Quality & Quantity
https://doi.org/10.1007/s11135-025-02441-y

The economic feasibility of sustainable and healthy diets: a price-based analysis in Italy

Ilaria Benedetti[1] · **Stefano Marchetti**[2] · **Haoran Yang**[2] · **Mathias Silva Vazquez**[3]

Received: 27 April 2025 / Accepted: 3 October 2025

Abstract
This study focuses on the economic dimension of food poverty and aims to develop a replicable methodology for estimating the cost of adhering to Healthy and Sustainable Diet (HSD) patterns in Italy. Using a novel dataset built through web scraping from the Osservatorio Prezzi e Tariffe, we estimate diet costs across different population groups in line with national nutritional guidelines. The dataset provides detailed price information for key food categories across Italian provinces, enabling an assessment of local disparities in food affordability. To address missing data, we apply an imputation strategy that leverages spatial and temporal correlations. Results reveal substantial variation in HSD costs across provinces, with important implications for food poverty and social inequality. This work contributes to the broader discussion on the economic accessibility of sustainable diets and offers a methodological framework for estimating food costs using publicly available data. The findings have practical relevance for policymakers seeking to enhance access to nutritious and environmentally sustainable food options.

Keywords Food prices · Sustainable diet · Food affordability · Web scraping

✉ Ilaria Benedetti
i.benedetti@unitus.it

Stefano Marchetti
stefano.marchetti@unipi.it

Haoran Yang
haoran.y@outlook.com

Mathias Silva Vazquez
mathias.silva.vazquez@uniroma2.it

[1] Department of Economics, Engineering, Society and Business Organization, University of Tuscia, Via del Paradiso, 47, 01100 Viterbo, VT, Italy

[2] Department of Economics and Management, University of Pisa, Via Ridolfi n.10, 56124 Pisa, PI, Italy

[3] Department of Economics and Finance, Tor Vergata University of Rome, Via Columbia, 2, 00133 Rome, RM, Italy

Published online: 07 November 2025

1 Introduction

During the last decades the issue of food poverty, defined as "the inability to acquire or consume an adequate quality or sufficient quantity of food in socially acceptable ways, or the uncertainty that you will be able to do so" (Dowler et al. 2001), has gained increasing attention, particularly in the context of rising food prices and economic instability. The escalating cost of food has exacerbated the risk of food poverty, disproportionately affecting the most vulnerable groups. Food price inflation imposes a significant economic burden on low-income households, restricting access to essential nutrition and further widening health and social disparities (Sanderson Bellamy et al. 2021). Recognizing these challenges, the United Nations' Sustainable Development Goals (SDGs), particularly SDG 2, emphasize the importance of ensuring universal access to adequate and nutritious food by 2030. Food insecurity, commonly measured through the prevalence of undernourishment (SDG indicator 2.1.1), remains a persistent issue even in high-income countries, including those within the European Union (Beacom et al. 2021; Marchetti and Secondi 2022; Penne and Goedemé 2021). The ability to afford nutritious food is particularly crucial for vulnerable populations, as economic constraints often limit dietary choices (Principato et al. 2022; Schneider et al. 2023). Households allocate a considerable proportion of their budgets to food, and this share tends to increase for lower-income families, who dedicate a higher percentage of their expenditures to essential goods. Within the European Union, food and non-alcoholic beverages account for an average of 14.3% of household expenditure, although this proportion varies significantly across Member States, ranging from 8.3% in Iceland to 24.8% in Romania, reflecting substantial differences in income levels and consumption patterns (Adam et al. 2022; Zhu et al. 2022). In Italy, household spending on food and non-alcoholic beverages represents approximately 18.4% of total consumption, underscoring the central role of food affordability in economic well-being (ISTAT 2023). In light of these pressing challenges and the pivotal role of dietary affordability in shaping public health and social equity, a healthy and sustainable diet (HSD) is defined as a dietary model that simultaneously promotes human health and well-being while minimizing environmental impact. This approach prioritizes foods that are both nutritionally beneficial and sustainably produced, characterized by low greenhouse gas emissions, efficient land and water use, and biodiversity conservation (Principato et al. 2025). However, the economic feasibility of adopting such diets remains a crucial issue, particularly for lower-income households, whose food choices are more constrained by price fluctuations. Against this backdrop, this study examines the affordability of HSD in Italy, defined according to the nutritional recommendations from the Italian CREA CREA (2018) Guidelines and the LARN (Livelli di Assunzione di Riferimento di Nutrienti ed energia). Importantly, the paper adopts an innovative approach by leveraging web-scraped price data and applying spatial-temporal imputation techniques, offering a replicable methodological contribution that extends beyond the Italian case. Price data was collected through web scraping techniques from the Italian Osservatorio Prezzi e Tariffe, which is carried out by the Ministry of Enterprises and Made in Italy. This data allowed for the estimation of the minimum, average, and maximum costs of the proposed diets, providing a comprehensive understanding of their affordability across different socio-economic groups.

A key challenge in compiling the dataset was handling missing values, which arose either due to the absence of certain products in specific provinces or total lack of data in some

provinces in the Osservatorio Prezzi e Tariffe data. To address this issue, a comprehensive strategy for data imputation was employed. Specifically, missing values were assessed in relation to seasonal trends, spatial correlations between neighboring provinces, and historical price patterns. This approach ensured that the dataset remained as complete as possible, minimizing gaps that could otherwise compromise the robustness of subsequent analyses. The integration of spatial and temporal interpolation methods further strengthened the dataset's reliability, allowing for more precise estimates in cases where direct price observations were unavailable.

By focusing on the economic dimension of food poverty, this study aims to contribute to the development of replicable methodologies for evaluating the cost of a healthy and sustainable dietary pattern. Specifically, it aims to quantify the financial cost of HSD for different groups of individuals (adult male and female, elderly people, children and adolescents) with a focus on the seasonality of products. Additionally, the study monitors cost trends in the Italian context, identifying regional disparities that may indicate inequalities in access to healthy and sustainable dietary pattern. Finally, it assesses whether price increases for healthy and sustainable dietary pattern are evenly distributed across all price ranges (minimum, average and maximum). By considering the economic dimensions of food poverty in term of sustainable and healthy dietary pattern, this research contributes to the broader discourse on mitigating food insecurity while promoting dietary choices that support both public health and environmental sustainability.

The structured dataset obtained through web scraping provides a rich empirical foundation for investigating food price dynamics in Italy. By leveraging advanced data extraction techniques, this study ensures that publicly available information is systematically collected, standardized, and made accessible for rigorous economic analysis. The resulting dataset not only supports research on food poverty and sustainability but also serves as a benchmark for future studies seeking to understand the intersection of economic constraints and dietary choices in a rapidly changing food landscape.

The remainder of this paper is structured as follows: Sect. 2 outlines the data, with an overview of prices dataset and missing values. Section 3 reports the methodological approach followed for the treatment of missing values. Section 4 presents the results, focusing on the economic feasibility of the proposed diets. Finally, Sect. 5 concludes with recommendations for future research and policy interventions aimed at reducing food poverty through the promotion of HSD. Moreover, this section discusses the implications of these findings for policy and practice. By addressing these dimensions, the study not only sheds light on the affordability of sustainable diets in Italy but also contributes methodologically to the broader discourse on food poverty, offering tools that may inform both academic research and policy design in other contexts.

2 From definition of HSD to the data: the osservatorio prezzi dataset

To assess the costs of HSD for specific groups of people, in this paper we defined weekly meal plans that adhere to evidence-based dietary guidelines and assessed the economic feasibility through a detailed cost analysis using structured price data. This dual focus on health and sustainability guided the development of meal plans that are not only nutritionally adequate but also environmentally sound. The methodological approach used to develop these

meal plans drew on several key sources of evidence. First, nutritional recommendations from the Italian CREA (2018) Guidelines and LARN (Livelli di Assunzione di Riferimento di Nutrienti ed energia) nutrient reference values were used to ensure that the dietary plans met the nutritional needs of different population groups CREA (2018). Secondly, systematic reviews and meta-analyses linking dietary intake to the risk of chronic diseases, in particular cardiovascular disease and type 2 diabetes, were taken into account to maximise the health benefits of the proposed diets (Giosuè et al. 2022; Riccardi et al. 2022). Finally, the meal plans were differentiated by gender and age group to ensure that the specific nutritional needs of different populations were met. The weekly diet model set consumption frequencies for different food groups to maximise health benefits and minimise disease risk. Foods associated with reduced disease risk, such as whole grains, low glycaemic index (GI) carbohydrates, fruits, vegetables, olive oil, nuts, yoghurt, legumes and fish, were included more frequently. Conversely, high-risk foods such as processed meats and high-GI carbohydrates were limited to occasional consumption. This approach is in line with the findings of Principato et al. (2022), which emphasise the importance of dietary patterns that optimise the prevention of cardiovascular disease while mitigating climate change. Unlike most previous studies relying on household survey data, our framework directly integrates high-frequency, province-level price information, allowing for a more granular and timely assessment of food affordability. This represents a key methodological innovation of the study.

