The Influence of City Size on Dietary Choices

Hao Cheng
Institut für Kartographie und Geoinformatik
Hannover, Germany
cheng@ikg.uni-hannover.de

Markus Rokicki
L3S Research Center
Hannover, Germany
rokicki@l3s.de

Eelco Herder
L3S Research Center, Germany & Radboud Universiteit Nijmegen, the Netherlands
eelcoherder@acm.org

ABSTRACT
In the past decades, the process of urbanization has shaped general socio-economic aspects of cities with different population sizes. Among them, food consumption is a good indicator to reflect the quality of life. In this paper, we study the impact of city size on food preferences, as shown by users of a large German food sharing community. We quantitatively and qualitatively analyze differences in dietary choices made by users who indicate to live in cities of different sizes, from metropolises and big cities to medium and small towns. Further, we demonstrate that the city size of the creators of online recipes can be predicted with a good accuracy of 86%, using predictors based on recipe authors’ profiles, recipe popularity, season, and recipe complexity and contents. The findings indicate that city size is a useful feature to take into account in various other domains.

KEYWORDS
Online food; city size differences; classification; food consumption

1 INTRODUCTION
In the past decades, urbanization has taken place around the world, with increasing numbers of people living in cities. Cities are believed to be focal points for economic growth, innovation, and employment [6]. Researchers have found that socio-economic characteristics are largely shaped by a city’s population size [3]. There has been some recent interest in investigating whether people eat differently and show different culinary activities across city sizes. For example, [17] shows differences in Asian cities of different sizes in the process of urbanization, diet change, and transformation of food supply chains.

Various studies on the influence of urbanization on food consumption have been conducted, albeit usually based on questionnaires and interviews [7, 23]. In addition, differences in eating habits between countries, states, and cities have been observed in quantitative studies [2, 8, 13, 19, 22, 24].

One challenge in studying the relation between food consumption and urbanization lies in collecting large amounts of data across cities. Nowadays, recipes and cooking information is readily available and easier to access than before. Online food communities, populated by users with various demographic backgrounds, provide a rich source of information for learning culinary patterns and predicting personal preferences.

In this paper, we investigate how city size captures many individual differences in food preferences and thus can serve as a meaningful addition to more traditional spatial features, such as geographical coordinates and country. To the best of our knowledge, we are the first to use online recipe data for quantitative and qualitative analysis of eating habits and preferences in relation to city size. Further, understanding food preferences across city sizes can be leveraged to improve food recommendation performance, which we confirmed in [5] by comparing different spatio-temporal contexts for context-aware food recommenders.

Contributions. In this paper, we explore the impact of city size in the large German online food community Kochbar1. We conduct a two-fold study on differences in food preferences. First, we perform statistical and qualitative analyses to investigate the nature of these differences between different city sizes. Further, we perform a classification experiment to investigate to what extent features related to these differences allow for predicting city size categories for individual recipes and to analyze which of these differences are most meaningful in this context. This way, we aim to provide insights into the nature as well as the impact of city size on differences in the field of cooking and food preferences.

2 RELATED WORK
Influence of City Size. The past decades have witnessed increasing numbers of people moving into cities from rural areas and the expansion of cities of all sizes. Cities provide significant opportunities for economic and social development [6]. City people usually cannot produce their own food, not only due to lack of space, but also due to lack of (spare) time, and therefore are dependent of the city’s food chains and food offerings [21]. Bettencourt et al. presented empirical evidence for relations between the population size of cities and a wide range of characteristics, including energy consumption, economic activity, demographics, infrastructure, innovation, employment, patterns of human behavior, using extensive data collected from US metropolitan statistical areas, European larger urban zones, and Chinese urban administrative units [3]. Wealth and prices scale superlinearly with city size, while individual human needs (job, house, household water consumption) scale linearly and material quantities associated with infrastructure scale sublinearly. Sarkar et al. used scaling indicators to analyze income inequality in Australia [20]. They found that a lower-income

1http://www.kochbar.de
were carried out using online data. Ahn et al. clustered recipes in Austria, and Switzerland [22]. West et al. analyzed web usage logs in German-speaking countries (Germany, Austria, and Switzerland [22]. West et al. analyzed web usage logs to discover nutrient patterns of different American states [24].

