables we considered and C be the set of all the categorical embedding vectors in the embedding space S, such that each $c_i \in C$ corresponds to a $v_n \in V$.
(i) Iterate over every $c_i \in C$.
(ii) For each $v_n \in V$, choose among its corresponding categories the one that is closest to c_i in S, using cosine similarity as a proximity metric.
(iii) Compute a household vector h_i using all the chosen categories and equation (1).

After this repeating this process for each 2-year subset we ended up with a set of household vectors for each period. These household vectors are plotted in Figure 6 for the period 2010-2011, together with the vectors of the observed households for this period.

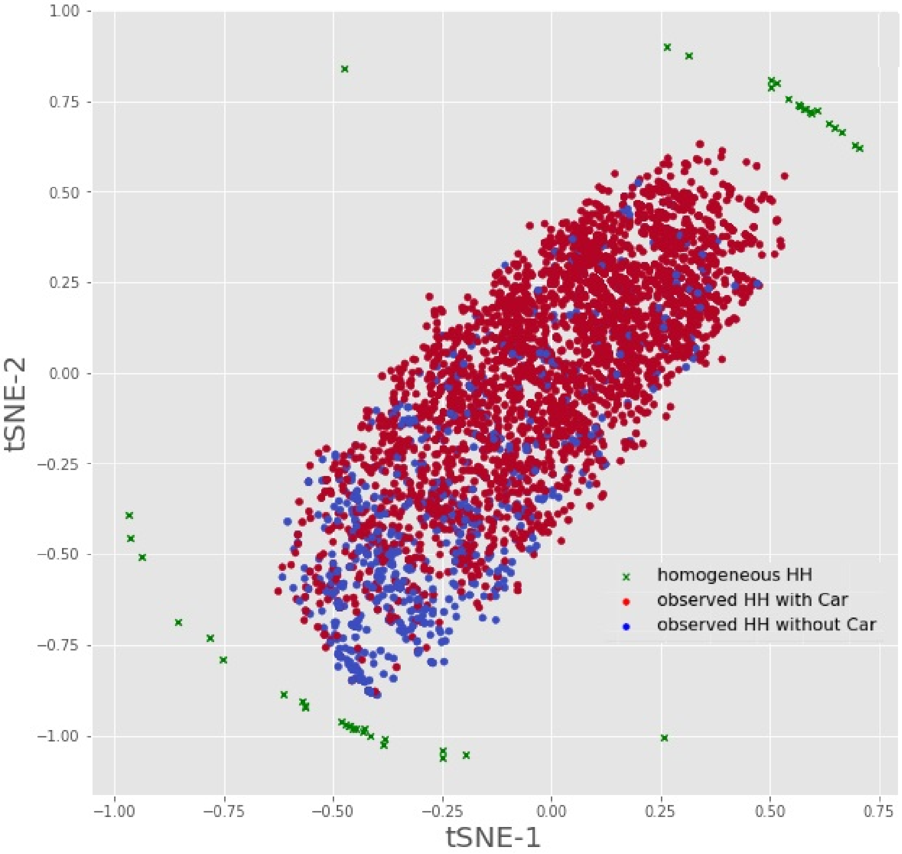

Figure 6: *2-D tSNE projections of the homogeneous household embeddings and the observed household embeddings for the years 2010-2011.*

Step 2: finding a subset of households $P \subseteq U$ *that are composed by categories that exhibit a pronounced behavior towards (1) car ownership and (2) not owning a car.*
(i) Apply cosine similarity clustering on the embedding vectors of U using a similarity threshold

t[9].
(ii) From the emerging clusters in (i), choose the two major ones that are associated with low and high probability of car ownership in the embedding space.

The resulting clusters are composed of prototypical household vectors with respect to our classes, i.e. prototypical households for car ownership and prototypical households for not owning a car. Thus P can be divided into 2 mutually exclusive clusters of prototypical household embeddings: P_{Car} and P_{noCar}. The prototypical household vectors are plotted in Figure 6 for the period 2010-2011, together with the observed household vectors corresponding to the same time period.

