Conditional Topic Allocations for Open-Ended Survey Responses

Tobias Wekhof[*]

September 8, 2022

Abstract

This paper develops a novel topic model for text data by conditioning on observables, named the "Conditional Topic Allocation" (CTA). This data-driven method allows identification of latent topics that explain other observable variables. It is particularly suited for small-scale text data, such as open-ended survey responses. First, CTA is used to extract topics from open-ended text answers that explain single observable variables. Then, in empirical models where text is a control variable for unobservable characteristics, CTA's flexible scope of applications allows uncovering latent variables from the text. As a proof-of-concept, this approach is used to analyze the sentimental value homeowners place on their homes and how this relates to energy-efficiency valuation. Specifically, responses from open-ended survey questions are used as control variables in a hedonic regression, and CTA serves to identify latent preferences associated with nearly 50% of the valuation of energy efficiency for single-family houses.

JEL Classification: C83, Q50, Q40, D12, D91.
Keywords: topic model, natural language processing, open-ended questions, energy-efficient housing

[*] ETH Zurich, Center of Economic Research, Zurich, Switzerland, twekhof@ethz.ch.

I want to thank Elliott Ash, Adan Martinez Cruz, Anne-Kathrin Faust, Massimo Filippini, Sébastien Houde, Christopher Roth, Dominik Rothenhäusler, and Michael Stiefel, the participants of the internal CER-ETHZ seminar, the 1st Monash-Warwick-Zurich Text-as-Data Workshop, the C-SEB seminar, and the 9th AWEEE. The Swiss Federal Office of Energy (SFOE) provided funding for this research under the grant SI/501886-01. The SFOE was not responsible for the study design, the collection, analysis, and interpretation of data, or writing this paper. The content does not necessarily represent the official views of SFOE. All omissions and remaining errors are my responsibility.

1 Introduction

Current text-analysis methods from artificial intelligence and its application in Natural Language Processing (NLP) are designed for large samples and often perform poorly on smaller data sets. In this context, the recent attention on survey data with only several hundreds or thousands of respondents provides a resourceful use case. Such a setting allows for collecting text data from respondents in combination with other variables from standard closed-ended questions. These text answers (i.e., written narratives) can be a powerful way to learn about individual preferences and characteristics. The unique nature of survey responses, which allows combining text answers with additional survey variables, impedes the development of new ways to conduct topic analysis.

This paper develops a method to explain numerical outcome variables with a moderate sample of text data and thereby identify topics. I call this method "Conditional Topic Allocation" (CTA), which allows topic identification in open-ended text answers that are correlated, and hence conditional, with other survey variables. These topics could therefore indicate latent variables contained in the text. This data-driven method enables applications in a wide range of settings where text data and other characteristics of each text sample are available. It is particularly suited for survey data because this type of data consists of small samples with short and context-specific text answers combined with other observable variables. Survey data provides a context where the link between both data types is straightforward: respondents answer open-ended questions that provide the text data used to allocate topics. Further, other standard closed-ended questions provide information about the respondents, as is typical in conventional surveys. Concisely, CTA consists of performing a topic extraction by conditioning on observables.

In its primary application, CTA allows an explanation of a single-outcome variable. Further, CTA's flexibility enables an explanation of text when it serves as a control variable in a regression. In this second application, the information from text data is used as a covariate to address an omitted variable bias (OVB). If not included in the model, the information in the text could bias the explanatory variable. Here, CTA allows identifying topics extracted from the text data that are correlated with the missing latent variables. These topics can be associated with either a negative or a positive bias on an explanatory variable. It is hence necessary to include the information captured in the text as a set of numerical control variables to address the OVB. Once this information from the text is converted to numerical data, it is of considerable size; To analyze the data, I therefore use a method called double Machine Learning (double ML), which is specifically designed for this purpose (Chernozhukov et al. 2018).

In this study, I apply CTA to explain homeowners' valuation of energy-efficient housing in Switzerland with open-ended text answers. Housing decisions involve numerous factors, only some of which are readily observable. For example, a homeowner's often emotional attachment to a house is challenging to quantify when performing an economic analysis. These feelings are essential to how homeowners perceive their houses, but their effect often remains latent in econometric analysis. Text data can capture subtle variations related to emotions important in the housing decision and a potential source of bias if left unaccounted for. Open-ended survey questions provide a new opportunity to capture the complex individual narratives associated with the sentimental value of a home.

This paper contributes to the literature along three dimensions. First, it adds to the range of methods for topic analysis with text data, an important field of NLP. Usually, topic analysis methods rely on unsupervised models. In contrast, CTA uses the particular data structure of surveys; this provides additional variables and allows a supervised approach.[1] In Machine Learning, a supervised approach describes a setting where an algorithm is trained

[1] The majority of the literature that analyzes text data in social sciences relies on large corpora of existing text such as newspaper articles, congressional speech, or Twitter feeds (e.g., Baylis 2020, Belmonte and Rochlitz 2019, Benesch et al. 2019, Gentzkow et al. 2019, or Morales 2021).

with a dataset (e.g., text data) to predict another variable. Hence, CTA identifies topics that are predictors for an outcome variable.[2]

In contrast to the supervised approach, in an unsupervised approach, the algorithm identifies topics solely based on the text and without conditioning any other variables. One of the main unsupervised text-analysis tools used on larger text-corpora in social sciences is the Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003). Building on the LDA, the structural topic model (STM) by Roberts et al. (2014) is widely used in text-topic analysis. In addition to LDA, the STM allows for conditioning those topics on covariates such as characteristics of survey respondents or experimental treatments. In both models, the retrieved topics are inherent to the text and not directly connected to a specific outcome.[3] A third related model consists of the supervised Indian Buffet Process (sIBP) by Fong and Grimmer (2016). This model infers latent treatments from text samples that correlate with an outcome variable, similar to CTA, where latent topics from the text are correlated with an outcome. There are several differences in the applications: sIBP uses text as an independent treatment that affects individuals, and the number of text samples can be smaller than the number of individuals. CTA, in contrast, was designed for applications from short text samples as they occur in survey data, where respondents write the text, and the text contains latent characteristics about the respondent

Finally, CTA allows for very flexible applications in inferring topics: it can explain a single-outcome variable or a more complex setting where text is a covariate, such as with double ML or conditional average treatment effects.

[2] In the literature, some examples use text to explain an outcome variable in a similar setting to CTA. Like the basic setup of CTA, words are associated with a positive or negative effect on an outcome. However, these studies do not use open-ended survey responses and allocate words manually into topics (e.g., Netzer et al. 2019)

[3] Similarly, other types of topic models based on LDA incorporate covariates as labels in their algorithm (McAuliffe and Blei 2007, Ramage et al. 2009). Conceptually, these models differ from CTA because the covariates are used as additional information to find topics. However, the identification strategy of these topics does not explicitly aim to explain the outcome variable. Moreover, these models depend on the LDA, while CTA is model agnostic.

Second, this paper contributes to the literature on open-ended survey questions. CTA makes this type of survey response more accessible and facilitates its analysis. Recognizing the increasing importance of text as data in social sciences highlights that little research has sought to explicitly ask people to speak or write their minds and then use this information to elicit preferences. In the context of surveys, open-ended questions provide a method for this purpose.[4] Historically, open-ended questions have been used in social sciences only on a small scale mainly because researchers had to perform the coding manually (Krosnick 1999, Roberts et al. 2014). However, in contrast to closed-ended questions, the open-ended format does not suffer from priming respondents; rather, it encourages them to write what is on their minds. Recently, with advances in artificial intelligence and text analysis, open-ended survey questions have been used to address important questions in social sciences, such as on immigration (Bursztyn et al. 2020, Egami et al. 2018), climate change (Tvinnereim and Fløttum 2015), macroeconomic shocks (Andre et al. 2021, 2022), and sustainable finance (Filippini et al. 2022), and also to elicit policy preferences (Ferrario and Stantcheva 2022, Houde and Wekhof 2021).

Third, applying CTA to the housing market contributes to the literature on the so-called green premium of buildings. This topic has attracted significant attention in the academic and policy spheres. Usually, it consists of estimating the willingness to pay for a green certification that buildings obtain if they meet specific ecological standards. A large body of evidence suggests these certifications lead to a substantial premium and are thus valued highly by market participants (Brounen and Kok 2011, Kahn and Kok 2014). The majority of the evidence, however, relies on cross-sectional regressions in which the certification is only one observable among many others (Eichholtz et al. 2010, Fuerst and McAllister 2011, Koirala et al. 2014). In this paper, I show how CTA can use narratives to uncover new insights about individual preferences with a focus on the sentimental value of housing,

[4] An open-ended question describes a question that respondents cannot answer with a static response (i.e., yes/no or with a single word). Hence, open-ended questions require writing an initial response.

and I revisit the green premium. My approach uses double Machine Learning combined with open-ended survey text responses to estimate how households value an environmental certification.

The organization of this paper is as follows: section 2 explains the idea behind CTA and the necessary steps for its implementation. Section 3 describes the empirical strategy that uses text to address OVB and explains the bias with CTA. Section 4 explains the data, and section 5 presents the results of CTA to explain single observable variables. Section 6 describes the results with text to address an OVB in the hedonic regression and explains the bias with CTA. Section 7 contains concluding remarks.

2 CTA - basic concept

CTA facilitates explaining observable variables with written text samples by identifying relevant topics within the text. The general setting for this application consists of text data combined with one or several observables using a relatively small sample size. Performing CTA involves three steps: first, the text data is preprocessed and used to predict an observable variable. Second, each word's positive or negative importance for the prediction is computed using Shapley additive explanations (SHAP) values, an approach developed by Lundberg and Lee (2017), which is explained in section 2.2. The last step allocates words with a positive or negative predictive power separately into topics according to their semantic similarity.[5]

2.1 Step 1: predict an outcome variable with text data

The first step transforms the text answer into numerical data to predict an outcome variable; this is done in the data preprocessing. First, preprocessing considers the words of each text

[5] The semantic similarity between two words describes how closely two words relate to each other for the content they describe. For example, the words pen and paper are very similar in semantic terms; however, pen and monkey are less similar.

response individually and retains only nouns, adjectives, verbs, and adverbs from the text. This procedure, known as part-of-speech (POS) tagging, was implemented with the Python-based spaCy library (Honnibal et al. 2020).

Next, the spaCy library lemmatizes all remaining words, replacing each word with its basic root form as it would exist in a dictionary. This canonical form of a word is called a lemma (e.g., the lemma of the word better is good.) In the next step, the most-frequently used words are retained.[6]

Finally, a feature matrix comprised of text covariates ($Text_i$) is constructed as a so-called bag of words. In this matrix, each column designates a feature that indicates with a dummy variable if a word is present in an individual response.

After transforming the text answers into a numerical feature matrix, an ML algorithm predicts other numerical variables from the survey. The underlying assumption is that the text response from respondent i correlates with a specific characteristic of that respondent, $Text_i \sim Y_i$.

The text answer, $Text_i$, is a high-dimensional matrix, and the correlation structure between the $Text_i$ and the outcome Y_i is likely to be non-linear because the data originates from unstructured text. The function $F()$ approximates that correlation structure by mapping the $Text_i$ to the outcome Y_i. Using the text answers and the outcome Y, I estimate $\widehat{F()}$.

With $Text_i$ as input, $\widehat{F()}$ returns γ_i, an approximation of Y_i. Hence,

$$\widehat{F(Text_i)} = \gamma_i \tag{1}$$

[6] Depending on the ML algorithm used for the prediction step, there may be different thresholds for the minimum frequency of these words because words appearing with a low frequency are a source of noise. Hence, checking the sensitivity of the prediction to different thresholds may provide additional robustness to the analysis. However, CTA's main objective is identifying important topics with predictive power, not optimizing the prediction alone.

where $\gamma_i \sim Y_i$.

This prediction step is model agnostic and can use any machine learning algorithm. The small sample size of survey data favors using the entire sample to train the algorithm and predict the outcome variable. To avoid overfitting, the estimation of the propensity score γ_i employs a 10-fold cross-validation. First, cross-validation partitions the data into 10 randomly chosen bins. Then, the algorithm uses nine of the 10 bins to train the model and estimate γ_i for the tenth bin. This procedure repeats 10 times until the algorithm estimates γ_i for the entire sample.

2.2 Step 2: compute feature importance

The second step consists of computing the feature importance (i.e., the importance of each word for the prediction). This phase uses the numerical SHAP method developed by Lundberg and Lee (2017) and is based on the Shapley values from cooperative game theory (Shapley 1953). The underlying idea is to treat the model $F()$ as a black box that produces an output (here, γ_i) using a specific input of features (here, $Text_i$). In the game-theoretical approach, each word represents a player contributing to a joint outcome, the predicted propensity score. A permutation-based approach computes the individual contributions, which results in the marginal contribution of each word to each prediction. The SHAP value for an individual word shows the word's contribution to the predicted propensity score for each individual.[7]

After computing the SHAP values, the resulting output matrix will contain a different SHAP value for each word in each response. In this new matrix, each column represents a word from the corpus. Within a column, each observation consists of a SHAP value, including words not used in the individual answer. As a result, a word vector contains an array of SHAP values that describe the individual effect of the particular word in predicting

[7] Note those values do not show a causal relationship. Moreover, the SHAP values describe a word's contribution on an individual level and conditional on the other specific words the individual answer contained.

the outcome variable for every respondent. Based on these SHAP values, it is possible to calculate the average SHAP value for each word; this can be either a positive or a negative value. Further, a t test on the average SHAP value determines if a word significantly impacts the prediction. This step results in two sets of words that have either a positive or a negative impact on the prediction.