Regional Differences in Culinary Activities. In addition to studies on urbanization, there are culinary analyses based on location information. In contrast to the former, most of the following works were carried out using online data. Ahn et al. clustered recipes by their flavors and constructed flavor networks to uncover the ingredient preferences of different cuisines worldwide [2]. Similarly, Sajadmanesh used web data to explore worldwide culinary habits [19]. Howell et al. analyzed taste preferences for different countries [8]. Laufer et al. studied Wikipedia data to analyze European food cultures [13]. Wagner et al. used server logs to reveal ingredient preferences in German-speaking countries (Germany, Austria, and Switzerland) [22]. West et al. analyzed web usage logs to discover nutrient patterns of different American states [24].

As indicated above, location features were mostly extracted in terms of individual countries, states, or cities. However, cities with similar sizes are presumed to share several common characteristics: for instance, Readon et al. [17] analyzed diet changes and transformations in food supply chains in Asian cities in the process of urbanization, finding differences over time, as well as between rural and urban areas. Based on these insights, we propose to group cities into subsets according to their population size and to investigate differences in culinary habits and preferences between these groups.

<table>
<thead>
<tr>
<th>name</th>
<th>#entities</th>
<th>name</th>
<th>#entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>recipes</td>
<td>405864</td>
<td>categories</td>
<td>246</td>
</tr>
<tr>
<td>ingredients</td>
<td>1485</td>
<td>category classes</td>
<td>7</td>
</tr>
<tr>
<td>cooks</td>
<td>18212</td>
<td>≥ 10 recipes</td>
<td>4976</td>
</tr>
<tr>
<td>ratings</td>
<td>7796004</td>
<td>5-star-ratings</td>
<td>7724641</td>
</tr>
<tr>
<td>raters</td>
<td>19444</td>
<td>≥ 10 ratings</td>
<td>6231</td>
</tr>
<tr>
<td>comments</td>
<td>2751820</td>
<td>≥ 10 comments</td>
<td>4922</td>
</tr>
<tr>
<td>commenters</td>
<td>21951</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Summary of the dataset

3 METHODOLOGY

3.1 Dataset

Our study is based on a large-scale crawl from Kochbar.de, provided by Kusmierczyk et al. [12]. Kochbar.de is one of the most popular German online food communities, where users can upload, search, rate, and comment on food recipes. User profiles contain demographic information and the uploaded recipes contain specific information about ingredients, cooking directions, nutritional data, comments, user views, and ratings.

The dataset consists of over 400 thousand recipes in more than 200 categories, published between March 2008 and November 2014 (see Table 1 for an overview). Users provided more than 7 million ratings on the recipes and out of the active raters (those who have rated at least 10 recipes), more than 2 thousand provided location information in their profiles. The ratings are overwhelmingly positive, with over 99% of the ratings being a 5 star rating. Cooks are those who at least uploaded one recipe, over one-fourth of them are active ones (those who have uploaded 10 recipes).

City Sizes. City size is not an explicit feature provided by users. Therefore, we use Geonames location data (latitude and longitude) to find the closest city for the users. We adapt a settlement hierarchy to categorize cities based on their populations according to Geonames city population data. In particular, we group cities into five different city sizes: metropolis (≥ 1m), big city (500k, 1m), medium city [100k, 500k], small city [50k, 100k], and town [15k, 50k].

3.2 Alternative Rank Normalization

In our analysis of differences in recipe content for different city size conditions, recipe title terms, ingredients, and categories that are peculiar or specific to a particular city size are of particular interest. To this end, we make use of techniques for (text) corpus comparison. In their work on termhood extraction via corpus comparison [10], Kit and Liu explored the usefulness of different ranking approaches to describe a given corpus via comparison with a background corpus. In our work, we compare multiple context conditions, such as city sizes, with each other using Alternative Rank Normalization (ARN): to characterize recipes in one condition (e.g. to find ingredients used in metropolises but not in others) we construct two ranked lists of the items of interests (e.g. ingredients), one for the corpus of interest (e.g. metropolis) and the background corpus (all other cities). We then calculate the (normalized) rank difference in the item’s rank between both corpuses and sort the items by the rank difference; the most salient items for the corpus are the ones with the largest rank difference (i.e. ranked high in the corpus of interest and ranked low in the background corpus).