To estimate the cost of a HSD in line with nutritional recommendations and recent nutritional evidence in the literature, the dataset used in this study was compiled through a systematic web-scraping approach applied to the Osservatorio Prezzi e Tariffe, an official platform maintained by the Italian authorities that provides detailed information on consumer prices in different provinces and product categories. The dataset covers the period from August 2021 to March 2024, providing a comprehensive time span for analysing price dynamics. The final dataset comprises 326,721 observations, each corresponding to a specific product at a specific time and place. For transparency, the scraping procedure systematically retrieved prices at the minimum, maximum, and average levels for each product, ensuring comparability across provinces and time. The data collection process was designed to ensure the systematic retrieval of price information for multiple products across all Italian provinces. This involved defining input parameters such as year, month, province, and product type, allowing for the automatic generation of all possible combinations to ensure that missing data could be identified and appropriately handled.[1]

The scraping process followed a structured pipeline that involved multiple steps. First, the script dynamically generated URLs corresponding to all product categories for each province and month covered in the dataset. These URLs were accessed programmatically, and their HTML content was parsed to extract relevant tabular data. Specifically, the script targeted the minimum, maximum, and average prices recorded for each product, ensuring that all available price points were captured. Where no data were available for a given combination, missing values were explicitly recorded to maintain data integrity and facilitate subsequent imputation techniques.

[1] Given the complexity of the dataset and the need for structured information, Python 3.12.2 was used for implementing the scraping procedure. Several libraries were employed to facilitate data extraction and processing, including BeautifulSoup for parsing and navigating HTML content, requests for retrieving web pages, and pandas for organizing and manipulating the extracted data. Additionally, tqdm was integrated to track the progress of the extraction process, ensuring efficiency and transparency in data collection.

The dataset includes six main product categories, namely food products, fruits and vegetables, seafood, personal and home care items, energy, and services. However, the web-scraping procedure was specifically applied to the food, fruits and vegetables, and seafood categories, given the research focus on food affordability and sustainable diets. Within these categories, a total of 167 distinct food products were tracked, including staple items such as grains, dairy products, meat, fish, and fresh produce. All products were classified following the COICOP (Classification of Individual Consumption According to Purpose) standard, ensuring international comparability and alignment with official statistical frameworks.

A key challenge in dataset construction was the presence of missing values, stemming from regional differences in food availability and incomplete reporting. These issues are addressed through the imputation strategy described in Sect. 3.

3 Methodology

For each province in Italy, indexed by $j = 1, \ldots, 107$, defining the cost of HSD requires devising a summary retail selling price $p^{r;*}_{i,j,t}$ for each *i*-th item in the basket of food items $i = 1, \ldots, 167$ measured in standardized units at some or all of the $t = 1, \ldots, 39$ monthly periods covered in the data. Let there be presumably $K_{i,j}$ selling points within each province for each item *i* and each with its retail price $p^{r}_{i,j,t,k}$, $k = 1, \ldots, K_{i,j}$ following a distribution denoted $p^{r}_{i,j,t,k} \sim f_{p^{r}_{i,j,t}}$.[2]

From a sample of size $N_{i,j,t}$ of selling points' prices $\{p^{r}_{i,j,t,k}\}_{k=1}^{N_{i,j,t}}$ across 64 provinces and covering the period spanning from January 2021 up to and including March 2024, the prices data is comprised mainly of the sample average price, $\bar{p}_{i,j,t}$, the sample maximum price, $p^{max}_{i,j,t}$, and the sample minimum price $p^{min}_{i,j,t}$, respectively at the province-item-month level.

Our data is therefore of areal-temporal format with units defined as $(p^{min}_{i,j,t}, \bar{p}_{i,j,t}, p^{max}_{i,j,t})$, where *i* indexes food items, *j* indexes provinces, and *t* indexes months covered in the underlying sample.

Missing data limitations hinder the analysis at several levels within this framework. Firstly, within any of the 64 represented provinces it could be the case that non-sampling or non-reporting issues of retail selling prices at the selling point level are non-random. If the presence of these issues is concentrated around specific segments of the corresponding retail price distribution $f_{p^{r}_{i,j,t}}$ then the sample representativeness at the province-item level can be poor. Given the unavailability of the underlying micro-data sample $\{p^{r}_{i,j,t}\}_{k=1}^{N_{i,j,t}}$ little evidence is available on the incidence of this issue in the data.

Secondly, selling price statistics $(p^{min}_{i,j,t}, \bar{p}_{i,j,t}, p^{max}_{i,j,t})$ can be missing altogether for several province-item-month units in the data. In these cases, information on selling prices at other periods in time, other nearby provinces, and other related items' prices contained in the data can be exploited for the purpose of inferring and imputing these missing quantities.

Finally, at the most aggregate level, the unavailability of data on selling prices for any and all items for all provinces not included in the sample hinders the definition of the required

[2] For simplicity, it is assumed that the total number of selling points $K_{i,j}$ for each and any item *i* in each and any province *j* is time-constant.

basket cost of HSD in these provinces. One possibility in overcoming this limitation is to impute missing prices for any given item as an extrapolation based on the spatio-temporal autocorrelation of prices within the observed data.

As the task of extrapolating prices for provinces beyond the 64 covered in the sample can be facilitated after dealing with missing price issues within the observed sample, devising an imputation strategy for the latter is prioritized firstly.

A first main characteristic of the available sample of prices data is the strong heterogeneity in the incidence of missing observations across items, provinces, and periods. Roughly 50% of the 167 items covered in the data have missing values for more than half of the province-month units in the panel. Additionally, price information is missing for about 80 items on average across all province-month units, with the most complete observation missing price information for 45 items and the 25% of least complete observations missing information for at least 90 items.

To reduce the incidence of missing prices while still allowing for the definition of HSD varying across seasons, information can be aggregated across time periods into seasonal averages, minimum and maximum prices. Under this aggregation time variation in the data is reduced to variation across the $s = 1, \ldots, 12$ seasons covered in the sample. Each seasonal average price $\bar{p}_{i,j,s}$ is then computed at the item-province level as the average observed mean price across any and all values $\bar{p}_{i,j,t}$ for months corresponding to the s-th season. The corresponding overall minimum and maximum prices, denoted $p_{i,j}^{min}$ and $p_{i,j}^{max}$ respectively, are computed as the lowest minimum and highest maximum observed prices for the item across all months and seasons in the data. After aggregation at the seasons level, the resulting prices panel spans across 768 province-season units and the incidence of missing values is reduced by about 4 percentage points on average compared to those computed on the monthly panel.

3.1 Imputation method for observed provinces

As all variables in the data are strictly positive and continuous, imputing plausible prices for missing observations in the prices dataset can be done informatively by exploiting sample correlations in items' prices across items, provinces, and seasons. A first possible approach to study all correlations across prices given the large number of items covered in the panel is to exploit Principal Components Analysis (PCA) to identify the main latent axes of variation across observations in the data. These axes may then be taken as a fitted model from which to predict values for the missing observations (e.g., Josse et al. 2011). However, given the high heterogeneity in the incidence of missing values across items, provinces, and seasons, a method allowing for analysing correlations across variables observed over different sample sizes is required for this purpose.

One possible Principal Components method feasible in this setting is that of non-linear estimation of these components by iterative partial least squares (NIPALS) (Wold 1966) as in Martens and Martens (2001, p. 381). The `nipals R` package (Wright 2024) provides efficient implementations of NIPALS-like algorithms for PCA which allow for different number of missing observations across variables in the dataset and for producing single imputations of these from the estimated Principal Components.

For PCA applications it is important to scale all variables to a common range to properly identify correlations. In this case, as only seasonal average prices can variate across time

for any item and province, a feasible scaling is that of min-max normalization of seasonal average prices. Denoted $\tilde{\bar{p}}_{i,j,s}$, the normalized seasonal average price for item i in province j at season s is computed as:

$$\tilde{\bar{p}}_{i,j,s} = \begin{cases} \frac{\bar{p}_{i,j,s} - p_{i,j}^{min}}{p_{i,j}^{max} - p_{i,j}^{min}}, & \text{if } p_{i,j}^{max} \neq p_{i,j}^{min} \\ \frac{\bar{p}_{i,j,s} - p_{i,j}^{min}}{p_{i,j}^{min}}, & \text{if } p_{i,j}^{min} = p_{i,j}^{max} \end{cases}$$

and is defined on the [0, 1] interval.

