Culinary Map of Turkey

Ezgi Cinkılıç
Industrial Engineering and Artificial Intelligence Department
TOBB University of Economics and Technology
Ankara, Türkiye
ezgi.cinkilic@etu.edu.tr
Culinary Map of Turkey – GitHub Repository
Culinary Map of Turkey Dataset – Kaggle

***Abstract*—This study aims to investigate the geographic distribution of recipes using a large dataset of user-generated content collected from an online recipe-sharing platform. The primary objective is to identify regional culinary zones within Turkey by analyzing recipe distributions across provinces. This approach offers an innovative, data-driven alternative to traditional survey-based methods, providing a novel perspective on Turkey's rich and diverse food culture through large-scale behavioral data.**

The methodology involved extensive data collection and preprocessing, including web scraping, data cleaning, and normalization to address significant imbalances among recipe categories and city-level contributions. Clustering analyses were conducted using K-Means and spatially constrained hierarchical clustering (AZP algorithm), with the optimal number of clusters determined through Elbow and Silhouette criteria. Z-score analyses were employed to evaluate prominent recipe categories within each cluster. The final clusters were visualized using interactive geographic maps to reveal meaningful regional culinary patterns.

This study demonstrates how large-scale user data combined with advanced clustering and spatial analysis can provide culturally meaningful insights beyond traditional survey-based approaches.

***Index Terms*—Data Mining, Web Scraping, Geographical Analysis, Clustering, K-Means, Hierarchical Agglomerative Clustering**

I. Introduction

This project aims to construct a data-driven digital culinary map of Turkey by analyzing user-generated recipe data obtained from a popular recipe-sharing platform. By combining recipe categories (e.g., desserts, meat dishes, vegetable-based dishes) with the location information of users, regional culinary patterns are uncovered and visualized across Turkish provinces. This approach offers an alternative to traditional survey-based studies by leveraging real-world behavioral data at scale.

Understanding regional culinary identities is not only essential for preserving intangible cultural heritage but also provides valuable insights for fields such as gastronomy tourism, regional branding, and food marketing. Traditional methods such as field surveys or expert interviews are often time-consuming, limited in scale, and prone to bias. In contrast, online platforms where users organically share everyday recipes present a rich, underutilized source of culinary data that reflects actual cooking behavior rather than self-reported preferences or idealized regional cuisines.

In particular, this project assumes that the recipes uploaded by users—especially those beyond traditional or ceremonial dishes—are indicative of meals they frequently prepare in daily life. Under this assumption, we aim to uncover what people are likely consuming on a regular basis and to detect subtle differences in commonly consumed food categories across regions. This allows us to investigate whether culinary habits change as geographic location changes, and whether new, data-driven culinary zones can be identified based on these behavioral patterns.

The project followed a multi-stage methodology:

- **Data Collection:** Recipes and user profiles were scraped using Python's BeautifulSoup, with batch control and rate limiting.
- **Preprocessing:** Categories were cleaned, duplicates removed, and categorical variables label-encoded; one-hot encoding was considered for clustering stages.
- **Geo-Analysis:** GeoPandas and Folium were used to visualize province-level user and recipe distributions via choropleth maps.
- **Clustering:** K-Means clustering identified culinary zones by grouping provinces with similar recipe category patterns.

The main objectives of this project are outlined below:

- Analyze the distribution of recipe categories across Turkish provinces using real user data.
- Visualize regional densities of specific food types and identify location-specific culinary patterns.
- Define data-driven culinary zones through clustering and construct a digital map reflecting regional food trends in Turkey.

By combining techniques from data mining, spatial analysis, and unsupervised learning, this project provides an innovative framework for understanding how food culture is expressed in digital environments and how it may diverge from traditional regional narratives.

II. Literature Review

Previous studies on Turkey's culinary regions show methodological differences. For example, Yayla et al. (1) and Yayla and Aktaş (2) identified culinary zones by analyzing ingredient compositions of recipes sourced from experts, focusing on detailed recipe content and ingredient distributions. These studies emphasize flavor regions based on ingredient spatial patterns.

On the other hand, Yayla (3) proposed culinary regions by considering broader geographical, climatic, migratory, and agricultural factors without detailed recipe data.