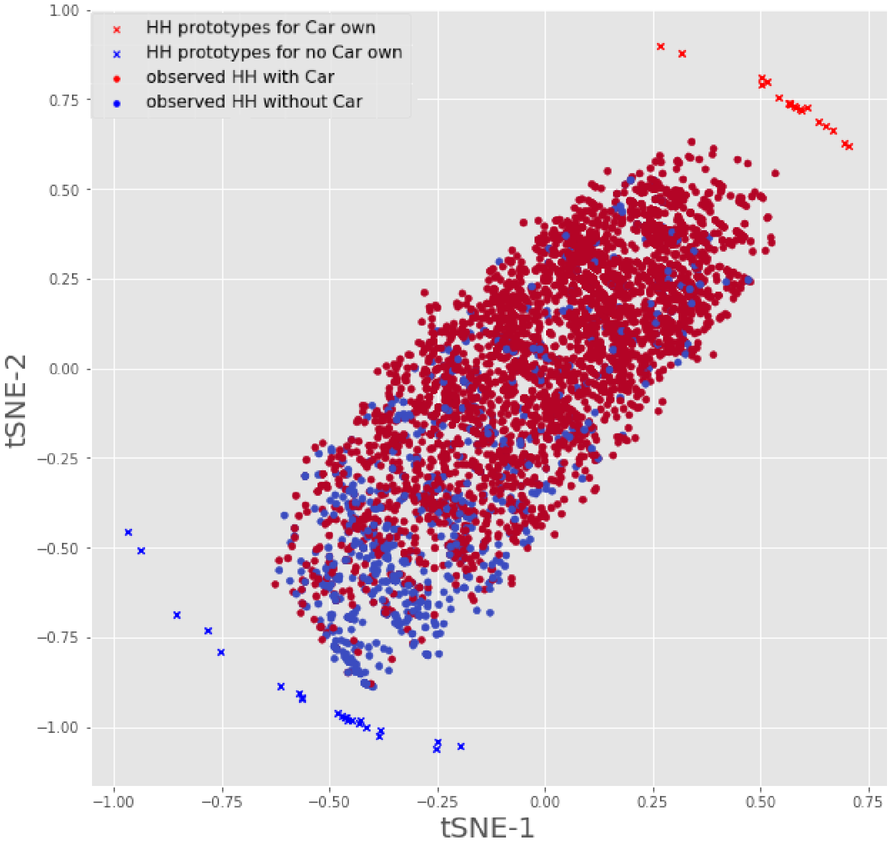

Figure 7: *2-D tSNE projections of the prototypical household embeddings and the observed household embeddings for the years 2010-2011.*

In order to provide an intuitive and short description of P_{Car} and P_{noCar}, we created word clouds[10] (shown in Figure 8) highlighting the most representative categories in each prototypical cluster. We can observe that the 2 prototypical clusters are composed of households with differing and distinct features. The main prototypical features of P_{noCar} are : living in Copenhagen area, being single or single with children, renting a flat, having lower levels of household income and being student or pensioner. On the other hand, the prototypical features P_{Car} can be summarized into: families with children, higher levels of household income, owning a house, living in a farm and having relatively low educational level. Such insights can be helpful for identifying target groups for policy initiatives. For example, if the objective is to reduce car ownership and encourage other modes of transport, the policy efforts should be mainly focused

[9] For the purposes of this study we set $t = 0.95$

[10] using tf-idf weighting scheme

on groups of the population that are characterized by the socioecomonic features of P_{Car} rather than P_{noCar}.