Although NIPALS methods can handle a large number of scalar variables, they can be sensitive to a large incidence of missing values. To avoid this possible limitation, a sequential imputation scheme can be devised to first exploit information on price variables with a lower incidence of missing values for producing imputations and then exploit their resulting imputed versions in fitting a predictor for further price variables with higher incidence of missing values. This logic of sequential imputations of different subsets of variables at a time belongs to the class of Multivariate Imputation by Chained Equations (MICE) methods and has been explored in the context of PCA methods in, e.g., Costantini et al. (2024).

Once the normalized seasonal average price variables have been completed through a single imputation from the NIPALS algorithm, these must be re-scaled into the corresponding range for each item's prices. This is straightforward for item-province units with observed $p_{i,j}^{min}$ and $p_{i,j}^{max}$ prices, but requires an additional imputation for those units lacking information on the item's price in all seasons. Aside from the constraint imposed by the definition of the price variables, $p_{i,j}^{min} \leq \bar{p}_{i,j,s} \leq p_{i,j}^{max}$, the relationships across minimum and maximum prices of different items or provinces can follow highly complex processes. In the interest of a non-parametric imputation method for these price range variables, Random Forests can be exploited (e.g., see Tang and Ishwaran 2017).

As implemented in the `randomForest R` package (Liaw and Wiener 2002), a Random Forest can be trained as a predictor of the values of any given item's prices $\tilde{\bar{p}}_{i,j,s}$ allowing for missing values only in the covariates used as predictors. This is first overcome by imputing any and all missing observations by the sample mean of the corresponding variable and fitting a Random Forest on the imputed data. A proximity measure can then be devised to quantify similarities across any and all pairs of observations $(\tilde{\bar{p}}_{i,j,s}, \tilde{\bar{p}}_{i,j',s'})$, $j \neq j'$, $s \neq s'$ in the data in terms of their predictability by the Random Forest. Finally, a single imputation for $p_{i,j}^{min}$ and $p_{i,j}^{max}$ is obtained as a proximity-weighted average of all observed values for these variables in the data. The procedure can also be applied iteratively by taking the resulting imputed dataset as the initial step to fit the Random Forest.

An important property of the Random Forest imputation approach in this setting is that by producing imputations for a given item's price range bounds as a weighted average of observed values, no imputed value will be placed outside the sample range for these variables. Additionally, the $p_{i,j}^{min}$ and $p_{i,j}^{max}$ bounds for the i-th item prices are imputed from the last Random Forest fitted to its corresponding normalized seasonal price variable $\tilde{\bar{p}}_{i,j,s}$ which requires fitting as many Random Forests as iterations run for each item for which prices should be imputed. This can introduce important computational time costs to this approach. The `rfImpute` implementation within the `randomForest R` package provides an efficient algorithm for these computations.

Figure 1 provides a schematic overview of this imputation procedure, summarizing the sequential steps from data preprocessing and NIPALS PCA to Random Forest imputation and the construction of the completed seasonal panel. "Appendix 1.1" provides further details on the implementation of this Random Forest imputation strategy alongside the NIPALS step. At the end of this first imputation process we obtain minimum, average and maximum price for each item and each season for the 64 observed provinces.

In this step we adopt a single imputation strategy rather than multiple imputation. The objective is to construct a complete dataset for the surveyed provinces that can serve as the empirical basis for subsequent extrapolation to the unsurveyed provinces. Since the surveyed provinces provide both current and historical price information as well as relevant covariates (such as neighboring prices, income, and season), a single imputation is sufficient to recover a consistent and coherent dataset. In contrast, the unsurveyed provinces lack both current and historical prices, so the same imputation logic cannot be directly applied and requires a different extrapolation strategy. For our purposes, single imputation ensures stable point estimates with considerably lower computational costs, while multiple imputation would not materially alter the results relevant to the construction of this baseline dataset. The next step is thus to impute prices in the unobserved provinces.

3.2 A two-stage imputation method for unobserved provinces

We have imputed any and all missing observations of food product prices for the surveyed 64 provinces across the 12 survey seasons. To predict food prices in the 43 unsurveyed provinces, we need to consider several key factors that influence food prices, as well as the

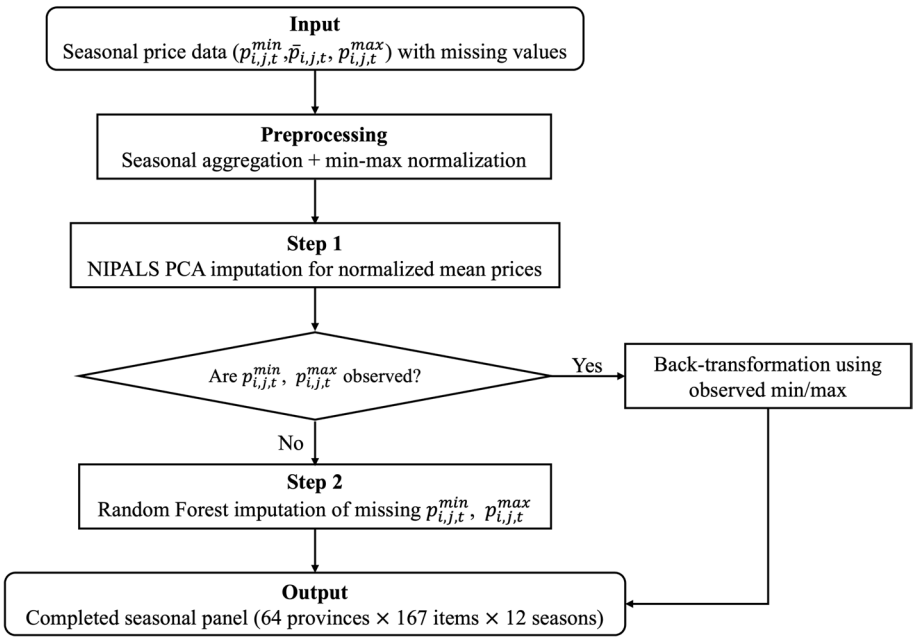

Fig. 1 Flowchart of the imputation strategy for missing prices in surveyed provinces using NIPALS PCA and random forests

variables available in the dataset. First, we assume that there are no structural price differences between the surveyed and unsurveyed provinces because we assume the selection of provinces for the survey was random. This helps mitigate any bias that might affect the imputation process.

Secondly, however, a challenge arises from the fact that we do not have any historical price data for the unsurveyed provinces. This absence of data makes it impossible to use time-series methods to impute missing values, even though economists usually assume that prices are time-series dependent. Instead, we turn to taking the advantage of the available geographic information in neighboring provinces of each unsurveyed province. Given the randomness in the survey design, we can also infer that there should be no systematic price differences between unsurveyed provinces and their adjacent counterparts. Furthermore, neighboring provinces often share similar socioeconomic characteristics and have close economic interactions through land transport, which typically results in stronger correlations in food products' prices.

Another important factor in predicting food prices is the relationship between local income levels and local prices (Bekkers et al. 2017; French et al. 2010). Typically, a province's residents' income levels are strongly correlated with the price levels of goods in that province.

Moreover, food prices are also affected by the characteristics of the food products themselves, such as the conditions required for crop cultivation or livestock farming, transportation, storage, and packaging, also play significant roles in determining their prices.

Finally, considering the special timespan of our dataset from 2021 to 2024, during which some significant regional and worldwide events, such as the Covid-19 pandemic and wars in middle east and eastern Europe, happened, we also take the effects of years into account.

Including all of these elements, we aim to construct a predictive model for imputing the minimum, mean, and maximum prices, respectively, using correspondingly the average of minimum, mean, and maximum prices of each food product in neighboring provinces as a primary factor, while also gradually introducing additional variables. Specifically, we include the average (per capita) after-tax annual income of residents in each province (obtained from the Italian Tax Agency), which serves as a proxy for the local income levels, as well as the season of survey wave to account for seasonal price fluctuations. Our regression models also consider product-specific characteristics and the effects of the survey year.

For imputing food prices in unsurveyed provinces, we construct five regression models. Model 1 is a baseline OLS regression, where the price of a food product is explained by the average price in neighboring provinces, local residents' after-tax income, and the survey season. The specification of Model 1 is:

$$p_{i,j,t} = \beta_0 + \beta_1 p_{i,-j,t} + \beta_2 INC_{p,t} + \beta_3 Season_t + \varepsilon_{i,j,t} \tag{1}$$

Model 2 adjusts for differences in variable magnitudes to avoid unstable estimates. Model 3 extends Model 1 by introducing product-level random effects to capture unobserved product-specific characteristics. Model 4 further incorporates year fixed effects to control for survey-year variations. Finally, Model 5, after confirming the existence of product-specific effects, estimates separate regressions for each food product and uses them for the final imputations. The detailed procedure of model construction and selection is reported in the "Appendix 1", and the final regression specification is:

$$p_{i,j,t} = \beta_0 + \beta_1 p_{i,-j,t} + \beta_2 INC_{p,t} + \beta_3 Season_t + \gamma_t + \varepsilon_{i,j,t} \quad (2)$$

where $p_{i,j,t}$ denotes the price of product *i* in province *j* at time *t*, $p_{i,-j,t}$ is the average price in neighboring provinces, $INC_{p,t}$ is the average after-tax annual income of local residents, $Season_t$ is the survey season, and γ_t is the year fixed effect.