Most prior research relies on expert knowledge or survey data, lacking large-scale datasets that combine user-generated recipe content with user location information. Publicly available recipe datasets rarely include user profiles, which severely limits region-specific analysis. Moreover, although some datasets contain user interaction data (e.g., comments or ratings), these are mostly found in restaurant review platforms or recipe feedback sections and are not structured in a way that allows matching users to their original recipe contributions and associated locations.

To our knowledge, there is no existing dataset in Turkey—or globally—that includes detailed recipe texts along with verified user location data at the city or province level. Furthermore, in recipe-sharing platforms, city-level user metadata is either unavailable or shared very rarely due to privacy policies and platform design. When available, many shared recipes are uploaded by professional chefs or culinary influencers, leading to a dataset that is more reflective of aspirational or curated culinary presentations rather than everyday cooking behaviors of the general public.

This project addresses these gaps by creating a custom dataset via large-scale web scraping that combines user-uploaded recipe content with location and profile metadata. This enables the use of both supervised and unsupervised machine learning techniques for uncovering regional culinary patterns. The study offers a novel approach to mapping food culture in Turkey by leveraging behavioral data from ordinary users, rather than relying on traditional survey or expert-driven methodologies.

III. Data Set, Data Characteristics, and Features

A. Data Collection & Preprocessing Steps

The dataset was created using web scraping from the publicly accessible recipe-sharing website https://www.nefisyemektarifleri.com/tarifler/. Data collection was carried out using the BeautifulSoup library in Python. Both recipe category pages and user profile pages were targeted.

To prevent IP blocking and reduce server load, request intervals were randomized between 1 and 3 seconds. Additionally, after every 100 requests, a 60-second delay was introduced due to rate-limit errors (HTTP 429). A secure scraping pipeline was implemented, where previously collected profile URLs were tracked, and each batch was safely written to disk before proceeding. This ensured recovery in case of interruptions and avoided re-downloading already collected data.

Category and Subcategory Collection: The data collection process began by scraping all available recipe categories and subcategories. For each category, the category name, subcategory name, and category URL were recorded. Non-relevant or advertisement-related categories (e.g., baby foods or promotional items) were manually removed. After cleaning, 68 valid food category–subcategory pairs were identified and assigned unique numerical IDs for processing.

Recipe Data: For each of the 68 categories, the number of recipe pages was identified. Each category page listed 23 recipes per page. Using this structure, all pages under each category were scraped in batches to avoid server overload. Approximately 900,000 recipes were extracted, and after removing duplicates (recipes appearing in multiple categories), 800,140 unique entries remained. For each, the recipe title, category information, recipe URL, and the associated user profile URL were collected.

User Profile Data: After extracting 800,140 unique recipe entries, approximately 94,000 unique user profile URLs were identified. Due to deleted or renamed accounts, 74,892 profiles were successfully scraped in batches of 1,000 to avoid rate-limiting.

Each user profile included the username, membership duration, total number of recipes shared, number of followers and followings, and the city of residence. The membership duration was originally provided as a string in Turkish (e.g., “5 yıl 3 ay”). During preprocessing, this string was parsed and converted into a single integer value representing the total number of months.

Out of the 74,892 user profiles, 37,684 included non-empty city information and were considered for location-based analysis. To ensure data consistency, entries with missing values (e.g., NaN) and non-Turkish locations (e.g., ”yurtdışı”) were removed, leaving 35,436 valid profiles. City names were normalized using a custom function to handle inconsistencies in spelling and encoding. Fuzzy matching was applied to align cleaned names with official Turkish city names from a GeoJSON file, enabling accurate geographic analysis.

Final Dataset Construction: User profile and recipe data were merged using an inner join on profile names, retaining only recipes with matching user profiles. This resulted in a dataset of 21,807 unique users and 321,312 associated recipes. Main and subcategory names were added using the category ID to complete the dataset.

While the dataset is stored in tabular form, columns such as recipe name and URL contain semi-structured text, whereas the remaining columns are fully structured.

User profile and recipe data were merged using an inner join on profile names, retaining only recipes with matching user profiles. This resulted in a dataset of 21,807 unique users and 321,312 associated recipes. Main and subcategory names were added using the category ID to complete the dataset.

While the dataset is stored in tabular form, columns such as recipe name and URL contain semi-structured text, whereas the remaining columns are fully structured.