3.3 Random Forests

To predict recipe city size, we use Random Forests [4], a state-of-the-art classifier that is resistant to overfitting and that can also be applied to rank importance of features.

Breiman showed that the accuracy of random forests depends on the strength of individual trees and the correlation between the trees [4]. A modified bagging procedure is used to learn the
to the tree ensemble. In addition to repeated randomly sampling a new training set to grow a new tree for the ensemble, the tree learner also sub-samples the feature space randomly at each split, thus reducing correlation between individual trees. Breiman pointed out that bagging in this manner enhances the accuracy and can be used to estimate the generalization error, strength, and correlation of combined trees. Although individual trees are sensitive to overfitting, the average of the vote of all the combined trees is not, despite increasing model complexity by incorporating more trees.

4 DATA ANALYSIS

In this section, we analyze users and recipes from different city sizes. In order to reduce noise caused by exotic recipes and eccentric users, we focus on users who have rated at least 20 recipes and recipes with at least 10 ratings. Table 2 shows an overview of the resulting dataset used for our analysis. Recipes are assigned to city sizes based on the location information provided by their cooks.

<table>
<thead>
<tr>
<th>sizes</th>
<th>cities</th>
<th>pop.</th>
<th>cooks</th>
<th>recipes</th>
<th>raters</th>
<th>ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>metro.</td>
<td>3</td>
<td>6425862</td>
<td>339</td>
<td>12735</td>
<td>3632</td>
<td>386194</td>
</tr>
<tr>
<td>big</td>
<td>10</td>
<td>6028762</td>
<td>387</td>
<td>12358</td>
<td>3903</td>
<td>399695</td>
</tr>
<tr>
<td>med.</td>
<td>74</td>
<td>16352954</td>
<td>778</td>
<td>27014</td>
<td>4250</td>
<td>636914</td>
</tr>
<tr>
<td>small</td>
<td>107</td>
<td>8610911</td>
<td>509</td>
<td>14502</td>
<td>3950</td>
<td>433919</td>
</tr>
<tr>
<td>town</td>
<td>640</td>
<td>48314158</td>
<td>1876</td>
<td>89705</td>
<td>4574</td>
<td>219259</td>
</tr>
</tbody>
</table>

| sum    | 834    | 85732647 | 3889 | 156314 | 4718  | 4675981 |

* sum of distinct raters

4.1 Analysis of User and Recipe Attributes

Figure 1 depicts the percentages of the population, cooks, recipes, and ratings relative to the respective total sums for each city size. In terms of absolute numbers, most users live in towns. Medium-size cities have relatively more cooks than their population (19.1%). However, this somewhat larger user community provides fewer recipes (17.3%) and ratings (13.6%). By contrast, towns (cities of 50,000 and smaller) have relatively fewer cooks (48.2%), but these users are more active in terms of recipes (57.4%) and ratings (60.3%). Indeed, the average number of recipes per cook is statistically higher for towns ($M = 47.82$) than for small cities ($M = 28.49$, $W = 337150$, $p < .001$, $r = .07$) and big cities ($M = 31.93$, $W = 449950$, $p < .05$, $r = .05$).

Recipe uploading behavior differs significantly between city sizes during different days of week ($N = 156314$, $x^2 = 352.55$, $df = 12$, $p < .001$) and seasons ($N = 156314$, $x^2 = 63.165$, $df = 4$, $p < .001$). For example, cooks from small cities uploaded many more recipes on weekends, whereas cooks from metros uploaded fewer recipes on weekdays than expected. In Spring, cooks from medium cities were more motivated to upload recipes; in Autumn, cooks from small cities were less active.