A detailed discussion of the rationale behind constructing each regression model, as well as the full mathematical specifications of Models 1 to 5, is provided in "Appendix 1.2", while a concise summary and comparison of the models is presented in Table 1.

From Model 1 to Model 5, we use observations from the surveyed provinces as our sample. The imputed prices for each food product in the unsurveyed provinces are obtained by multiplying the estimated coefficients by the corresponding variables of those provinces and summing the results. Since the procedure relies heavily on neighboring provinces' average prices, a complication arises for two out of the 43 unsurveyed provinces, which have no surveyed neighbors. For these provinces, we implement a second-round imputation based on the predicted prices of their neighboring provinces.

The calculation of the basket costs for maintaining HSD depends on successfully imputing the missing prices for both surveyed and unsurveyed provinces. We calculate the minimum, average, and maximum basket cost, respectively, for each population profile for each semester (spring-summer and autumn-winter), with the cost calculated on a monthly basis. For each food category and each survey season in a province, we calculate the mean values of the minimum, average, and maximum prices across all food products in the category. These mean values represent the minimum ($p^{min}_{g,j,t}$), average ($\bar{p}_{g,j,t}$), and maximum ($p^{max}_{g,j,t}$) prices for the category *g* in that survey season and province. Therefore, the representative prices of each food category can be expressed as:

$$p^{min}_{g,j,t} = \frac{1}{n_g} \sum_{i \in g} p^{min}_{i,j,t}$$

$$\bar{p}_{g,j,t} = \frac{1}{n_g} \sum_{i \in g} \bar{p}_{i,j,t}$$

$$p^{max}_{g,j,t} = \frac{1}{n_g} \sum_{i \in g} p^{max}_{i,j,t}$$

Table 1 Summary and comparison of imputation models for unobserved provinces

Model	Equation	Description
1	Eq. (1)	Baseline OLS regression including neighboring provinces' average price, residents' after-tax income, and survey season
2	Eq. (1) (rescaled)	Same specification as Model 1, but rescaled to address differences in magnitudes between income and food prices, improving numerical stability
3	Eq. (3) (Appendix)	Extends Model 1 by adding product-level random effects (μ_i) to capture unobserved heterogeneity across food products
4	Eq. (4) (Appendix)	Builds on Model 3 by further including year fixed effects (γ_t) to control for survey-year variations
5	Eq. (2)	After confirming the presence of random effects, separate regressions are estimated for each product, allowing product-specific price determinants to be captured more accurately

where n_g is the number of food product types belonging to category g. Then, according to the HSD suggested by nutritionists, we respectively multiply the representative minimum, average, and maximum prices of each category by its corresponding suggested monthly consumption amount ($Q_{\lambda,t}$) and get corresponding cost of this category for the population profile λ. Summing up the monthly cost of each category, we obtain the minimum ($C^{min}_{\lambda,j,t}$), average ($\bar{C}_{\lambda,j,t}$), and maximum ($C^{max}_{\lambda,j,t}$) monthly basket cost of maintaining HSD for the population profile λ in the corresponding season t in province j, which can be expressed as:

$$C^{min}_{\lambda,j,t} = \sum_j p^{min}_{g,j,t} \times Q_{\lambda,t}$$

$$\bar{C}_{\lambda,j,t} = \sum_j \bar{p}_{g,j,t} \times Q_{\lambda,t}$$

$$C^{max}_{\lambda,j,t} = \sum_j p^{max}_{g,j,t} \times Q_{\lambda,t}$$

4 Results

This section presents the results in two parts. In the first part, we report the regression results (obtained by adopting the methodology explained in Sect. 3) used to determine the optimal model for imputing missing prices in unobserved provinces, and provide the summary statistics for each food category across the entire dataset (including both original and imputed data). In the second part, we present the basket costs, discussing its trend over time at the national level and its geographical distribution.

4.1 Results of regression and imputation of missing prices

Based on complete price data for all observed provinces, we further imputed the prices of unobserved provinces. To achieve this, we first tested five groups of regression models to determine the correlations between the prices of each food product and its potential influencing factors. Then, using these factors, combined with the estimated coefficients, we predict these missing prices. Therefore, we present the results for each of the two steps.

Table 2 presents the regression results for Models 1 through 4, based on the group using average price data. The regression results for the same models, based on the minimum and maximum price data, are presented in the Appendix Tables 3 and 4, respectively.

The first column in Table 2 shows a strong correlation between the average price of food products in neighboring provinces and the price level in the current province, with a high goodness of fit ($R^2 = 0.968$). After rescaling the magnitudes of the food product prices in neighboring provinces and the average income after tax in the current province, the results from Model 2 (second column) indicate identical explanatory power as Model 1. This suggests that disparities in variable magnitudes do not affect the regression results, and we therefore retain the original scale of the data in subsequent analyses.

The third column reports the results of Model 3, which introduces a random-effect term for food products. The coefficient for *Neighbor Price* decreases from 0.989 to 0.523, showing that both average income and the product-level random effect also significantly contrib-

Table 2 Regression results for data with mean prices

	OLS regressions		Mixed effect regressions	
	Model 1	Model 2	Model 3	Model 4
Neighbor price	0.989***	6.763***	0.523***	0.498***
	(0.001)	(0.004)	(0.003)	(0.004)
Average income	0.000***	0.012***	0.000***	0.000***
	(0.000)	(0.004)	(0.000)	(0.000)
Season	0.000	0.000	0.001	0.000
	(0.007)	(0.007)	(0.007)	(0.007)
Year_2022	–	–	–	0.102***
	–	–	–	(0.010)
Year_2023	–	–	–	0.258***
	–	–	–	(0.010)
Year_2024	–	–	–	0.289***
	–	–	–	(0.012)
Constant	0.020	4.945***	1.661***	1.711***
	(0.029)	(0.005)	(0.252)	(0.264)
N	116,228	116,228	116,228	116,228
R-squared	0.968	0.968	–	–
Adjusted R-squared	0.968	0.968	–	–
REML	–	–	360,719.9	359,790.8
Var. (product)	–	–	10.401	11.518
Var. (residual)	–	–	1.288	1.277
Std. Dev. (product)	–	–	3.225	3.394
Std. Dev. (residual)	–	–	1.135	1.130
Scaled	No	Yes	No	No
Product RE	No	No	Yes	Yes
Year FE	No	No	No	Yes

(1) Standard errors are shown in brackets. (2) Model 5 is not reported as a single regression equation because it involves estimating separate regressions for each food product. After confirming the presence of product-specific effects, we perform product-level regressions and use the resulting coefficients to impute missing prices. For brevity, these individual specifications are not listed here but are fully documented in "Appendix 1.2"

ute to price variation. The variance component of the random effect is 10.4 (SD = 3.2), far exceeding the residual variance of 1.3, suggesting that most of the variability in the data stems from systematic differences between food products rather than random noise.

Model 4, shown in the fourth column, extends Model 3 by including year fixed effects. The estimated coefficients for 2022, 2023, and 2024 (relative to 2021 as the baseline) are all positive and significant, capturing substantial year-to-year price increases. The inclusion of these effects slightly reduces the coefficients for *Neighbor Price* (from 0.523 to 0.498) and average income, indicating that part of their explanatory power overlaps with inter-year variation. The random-effect variance modestly increases, and the REML value decreases slightly, suggesting a marginal but consistent improvement in model fit. Overall, Model 4 provides a more comprehensive specification by controlling for both product- and time-level heterogeneity.

Having established both product-specific and year-specific heterogeneity, Model 5 estimates separate regressions for each food product. Figure 2 illustrates the distribution of p values for the *Neighbor Price* coefficient across these regressions. The histogram shows that *Neighbor Price* is a highly significant predictor for the vast majority of products. For the few cases where it is not significant, other explanatory variables (such as average income

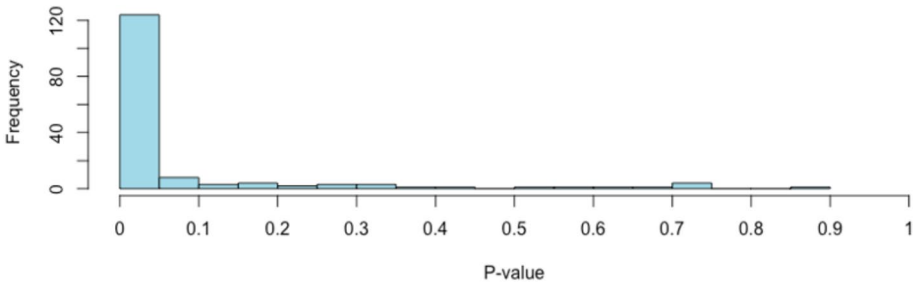

Fig. 2 Distribution of *p* values for the *Neighbor price* coefficient across product-level regressions (Model 5)

and season) still play an important role. Based on these regressions, we use the estimated coefficients to impute missing prices in unsurveyed provinces.