Usernames were anonymized as 6-digit profile IDs prior to public release to protect user privacy.[1]

[1] https://www.kaggle.com/datasets/ezgicinkilic/turkish-recipe-sharing-platform-dataset/data

B. Feature (Variable) Analysis

- **Handling Missing/Erroneous Data:** Users with missing username or city information were excluded from the dataset. There were no missing values in the following and follower count columns. Missing values in the recipe count column were imputed using the median to reduce the impact of outliers and maintain data integrity.

TABLE I
MEASUREMENT LEVELS OF DATASET VARIABLES

Variable	Description	Level
category_id	Category identifier	Nominal
recipe_name	Name of the recipe	Nominal
recipe_url	Link to the recipe page	Nominal
profile_name	Username of the recipe owner	Nominal
registration_month	Months since user joined	Ratio
recipe_count	Number of recipes by the user	Ratio
followers	User's follower count	Ratio
following	Number of users followed	Ratio
city	City name	Nominal
main_category	Top-level recipe category	Nominal
sub_category	Subcategory of the recipe	Nominal

- **Exploratory Data Analysis:** Exploratory data analysis was performed to understand relationships, distributions, and spatial patterns in the dataset.
  Positive correlations (0.58) were observed between follower counts and the number of recipes shared by a user. Cities with higher user counts tended to have higher aggregate recipe counts as well. Correlations among other feature pairs were close to zero, suggesting little to no linear association.
  When calculating p-values for the correlation matrix, all p-values were effectively zero due to the very large sample size. Therefore, practical significance was considered instead. A correlation of 0.58 indicates a strong positive relationship. Meanwhile, correlations of 0.13 and 0.15 observed between recipe count, follower count, and registration month indicate weak positive relationships.
  Numeric variables such as recipe_count, followers, and following showed right-skewed distributions, with most users having low counts and a few users having very high counts. Particularly for follower and following counts, a large portion of users had values at or near zero, indicating these features may not be strongly discriminative. Logarithmic transformations were applied to reduce the effect of extreme values before drawing box plots.

TABLE II
SUMMARY STATISTICS FOR NUMERIC VARIABLES

Statistic	registration_month	recipe_count	followers	following
Mean	42.05	90.10	117.94	803.54
Std Dev	18.51	31.22	167.33	3013.61
Min	1	0	1	0
25th Pctl	25	65	19	27
Median	41	92	52	92
75th Pctl	59	116	146	375
Max	68	213	1363	48908

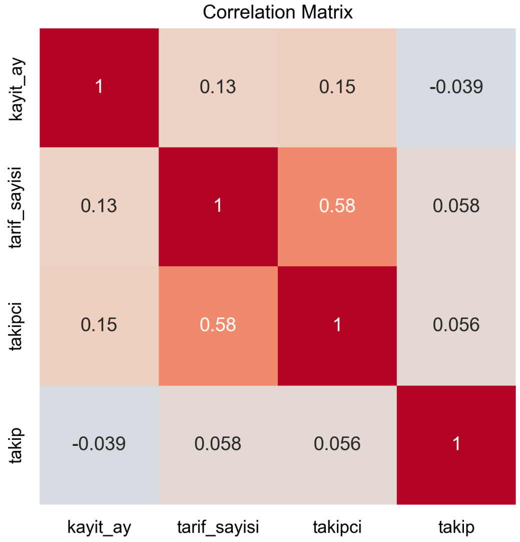

Fig. 1. Correlation Heatmap Between Key Features

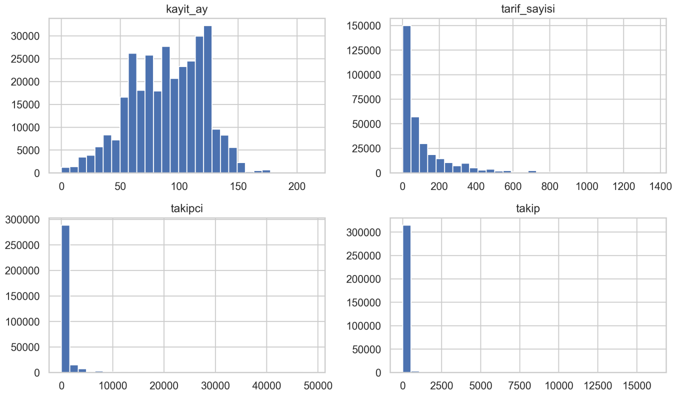

Fig. 2. Histogram of Numeric Variables

Categorical variables such as main_category and sub_category were analyzed with frequency plots. As shown in Fig. 3, dessert and bakery categories account for nearly half of all recipes, indicating class imbalance within the data. This imbalance is also reflected in the word cloud visualization included in Fig. 4.