2.3 Step 3: allocate words to topics

The relevant keywords for each prediction identified using the SHAP values in Step 2 are still too numerous to allow for an interpretation of the words. Therefore, this third step allocates keywords to topics according to their semantic similarity. In order to allocate the most important keywords to topics, I propose a data-driven method based on word embeddings, which is efficient on small text samples.

The semantic distance between the keywords relies on a word embedding trained on the main text corpus consisting of all the text answers.[8] The word embedding consists of a matrix that assigns an embedding vector to each word in the text corpus. Each vector measures the semantic distance to the other vectors in the matrix. This is obtained by computing the cosine similarity between two vectors. For example, the distance between the words pen and paper is smaller than between pen and monkey.

However, the small text corpus with its relatively few answers, short texts, and specific vocabulary makes it challenging to train well-performing word embeddings. For this reason, it is necessary to replace the embedding vectors for all words whenever possible with pretrained embeddings. This study uses the pretrained German fastText word embeddings (Grave et al. 2018). After that, the word2vec algorithm updates the new-word embeddings to the specific context of the text answers by retraining on the initial text corpus, but using the new embeddings as initial weights.

[8] I use word2vec with the continuous bag of words (CBOW) algorithm from the Python library Gensim (Řehůřek and Sojka 2010). The embedding size is 300 and is computed with 50 epochs.

Algorithm (1) outlines how to allocate the selected positive or negative keywords to topics. Based on the matrix with all distances between words, the allocation starts with the first topic by selecting the vector assigned to the word with the highest absolute SHAP value. Next, the closest word is selected based on its cosine similarity to the starting word. The function $h()$ returns either the cosine similarity between two words or the average similarity between one word and a set of words. This similarity between the first two words that compose a topic, $maxsim^{(k)}$, serves as a benchmark for future words to be included in the cluster. Consequently, this is the most prominent similarity between two words within that topic. However, these first two words only form a separate topic if their similarity is not below a minimum similarity, α, which the researcher defines.

To select the third word for the topic and all subsequent words, the distances to all other words are averaged over the word vectors already included to select the closest word. Here, a threshold is applied, which is labeled δ. This threshold defines as a fraction of the similarity between a topic's starting word and its closest word. The algorithm repeats this process and adds words to the topic until the distance from the newest word to all other words in the topic exceeds the threshold similarity value, δ. After adding a word to a topic, the initial word list omits the respective word. For the next cluster, the starting word is selected as the word with the highest SHAP value in the list of remaining words. After that, the algorithm repeats the steps from the first topic until it has allocated all the words to topics.

Finally, there are three possibilities to select a given number of topics from all topics: (a) the most important topics based on the highest SHAP value among all words in the topics; (b) a selection that depends on the highest average SHAP value per topic; (c) a choice based on the highest impact — the product of SHAP values with each word's overall frequency. The subsequent analysis uses the first option to select the four most important topics, which are ranked based on their most decisive word in terms of SHAP values.

Algorithm 1 Allocating words to topics

```
$w \leftarrow \{words \mid (pvalue(shap(word_i)) < 0.01)\}$
while $n^{(w)} \geq 2$ do
    $k \leftarrow newtopic$
    $word_i^{(k)} \leftarrow argmax_i(shap(word_i^{(w)}))$
    $word_j \leftarrow argmax_{w-i} h(word_{-i}^{(w)}, word_i^{(k)})$
    $maxsim^{(k)} \leftarrow h(word_i^{(k)}, word_j)$
    if $maxsim^{(k)} \geq \alpha$ then
        $word_j^{(k)} \leftarrow word_j$
        while $meansim^{(k)} \geq maxsim^{(k)} \times \delta$ do
            $word_m \leftarrow argmax_{w-i,j} h(word_{-(i,j)^w}, words^{(k)})$
            $meansim^{(k)} \leftarrow h(words^{(k)}, word_m)$
            $word_m^{(k)} \leftarrow word_m$
            $w \leftarrow w_{-m}$
        end while
    end if
end while
```

Note: This algorithm allocates keywords to topics based on their semantic similarity. It is necessary to compute the feature importance beforehand (e.g., with SHAP values). Moreover, the algorithm relies on pretrained word embeddings for all keywords. The parameters α and δ influence the number of topics, the selected keywords, and the number of keywords per topic. These parameters can be set either manually by the researcher or by optimizing the topic's semantic attributes, such as the interpretability score. The variable $n^{(w)}$ describes the number of words in the list of words w, and $word_i^{(k)}$ denotes word i in topic k.

Selecting the optimal parameters for the two thresholds involves calculating the topics' interpretability scores. Similar to Dieng et al. (2020), the interpretability score is based on two intermediate scores. First, the coherence score measures the average similarity of the words within a topic (intra-topic). Second, the diversity score describes the inter-topic difference between two topics by calculating the average similarity between all pairs of words between the two topics. The diversity is subtracted from one to compare it to the coherence score.

$$coherence_k = \frac{n^{(k)}(n^{(k)}-1)}{2} \sum_{i \neq j}^{i,j} h(word_i^{(k)}, word_j^{(k)})$$

$$diversity_{k,l} = 1 - \frac{1}{n_{k<l}^{(k)}!} \sum_{i \neq j}^{i,j} h(word_i^{(k)}, word_j^{(l)})$$

where $h()$ is a function that returns the cosine similarity based on the word-embeddings between two words. For the coherence score, the word pairs are within the same topic (word i in topic k, $word_i^{(k)}$, and word j in topic k, $word_j^{(k)}$). The word pairs in the diversity score are from two different topics (word i in topic k, $word_i^{(k)}$, and word j in topic l, $word_j^{(l)}$). The variable $n^{(k)}$ describes the number of words in topic k, $\frac{n^{(k)}(n^{(k)}-1)}{2}$ is the number of pairs of words in topic k and $\frac{1}{n_{k<l}^{(k)}!}$ is the number of pairs of words between topic k and topic l, n_k is the total number of topics.

Further, the quality measure between two topics is the product of the average coherence scores and their diversity score. Finally, the interpretability score consists of the average quality for all topics multiplied by the average absolute SHAP value from the words of the five major topics. This score provides a single number for all topics. Including the average SHAP values in the optimization considers the predictive power of the most important topics. For this step, the researcher decides on the number of topics to include in the interpretability score.

$$
\begin{aligned}
quality_{k,l} &= \frac{coherence_k + coherence_l}{2} \times diversity_{k,l} \\
interpretability &= \frac{1}{5}\sum_{k=1}^{5}\frac{1}{n_i}\sum_{i=1}^{i} abs(shap(word_i^{(k)})) \times \frac{1}{n_k}\sum_{k=1}^{k} quality_k
\end{aligned}
$$

Next, finding the optimal value for both thresholds, α and δ, is necessary. The optimal values result from maximizing the interpretability score with a grid search that considers an interval of values for both variables.

3 Application to hedonic regressions

I apply the idea of using CTA to explain an OVB when text serves as a control variable in a regression. NLP combined with narrative elicitation can be used to overcome an OVB and help explain the respondent's underlying reasoning with specific topics. Before applying CTA to explain what topics can reduce an OVB, I first describe how text can be used as a control variable to address bias. In a second step, I apply CTA to explain what topics from the text are associated with biasing an explanatory variable.

Specifically, I apply this idea to hedonic regressions (Rosen 1974). Hedonic regressions aim to recover households' preferences from equilibrium prices. This method plays a central role in economics, especially in environmental and urban economics because preferences for different environmental amenities can be critical determinants of housing decisions. There is, however, an inherent difficulty in estimating hedonic regressions, especially in the housing market. Each building has its own set of unique characteristics. Researchers do not observe all of them, and these characteristics are perceived differently by each market participant. In addition, consumers may have individual preferences that are not homogeneous. Due to unobserved preferences, hedonic regressions are prone to endogeneity. Several techniques are used to try to overcome this problem, which has led to the popularity of using quasi-experimental techniques when possible. In that case, repeated sales of the same property control for time-invariant unobservables. Alternatively, structural approaches are available (Bajari and Benkard 2005).

In this paper, I apply a third complementary approach by incorporating recent advances in machine learning and econometrics. Veitch et al. (2020) suggested using text as a control variable to address OVB in a general setting. Because of the high dimensional nature of text data, an approach called double Machine Learning by Chernozhukov et al. (2018) allows this type of analysis. However, until now, applying text data as a control variable has received little attention (e.g., Manzoor et al. 2020).

Equation (2) shows the basic setup of a hedonic regression for the valuation of the Minergie-label :

$$house\ value_i = \beta_0 + \beta_1 * Minergie_i + X_i + \epsilon_i \tag{2}$$

where X_i is a set of covariates that captures preferences. Omitted characteristics in the error term might correlate with *Minergie* and the house value.

In the following, I argue the information in the text answer can constitute an OVB. Further, partialling out that variation using the double ML method, advocated by Chernozhukov et al. (2018), can reduce that bias. Recently, Manzoor et al. (2020) used double ML with text as a robustness check to show the text does not influence their model. The intuition behind double ML is to control for many covariates in a regression setting.

Equation (3) shows the hedonic regression setup that now includes $Text_i$:

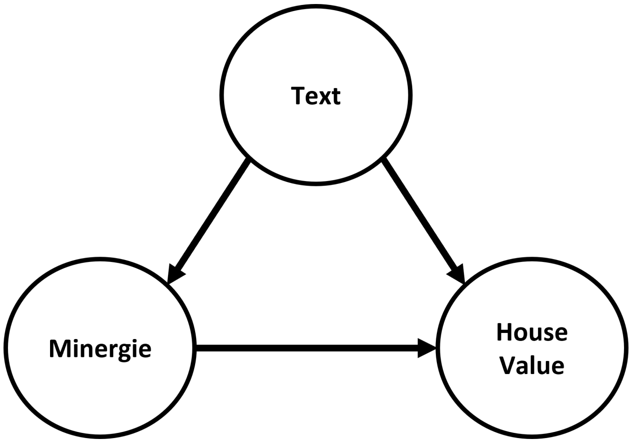
$$house\ value_i = \beta_0 + \beta_1 * Minergie_i + X_i + Text_i + \epsilon_i \tag{3}$$

where $Text_i$ is a large matrix that contains the information from the open-ended text answer in the form of a bag of words. Each column of the feature matrix represents a word as a dummy variable that takes the value of one when the word appears in the text answer, i, and zero otherwise.

The open-ended question aims to capture the sentimental value respondents attribute to their homes. In the context of the OVB, the purpose of the text responses is to capture elements correlated with both the housing value and the green certification. To compare the open-ended with closed-ended questions, respondents rated their importance of housing attributes on a scale between one and seven. In addition, respondents rated their happiness concerning their house in a closed-ended question.

Figure 1 shows my approach in a directed acyclic graph (DAG). The basic hedonic regression infers the effect of Minergie on house value. Both variables are, however, influenced by omitted variables. In this case, I assume one major component (among others) of that bias lies in an individual's emotional attachment to the person's house. Hence, I further assume it is possible to capture this attachment with open-ended survey responses. Depending on which elements of the house an individual is attached, the text will be different and so will the effect on both Minergie and house value. For instance, it could be possible individuals form emotional attachments to a home because they renovated it themselves. These respondents may also attribute a lower valuation to Minergie buildings because it is not possible to modify them without professional assistance. It is important to note not all elements of the narrative will correlate with *both* Minergie and the house value.[9] Figure 1 displays the relationship between the elements discussed.

Figure 1: DAG with the relation between Minergie, housing value, and the information captured in the text answer



Note: This directed acyclic graph (DAG) shows the relation between the Minergie certification, housing value, and emotional attachment captured in the text responses. The Minergie certification influences only house value. The information from the text, however, influences both the Minergie label and the house value.

[9] As with all consumer preferences, reverse causality could be an issue. Respondents wrote the text answer about their house while already living in it. Therefore, the responses may be a result of the specific type of housing instead of deeper personal preferences. However, this problem extends to most house features, irrespective of whether a closed- or open-ended question recovers the information. This study assumes most text features express deeper personal preferences.

Equation (4) shows the relationship depicted in the DAG, where the text acts as a confounding variable.

$$
\begin{aligned}
house\ value &= \beta_1 * Minergie + g_0(Text) + u \\
Minergie &= m_0(Text) + v
\end{aligned}
\tag{4}
$$

The house value depends on the Minergie-certification, a function of the text $g_0(Text)$, and the error term u. The Minergie certification is also dependent on the text. However, this is through a different function, $m_0(Text)$.