To validate the imputed values obtained from Model 5, we conducted a Monte Carlo hold-out exercise following standard practice in the imputation literature (e.g., Bertsimas et al. 2017). In each repetition, 5% of the observed province–item–season cells were randomly masked, and the imputation algorithm was applied to simultaneously recover both the masked and the originally missing entries. To avoid information leakage, covariates such as the Neighbor Price were recomputed using only the non-masked observations in each repetition.

The diagnostics indicate that the imputed values closely align with the observed truths. Across 20 repetitions, the mean absolute error (MAE) averaged 0.279, the root mean squared error (RMSE) averaged 0.635, and the mean absolute percentage error (MAPE) was approximately 9.7%. The correlation between observed and imputed values remained consistently high, with an average of 0.985. These results confirm that the imputation procedure provides accurate and stable estimates suitable for subsequent extrapolation to the unsurveyed provinces. The low absolute errors, relative error, and near-perfect correlation indicate that the imputation is both reliable and effective for preserving the underlying price patterns.

Finally, while the above discussion is based on regressions using mean prices, similar results are obtained when using minimum- and maximum-price data. Although the significance of control variables may vary across these specifications, the robust correlation between neighboring and local prices consistently emerges, reinforcing the use of *Neighbor Price* as a key predictor for imputations.

4.2 Basket costs: time trends and spatial distribution

Once the HSDs were defined in terms of suggested food-category types and the corresponding amount of each food category, we provided the estimation of the cost of these diets (basket cost) over time and space by considering three scenarios: the minimum level, the average level, and the maximum level, corresponding respectively to the lowest, average, and highest recorded prices for each food item included in the dietary pattern within a given category.

This section presents the dynamics of basket costs across time and space for the three core population profiles under analysis—females, adolescents, and babies. The analysis is

structured along two dimensions: first, the evolution of basket costs over time and their seasonal patterns; second, the geographic variation observed within each survey wave. Emphasis is placed on the female profile, followed by a comparative discussion of the adolescent and baby profiles, highlighting the most salient contrasts.

In the minimum-level basket cost, the most affordable combination of locally available items needed for an overall healthy and sustainable diet at each time and place is calculated (Herforth et al. 2020). This scenario, which replicate the metric developed by FAO, is predicated on the assumption that a household consistently manages to procure food products within each category at their lowest available price during the reference period. This approach enables the flexibility to substitute items within each category, based on the most economical combination of foods that aligns with each definition of diet quality (Mahrt et al. 2019). The objective is to ascertain the lowest cost at which essential nutrients and food groups necessary for each dietary standard can be obtained, thereby identifying the expenditure level required to afford that level of diet quality (Herforth et al. 2020). This reveals food system performance in bringing the required mix of foods within reach of low-income people. Indeed, several studies show that rising food prices have a disproportionate impact on poor households, imposing a greater burden on them than on non-poor households (Abay et al. 2023; Penne and Goedemé 2021; Shabnam et al. 2023).

This minimum-level basket cost serves an important practical function: it defines a theoretical lower bound for the necessary food expenditure under the HSD framework.

Due to space constraints, we present basket cost estimates only for adult women, adolescents, and babies, as these groups are generally considered more vulnerable in the food consumption market (Penne and Goedemé 2021). Basket costs for elders and adult males are presented in appendices. In this context, adult women, adolescents, and infants can be regarded as comparatively disadvantaged groups (Grimaccia and Naccarato 2022; Potsi et al. 2016). Gender is a salient factor in food insecurity, with women experiencing a higher frequency of food insecurity events than men in all regions of Europe (Grimaccia and Naccarato 2022). Adolescents and children, in particular, lack independent economic capacity and decision-making power, making their food choices largely dependent on parents or guardians. Consequently, they do not hold substantial autonomy in the food market. Diets and non-optimal eating habits for children can lead to a reduction in cognitive abilities and poorer school results (Palladino et al. 2024). Adolescents who experience insufficient access to food may encounter a range of adverse emotional responses and develop psychological distress. These responses may stem from concerns regarding their ability to acquire sufficient nourishment, the perceived financial burden on their parents, and the sense of social exclusion from food-centric social interactions (Leung et al. 2020).

The average-level basket cost is calculated based on the mean price of all available products within each food category. It reflects a more typical pattern of food acquisition, where households do not deliberately seek the lowest prices but instead purchase items at average market rates, without strong preferences for specific brands, varieties, or origins.

The maximum-level basket cost, in contrast, is derived by assuming that households purchase the most expensive item in each food category. While this scenario is highly unrealistic from a behavioral standpoint, it provides a meaningful upper bound for food expenditures within the HSD framework.

4.2.1 Temporal variation of basket cost

Figures 3, 4, and 5 respectively present the distributions of basket cost for adult females, adolescents, and babies—the three core demographic profiles under study. Due to space constraints, results for adult males and the elderly are displayed separately in Figs. 9 and 10. In each graph, the basket cost is computed at the minimum, average, and maximum levels for each of the twelve survey seasons. The x-axis uses a year–season format, where seasons are abbreviated as SP (spring), SM (summer), AT (autumn), and WT (winter). For example, the summer of 2021 is denoted as "21 SM". The y-axis represents the monthly estimated basket cost in euros.

For each survey season, we present the basket cost at the minimum, average, and maximum levels in blue, green and red, respectively. For each level in each survey season, the plot consists of three components. From left to right, they are: a scatter plot demonstrating the specific level of basket cost in each Italian province, and the provinces with the highest and lowest basket costs are marked next to the corresponding scatters with the standard two-character abbreviation of province name; a box plot showing the highest-, 75%-, median-, 25%-, and lowest-level of basket cost among all Italian provinces; and a violin plot presenting the kernel-density estimation result.

Across all three main profiles, a consistent seasonal pattern is observed: basket costs are generally higher in spring and summer than in autumn and winter. A notable exception is the baby group, whose basket costs display a reversed seasonal pattern—being higher in colder months. This contrast underscores the role of profile-specific dietary requirements in shaping time-based expenditure dynamics.

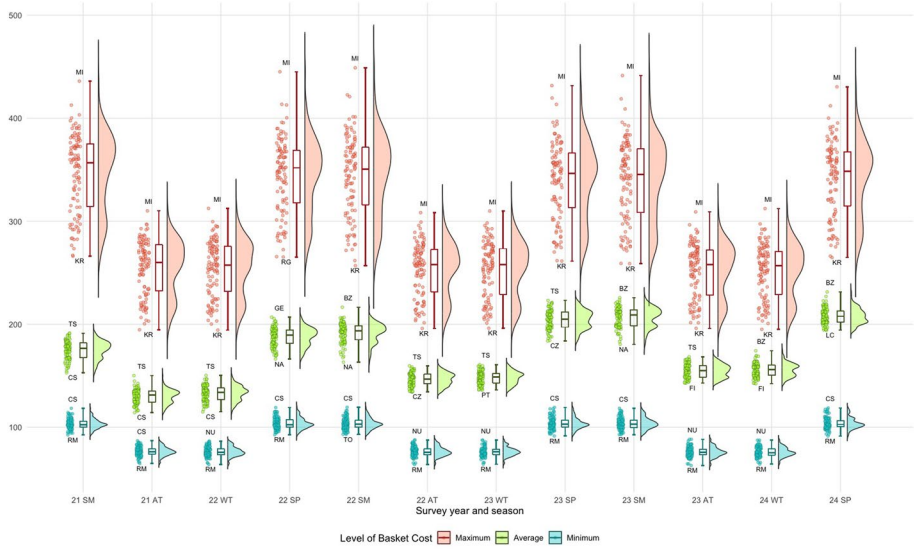

Fig. 3 Quantity and distribution of basket costs for adult females in each survey season at minimum, average, and maximum levels. *Note:* The two-character notations above and below each scatter plot follow the standard abbreviations of Italian provinces, whcih includes BN (for Benevento), BZ (for Bolzano), CS (for Cosenza), CZ (for Catanzaro), GE (for Genoa), KR (for Crotone), LC (for Lecco), MI (for Milan), NU (for Nuoro), RG (for Ragusa), RM (for Rome), TS (for Trieste), and TV (for Treviso)