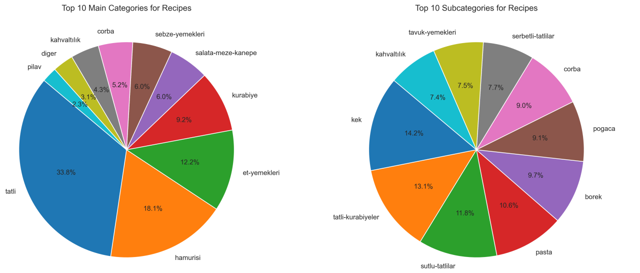

Fig. 3. Recipe Distribution by Main and Subcategories

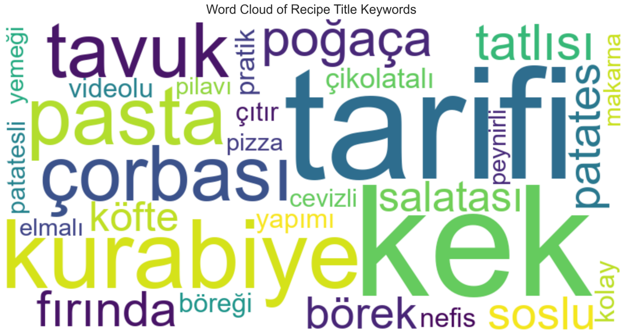

Fig. 4. Word Cloud of Recipe Title Keywords

Additional visualizations of city-level user counts relative to population, and recipe counts relative to user counts, demonstrate that user numbers are linked to city population, while recipe numbers depend on user counts. A Pearson correlation of 1.00 between user count and recipe count confirms a perfect positive linear relationship, meaning that more users directly translate into more shared recipes.

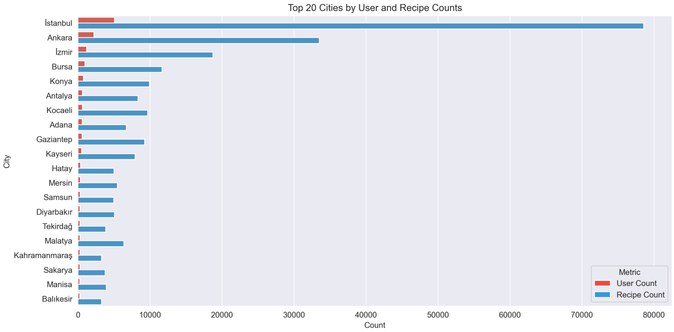

Fig. 5. Number of Users and Recipes in Top 20 Cities

Spatial patterns were explored by aggregating user and recipe counts at the city level. An interactive choropleth map of Turkey (Fig. 6) allows toggling between user density and recipe density views. In addition to user and recipe distributions, category-specific recipe maps were also generated to illustrate which food categories were shared in which cities.

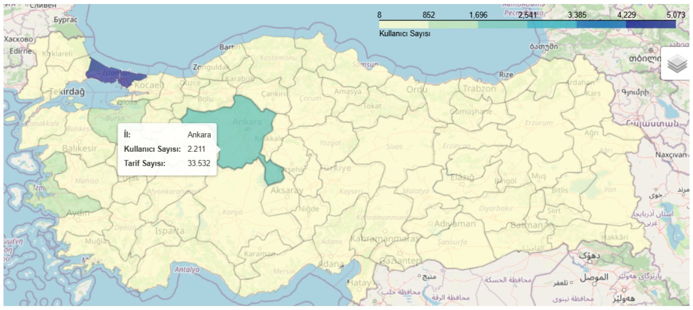

Fig. 6. Choropleth Map of User and Recipe Densities

Before applying clustering algorithms, normalization of numeric variables is important due to differing scales and distributions. Normalization helps reduce the impact of outliers and scale differences, improving clustering performance.

IV. Used Methodology

The methodology employed in this study combines normalization, categorical encoding, clustering, and spatial analysis to uncover culinary patterns across Turkey.

To address the significant imbalance in the number of shared recipes across cities and between recipe categories, a two-step normalization procedure was applied. First, inter-category normalization was performed to equalize the influence of overrepresented categories. Then, intra-city normalization was applied to standardize recipe distributions within each city, resulting in a normalized feature vector per city that reflects the relative prominence of each category locally.