Small but significant differences of cooks’ demographics in terms of age and gender are also found between city sizes. Medium city cooks are younger on average ($M = 39.3$) compared to the other city sizes ($M = 40.5$), $W = 266970$, $p < .001$, $r = .08$, while town cooks are slightly older ($M = 40.9$). In terms of gender, the distribution is unequal as well. Further, with respect to recipe popularity, differences between genders align with observed differences in the number of ratings between city sizes ($N = 4697785$, $x^2 = 145500$, $df = 4$, $p < .001$), as suggested by previous work on gender differences in online cooking [18].

4.2 Analysis of Recipe Contents

We continue our analysis by focusing on the features of the recipes in the category “main dishes”. Based on 31 identified red-meat ingredients [18], we observe a relationship between the use of red meat and city size ($N = 45497$, $x^2 = 20.34$, $df = 4$, $p < .001$). On
average, red meat is used more in big cities ($M = 38.4\%$) and less in metropolises ($M = 34.7\%$) and medium ($M = 36.4\%$) cities. In line with the expectation of more exotic dishes in larger cities, spices are more frequently used in big cities ($M = 94.3\%$) and metropolises ($M = 93.4\%$) than in medium ($91.3\%$) cities ($N = 45497, \chi^2 = 35, df = 4, p < .001$). This is illustrated nicely by the number of curry dishes decreasing with city size, as shown in Figure 3.

Figure 3: Curries used in main dishes for each city size.

Main dishes from metropolises ($M = 12.6$), big cities ($M = 12.5$), and small cities ($M = 12.3$) use more ingredients than the ones of the medium–cities ($M = 11.52$, e.g. metropolis vs. medium city: $W = 11288000, p < .001$, $r = .15$). Similarly, metropolises ($M = 45.7\%$), small city ($M = 43.9\%$), and big city ($M = 43.2\%$) main dishes take longer to prepare than medium city dishes ($M = 39.9\%$, e.g. metropolis vs. medium city: $W = 4453400, p < .001$, $r = .55$).

In terms of nutrients, small-city main dishes contain fewer calories ($M = 809\text{kJ}$) than dishes from the other city sizes (e.g. small city vs. metropolis $M = 927.7\text{kJ}$: $W = 5744900, p < .001$, $r = .13$), as well as less fat (e.g. small city $M = 15.9\text{g}$ vs. metropolis $M = 13.2\text{g}$: $W = 5666500, p < .001$, $r = .12$), but more protein (such as small city $M = 6.7\text{g}$ vs. medium city $M = 6.2\text{g}$: $W = 8610800, p < .001$, $r = .06$). The differences for carbohydrates are not that obvious, but still significant. For example, small-city ($M = 11.7\%$) main dishes contain more carbohydrates than big-city ($M = 10.7\%$) main dishes ($M = 6.2\text{g}$: $W = 4566900, p < .001$, $r = .13$).

In summary, metropolis main dishes are more complicated than the medium city main dishes. With regard to the nutrition, the small city main dishes are better in quality than the medium city main dishes.

4.3 Qualitative Analysis of Recipe Contents

Next, we analyze recipe titles\(^4\), categories and ingredients qualitatively by means of ARN (see Section 3.2). Again, we focus on recipes from the category “main dishes”, to avoid bias caused by the type of dishes.

Table 3 lists the 20 most peculiar title terms for the different city sizes. There are many foreign dishes in the metropolis and big city categories, some even in the medium city category. By contrast, the small city and town categories contain more traditional recipe titles. For example, *dim sum* and *sticky rice* are typical Asian foods, while *Baden* and *wild rice* are very local and traditional.

The foreign and traditional pattern is further confirmed by recipe categories (the corresponding table is not shown in this paper due to space limitations). Metropolises and big cities contain more exotic categories, like Orient, Turkey and Indonesia, whereas location categories for towns include the local East-Frisian, Swabian and Thuringian cuisines, as well as some nearby countries like Denmark and Switzerland. Some unexplainable patterns have been found as well, though, such as the presence of Romania in the town category. This may be caused by a large minority in a particular town or due to a nearby big city or metropolis, where people have far easier access to more exotic ingredients and foreign cuisines.