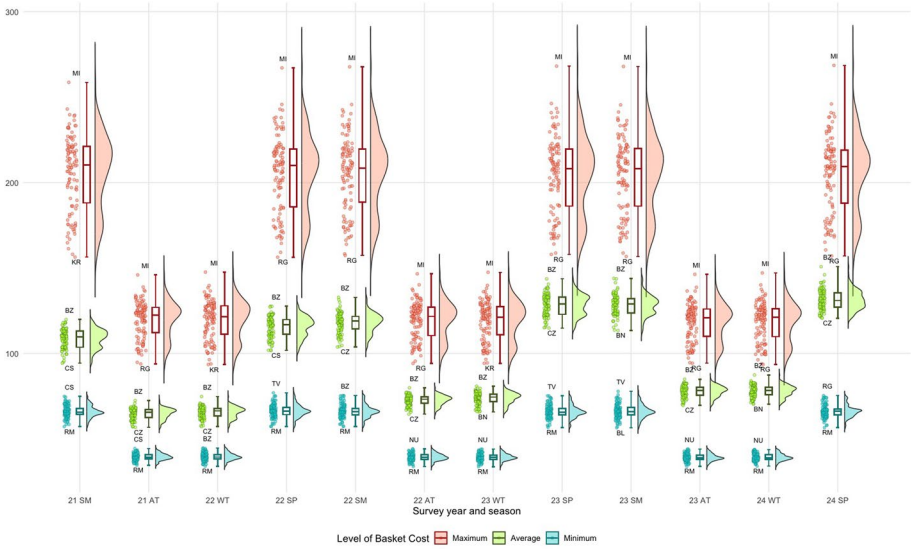

Fig. 4 Quantity and distribution of basket costs for adolescents in each survey season at minimum, average, and maximum levels. *Note:* The two-character notations above and below each scatter plot follow the standard abbreviations of Italian provinces, which include BN (for Benevento), BZ (for Bolzano), CS (for Cosenza), CZ (for Catanzaro), KR (for Crotone), MI (for Milan), NU (for Nuoro), RG (for Ragusa), RM (for Rome), and TV (for Treviso)

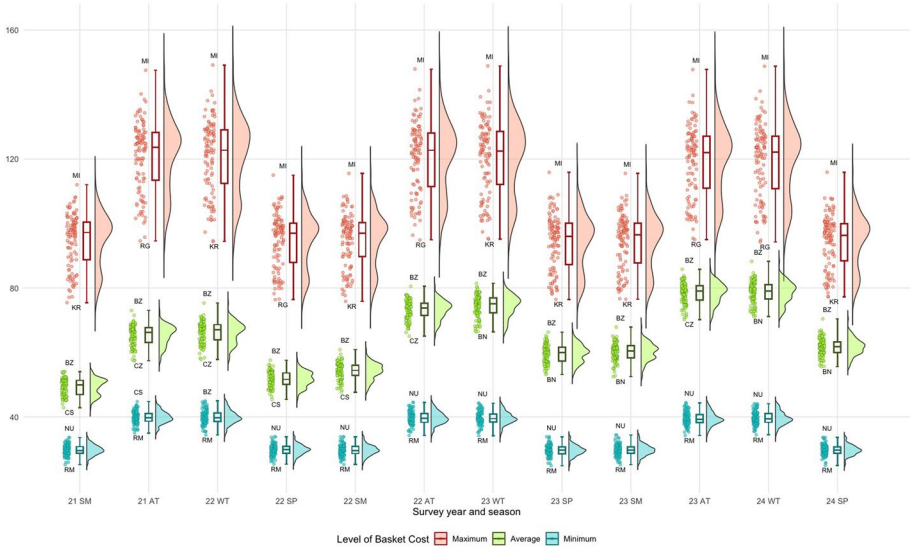

Fig. 5 Quantity and distribution of basket costs for babies in each survey season at minimum, average, and maximum levels. *Note:* The two-character notations above and below each scatter plot follow the standard abbreviations of Italian provinces, which include BN (for Benevento), BZ (for Bolzano), CS (for Cosenza), CZ (for Catanzaro), KR (for Crotone), MI (for Milan), NU (for Nuoro), RG (for Ragusa), and RM (for Rome)

At the minimum-price level, basket costs remain largely stable over time for all three groups. The standard deviation is relatively low—€14.5 for women, €13.7 for adolescents, and only €5.33 for babies—indicating limited temporal volatility. Seasonal means are distinct but consistent: for example, female basket costs average around €76 in spring–summer and €103 in autumn–winter, while adolescent and baby profiles maintain similarly stable patterns within their respective seasonal ranges. These results suggest that the lower bound of food expenditure under the HSD standard is buffered against short-term price shocks, potentially reflecting constraints in pricing strategies for low-end or outlet food items.

At the average-price level, a clear and sustained upward trend is observed across all profiles. From summer 2021 to spring 2024, average basket costs increased from €175.3 to €207.9 for women (an increase of 18.6%), from €108.6 to €131.4 for adolescents (21.0%), and from €49.3 to €61.7 for babies (25.1%) during the warmer seasons. Similar increases are evident in the colder months: for example, female basket costs rose from €130.1 in autumn 2021 to €155.7 in winter 2023 (19.7%), while adolescent and baby costs rose from €64.7 to €78.0 (20.5%) and from €65.3 to €78.6 (20.4%), respectively. The simultaneous rise in both the mean and lower bound of the cost distribution indicates a broad-based shift in mid-range food prices over time, likely reflecting underlying inflationary pressures and changes in supply chain dynamics. Among the three groups, the increase is most pronounced for babies, suggesting a particularly steep escalation in the cost of infant HSD.

At the maximum-price level, basket costs are also stable over time, though with pronounced seasonal contrasts. Female and adolescent profiles show higher basket costs in spring–summer (averaging €253 and €205, respectively), while for babies, the pattern is reversed, with higher costs in autumn–winter (around €120). The standard deviation is highest for women (€55.4), followed by adolescents (€46.5) and babies (€16.2), reflecting differences in the spread of premium-priced items across profiles and seasons. Unlike the average level, maximum-level basket costs appear less responsive to temporal shifts, possibly due to structural stickiness in high-end pricing.

To summarize the analysis of Figs. 3, 4 and 5, the distributional patterns across minimum, average, and maximum price levels reveal three key insights. First, all three levels display clear seasonal variation, with basket costs consistently higher in spring and summer than in autumn and winter, although adjacent seasons show limited within-pair variation. Second, only the average-level basket shows a marked upward trend over time followed by lower-level baskets, whereas maximum-level baskets remain relatively stable throughout the twelve observed seasons. This suggests that recent increases in household food expenditure are primarily concentrated in the mid-price range, while the upper extreme appears less sensitive to temporal shifts. Third, the dispersion of basket costs across provinces varies substantially across levels, with the narrowest range observed at the minimum-price level and the widest at the maximum-price level—suggesting potential structural differences in how food costs are distributed across regions and market segments.

These findings resonate with emerging evidence on inflation inequality and its underlying mechanisms. Goods in the low-to-medium price range exhibit heightened sensitivity to inflationary pressures, a phenomenon attributable to a confluence of supply- and demand-side factors. From a supply perspective, these goods are frequently contingent on fundamental raw materials and global supply chains, rendering them more susceptible to cost variations and interruptions (Faber and Fally 2022). Furthermore, given the lower profit margins characteristic of the retail sector, retailers have limited capacity to absorb cost

increases and tend to pass them on to consumers more quickly (Alvarez et al. 2024). On the demand side, during periods of high inflation, consumers often substitute toward more affordable options, which in turn experience faster price increases than premium goods (Argente and Lee 2021). This pattern has become especially pronounced during the most recent inflationary episode, where post-pandemic dynamics have accentuated disparities in consumption baskets and exposed lower-income households to disproportionate burdens. This inflation inequality—driven by divergent consumption behaviors and store-level price heterogeneity—has emerged as a defining feature of the current economic context, raising pressing concerns for policymakers. It is therefore essential to deepen our understanding of these dynamics, particularly for food categories, as lower-income households spend a larger share of their budgets on food and are more vulnerable to rising prices (Abay et al. 2023). Notably, most of the literature suggests that elevated inflation is typically accompanied by increased price dispersion across stores and product types (Sheremirov et al. 2021), further exacerbating the challenge of maintaining equitable access to essential goods.

The next section investigates the spatial distribution of basket costs across Italian provinces.

4.2.2 Spatial distribution of basket cost

We use choropleth maps to examine the spatial distribution of basket costs across Italian provinces and their evolution over time. Each figure is divided into twelve submaps, corresponding to the twelve survey waves. The submaps are arranged in a 3-by-4 grid: each column represents the same season across different years, while rows display summer, autumn, winter, and spring, respectively. The notation below each subfigure specifies the year and season (SM" for summer, AT" for autumn, WT" for winter, and SP" for spring). For example, "21 SM" refers to summer 2021. To ensure comparability, we adopt the same legend for adult females, adult males, elders, and adolescents, while the baby group has a separate scale reflecting its systematically lower basket costs. Across all three pricing levels, basket costs exhibit clear and systematic spatial variation, revealing structural differences in food affordability. For adult females, the minimum basket cost shows relatively limited dispersion across provinces (generally below €30). Nevertheless, some southern provinces, such as Naples and Cosenza, report the highest minimum costs, whereas northern metropolitan centers (e.g., Turin and Milan) display the lowest (Fig. 6). This pattern indicates that local retail structures, rather than overall price levels, strongly shape the lower bound of affordability.