It is necessary to control for the information in the text to recover an estimator for the Minergie valuation without bias. The high dimensionality of the text makes including this data as a covariate challenging. In a regular ordinary least squares (OLS) model, controlling for many covariates will lead to too-many degrees of freedom and, therefore, little statistical significance for the coefficients of interest. Alternative estimation strategies consist in using OLS, but by doing so with fewer variables. This strategy would not capture the bias correctly, but it could show a particular trend. A second alternative would be to reduce the dimensionality by clustering variables into groups. However, these clusters are sensitive to the type of clustering and the number of clusters.

A third alternative consists of partialling out the variation contained in the text. Although partialling out provides a strategy to control for the variation in $Text_i$, it removes the possibility of interpreting the coefficients for each covariate. In this specific setting, however, the coefficients of the covariates from the feature matrix $Text_i$ are not of interest. The variables from the text primarily serve as controls to obtain a more precise estimate of the variable of interest, $Minergie$. Hence, it becomes possible to use an alternative method to control for $Text_i$ by partialling out the effect of $Text_i$ from both the variable of interest *Minergie* and the dependent variable *house value*.

Algorithm 2 shows the basic steps for using double ML to partial out the information contained in the text according to the Frisch-Waugh-Lovell theorem:

Algorithm 2 Double Machine Learning

```
Estimate the function $F(Text)$, such that $house\ value = F(Text) + u$
Compute the residuals $u$: $\widehat{house\ value} = house\ value - F(Text)$
Estimate the function $G(Text)$, such that $Minergie = G(Text) + v$
Compute the residuals $v$: $\widehat{Minergie} = Minergie - G(Text)$
Regress $\widehat{house\ value}$ on $\widehat{Minergie}$ to obtain $\beta_1$
```

Note: This algorithm shows the basic steps of double ML using the Frisch-Waugh-Lovell theorem. Double ML allows to partial out variables having high dimensionality. For this reason, double ML uses machine learning algorithms for the prediction steps.

After partialling out the variation captured by the text, CTA allows an explanation of the topics that influence the bias on the explanatory variable. To perform CTA, I compute the SHAP values for the feature importance of the double ML estimation. The double ML estimation does not give an individual propensity score; it provides a single coefficient for all observations. For this reason, I compute the SHAP values on the individual ratio of $\widehat{house\ value}_i$ and $\widehat{Minergie}_i$. Given $\widehat{house\ value}_i$, $\widehat{Minergie}_i$ and ϵ_i from the double ML procedure, it is possible to express the Minergie-coefficient β_1:

$$\begin{aligned} \widehat{house\ value}_i &= \beta_1 * \widehat{Minergie}_i + \epsilon_i \\ \beta_1 &= \frac{\widehat{house\ value}_i - \epsilon_i}{\widehat{Minergie}_i} \end{aligned} \tag{5}$$

Using the SHAP values for each word allows for calculating the average SHAP value for each word and the standard error of the mean. Next, I select words with an average SHAP value significantly different from zero. As before, this step provides two sets of words: those with a significant positive effect and those with a negative effect on the bias. Like in the previous analysis, CTA allocates the individual words into topics using the pretrained word embeddings.

4 Survey-Data

This study is based on a survey of single-family homeowners in Switzerland. In addition to owners of conventional buildings, the sample included owners of energy-efficient buildings with the Minergie certification. In Switzerland, certain low-energy-consumption buildings are certified with the Minergie label, which was introduced in 2001. This certification is based on building standards and practices that ensure high energy-efficiency levels and comfort. Major features of Minergie-certified buildings are a high insulation level and a unique ventilation system.[10] In the Canton of Zurich, the targeted area for this study, 2,071 single-family buildings hold the Minergie certification (1.6% of all single-family houses in Zurich).

The household survey was implemented in cooperation with the Statistical Office of the Canton of Zurich. The Canton of Zurich sent out 16,700 personalized invitation letters by postal mail directly to households on February 3, 2020, on behalf of the Centre for Energy Policy and Economics at ETH Zurich. The invitation included a link to an online survey using the software SurveyMonkey. To incentivize participation, respondents could win one of 100 gift certificates, each worth about $200 USD, in a lottery. The Statistical Office only invited homeowners living in single-family homes that were constructed prior to 1990. In addition, the sample was stratified such that 50% of the invited households had applied for a renovation permit five years prior to the survey. In 2020, the Canton of Zurich comprised 127,950 single-family homes, and the owners of 10,737 of these buildings applied for a renovation permit between 2015 and 2019.

Respondents answered the online survey between February 5, 2020, and March 13, 2020. The response rate was very high: 3,471 respondents answered the survey out of

[10] The advanced building technology also means owners usually cannot perform renovations independently, as this would damage the insulation and the ventilation system.

16,700 invitations, a response rate of 20.7%. Invitation letters were sent to 14,629 owners of conventional single-family homes, which were stratified according to the described criteria; from these invited homeowners, 2,947 completed the survey. The remaining 2,071 letters were sent to all Minergie-certified single-family homes in the Canton of Zurich. Out of the Minergie sample, 524 households responded to the survey. The final sample comprised 2,694 completed responses.[11] For the total sample, 78% of respondents were male, and the sample's mean age was 58 years.

The following two subsections present the main variables from the survey, separately for closed-ended and open-ended questions.

4.1 Closed-ended variables

Table 1 presents the summary statistics for the final sample, and it is differentiated between conventional buildings and Minergie buildings. Out of the final sample with 2,694 respondents, 500 lived in a Minergie-certified building and 2,194 in a conventional building. The average owner age was higher for owners of non-Minergie homes: 58.9 years compared to 52.5 years for Minergie homeowners. On average, traditional houses were older than the Minergie houses (61 years compared to 21 years) and with smaller square meterage (168m^2 compared to 199m^2). The fraction of female respondents was 21% for non-Minergie owners and 22% for Minergie owners. Moreover, Minergie owners had a higher income than conventional homeowners (14,836 CHF compared to 12,545 CHF). Out of all respondents, 88% lived in a coupled (two-person) household. However, Minergie homeowners more frequently lived in a coupled household with children (68% compared to 45%), likely due to the younger age of Minergie owners.

[11] Out of the initial 3,471 responses, 777 were incomplete: 553 respondents did not answer the open-ended question; 35 stopped responding after answering the open-ended question; 176 respondents did not complete the question on the house value; and 13 respondents gave a house value of zero.

Survey respondents stated the monthly rental value they estimated they would receive if they rented their home; this allows approximating the building value.[12] Because the house value is self-reported, it is likely this valuation is prone to respondents' subjectivity and thus does not only reflect the actual market value. I assume the elements to which individuals give sentimental value will influence their overall housing valuation. Compared to Minergie-certified houses, conventional homeowners self-reported a lower rental value (3,852 CHF compared to 4,437 CHF).

The survey included two questions asking respondents to rate their personal happiness and their happiness concerning their homes. These closed-ended happiness questions aimed to capture emotional sentiment. Like the questions on general happiness by Lyubomirsky and Lepper (1999), respondents rated how happy they were with their home on a scale of 1 to 7. For the second question, respondents rated their happiness for their home compared to their peers. In both questions, the two groups of respondents gave similar ratings.

Further, respondents rated the importance of several elements of their satisfaction with their home in 13 closed-ended questions (on a Likert scale of 1 to 7). These answers provided a comparison to the open-ended question. Notably, these closed-ended questions appeared after the open-ended question in the survey. Including these closed-ended questions aimed to capture classic elements used in the literature on for housing preferences (e.g., Boumeester 2011, Brounen and Kok 2011, or Coolen 2011). The following factors were made available to respondents: aesthetics, location, floor space, garden, parking, the house's monetary value, maintenance, carbon footprint, energy costs, Minergie label, indoor air quality, thermal comfort, and noise protection. For most preferences, both groups gave an equal rating except for attributes associated with the Minergie technology. The most apparent difference is in the importance of the Minergie certification (2.4 for non-Minergie owners and 5.2 for

[12] Homeowners in Switzerland generally know the approximate monthly rent they would receive for their home because this value is relevant for tax purposes (the so-called Eigenmietwert).

Minergie owners). This is followed by indoor air quality (4.6 compared to 6.1) and the carbon footprint (5.2 compared to 6.0).

Table 1: Summary statistics for survey respondents

Variable	Conventional	Minergie	Variable	Conventional	Minergie
# Observations	2'194	500	*Housing preferences [/7]*		
Socio-economics			Aesthetics	5.27	5.44
Owner age	58.83	52.49	Location	6.10	6.32
Female [%]	21	22	Floor space	5.09	5.44
Couple, with children [%]	45	68	Garden	5.82	5.72
Couple, no children [%]	43	21	Parking	4.67	4.93
Income	12'545.27	14'836.36	Monetary value house	4.35	4.72
Building characteristics			Maintenance	5.23	5.87
Building age	61.48	21.24	Carbon footprint	5.16	5.98
Floor space [m^2]	168.46	198.72	Energy costs	5.22	5.93
Rental value [CHF/month]	3852.18	4436.78	Minergie	2.37	5.23
Happiness with house [/7]			Indoor air quality	4.60	6.17
Satisfaction (compared)	5.86	5.89	Thermal comfort	5.54	6.13
Satisfaction (own)	6.18	6.23	Noise protection	5.63	6.10

Note: This table presents the summary statistics separately for conventional homeowners and Minergie owners. Additional information on the variables can be found in the Appendix in Tables F.1 and F.2.

4.2 Open-ended responses

Open-ended questions rely on a cognitive process called *recall*, in which respondents actively reflect on their answers. In contrast, closed-ended questions are based on *recognition*, a different mental process by which a respondent identifies an answer among multiple options (Anderson and Bower 1972). The research on how respondents answer these two different question types is still scarce. For instance, Reja et al. (2003) compared open- and closed-ended questions in web-based surveys. They found responses from open-ended questions spread considerably beyond the predefined topics in closed-ended questions, and the ranking of topics was similar for those categories common to both types of questions. More recently, Houde and Wekhof (2021) found respondents tend to focus on a few but more important topics explaining their decision making with open-ended questions. Compared to answers to closed-ended questions, open-ended responses uncover a broader set of topics at the population-wide level.

The literature in social sciences that reflects on using open-ended questions yields encouraging insights on implementing this survey instrument. For instance, respondents are generally motivated to answer open-ended questions and have a low dropout rate (Geer 1988). Moreover, respondents' answers to open-ended questions are generally well reflected and express concerns noteworthy to the individual who responded (Geer 1991, Schuman et al. 1986). Further, open-ended answers do not depend on respondents' rhetorical abilities, but they do show their attitudes (Geer 1988).

The first step in working with open-ended questions is to motivate meaningful answers. However, a difficulty with open-ended questions is convincing respondents to undertake the effort of writing a text that genuinely represents their thinking. It is essential the researcher communicates several points including: Why a respondent's answer is important to the researcher? How are those answers processed, and what does the researcher hope to accomplish with them? Finally, it is crucial to set the tone to guide respondents into the topic of the answer. In this case, the emotions associated with their homes are sought. For this reason, I chose the following formulation for the open-ended question:

> *We are aware that your house has a monetary as well as a sentimental value for you. The meaning of a home is reflected in memories and feelings - positive and negative at the same time. These emotions influence the projects and investments you make in your house. For this reason, we are interested in your feelings towards your home. We would therefore like to ask you to write two short texts that will enable us better to understand the emotional relationships between you and your home.*
>
> *Describe the aspects of your home that give you positive feelings. Please write a short text of about 4 sentences.*[13]

Two survey answers are presented as examples to better understand the data:

> *Out of conviction, I bought a demolition-ready property with 1000m^2 of land and made it habitable in 4 months with craftsmen friends. Through this cooperation, I learned a lot because I was a construction manager and subworker at the same time for all craftsmen. As a trained mechanic in*

[13] The original question was in the German language. This question was followed by a second open-ended question that asked the respondent to describe the elements of the home that give negative feelings. However, these responses were less informative and therefore not used for this analysis.