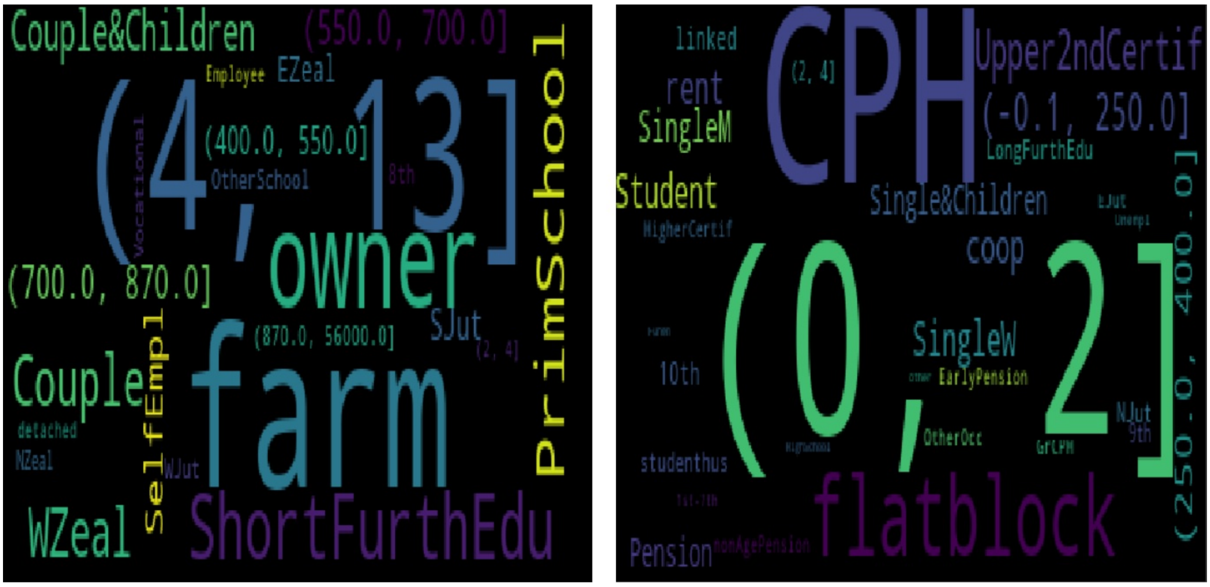

Figure 8: *Wordclouds for the prototypical household clusters P_{Car} and P_{noCar} associated with car ownership and no car ownership respectively.*

6 Conclusion and Future Perspectives

This paper presented a joint space approach of representing categorical variables as embedding vectors based the PyTre algorithm for traveling embeddings generation introduced in [1]. The resulting representations form meaningful clusters in the embedding space with respect to the target variable allowing us to detect areas with differing degrees of probability for car ownership in the embedding space, and identify groups of categories that exhibit similar behavior towards the study variable.

The main advantage of the joint space approach however, was that it allowed us to take advantage of the compositional capabilites of the categorical vectors and use them to represent households, extending the representations to a higher hierarchical level and change our focus from the level of variables to that of observations. This was achieved by introducing the centroid-based method, an NLP inspired method originally used for document representation. The new household embeddings encoding was compared to the dummy variables encoding and managed to outperform the traditional encoding method in out-of-sample evaluation, proving the effectiveness of the proposed method to produce meaningful and high-quality respresentations. However we should highlight that the mapping of individual categories back to the “dummy" space is not as straightforward as in the separate space approach presented in [1]. Since there is no meaning for the value of the beta coefficients in an embeddings encoding [1], further work is needed towards this direction to infer such mappings that will allows us to analyse the individual coefficients and the effect of a given category on the study variable.

The centroid-based method not only allowed us to represent households *observed* in the data but also to define *prototypical* households in respect to the two classes (car ownership - no car ownership), identify their key characteristics and their position in the embedding space. In the future, we intend to use the prototypical households vectors to define points of reference in the embedding space in order to measure the prototypicality of the households observed in the data. This process will be useful for ranking the households according to how prototypical they are with respect to car ownership but also to obtain numerical values (measurements) of this property and introduce prototypicality as a new continuous variable.

Lastly we intend to use the embedding representations that we generated for consecutive time periods in order to capture temporal behavioral dynamics, and understand how the study behavior evolves over time for different subgroups of the population. For this purpose, we intend to use NLP-based methods - originally used for automatic translations between different languages spaces- in order to acquire linear mappings between the embedding spaces of different time periods that will ultimately allow us to compare their elements.

References

[1] Pereira, F. C. (2019). Rethinking travel behavior modeling representations through embeddings. arXiv preprint arXiv:1909.00154.

[2] Son, L. H., Allauzen, A., Wisniewski, G.,& Yvon, F. (2010, October). Training continuous space language models: Some practical issues. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 778-788). Association for Computational Linguistics.

[3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3

[4] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

[5] Perone, C. S., Silveira, R., & Paula, T. S. (2018). Evaluation of sentence embeddings in downstream and linguistic probing tasks. arXiv preprint arXiv:1806.06259.

[6] Camacho-Collados, J.,& Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research, 63, 743-788.

[7] Feng, S., Cong, G., An, B., & Chee, Y. M. (2017, February). Poi2vec: Geographical latent representation for predicting future visitors. In Thirty-First AAAI Conference on Artificial Intelligence.

[8] Zhao, S., Zhao, T., King, I., & Lyu, M. R. (2017, April). Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In Proceedings of the 26th international conference on world wide web companion (pp. 153-162).

[9] Yan, B., Janowicz, K., Mai, G., & Gao, S. (2017, November). From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 1-10).

[10] Wang, Z., Li, H., & Rajagopal, R. (2020). Urban2Vec: Incorporating Street View Imagery and POIs for Multi-Modal Urban Neighborhood Embedding. arXiv preprint arXiv:2001.11101.

[11] Yabe, T., Tsubouchi, K., Shimizu, T., Sekimoto, Y., & Ukkusuri, S. V. (2019, November). City2City: Translating Place Representations across Cities. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 412-415).

[12] Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing Management, 40(6), 919-938.

[13] Rossiello, G., Basile, P., Semeraro, G. (2017, April). Centroid-based text summarization through compositionality of word embeddings. In Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres (pp. 12-21).

[14] Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 746-751).