A similar configuration is observed for adolescents: minimum basket costs remain higher in selected southern provinces, while northern urban areas tend to show lower values. The overall dispersion is limited, suggesting that, although regional inequalities persist, the minimum attainable cost of a sustainable diet is relatively stable across the country (Fig. 7).

For infants, the pattern differs slightly. Southern provinces again record higher minimum costs, but overall levels are substantially lower and less dispersed compared to other demographic groups. This reflects both the narrower composition of the infant dietary basket and the reduced product differentiation in this category (Fig. 8).

Taken together, the results reveal a dual dynamic. While average and maximum basket costs (not shown here) are consistently higher in northern metropolitan areas, minimum costs often peak in the South, where the concentration of large-scale retailers is weaker.

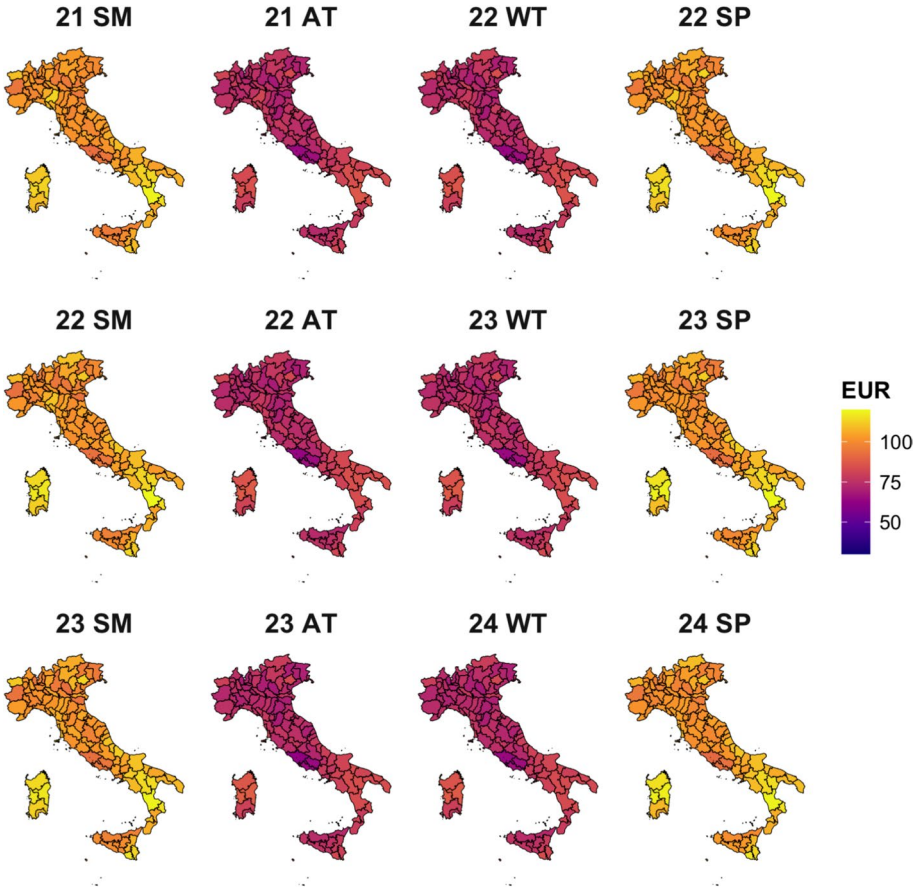

Fig. 6 Minimum basket cost for adult females in each survey period

These spatial patterns are partly explained by the structure of the Italian retail system. Minimum basket costs, which are unexpectedly higher in southern provinces such as Naples and Cosenza while lower in urban centers like Rome and Turin, reflect the uneven distribution of large-scale retail chains (GDO—Grande Distribuzione Organizzata). GDO networks are more densely concentrated in central and northern regions, where competitive intensity, high consumer turnover, and an emphasis on product freshness foster frequent discounting practices and price promotions, particularly for perishable or near-expiry goods. By contrast, in areas where the GDO presence is weaker and local retailers prevail, such mechanisms are less widespread, contributing to persistently higher minimum-level basket costs (Tiberti and Tiberti 2018). Spatial dispersion also varies substantially across basket types. Minimum costs show the narrowest cross-provincial range (typically below €30), while maximum costs can differ by more than €200, pointing to substantial inequality in access to high-end products. Average costs fall in between, with moderate but significant variation, making them a sensitive indicator of regional disparities in standard food access. Overall, these findings suggest that geography plays a critical role in shaping both the lower and upper bounds of affordability under the HSD. While northern regions are generally more

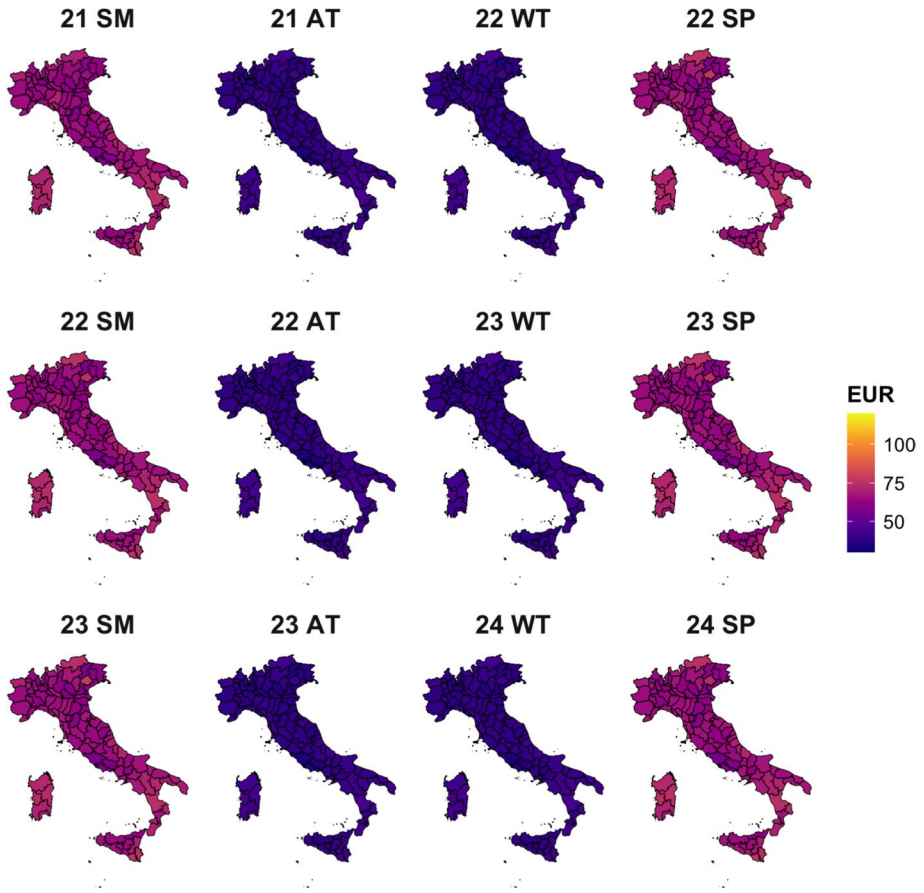

Fig. 7 Minimum basket cost for adolescents in each survey period

expensive at mid- and high-price levels, their large cities often provide more affordable options at the lower end. Conversely, in the South, higher minimum costs—despite lower average prices—may represent a disproportionate burden for price-sensitive households. This divergence highlights the importance of considering regional price structures together with household vulnerability when assessing food access and affordability in Italy.