> *mechanical engineering, the basic requirements were there, and I did 50% of all work myself. Thus I knew everything about each electricity line, water pipe, wall, and ceiling condition, and the house became a child of mine!*[14]

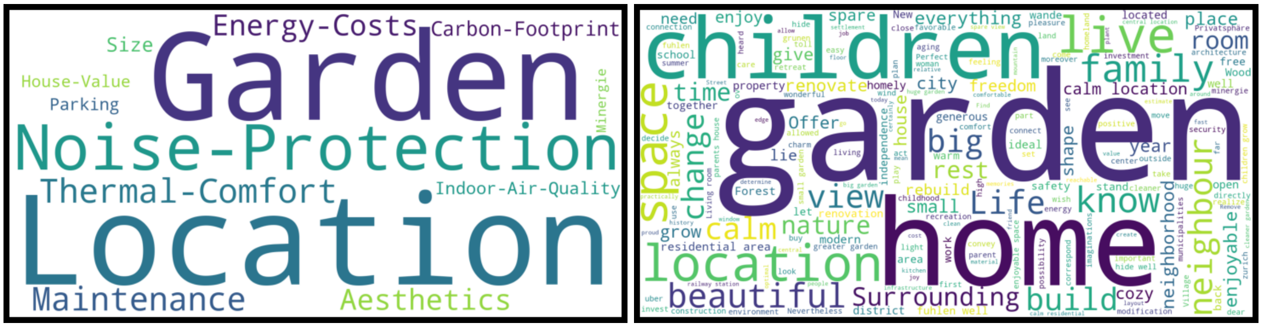
> *Small cozy house with a small garden. Cozy sun-flooded rooms. Place where all children were born and grew up. Home for our family.*[15]

Both example responses are well written and contain different topics that exceed the scope of multiple-choice answers. This difference to closed-ended questions becomes observable when comparing both answer types. Figure 2 compares the elements from the closed-ended questions and the most-used words from the open-ended answer. The left panel presents the results of the closed-ended questions on the importance of different housing features with a word cloud. If an element received a higher average score, the word appears larger. In comparison, the word cloud on the right panel shows the most important words from the answers to the open-ended question. Here, a word appears larger if it occurs more frequently throughout all responses. There is an overlap between the words in both word clouds. For instance, *garden* and *location* are critical in both word clouds. The open-ended answer allows for a much more complex scope of responses. For example, the words *family* and *children* are vital to respondents in their open-ended answers.

[14] Original German text: *Aus Überzeugung kaufte ich eine Abbruchreife Liegenschaft mit 1000m^2 Land und machte sie in 4 Monaten bewohnbar zusammen mit befreundeten Handwerkern . Durch diese Zusammenarbeit lernte ich sehr viel, da ich für alle Handwerker Bauleiter und Handlanger zugleich war. Als gelernter Mechaniker Maschinenbau waren die Grundvoraussetzungen vorhanden, und ich leistete 50% aller Arbeiten selber. Dadurch wusste ich über jede Strom, Wasserleitung, Wand und Deckenbeschaffenheit alles und das Haus wurde zu einen Kind von mir!*

[15] Original German text: *Kleines gemütliches Haus mit kleinem Garten. Gemütliche sonnendurchflutete Räume. Ort an dem alle Kinder zur Welt kamen und gross wurden. Heimat für unsere Familie.*

Figure 2: Word cloud of central elements for home satisfaction



(a) closed-ended

(b) open-ended

Note: The left panel shows a word cloud with the relative importance of 13 closed-ended questions. Respondents could state the importance of each element for their satisfaction with their home on a scale from 1 to 7. More prominent words indicate topics with a higher average score. The right panel shows a word cloud with respondents' most commonly used words to describe elements related to their home that evoke positive feelings. Words in a larger font occurred more frequently. All text answers were initially in the German language; they were lemmatized and then translated into English using DeepL Translate.

Table 2 presents the summary statistics for the open-ended responses. Out of the initial 3,471 respondents, 553 did not answer the open-ended question, which results in an attrition rate of 15.9% (224 respondents were excluded because other variables were incomplete, see footnote see footnote 11). On average, an answer consisted of 20 words, with a standard deviation of 16 words, and the 90th percentile at 42 words. The raw corpus, prior to the pre-processing, consisted of 8,235 unique words.

Next, the raw text corpus was pre-processed to convert it into numerical data. First, the pre-processing retained only the text's nouns, verbs, adjectives, and adverbs. Then, the pre-processing lemmatized the text corpus to use for numerical analysis. For instance, once lemmatized, the text from the second example changed to the following sequence of words: *small, house, small, cozy, sun, flood, room, place, child, bear, grow, home, family.* The next step reduced the dimensionality of the original text corpus by selecting only words with at least five characters. Further, words that occurred less than eight times throughout the entire data set were omitted. The final feature matrix contained 584 words in which each column represents a word in the form of an indicator variable that takes the value of one if a response contains that word and zero otherwise.

Table 2: Summary statistics for open-ended responses

Variable	
# Complete answers	2,694
Attrition rate	15.9%
Mean # words	20
90 Percentile # words	42
SD # words	16
Total # unique words	8,235
# Words after pre-processing	584

Note: This table presents the summary statistic for the open-ended survey responses for the entire sample.

5 Results: CTA and single-outcome variables

In this section, I use CTA to explain two observable survey variables, house value and Minergie ownership, with topics exacted from the open-ended survey question. To perform CTA, I first demonstrate the prediction using the text is not random. Next, I use SHAP values to show the importance of certain words and then allocate these keywords to topics.

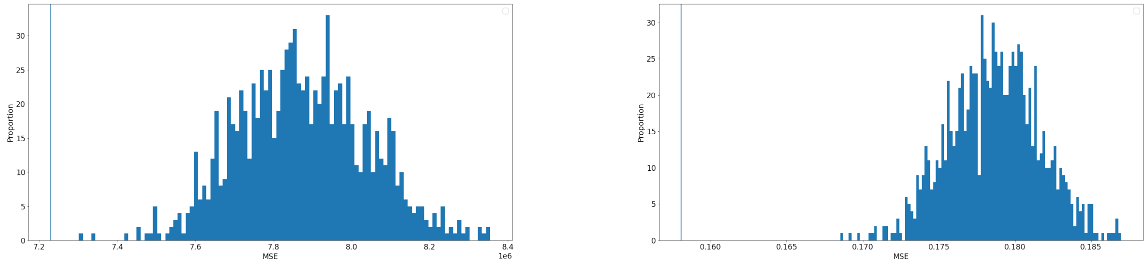
First, I predict the self-reported rental value using the text answers people use to describe the aspects of their home that give them positive feelings.[16] A permutation test with 1,000 draws shows the prediction was not random. In each permutation, the simulation shuffles train and test data to disconnect the text responses from the correct outcome variable. The machine learning algorithm then trains on the training set and predicts γ_i from equation (1) with the test data. Next, the permutation test computes the mean squared error based on γ_i and the original test data, Y_i. Because the rental value is a continuous outcome, I chose a lasso (least absolute shrinkage and selection operator) regression as the most appropriate model (with an alpha parameter of 0.1). The Minergie-ownership variable, unlike the rental

[16] Because the analysis is about house value, it is crucial to take a closer look at income distribution. Higher house values for Minergie buildings could possibly be due to a higher income for this group of homeowners. However, this is not the case in this specific survey data. A regression in Table C.1 (Appendix) shows the interaction term of income and Minergie ownership has no statistically significant effect on the rental value.

value, is a binary indicator that requires a classification method for the prediction. For this reason, I chose a random forest classifier to predict the Minergie-ownership variable.

Figure 3 shows the distribution of the mean squared error (MSE) from 1,000 permutations with randomly shuffled text samples for the predictions of the rental value and Minergie ownership. The MSE obtained with the original text in each plot is indicated with a vertical line. For the prediction of the rental value with the original data, the MSE is at 7,228,713.89. The distribution of the random text data is clearly to the right of this value with a mean of 7,875,488.54. The Minergie label shows similar results with an original MSE of 0.158 and a mean MSE of the random permutations of 0.178. As with the rental value prediction, the original MSE is outside the distribution from the permutations.

Figure 3: MSE prediction of rental value and Minergie ownership with text



(a) House Value

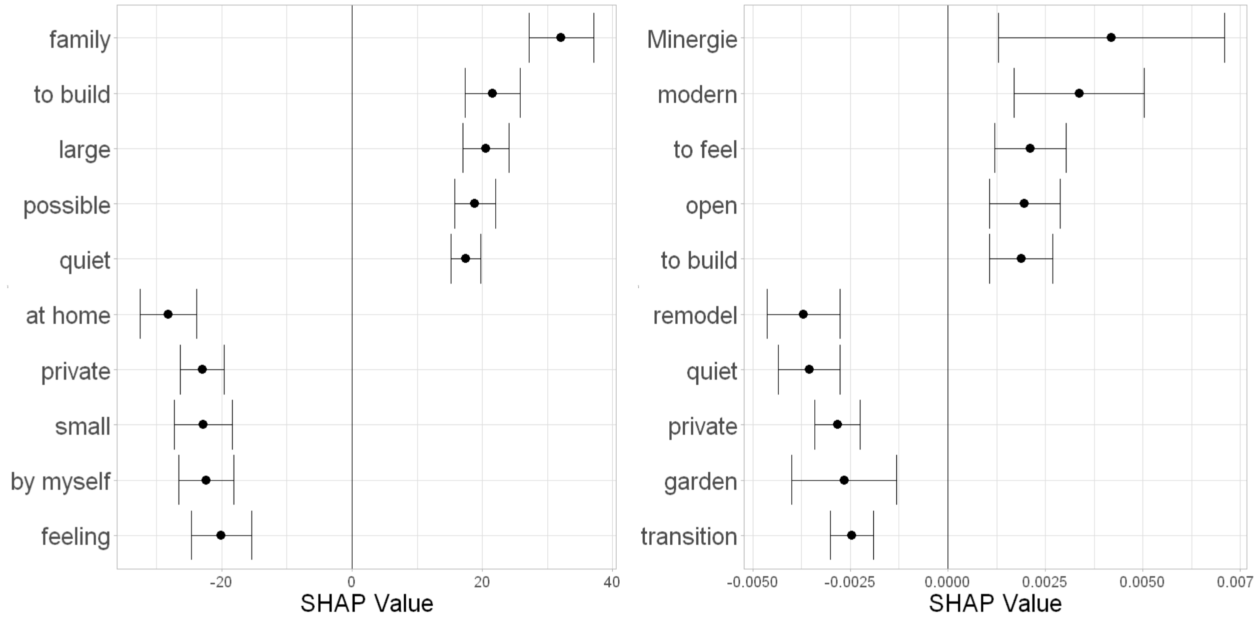
(b) Minergie

Note: This figure shows the distribution of the mean squared error (MSE) from a permutation test with the prediction of the house value (left panel) and Minergie ownership (right panel) using the open-ended text answer with 1,000 permutations. Each time, a lasso regression for the house value or a random forest classifier for Minergie is trained and validated with 10-fold cross-validation. The draws for each of the 10 folds in each permutation randomly shuffle the order of the text answers. This means the text and the outcome variable are no longer connected. The distribution shows the MSE with the randomly shuffled text. A vertical line indicates the original MSE in each plot. The sample consists of 2,694 observations.

The next step computes the SHAP values for both models. Figure 4 shows the average feature importance based on the SHAP values for the five most-significant positive and negative words separately for house value and Minergie ownership. Each dot marks the average SHAP value for all observations, and the bars represent the 95% confidence interval. Words on the upper half of both panels positively affect the propensity score. The presence of these words in a response increases the probability of the respondent owning a house with the Minergie label or a high rental value. Results with the SHAP values show that the words *build*, *large*, *family*, and *quiet* have a positive impact on housing prices. The words *small* and *by myself* (here, often used in the sense of performing renovations without assistance) have a negative effect. This pattern implies respondents who place a high emotional value on performing renovations by themselves and enjoy living in small buildings report a lower monetary house value. As for the house value, the algorithm computes the feature importance for the Minergie label. Results from the feature importance according to the SHAP values give a clear picture: the word *Minergie* has the highest predictive power for owning a Minergie building,[17] as do the words *modern* and *build.* The words *remodel*, *quiet*, and *garden* negatively affect the probability of owning a Minergie building.

[17] Out of 500 Minergie owners, 47 explicitly mentioned the word Minergie in their answer. Therefore, the word Minergie is not a perfect predictor for the Minergie label. Moreover, to influence the valuation of the green label, the word Minergie also must be highly correlated with the house value.

Figure 4: SHAP Values: top five words with positive and negative predictive power for rental value and Minergie ownership



(a) House Value

(b) Minergie

Note: This figure shows the SHAP values for the top 10 words with the highest predictive power for the prediction of house value (left panel) and Minergie ownership (right panel) with the open-ended text answer. The SHAP values indicate the predictive power for a given word, but they do so only for an individual text answer (ceteris paribus all the other words included in that answer). For each word, the dot indicates the mean SHAP value for all observations; the bars show the 95% confidence interval.

The last step of CTA allocates the significant positive and negative keywords separately to topics. Table 3 presents the four major positive and negative topics for both predictions. For a high house value, the most relevant topics describe the planning and construction of the building. Further, these topics point out building characteristics, such as modern architecture and generous floor space, and a quiet location. Moreover, words in the topic *family* are also positively associated with a higher stated rental value. The major topics for a lower house value point out a smaller dwelling size, refer to the owner's independence, and emphasize the respondent's emotional attachment. In contrast to the high-value topic that describes constructing a new building, owners of lower-priced houses frequently mentioned retrofits.

For the Minergie certification, the most positively associated topics consist of energy efficiency and building characteristics that emphasize modern design and comfort. Other crucial topics include a general expression of emotions and the construction of the building. In contrast, two major topics negatively correlated with owning a Minergie house describe renovations and, interestingly, a large building size. Further, respondents who mentioned their autonomy and independence are less likely to live in a Minergie house. This occurs mainly because for many respondents, autonomy involves modifying the building, which is more challenging with Minergie buildings.