The same pattern can also be found in the selection of ingredients used. For example, metropolis recipes use *soja, quinoa, carmelli, aioli*, and *tellicherry pepper* distinctively. These ingredients are either from foreign countries, exclusive or expensive. Many of these ingredients are also spice ingredients: we observed a descending use of spices from the metropolis to town categories (see Section 4.2).

We further break down recipes to different genders across city sizes. Women show a preference for vegetarian and sweet dishes in the metropolis and the big city categories. This preference becomes weaker in the smaller city size categories. On the other hand, men show a preference for hearty, spicy, and foreign cuisines.

When we closer inspect the three major metropolises (Berlin, Hamburg, and Munich), we observe that – apart from the foreign cuisine cooked in all three – each city has its own distinct characteristics. Vegetarian and vegan food is most popular in Berlin. The harbour city of Hamburg shows a slight preference for seafood. Munich is the city where you find Bavarian food and southern European cuisines, especially Italian cuisine, and of course a “beer culture”: *drink, autumn*, and *with alcohol* are among the 20 most peculiar terms or categories list in Munich.

4.4 Summary of Findings

Users from towns are more active than users from other, larger cities. Smaller city sizes are associated with fewer calories and fat, but also with less exotic and more traditional food, as well as with fewer spices. Medium-city recipes use fewer ingredients and take less time - which might indicate that inhabitants of medium cities

---

\(^4\)Recipe titles are preprocessed by applying tokenization and stemming with NLTK (http://www.nltk.org) and manually removing noisy terms.
experience more time pressure - and consequently also receive fewer views and ratings. At several points in our analysis, we found indicators in recipes from medium cities that hint at a lower average income and a larger proportion of immigrants than in both metropolises or towns – which is in line with social-economic literature [11]. The distinctive terms, categories and ingredients further illustrate the nature of these differences.

5 CITY SIZE CLASSIFICATION

In the previous section, pronounced differences in terms of recipe content and users have been shown across city sizes. To find out which of those differences are most discriminative, we perform a classification experiment to predict city size of recipes.

5.1 Setup

Based on our analysis, we select 53 features related to cooks and recipes, as shown in Table 4. For the cook properties, we consider demographic information, as well as some measures of activity and popularity. Recipe content is represented on a high level by recipe categories. In addition, we include features on recipe popularity, recipe structure and complexity, nutrition, and temporal context.

Table 4: Features for recipe city size classification

<table>
<thead>
<tr>
<th>Cook (6 features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>guest-book messages, age, points, active-days, uploaded recipes, gender.</td>
</tr>
<tr>
<td>Recipe popularity (5 features)</td>
</tr>
<tr>
<td>comments, ratings, favorites, views, average-rating.</td>
</tr>
<tr>
<td>Recipe structure (6 features)</td>
</tr>
<tr>
<td>duration, servings, ingredients, price, difficulty, spice-dish.</td>
</tr>
<tr>
<td>Recipe nutrients (4 features)</td>
</tr>
<tr>
<td>kj, carbohydrates, protein, fat.</td>
</tr>
<tr>
<td>Recipe time (2 features)</td>
</tr>
<tr>
<td>season, day-of-week.</td>
</tr>
<tr>
<td>Category (30 features)</td>
</tr>
<tr>
<td>occasions, special, international, Europe, main-dish, lunch, supper, summer, cheap, milk-products, spring, party, autumn, quick-easy, regional, winter, coffee-cake, vegetarian, meat, cake, intermezzo, snack, healthy-diet, dessert, starter, gluten-free, lactose-free, casserole, no-wheat, allergy.</td>
</tr>
</tbody>
</table>

In order to avoid biases introduced by a small number of very active users who uploaded hundreds or thousands of recipes, we randomly sample ten recipes per user. Furthermore, we perform under-sampling to balance the dataset across classes, which results in 718 recipes per city size. In the following, we first evaluate the discriminative power of features in terms of decrease in Random Forest (RF) accuracy (RF-MDA) [14]. Afterwards, we evaluate classification performance in terms of accuracy using 10-fold cross-validation.