Additionally, a binary encoding scheme was implemented to indicate the presence or absence of each recipe category in a given city. For each city, if a recipe was shared in a category, a value of 1 was assigned; otherwise, 0. These vectors served as input for the clustering algorithms.

Initial clustering was performed using the K-Means algorithm due to its efficiency and effectiveness in handling large datasets and producing well-separated clusters. The optimal number of clusters was determined to be 4, based on Elbow and Silhouette scores (see Fig. 7). After clustering, feature vectors were re-examined using z-scores and minimum scores to identify distinctive regional culinary patterns. Despite normalization, cake and pastry-related categories remained disproportionately dominant, masking variation in other categories. To counter this, a TF-IDF inspired weighting scheme was employed to emphasize underrepresented categories.

An initial map (Fig. 8) was generated using the TF-IDF weighted vectors, clustering cities into 4 groups without considering geographic proximity. Interestingly, even in this spatially agnostic setup, geographically coherent clusters emerged.

In the subsequent step, spatial relationships were explicitly incorporated through hierarchical agglomerative clustering. This method was preferred due to its ability to consider spatial constraints and form clusters based on proximity, which K-Means does not inherently provide. To model spatial adjacency, a contiguity matrix was constructed using the Queen contiguity rule via the libpysal library. This matrix was then converted into a graph structure using the networkx library to facilitate neighborhood analysis. The AZP (Automatic Zoning Procedure) algorithm was applied on this spatial graph with a constraint ensuring at least 10 cities per cluster. This adjustment was necessary because earlier attempts using 7 clusters (reflecting Turkey's administrative regions) resulted in highly unbalanced clusters, with some containing only a single city.

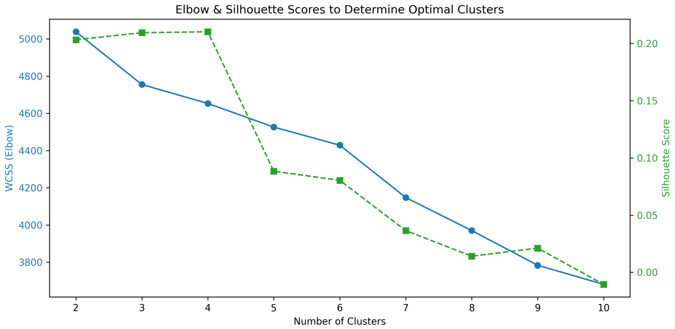

Fig. 7. Elbow and Silhouette Scores Graph

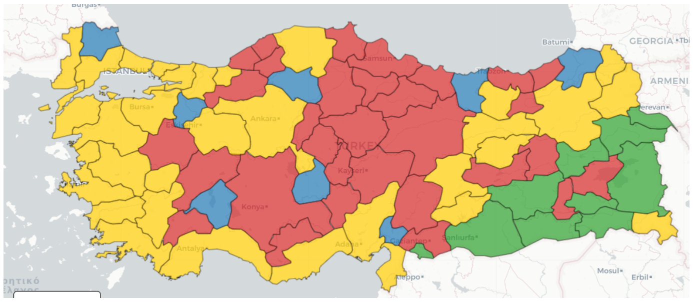

Fig. 8. 4 Culinary Regions (TF-IDF / Non-spatial)

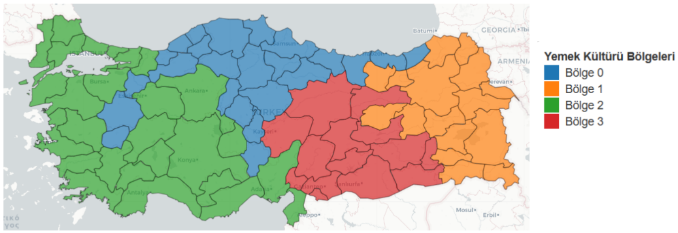

Fig. 9. 4 Culinary Regions (TF-IDF / Spatial)

V. Results

The final clustering revealed meaningful regional culinary patterns (Fig. 9).

Cluster 0, primarily covering the Black Sea and Central Anatolia regions, showed a dominance of traditional main courses such as pilaf, meatballs, and chicken dishes, with minimal preference for pastries or sandwiches.

Cluster 1, representing Eastern Anatolia, was characterized by practical and traditional snack-oriented recipes, including raw meatballs (çiğ köfte), savory cakes, stuffed grape leaves (sarma), and stews, whereas pilaf and red meat dishes were less common.