Figures B.1 and B.2 (Appendix) show the words that define the 10 most-important positive and negative topics for each prediction. The words are grouped according to their topic using wordzones (Hearst et al. 2019). In comparison, the ungrouped words for each prediction are presented in word clouds in Figures A.1 and A.2 (Appendix).

Table 3: Topics for house value and Minergie ownership based on predictive keywords

House Value	**Minergie**
Positive Predictive Keywords and Topics	
Family family, spend, experience, grow up, grow, visit, celebrate, at any time, play, wonderful, gladly, going, vacations	Energy Efficiency Minergie, sustainable, power
Construct to build, architectural, to plan, largely	Building Characteristics modern, convenient, snug, comfortable
Building Characteristics quiet, comfortable, convenient, modern, generous	Emotions to feel, feels, grow
Comfort comfort, security, protection	Construction to plan, furnishing, configuration
Negative Predictive Keywords and Topics	
Home size small, cozy, snug, homely, open, homelike, at the same time	Renovation remodel, renovate, rebuilt
Independence by myself, wholeheartedness, natural, knowledge, individual	Independence freedom, independence, regarding
Emotional feeling, to feel, perceived	Building Characteristics big, large, beautiful
Renovation to make, to modify, to ask, decide, determine, dependent, furnishing, owner	Autonomy by myself, owner, changes, investments, determine, largely, decide, to modify, to ask, modernize, owner, to meet

Note: This table presents the topics obtained from the most important keywords for each prediction based on their SHAP values and the word embeddings. The positive topics for the house value had an optimal threshold δ of 0.9 and a minimum similarity α of 0.4. The total number of topics was 33, and from the 202 relevant words, 38 could not be allocated to a word embedding. For the negative house value topics, the threshold δ was 0.9, the minimum similarity α was 0.5, the total number of topics 36, and out of 255 relevant words, 34 could not be allocated. For the positive Minergie topics, the threshold δ was 0.9, the minimum similarity α was 0.5, included a total of eight topics, and out of 69 relevant words four were not allocated. The negative Minergie topics had a threshold δ of 0.9, with a minimum similarity α of 0.5, and a total of 38 topics; out of 248 relevant words, 36 could not be allocated to the embedding matrix.

In this section, I employed CTA to explain house value and Minergie ownership with topics extracted from open-ended text answers. To do so, I linked the open-ended text responses to other outcomes from the survey, and I showed this correlation was not random.

Furthermore, specific words from the text answer were associated with either the house value or the Minergie certificate. Until now, both outcomes have been considered independently. In analyzing OVB, it is essential to link the correlation from the text to house value to the correlation from the same text to the Minergie certificate.

6 Results: CTA and omitted variable bias (OVB)

In this section, CTA is used to derive topics that explain an OVB when text functions as a control variable in a hedonic regression. Before applying CTA, I first show the text data is a covariate that reduces an OVB in a hedonic regression for the valuation of the Minergie label. Specifically, I estimate the model in equation (3) with and without the text-covariates $Text_i$.[18] In a second step, I explain which topics influence the bias by applying CTA.

6.1 Text as control variable

The double ML procedure with the information from the text uses a lasso regression for the house value and a random forest classifier for the Minergie label.[19] To perform this procedure, the double ML implementation in Python (Bach et al. 2022) trains the lasso regression and the random forest algorithm on the training dataset and predicts each outcome with the test data. To avoid overfitting, the algorithm uses a 10-fold cross-fitting to compute the double ML estimates. Cross-fitting, similar to cross-validation, splits the dataset into several random partitions; in this case, it is split into 10 equal parts. After that, cross-fitting uses nine out of the 10 partitions as a training set to train a machine learning algorithm, and it performs the prediction and regression step using the remaining 10% of the data. This procedure repeats for 10 partitions and reports the average Minergie coefficient over the 10 folds. Further, the algorithm repeats the entire estimation 10 times and reports the average estimate. These additional repetitions increase the robustness to the random sampling in cross-fitting.

[18] In a series of robustness checks in section E in the Appendix, I show including the text data in a classic OLS regression gives less meaningful results due to the text data's high dimensionality.

[19] I chose the hyperparameters in order to optimize the algorithm's fit (1,000 trees and no bootstrapping).

Table 4 presents the basic model before and after partialling out the text data without any covariates other than the text. Column (1) shows the baseline model, where Minergie has a coefficient of 584.6 and is statistically significant; this implies a monthly premium for Minergie houses of 584.6 CHF. After partialling out the text in column (2), the coefficient decreases by almost half to 306.9 and remains statistically significant. The valuation of 306.9 CHF corresponds to 6.7% of the monthly rental value of Minergie buildings. This result is in line with previous research on Minergie buildings in Switzerland and indicates a similar range, between 4% and 12% of the rental value for the premium (Banfi et al. 2008), and 5% to 10% higher construction costs (Salvi et al. 2008).

Table 4: Hedonic Regression: double Machine Learning results

	Dependent variable: Rental Value	
	(1)	(2)
Minergie	584.601***	306.913**
	(119.641)	(141.611)
Intercept	3852.175***	
	(51.543)	
Double ML on "Minergie"	No	Yes
Observations	2,694	2,694
R^2	0.009	

Note: *p<0.1; **p<0.05; ***p<0.01
This table presents the results from the basic hedonic regression setup from equation (3). Column (1) shows the results without considering the text as covariates; Column (2) includes the text from the open-ended question as a covariate that is partialled out using double Machine Learning.

Next, I estimate several model specifications that contain different sets of covariates. Table 5 presents the results of these estimations. Compared to the results in Table 4, which do not include any covariates except for the text, the additional covariates reduce the sample size because not all respondents provided complete responses for all variables. For this reason, the sample from the model in Table 5 decreased from the original 2,694 to 2,075 observations.

Column (1) in Table 5 shows the baseline regression containing only the Minergie dummy and the house value. This specification results in a Minergie coefficient of 587.9 and a high statistical significance. Column (2) controls for the text answers in the same way as in Table 4; there are no other covariates. Compared to the specification in column (1), double Machine Learning decreases the coefficient of the Minergie dummy variable to 283.7. This reduction indicates the text captured variation with a positive bias on the Minergie coefficient in column (1). In this case, the open-ended question on respondents' positive emotions associated with their homes captures information that, if not included in the model, would lead to an overvaluation of Minergie homes compared to non-Minergie buildings.

Next, in column (3), I introduce the 13 closed-ended preference measures in addition to the text data. Interestingly, the Minergie coefficient increases compared to column (2) to 396.5. Column (4) introduces the socioeconomic variables to the closed-ended preference measures and the text answers. Compared to controlling only for text, as in column (2), the coefficient slightly increases with higher standard errors, making the estimate insignificant at the 10% level. Finally, in column (6), I add the building characteristics to the model (i.e., building age and floor space). Compared to the specification in column (5) with socioeconomics, adding building characteristics decreases the coefficient and increases the standard error, which makes the estimate statistically insignificant.

Table 6 provides a benchmark by estimating a similar specification that only contains the closed-ended preference measures and using OLS, as described in equation (2).[20] As in Table 5, the first column contains the baseline regression with no covariates. Column (2) introduces all 13 housing-preference variables from the closed-ended questions. Moreover, this specification contains two happiness indicators. A first indicator describes the respondent's happiness with the individual's house on a scale of 1 to 7. A second variable indicates

[20] Table 6 only reports the Minergie coefficient. Table D in the Appendix shows all coefficients from the same model specification.

the respondents' happiness with their houses compared to their peers. This specification is crucial to compare the effect of closed-ended questions with the effect of the narratives obtained from the open-ended question. Respondents could give a rating on each of the 13 closed-ended questions on a scale from 1 (not important) to 7 (very important). Introducing these covariates increases the Minergie coefficient to 629.3, with a statistical significance at the 1% level. Compared to the similar specification using the open-ended text responses in Table 5, column (2), the coefficient is larger and increased. This difference indicates that the closed-ended questions do not measure the same preferences as the open-ended responses.

In column (3), I add socioeconomic variables about the survey respondents, namely their ages, genders, incomes, and household sizes. These additional control variables decrease the Minergie coefficient to 483.9, which is statistically significant. Finally, column (4) adds building age and floor space, which reduces the coefficient to 307.3 while still being statistically significant at the 10% level.

The magnitude of the estimation from column (4) with all closed-ended variables is similar to the results obtained with the open-ended text responses in Table 5, column (2). This similarity could indicate that the open-ended text responses capture variation beyond the 13 closed-ended preference questions but also contain information represented in the socioeconomic variables and the building characteristics. Moreover, the magnitude from the estimation with closed-ended variables from Table 6 column (4) is similar to the estimations with text and closed covariates in Table 5 column (5). Even though the coefficients obtained with text in Table 5 are not statistically significant, the similar magnitude could suggest that both models capture a similar variation in the data but with different noise levels.

Table 5: Hedonic regression: Double Machine Learning estimation results

	(1)	(2)	(3)	(4)	(5)
Minergie	587.906***	283.744*	396.576*	305.461	275.015
	(133.560)	(168.335)	(208.424)	(216.909)	(298.859)
Preferences (Text)	No	Yes	Yes	Yes	Yes
Preferences (closed)	No	No	Yes	Yes	Yes
Socioeconomics	No	No	No	Yes	Yes
Building Characteristics	No	No	No	No	Yes
Double ML on "Minergie"	No	Yes	Yes	Yes	Yes
Observations	2,075	2,075	2,075	2,075	2,075

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents the results from the hedonic regression from equation (3) in different specifications. Column (1) shows the hedonic regression with OLS and no covariates. The other columns use double ML to partial out different sets of covariates, including text data. All double ML models were estimated with a 12-fold cross-fitting and 10 repetitions, each using a different random sampling for the cross-fitting.

Table 6: Hedonic regression: OLS with closed-ended variables

	(1)	(2)	(3)	(4)
Minergie	583.292***	629.312***	483.961***	307.332*
	(133.476)	(167.856)	(168.213)	(180.268)
Preferences (closed)	No	Yes	Yes	Yes
Socioeconomics	No	No	Yes	Yes
Building Characteristics	No	No	No	Yes
Observations	2,075	2,075	2,075	2,075
R^2	0.009	0.064	0.133	0.171

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents the results from the hedonic regression from equation (2) with different sets of covariates. Column (1) shows the hedonic regression with OLS and no covariates. The other columns introduce different sets of closed-ended variables. Table D in the Appendix shows the same specification and reports the coefficients to all variables.

6.2 CTA – to explain OVB from text with topics

To show which topics influence the bias on the Minergie estimator, I apply CTA, as described in section 3. The first step consists of computing each word's importance using SHAP values.[21] The SHAP algorithm will compute the effect of the individual words on β_1 by estimating the effect of each word for each individual on $\widehat{house\ value}_i$ and $\widehat{Minergie}_i$, using equation (5).

[21] To ensure train and test data have the same proportion of Minergie owners and high-/low-valued homes, I stratify both datasets when partitioning the data using a k-means algorithm with eight clusters. Stratifying the train and test set based on those clusters implies keeping the same proportion of each cluster in the train and test set.

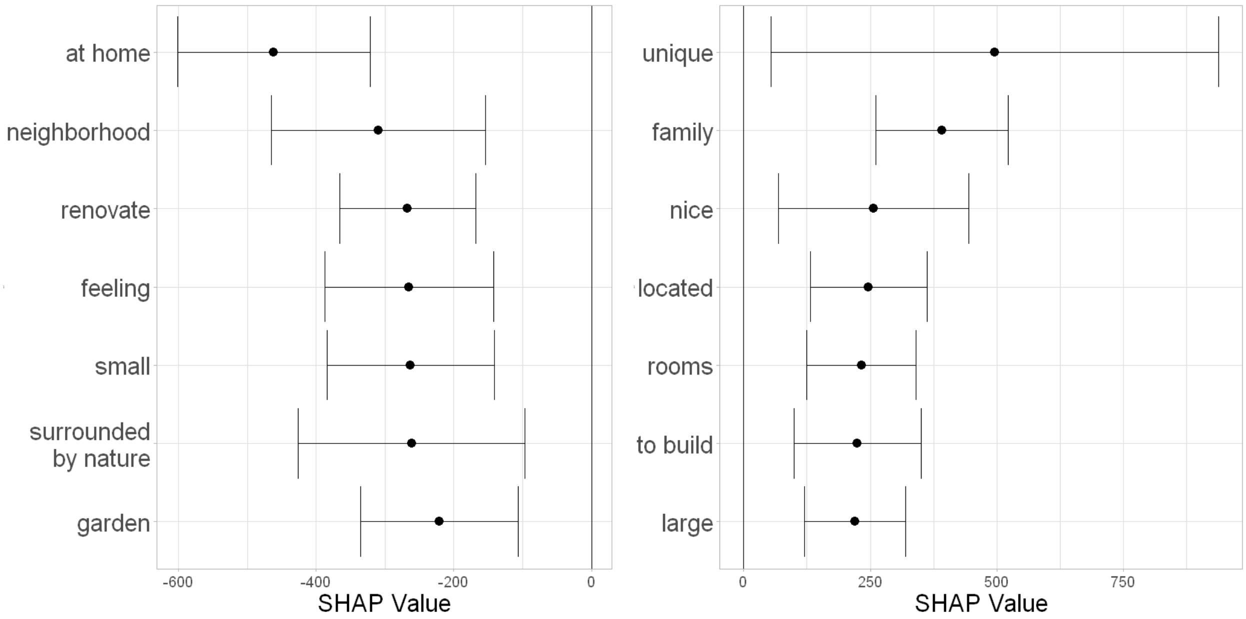
When controlled for, on the one hand, words associated with a decrease in β_1 induce a positive bias on the Minergie valuation. If the model did not include these words, the Minergie coefficient would be larger, and these buildings are overvalued. This overvaluation occurs because the lower housing price for certain non-Minergie dwellings can be explained partly by the respondents' preferences instead of the absence of the Minergie label. Controlling for those words (or doing so by partialling them out with double ML) will decrease the Minergie coefficient because the topics mentioned by these respondents differentiate low-value houses from high-value houses while also being correlated with the Minergie label. On the other hand, words correlated with the Minergie label and also a high house value will lead to an undervaluation of the Minergie label or a negative bias. In this case, some conventional homes might exhibit a similarly high price as Minergie buildings, even though the absence of this certification should lead to a lower price. However, this similarity does not necessarily occur because these homeowners did not value the Minergie technology. Rather, the higher price can be explained by a preference for certain building characteristics (e.g., modern architecture) that occur with Minergie buildings and high-priced conventional houses.