5.2 Results

Feature importance in terms of decrease in Random Forest accuracy is shown in Figure 4. Although we observe clear differences in feature quality between – but also within – feature groups, the feature importance scores show that almost all of the features groups are meaningful. Except for nutrient features, each of the feature groups is evenly represented among the 13 best features. It is worthwhile to note that nutrient values are particularly subject to noise, as some users provide inaccurate values.

Ranked most highly within the eight most useful features are all features that characterize recipe cooks in terms of their demographics, activity and popularity. Surprisingly, the temporal context – recipe season in particular – is also ranked quite highly. The remaining highly ranked feature groups relate to recipe popularity, content, and structure. The importance of popularity features – such as number of comments, ratings, favorites, and views – reinforce the results on differences in terms of recipe popularity shown in the previous section. Further, although the content features are coarse-grained and overall lower-ranked, our findings with regards to differences in food preferences are also reflected in the feature ranking. Categories for special and international dishes are ranked particularly high and categories that encode the other main differences found in the previous section, such as vegetarian and regional dishes, achieve overall similar feature importance scores. In addition, differences in recipe complexity in terms of duration, servings and number of ingredients appear to be useful as well.

Using all of the features, the Random Forest classifier predicts recipe city size with an overall high accuracy of 78%. Restricting the feature set to the top 20 features according to RF-MDA further improves classification accuracy to 86%. The confusion matrix for the latter results, shown in Table 5, further shows that the classification performance is robust across the city sizes, with only minor variations.

Table 5: Confusion Matrix using the top 20 features

<table>
<thead>
<tr>
<th>classified as</th>
<th>a = metropolis</th>
<th>b = big city</th>
<th>c = medium city</th>
<th>d = small city</th>
<th>e = town</th>
</tr>
</thead>
<tbody>
<tr>
<td>a = metropolis</td>
<td>628</td>
<td>37</td>
<td>14</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>b = big city</td>
<td>28</td>
<td>610</td>
<td>24</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>c = medium city</td>
<td>25</td>
<td>33</td>
<td>603</td>
<td>30</td>
<td>27</td>
</tr>
<tr>
<td>d = small city</td>
<td>32</td>
<td>33</td>
<td>22</td>
<td>617</td>
<td>14</td>
</tr>
<tr>
<td>e = town</td>
<td>19</td>
<td>29</td>
<td>27</td>
<td>15</td>
<td>628</td>
</tr>
</tbody>
</table>

6 DISCUSSION AND CONCLUSION

In this paper, we have shown that food preferences depend on the size of city that people are living in and discussed the nature of these differences. Among others, in Germany, people in metropolises eat more foreign food and people in smaller cities and towns eat more traditionally. Medium-city recipes contain less protein but more calories and fat than recipes from other city sizes. These features are sufficient for reliably predicting a user’s city size.

The findings of the influence of city size on dietary choices provide meaningful information for food recommendation. In [5], we show that using city size as a feature has a positive impact on context-aware recipe recommendation [1]: among several spatio-temporal contexts (day-of-week, season, and inner-border), city size turned out to give the best performance for food recommendation.

For our analysis, we relied on the user’s self-reported location, without taking any nearby larger cities into account. In future work,
it may be useful to take, in addition to the population, distances to adjacent cities into account - in order to distinguish between isolated towns and towns close to big cities or metropolises.

The influence of city size on food preferences, as discussed in this paper, is in line with the related work as discussed in Section 2. Knowledge on the impact on city size can be effectively translated into measures to reinforce or counteract such effects [16]; for example, it has been found that as city size grows, people tend to be less socially connected and less interested in local politics and affairs [15]; these findings have been used for improving the configuration of metropolitan institutions [9]. In sum, the size of a city has a clear impact on the habits, possibilities, interests and preferences of its inhabitants and therefore is expected to be a useful feature in context-based recommendation in general.

REFERENCES