Cluster 2 encompassed parts of the Marmara, Aegean, Mediterranean, and Central Anatolia regions, featuring practical and sweet snack recipes such as syrup-based desserts and sandwiches. Olive oil-based dishes were underrepresented in this cluster.

Finally, Cluster 3 exhibited less distinctive culinary features, with hot beverages being the only notable category. Due to insufficient data, categories such as kebabs, mantı (Turkish dumplings), and seafood were excluded from the z-score analysis. Although these categories might have influenced cluster formation to a limited extent, their low frequency prevented meaningful inclusion in the final interpretation.

When interpreting the results, it is notable that the detected culinary clusters resemble those identified in traditional region-based survey studies. The regions corresponding to the Black Sea and Eastern/Southeastern Anatolia were distinctly separated, aligning with known culinary traditions. However, the expected differentiation between the Aegean-Mediterranean and Central Anatolia was not clearly observed. This suggests that user-generated recipe data may not fully represent traditional regional cuisines in some areas. Instead, users appear more inclined to share commonly prepared everyday meals rather than region-specific traditional dishes.

Another important consideration is the potential influence of migration. Users may reside in cities different from their hometowns and share recipes reflecting their native culinary culture, rather than the region they currently live in. This can reduce spatial coherence, particularly in regions with fewer users or where geolocation metadata is missing. Despite these limitations, the clusters still reflect culturally meaningful patterns and demonstrate the feasibility of culinary regionalization using user-contributed web data.

VI. Conclusions and Discussion

A large dataset of user-shared recipes from a popular online platform was analyzed to uncover regional culinary patterns across Turkey. Data cleaning, normalization, categorical encoding, and clustering methods were applied, primarily using K-Means and spatially constrained hierarchical clustering. K-Means was selected for its efficiency and interpretability in high-dimensional data, while spatial clustering with the AZP algorithm incorporated geographic contiguity, resulting in more coherent regional groupings.

Four distinct culinary zones were identified, characterized by different recipe category preferences, reflecting both traditional and modern food culture variations. The clusters aligned well with known geographic and cultural regions, demonstrating the effectiveness of combining data-driven and spatial approaches in culinary analysis.

The results also suggest that recipe-sharing behavior is influenced by daily food habits and user preferences rather than strict adherence to traditional regional cuisines. Especially in regions like the Aegean and Central Anatolia, expected culinary distinctions were not strongly evident in the data. This could stem from a tendency among users to post widely consumed or convenient meals instead of local specialties. Furthermore, the absence of consistent location metadata and the presence of users living outside their native regions may introduce additional noise, particularly in regions with low representation.

Due to time constraints and the dataset size (over 300,000 recipes), detailed recipe content such as ingredients, cooking methods, and preparation times was not included in the clustering analysis. Collecting and processing these data would require substantial additional time and resources. This limitation restricts the granularity of insights but does not undermine the validity of the discovered regional patterns based on recipe category distributions.

For future work, the analysis will be extended by incorporating recipe content using natural language processing techniques. By examining ingredient usage, cooking techniques, and other textual features, the clustering process can be enriched to better capture subtle regional culinary distinctions. Additionally, temporal dynamics and socio-demographic variables may be integrated to analyze how culinary preferences evolve over time and across population segments.

Overall, this study provides a foundation for data-driven culinary regionalization in Turkey and demonstrates the value of combining normalization, clustering, and spatial analysis. The methodology and findings can inform further gastronomic studies, cultural heritage projects, and applications in regional food tourism or marketing.

References

[1] Önder Yayla and S. G. Aktaş, "Türk mutfağında lezzet bölgelerinin belirlenmesi: Adana-osmaniye-kahramanmaraş Örneği," in *1st International Congress on Future of Tourism: Innovation, Entrepreneurship and Sustainability*, (Turkey), 2017.

[2] Önder Yayla and S. G. Aktaş, "Mise en place for gastronomy geography through food: Flavor regions in turkey," *International Journal of Gastronomy and Food Science*, vol. 26, p. 100384, 2021.

[3] Önder Yayla, "Türkiye'nin lezzet coğrafyası Üzerine bir değerlendirme," *Osmaniye Korkut Ata Üniversitesi, Kadirli Uygulamalı Bilimler Fakültesi Dergisi*.