The average SHAP value and the standard error of the mean for each feature allow for selecting the features with the highest impact. Based on those values, a t test with the null hypothesis that the mean is differed from zero identifies and drops all words with a t-value smaller than 1.96. After this preselection, 77 words with a positive effect and 84 words with a negative effect on the double ML coefficient remain.

Figure 5 shows the mean SHAP values with a 95% confidence interval for the seven most-important positive and negative features. Each dot represents the word's effect relative to the baseline, the original double ML coefficient for Minergie ownership. A positive SHAP value means for an individual who used a specific word that the presence of the word positively affects the individual's coefficient (conditional on all the other words in the individual's response).

Words that decrease the individual coefficient imply if that particular word is present in an individual's response, the ratio between house value and Minergie ownership changes accordingly. Hence, the presence of these words is associated with a lower valuation for the Minergie label. Conversely, words that are associated with a higher β_1 coefficient imply if a respondent uses those specific words, the individual Minergie valuation is higher.

Figure 5: SHAP Values for the top 14 Text features from double Machine Learning



(a) negative

(b) positive

Note: This figure shows the SHAP values for the top 14 words with the most-significant influence on the double ML estimator in the hedonic regression for the value of the Minergie certification. The SHAP value indicates the predictive power for a word, but it does so only for an individual text answer (ceteris paribus all the other words included in that answer). For each word, the dot indicates the mean SHAP value for all observations with the bars showing a 95% confidence interval. Dots with negative SHAP values indicate a negative influence on the Minergie coefficient, which indicates a lower valuation of the Minergie label.

The allocation of keywords into topics follows from the predictions of the outcome variables. Based on the average SHAP values, CTA allocates words with either a positive or negative effect separately to topics. Then, it selects the four major topics based on the maximum SHAP value among the defining words. These topics describe an OVB for the valuation of Minergie-certified houses because they result from the double ML procedure. Therefore, the information in these topics is associated with lower-valued houses without

Minergie certification for the negative bias. The positive bias on the valuation implies the topics correlate with a high house value and, simultaneously, with the Minergie certification.

For the Minergie valuation, four major positive and negative topics are presented in Table 5. In addition, two wordzones present the entire set of topics in Figure B.3 (Appendix). For comparison, the unsorted words for the positive and negative valuation are presented in word clouds in Figure A.3 (Appendix).

The different topics indicate building characteristics do play a central role. High-valuation respondents described their homes as appealing, emphasized the light in the buildings, and pointed out the comfort their homes provide. Moreover, a positive bias is associated with a central location that includes schools, well-connected public transport, and an emphasis on the view and the landscape. In contrast, low-valuation respondents described their location as rural and mentioned renovation projects. Further, a negative bias links to the respondents' childhood memories. Finally, the general emphasis on emotions by explicitly mentioning the words *feeling* or *safe* is linked to a lower valuation.

Table 7: Topics based on keywords from double Machine Learning

Double ML - effect on Minergie valuation
Positive
Building characteristics rooms, light, appeal
Location city, schools, public, connectivity, municipality
View view, landscape, space
Comfort comfort, recreation, yield
Negative
Renovation renovate, adapt, reconstruct
Emotions to feel, feeling, safe
Location surrounded by nature, rural, ground level
Memories parents, grandfather, childhood

Note: This table presents the topics obtained from the most important keywords for double ML based on their SHAP values and the word embeddings. The positive topics had an optimal threshold δ of 0.9 and a minimum similarity α of 0.4. The total number of topics was 12, and from the 77 relevant words, 14 could not be allocated to a word embedding. For the negative topics, the threshold δ was 0.9. The minimum similarity α was 0.5, the total number of topics was 11, and out of 84 relevant words, 13 could not be allocated.

7 Conclusion

The increasing use of text data in social sciences with a too-small sample size to accommodate the standard AI analysis tools opens the door to new analytical methods. Open-ended survey questions especially provide many opportunities for text analysis. To gain more insights into this new form of data, I propose a novel topic model, the conditional topic allocation (CTA). This method allows the identification of topics in a text that correlate with an observable variable; this makes these topics conditional on the respective variable. In the first step, I used CTA to explain individual observable variables from a survey. Then, I applied CTA to explain text data that functions as control variables in a regression.

This study used data from a survey that asked respondents to answer an open-ended question about aspects of their homes that elicited positive emotions. To apply CTA, using Natural Language Processing, I first demonstrated the information from the text answers correlates with house value and the energy-efficient Minergie housing certification. Using a novel approach based on the SHAP values and word embeddings, I identified major words and topics related to high and low house value and the Minergie certification.

In a second step, I applied CTA to identify topics in a setting where text data is a control variable in a hedonic regression. Specifically, I addressed the problem of OVB, in which omitted variables are correlated with both dependent and independent variables, which leads to a biased estimate. To control for the information contained in the text data, I applied double Machine Learning to partial out an OVB in a hedonic regression. Results indicate the premium associated with the Minergie certification decreases by almost half from 585 CHF to 307 CHF after partialling out the information captured in the text. The magnitude of this coefficient only changes little when adding classical covariates. However, it loses its statistical significance due to higher standard errors. Based on the effect on the bias indicated by the SHAP values, it is possible to allocate the words into topics that provide a better understanding of which omitted topics contribute to the bias. A negative bias is associated with emphasizing renovations, a rural location, and childhood memories. A positive bias, and hence overvaluation, is associated with building characteristics such as the light and the general appeal of the home. Moreover, a positive bias is connected to a central location and emphasizes the view.

From a general perspective, the results provide an encouraging use case to employ CTA for analyzing open-ended survey responses in social sciences. Further, the findings from this study suggest open-ended questions constitute a valuable improvement for surveys and capture meaningful information about respondents. In this context, CTA allows for automatic and data-driven analysis of this type of data. For future research, the potential of CTA lies in facilitating the analysis of text data, especially open-ended survey responses.

References

Anderson, J. R. and Bower, G. H.: 1972, Recognition and retrieval processes in free recall., *Psychological Review* **79**(2), 97—-123.

Andre, P., Haaland, I., Roth, C. and Wohlfart, J.: 2021, Inflation narratives, *CEPR Discussion Papers 16758*.

Andre, P., Pizzinelli, C., Roth, C. and Wohlfart, J.: 2022, Subjective models of the macroeconomy: Evidence from experts and a representative sample, *The Review of Economic Studies* .

Bach, P., Chernozhukov, V., Kurz, M. S. and Spindler, M.: 2022, DoubleML – An object-oriented implementation of double machine learning in Python, *Journal of Machine Learning Research* **23**(53), 1–6.

Bajari, P. and Benkard, C. L.: 2005, Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach, *Journal of Political Economy* **113**(6), 1239–1276.

Banfi, S., Farsi, M., Filippini, M. and Jakob, M.: 2008, Willingness to pay for energy-saving measures in residential buildings, *Energy Economics* **30**(2), 503–516.

Baylis, P.: 2020, Temperature and temperament: Evidence from twitter, *Journal of Public Economics* **184**, 104161.

Belmonte, A. and Rochlitz, M.: 2019, The political economy of collective memories: Evidence from russian politics, *Journal of Economic Behavior & Organization* **168**, 229–250.

Benesch, C., Loretz, S., Stadelmann, D. and Thomas, T.: 2019, Media coverage and immigration worries: Econometric evidence, *Journal of Economic Behavior & Organization* **160**, 52–67.

Blei, D. M., Ng, A. Y. and Jordan, M. I.: 2003, Latent dirichlet allocation, *Journal of Machine Learning Research* **3**(Jan), 993–1022.

Boumeester, H. J. F. M.: 2011, Traditional housing demand research, *in* S. J. Jansen, H. C. Coolen and R. W. Goetgeluk (eds), *The Measurement and Analysis of Housing Preference and Choice*, Springer Netherlands, Dordrecht, pp. 27–55.

Brounen, D. and Kok, N.: 2011, On the economics of energy labels in the housing market, *Journal of Environmental Economics and Management* **62**(2), 166–179.

Bursztyn, L., Haaland, I. K., Rao, A. and Roth, C. P.: 2020, Disguising prejudice: Popular rationales as excuses for intolerant expression, *Working Paper 27288*, National Bureau of Economic Research.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J.: 2018, Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal* **21**(1), C1–C68.

Coolen, H. C. C. H.: 2011, The meaning structure method, *in* S. J. Jansen, H. C. Coolen and R. W. Goetgeluk (eds), *The Measurement and Analysis of Housing Preference and Choice*, Springer Netherlands, Dordrecht, pp. 75–99.

Dieng, A. B., Ruiz, F. J. and Blei, D. M.: 2020, Topic modeling in embedding spaces, *Transactions of the Association for Computational Linguistics* **8**, 439–453.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E. and Stewart, B. M.: 2018, How to make causal inferences using texts, *arXiv preprint arXiv:1802.02163* .

Eichholtz, P., Kok, N. and Quigley, J. M.: 2010, Doing well by doing good? green office buildings, *American Economic Review* **100**(5), 2492–2509.

Ferrario, B. and Stantcheva, S.: 2022, Eliciting people's first-order concerns: Text analysis of open-ended survey questions, *AEA Papers and Proceedings* **112**, 163–69.

Filippini, M., Leippold, M. and Wekhof, T.: 2022, Sustainable finance literacy and the determinants of sustainable investing, *Swiss Finance Institute Research Paper 2-02*.

Fong, C. and Grimmer, J.: 2016, Discovery of treatments from text corpora, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, pp. 1600–1609.

Fuerst, F. and McAllister, P.: 2011, Green noise or green value? measuring the effects of environmental certification on office values, *Real Estate Economics* **39**(1), 45–69.

Geer, J. G.: 1988, What do open-ended questions measure?, *Public Opinion Quarterly* **52**(3), 365–367.

Geer, J. G.: 1991, Do open-ended questions measure “salient” issues?, *Public Opinion Quarterly* **55**(3), 360–370.

Gentzkow, M., Shapiro, J. M. and Taddy, M.: 2019, Measuring group differences in high-dimensional choices: Method and application to congressional speech, *Econometrica* **87**(4), 1307–1340.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A. and Mikolov, T.: 2018, Learning word vectors for 157 languages, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Hearst, M. A., Pedersen, E., Patil, L., Lee, E., Laskowski, P. and Franconeri, S.: 2019, An evaluation of semantically grouped word cloud designs, *IEEE transactions on Visualization and Computer Graphics* **26**(9), 2748–2761.

Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A.: 2020, spaCy: Industrial-strength Natural Language Processing in Python.

Houde, S. and Wekhof, T.: 2021, The narrative of the energy efficiency gap, *CER-ETH Economics Working Paper Series 21/359*.

Kahn, M. E. and Kok, N.: 2014, The capitalization of green labels in the california housing market, *Regional Science and Urban Economics* **47**, 25–34.

Koirala, B. S., Bohara, A. K. and Berrens, R. P.: 2014, Estimating the net implicit price of energy efficient building codes on u.s. households, *Energy Policy* **73**, 667–675.

Krosnick, J. A.: 1999, Survey research, *Annual Review of Psychology* **50**(1), 537–567.

Lundberg, S. M. and Lee, S.-I.: 2017, A unified approach to interpreting model predictions, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Curran Associates Inc., p. 4768–4777.

Lyubomirsky, S. and Lepper, H. S.: 1999, A measure of subjective happiness: Preliminary reliability and construct validation, *Social Indicators Research* **46**(2), 137–155.

Manzoor, E., Chen, G. H., Lee, D. and Smith, M. D.: 2020, Influence via ethos: On the persuasive power of reputation in deliberation online, *arXiv preprint arXiv:2006.00707* .

McAuliffe, J. and Blei, D.: 2007, Supervised topic models, *Advances in neural information processing systems* **20**, 121–128.

Morales, J. S.: 2021, Legislating during war: Conflict and politics in colombia, *Journal of Public Economics* **193**, 104325.

Netzer, O., Lemaire, A. and Herzenstein, M.: 2019, When words sweat: Identifying signals for loan default in the text of loan applications, *Journal of Marketing Research* **56**(6), 960–980.

Ramage, D., Hall, D., Nallapati, R. and Manning, C. D.: 2009, Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 248–256.

Řehůřek, R. and Sojka, P.: 2010, Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.

Reja, U., Manfreda, K. L., Hlebec, V. and Vehovar, V.: 2003, Open-ended vs. close-ended questions in web questionnaires, *Developments in Applied Statistics* **19**(1), 159–177.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. and Rand, D. G.: 2014, Structural topic models for open-ended survey responses, *American Journal of Political Science* **58**(4), 1064–1082.

Rosen, S.: 1974, Hedonic prices and implicit markets: product differentiation in pure competition, *Journal of Political Economy* **82**(1), 34–55.

Salvi, M., Horehájová, A. and Müri, R.: 2008, Der nachhaltigkeit von immobilien einen finanziellen wert geben–minergie macht sich bezahlt [giving a financial value to the sustainability of real estate–minergie pays off], *Center for Corporate Responsibility and Sustainability, University of Zurich, Switzerland.* .

Schuman, H., Ludwig, J. and Krosnick, J. A.: 1986, The perceived threat of nuclear war, salience, and open questions, *Public Opinion Quarterly* **50**(4), 519–536.

Shapley, L. S.: 1953, A value for n-person games, *Contributions to the Theory of Games* **2**(28), 307–317.

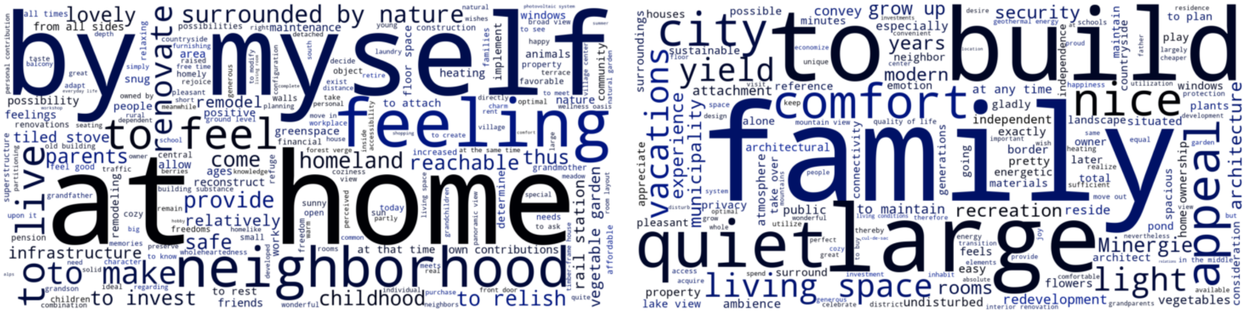
Tvinnereim, E. and Fløttum, K.: 2015, Explaining topic prevalence in answers to open-ended survey questions about climate change, *Nature Climate Change* **5**(8), 744–747.

Veitch, V., Sridhar, D. and Blei, D.: 2020, Adapting text embeddings for causal inference, *Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research, vol. 124, pp. 919–928.

Appendix

A Word clouds - Feature importance

Figure A.1: Word cloud, positive and negative words for housing-value prediction

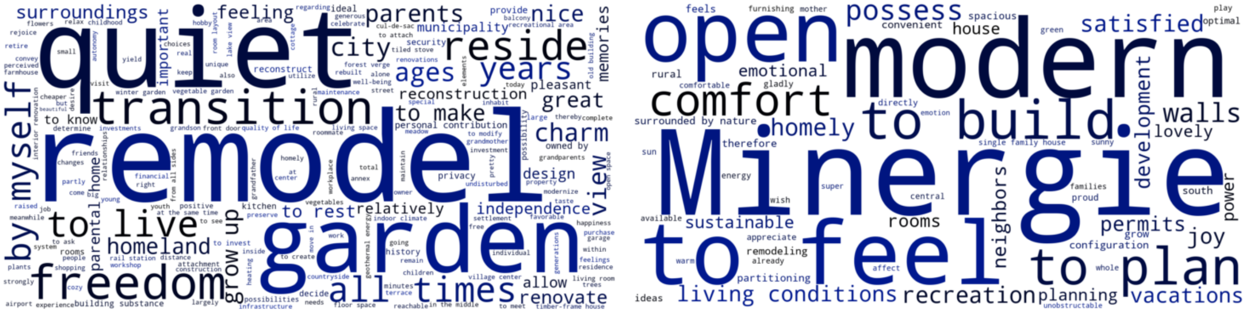


(a) low value

(b) high value

Note: The two panels show word clouds with the relative importance of each word to the prediction of a high or low house value. The size of each word depends on its SHAP value for the prediction.

Figure A.2: Word cloud, positive and negative words for Minergie-ownership prediction

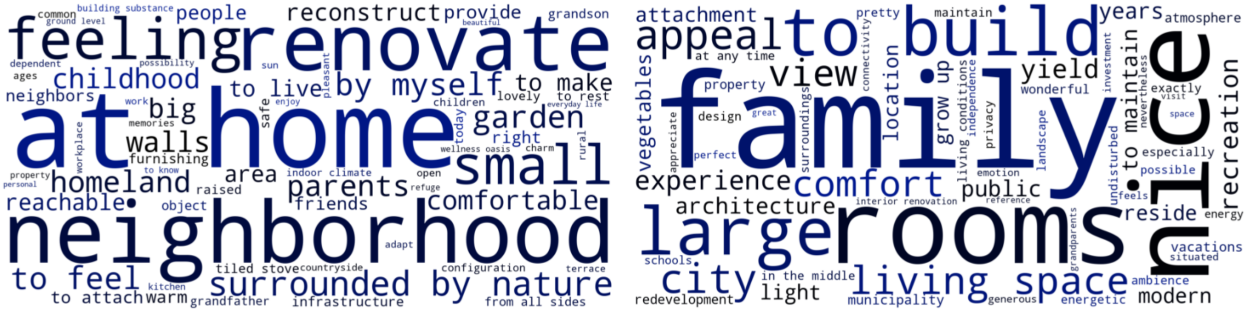


(a) non-Minergie

(b) Minergie

Note: The two panels show word clouds with the relative importance of each word to the prediction of Minergie ownership. The size of each word depends on its SHAP value for the prediction.

Figure A.3: Word cloud, positively and negatively associated words with Minergie valuation



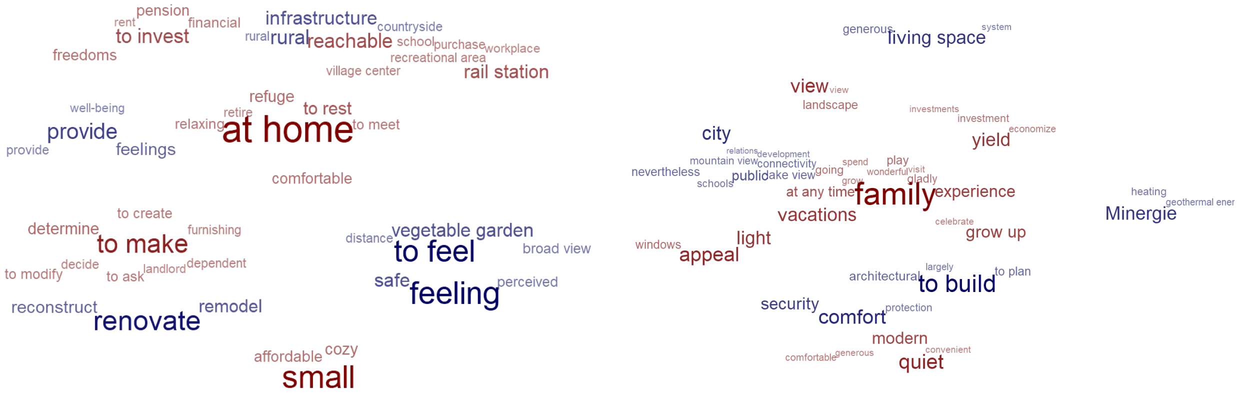
(a) negative

(b) positive

Note: The two panels show word clouds with the relative importance of each word to the prediction of Minergie valuation. The size of each word depends on its SHAP value for the prediction.

B Wordzones

Figure B.1: Wordzones, top-10 positive and negative topics for house value prediction

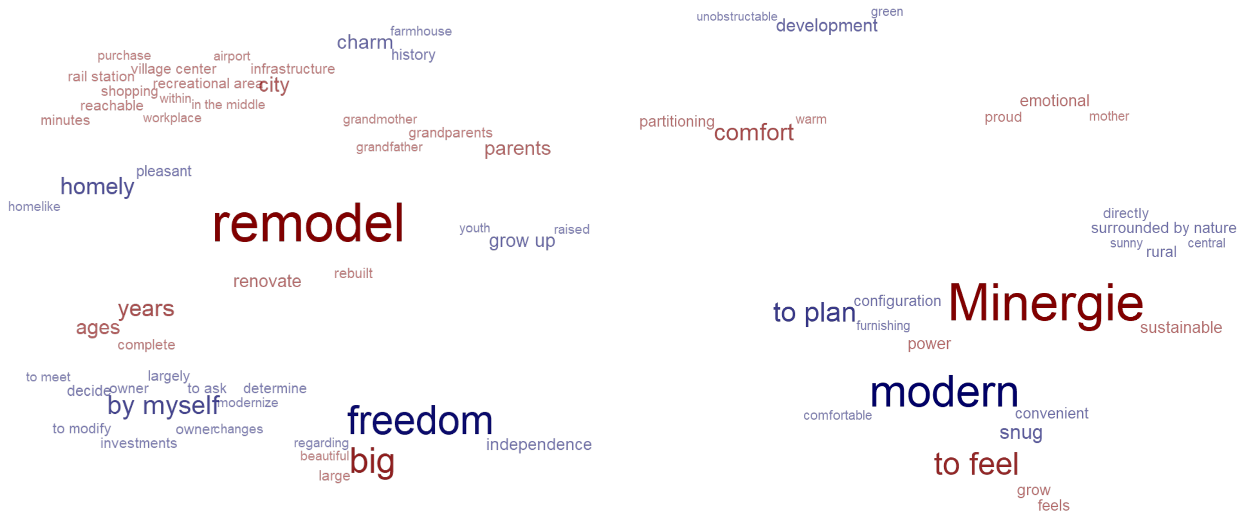


(a) low value

(b) high value

Note: The two panels show wordzones with the relative importance of the 10 most-important topics associated with a lower (left panel) or higher (right panel) house value. The size of each word depends on its SHAP value for the prediction. The wordzones followed Hearst et al. (2019) and were compiled with the corresponding code and is available at https://observablehq.com/@mahviz/wordzones-usable-alternative-toword-clouds

Figure B.2: Wordzones, top-10 positive and negative topics for Minergie-ownership prediction

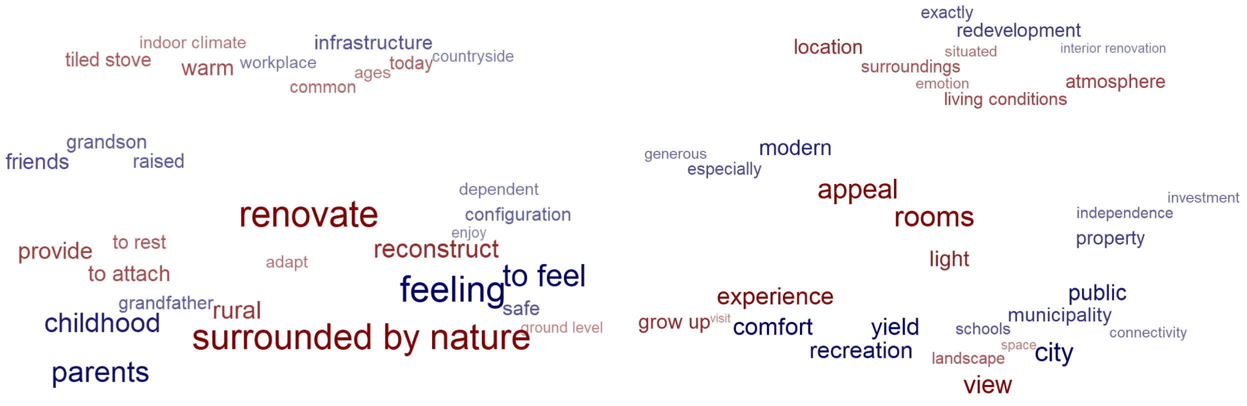


(a) non-Minergie

(b) Minergie

Note: The two panels show wordzones with the relative importance of the 10 most-important topics associated with non-Minergie ownership and Minergie ownership. The size of each word depends on its SHAP value for the prediction. The wordzones followed Hearst et al. (2019) and were compiled with the corresponding code, available under https://observablehq.com/@mahviz/wordzones-usable-alternative-to-word-clouds

Figure B.3: Wordzones, positive and negative effects on Minergie valuation



(a) negative

(b) positive

Note: The two panels show wordzones with the relative importance of the topics associated with a lower or higher valuation of the Minergie label. The size of each word depends on its SHAP value for the prediction. The wordzones followed Hearst et al. (2019) and were compiled with the corresponding code, available under https://observablehq.com/@mahviz/wordzones-usable-alternative-to-word-clouds

C Support of rental value with respect to Minergie ownership

Table C.1: OLS: Income and Minergie on rental value

	Rental Value
	(1)
Income	0.153***
	(0.013)
Minergie	408.990
	(419.010)
Income x Minergie	-0.010
	(0.028)
Intercept	1914.883***
	(169.923)
Observations	2,075
R^2	0.088

Note: *p<0.1; **p<0.05; ***p<0.01
This table presents an OLS regression of the rental value on household income, Minergie certification, and the interaction of both variables. Results indicate that neither the Minergie certification nor the interaction for the certification with income is associated with a higher rental value.

D OLS Regression - full table

Table D.1 presents the same estimation results as in Table 6 and reports the coefficient of all covariates. Column (1) shows the baseline regression, with a Minergie coefficient of 583. Column (2) introduces the 13 closed-ended questions on housing preferences and two questions on the general satisfaction with the house. Among the preference measures, aesthetics, location, and floor space positively impact house value, with statistical significance. The importance of a garden has a negative impact on the reported house value. The Minergie coefficient increases in this specification to 629.

In column (3), I add socioeconomic variables to the model. Among these additional variables, owner age and living in a couple-household with children are positively associated with a higher house value. The coefficients for the preference variables remain largely unchanged, except for the importance of the garden, which is no longer significant, and the importance of the carbon footprint, which has a negative and significant impact. Compared to the specification in column (3), the Minergie coefficient decreases to 484.

Finally, column (4) includes building age and floor space, decreasing the Minergie coefficient to 307. While building age has no statistically significant coefficient, the floor space coefficient is positively associated with a higher house price and is statistically significant. Compared to column (3), the other coefficients remain of similar size and statistical significance, except for the importance of floor space, which decreases by half while still being statistically significant.

Table D.1: Hedonic regression: OLS with all covariates

	(1)	(2)	(3)	(4)
Minergie	583.292***	629.312***	483.961***	307.332*
	(133.476)	(167.856)	(168.213)	(180.268)
Floor space (real value)				10.527***
				(1.087)
Building Age				1.426
				(2.030)
Female			-34.013	8.289
			(126.744)	(124.140)
Owner Age			30.780***	26.580***
			(5.471)	(5.373)
Couple, with children			352.001***	210.832*
			(129.201)	(127.301)
Income			0.139***	0.105***
			(0.012)	(0.012)
Own House Satisfaction		16.098	18.711	-22.951
		(73.445)	(70.931)	(69.534)
Comp. House Satisfaction		62.608	58.486	63.912
		(50.795)	(49.000)	(47.945)
Aesthetics		196.577***	162.512***	111.925***
		(44.232)	(42.943)	(43.027)
Location		193.252***	131.204**	140.816**
		(58.481)	(56.622)	(55.415)
Floor space (importance)		273.550***	204.324***	110.201**
		(51.364)	(50.166)	(50.122)
Garden		-110.670**	-75.017	-66.262
		(48.735)	(47.082)	(46.166)
Parking		-23.973	-42.367	-55.546*
		(34.600)	(33.577)	(33.608)
Monetary value house		48.348	-1.274	-4.311
		(40.193)	(39.013)	(38.171)
Maintenance		4.469	49.505	63.363
		(51.767)	(50.082)	(49.388)
Carbon-Footprint		-69.442	-130.729***	-130.442***
		(51.001)	(49.584)	(48.548)
Energy Costs		-63.035	45.790	49.744
		(52.359)	(51.218)	(50.153)
Minergie (importance)		-45.958	-38.940	-33.947
		(37.164)	(36.444)	(35.766)
Indoor Air Quality		28.729	-6.079	1.957
		(39.757)	(38.561)	(37.785)
Thermal-Comfort		-89.704	-75.823	-65.992
		(66.140)	(63.760)	(62.442)
Noise-Protection		31.769	26.430	25.343
		(55.537)	(53.850)	(52.739)
Intercept	3855.100***	1280.918***	-1805.156***	-2190.285***
	(59.041)	(492.377)	(593.148)	(604.731)
Observations	2,075	2,075	2,075	2,075
R^2	0.009	0.064	0.133	0.171

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents an OLS estimation of equation (2), the same estimation as in Table 6 but with all covariates reported. The model only contains closed-ended variables.

E Robustness Checks

An approach to using text in a regression could include a dummy variable for each word as a control variable. However, the substantial number of words makes the data too highly dimensional for a standard OLS model. To show the intuition, I included 13 closed-ended questions on housing preferences and the 13 most essential words as covariates in an OLS regression.

Table E.1 presents the results: compared to the baseline in column (1), the coefficient for Minergie increases when including the closed-ended questions as covariates in column (2). Including the 13 most important words in column (3) decreases the coefficient. Further, including the words and closed-ended questions in column (4) increases the coefficients, similar to column (2), which does not contain any words. However, these models only include a fraction of all the possible words (13/584) because although including all words as control variables is possible, that would lead to a dimensionality problem.

In Table E.2, I illustrate the idea by increasing the number of words as covariates. The words are sorted by their relative frequency. Hence, the first 100 words occur more often than the following one-hundred words. More words should capture more information from the text and hence give a more precise estimate of the Minergie coefficient. Generally, the results in Table E.2 show a larger number of words tends to decrease the Minergie coefficient, although this tendency stops once a certain threshold of words is reached.

In Table E.3, I repeat the previous specification, but I also include the other closed-ended covariates. As a result, the Minergie coefficient decreases further, but it is still larger than the main specification that uses double Machine Learning. Another possibility to reduce the dimensionality of the text-covariates is to cluster the variables and use the clusters as controls.

Table E.4 presents the results for this strategy. A principal component analysis clustered all words into a small number of clusters. Each column represents the same model, but with more clusters (a higher number of clusters implies the clusters can capture more heterogeneity). As in column (2), five clusters already reduce the Minergie coefficient. With more clusters, the coefficient decreases, but the trend seems to stop at 50 clusters, and there is almost no difference in column (6) with 100 clusters. All four models show the intuition that including the text answer as a covariate decreases the Minergie coefficient, which suggests the text can control for unobserved heterogeneity. However, the considerable number of words creates a high-dimensional dataset unsuitable for standard OLS.

Table E.1: Linear probability model with responses to open- and closed-ended questions

	(1)	(2)	(3)	(4)
Minergie	583.292***	628.890***	553.224***	624.278***
	(133.476)	(167.830)	(134.625)	(169.004)
Closed Questions (n=13)	No	Yes	No	Yes
Open Keywords (n=13)	No	No	Yes	Yes
Observations	2,075	2,075	2,075	2,075
R^2	0.009	0.063	0.023	0.072

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents a OLS model for the basic hedonic regression with house value as the dependent variable and Minergie as the independent variable. Column (2) adds housing preferences from 13 closed-ended questions as covariates. Column (3) adds the 13 most-frequent keywords from the related open-ended question. The keywords are represented as dummy variables that take the value of 1 if the keywords occur in the answer and 0 otherwise. Column (4) includes both the closed-ended question and the 13 keywords.

Table E.2: Linear probability model with different number of keywords

	(1)	(2)	(3)	(4)	(5)	(6)
Minergie	583.292***	434.303***	383.633**	379.709**	406.566**	342.448**
	(133.476)	(145.904)	(150.234)	(153.198)	(159.874)	(161.721)
# keywords	0	100	200	300	400	500
Observations	2,075	2,075	2,075	2,075	2,075	2,075
R^2	0.009	0.068	0.119	0.177	0.213	0.276

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents a OLS model for the basic hedonic regression with house value as the dependent variable and Minergie as the independent variable. Columns (2) to (6) add the 100 to 500 most-frequently occurring keywords from the open-ended question as control variables. The keywords are represented as dummy variables that take the value of 1 if a keyword occurs in the answer and 0 otherwise.

Table E.3: Linear probability model with different number of keywords, including all covariates

	(1)	(2)	(3)	(4)	(5)	(6)
Minergie	483.961***	408.723**	381.798**	400.304**	392.351**	369.921*
	(168.213)	(178.717)	(184.479)	(188.366)	(196.565)	(199.866)
# keywords	0	100	200	300	400	500
Observations	2,075	2,075	2,075	2,075	2,075	2,075
R^2	0.133	0.172	0.213	0.260	0.286	0.342

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents a OLS model for the basic hedonic regression with house value as the dependent variable and Minergie as the independent variable. Columns (2) to (6) add the 100 to 500 most-frequently occurring keywords from the open-ended question as control variables. The keywords are represented as dummy variables that take the value of 1 if a keyword occurs in the answer and 0 otherwise. All models control building characteristics, demographics, and preferences with closed-ended questions, analogously to column 5 in Table 5.

Table E.4: Linear probability model with text cluster with Principal Component Analysis (PCA)

	(1)	(2)	(3)	(4)	(5)	(6)
Minergie	583.292*** (133.476)	561.441*** (134.738)	561.044*** (134.785)	516.000*** (135.688)	458.495*** (140.059)	400.393*** (145.552)
# clusters	0	5	10	20	50	100
Observations	2,075	2,075	2,075	2,075	2,075	2,075
R^2	0.009	0.017	0.019	0.029	0.044	0.077

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents a OLS model for the basic hedonic regression with house value as the dependent variable and Minergie as the independent variable. Columns (2) to (6) add an increasing number of cluster indicators as covariates. The clusters are dummy variables that take the value of 1 if a respondent's text answer was classified into the respective cluster and 0 otherwise. Clustering was performed with the bag-of-words matrix and principal component analysis.

F Variable Description

Table F.1: Variable Definition (1/2)

Variable	Description
	Building Characteristics
Minergie	A dummy variable that takes the value of 1 if the respondent stated that they live in a Minergie-certified building.
Building Age	The respondent's building's age in years. Respondents could choose between the following brackets: 2010 or later, for 1940 to 2010, ten-year brackets were available, and a bracket for buildings before 1940. We converted 2010 or later to 2015, and the ten-year brackets from 1940 to 2010 were converted to the average value (e.g., "Between 1941 and 1950" was converted to 1945). The bracket "Before 1940" was converted to 1940.
Floor space	The floor space in square meters.
Rental Value	The self-estimated monthly rental value respondents would obtain for renting their houses on the market. Respondents usually have a proxy for that rental value because it is important in Switzerland for tax purposes.
Heating	Respondents were asked what primary heating source they use for their houses. They could choose between four options: oil, gas, heat pump, and other. Oil and gas were taken together as one variable.
Solar PV	A dummy variable that takes the value of 1 if the respondent's house has solar panels.
	Demographics
Income	The respondent's gross household income. Respondents could choose between the following brackets: below 8 000 CHF, 8 000 to 12 000 CHF, 12 000 to 16 000 CHF, 16 000 to 20 000 CHF, above 20 000 CHF, and no answer. Respondents with no answers were omitted from the dataset. We converted below 8 000 CHF to 8 000 CHF and above 20 000 CHF to 22 000 CHF. For all other brackets, we chose the average number between the two bounds (10 000, 14 000, and 18 000 CHF, respectively).
Owner Age	The respondent's age in years.
Female	A dummy variable that takes the value of 1 if the respondent's gender is female and zero otherwise.
Couple, with children	A dummy variable that takes the value of 1 if the respondent lives in a couple-household with children.

Note: This table describes the closed-ended variables that were collected in the survey.

Table F.2: Variable Definition (2/2)

Variable	Description
Housing Preferences	
aesthetics, location, floor space, garden, parking, monetary value house, maintenance, carbon footprint, energy costs, Minergie label, indoor air quality, thermal comfort, and noise protection	For each of the 13 elements, respondents answered the following question: “How important are the following factors for your satisfaction with your home? (1 = ”not important“ to 10 = ”very important“)”
Satisfaction (compared)	Similar to Lyubomirsky and Lepper (1999) respondents rated their happiness with their home on a scale of 1 to 7.
Satisfaction (own)	Similar to Lyubomirsky and Lepper (1999) respondents rated their happiness with their home relative to their peers on a scale of 1 to 7.

Note: This table describes the closed-ended variables that were collected in the survey.