5 Conclusion

This study provides an in-depth assessment of the affordability of healthy and sustainable diets (HSDs) in Italy, leveraging a novel dataset built through large-scale web scraping from the Osservatorio Prezzi e Tariffe and applying robust spatial-temporal imputation techniques. By estimating the minimum, average, and maximum basket costs across all Italian provinces and survey seasons, the analysis offers key insights into temporal trends, regional disparities, and inflationary impacts on diet affordability for different demographic groups. The results reveal three core findings. First, all basket cost levels exhibit pronounced sea-

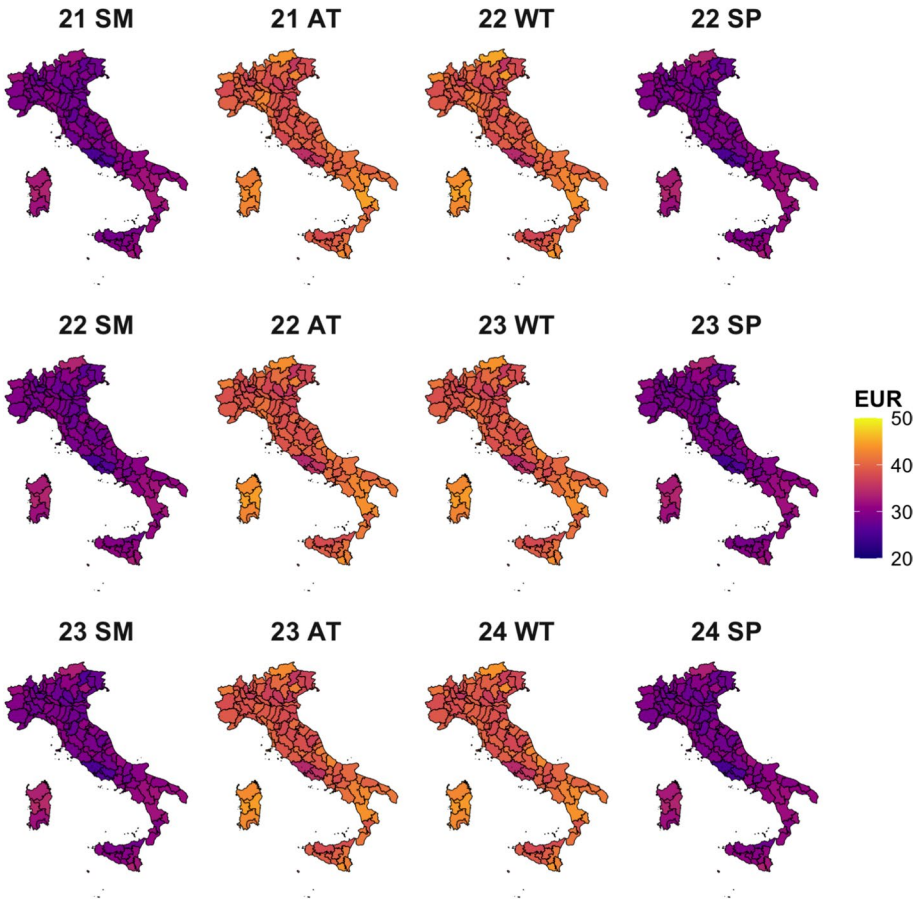

Fig. 8 Minimum basket cost for babies in each survey period

sonal variation, with spring and summer generally associated with higher expenditures—except in the case of infants, whose dietary costs peak during colder months. Second, while minimum and maximum basket costs remain relatively stable over time, a marked and sustained increase is observed in average-level basket costs, particularly for infants, suggesting that food price inflation has predominantly affected the mid-price range. Third, spatial disparities in basket costs reveal a dual logic: average and maximum costs are higher in northern provinces, whereas, unexpectedly, minimum basket costs are often elevated in southern regions. This counterintuitive pattern is linked to the uneven distribution of large-scale retail chains (GDOs), whose stronger presence in the North facilitates greater access to discounted and near-expiry items. These findings have direct implications for public policy. The upward trend in average HSD costs underlines the urgent need for inflation-sensitive policy tools, such as targeted food subsidies or voucher schemes, particularly for nutritionally vulnerable groups like women, adolescents, and infants. At the same time, the observed spatial inequalities highlight the importance of geographically differentiated strategies. In regions with persistently high minimum basket costs and limited GDO presence, support

mechanisms should focus on enhancing local food access infrastructure and encouraging price transparency in retail markets. Finally, the relative stability of minimum basket costs suggests that affordability can be preserved through policies that promote low-cost dietary options and reduce barriers to accessing essential, nutritious foods. Yet, economic affordability alone represents only one of the multiple dimensions shaping dietary practices. A growing body of literature shows how cultural values, social contexts, and behavioural factors strongly affect food choices. For example, Palladino et al. (2024) demonstrate that adolescents' lived experiences of food poverty in Italy are influenced by social stigma, family dynamics, and the role of food aid—illustrating the multidimensional nature of food access. Likewise, Castellini et al. (2020) emphasize that identity, cultural preferences, and sensory cues significantly influence dietary behaviour and sustainability transitions. Evidence from Italy further indicates that intentions and actual adoption of a sustainable diet are mediated by behavioural attitudes, normative beliefs, and perceived control (Biasini et al. 2023). In parallel, Nicholls and Drewnowski (2021) highlight the importance of integrating cultural acceptability and socio-economic equity into sustainable diet frameworks. Altogether, these contributions reinforce the need for integrative policy approaches that combine affordability with cultural, behavioural, and social drivers of diet. In a broader perspective, this study contributes to the growing discourse on inflation inequality and food poverty in high-income countries. The methodology presented offers a scalable approach for monitoring food affordability in real time, supporting more agile and evidence-based policymaking. As Italy and other EU countries strive to meet the targets of the Sustainable Development Goals—particularly SDG 2 on zero hunger—ensuring the economic feasibility of HSDs must remain a central policy priority. Future research should therefore explore behavioural responses to rising food costs and assess the effectiveness of interventions aimed at mitigating both the nutritional and economic consequences of inflation for at-risk populations.

Appendix 1

Appendix 1.1: Imputation algorithm for prices in sampled provinces

Algorithm 1 below describes the single imputation method implemented to obtain a complete imputed dataset of seasonal average prices, and minimum and maximum prices, for all seasons and provinces covered in the sample. In defining the sequential order of the imputations, three subsets of normalized seasonal average prices variables were defined based on the incidence of missing values on these in the dataset. A first subset, denoted $\boldsymbol{X}_1$ contains price variables $\tilde{\bar{p}}_{i,j,s}$ for all items with no more than 50% of missing observations, and constitutes the group of most complete variables. Consequently, a second subset, denoted $\boldsymbol{X}_2$ contains price variables for all items with at least 50% of missing observations but no more than 75%, and a third subset, denoted $\boldsymbol{X}_3$ contains price variables for all items with more

than 75% of observations missing. Their corresponding subsets of price range bounds variables are denoted as $\boldsymbol{X}_1^{bounds}$, $\boldsymbol{X}_2^{bounds}$, and $\boldsymbol{X}_3^{bounds}$ respectively. In the interest of reducing the variance of the imputations while constraining computational times, each Random Forest is fitted with 300 regression trees and for a total of 5 iterations. Finally, 64 province and 12 season dummies are also exploited as additional covariates in the Random Forest.

Algorithm 1 Multivariate sequential imputation using NIPALS and random forests

1: Set $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$.
2: Define $\tilde{\boldsymbol{X}}_1$, $\boldsymbol{X}_2^{(0)}$, and $\boldsymbol{X}_3^{(0)}$ from imputing all missing observations in $\boldsymbol{X}_1$,$\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ using `nipals` fit on $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ altogether.
3: Impute each item's missing bounds in $\boldsymbol{X}_1^{bounds}$ using `rfImpute` to fit Random Forests predicting its $\tilde{\bar{p}}_{i,j,s}$ values using $\tilde{\boldsymbol{X}}_1$, $\boldsymbol{X}_2^{(0)}$, $\boldsymbol{X}_3^{(0)}$.
4: Define $\tilde{\boldsymbol{X}}_2$ from imputing all missing observations in $\boldsymbol{X}_2$ using `nipals` fit on $\tilde{\boldsymbol{X}}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$.
5: Impute each item's missing bounds in $\boldsymbol{X}_2^{bounds}$ using `rfImpute` to fit Random Forests predicting its $\tilde{\bar{p}}_{i,j,s}$ values using $\tilde{\boldsymbol{X}}_1$, $\tilde{\boldsymbol{X}}_2$, and $\boldsymbol{X}_3^{(0)}$.
6: Define $\tilde{\boldsymbol{X}}_3$ from imputing all missing observations in $\boldsymbol{X}_3$ using `nipals` fit on $\tilde{\boldsymbol{X}}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$.
7: Impute each item's missing bounds in $\boldsymbol{X}_3^{bounds}$ using `rfImpute` to fit Random Forests predicting its $\tilde{\bar{p}}_{i,j,s}$ values using $\tilde{\boldsymbol{X}}_1$, $\tilde{\boldsymbol{X}}_2$, and $\tilde{\boldsymbol{X}}_3$.
8: Compute re-scaled seasonal price average variables as $\bar{p}_{i,j,s} = p_{i,j}^{min} + (p_{i,j}^{max} - p_{i,j}^{min}) \times \tilde{\bar{p}}_{i,j,s}$

Appendix 1.2: Construction and selection the predicting model for unsurveyed provinces

This appendix provides the detailed process of constructing and selecting the regression models (Model 1 to Model 5) used for imputing food prices in unsurveyed provinces. While the main text only reports the final specification and summarizes the models, here we document the step-by-step development.

Model 1: baseline OLS

We begin with a baseline ordinary least squares (OLS) model. For food product *i* at time *t*, the price in province *j*, $p_{i,j,t}$, is regressed on the average price among neighboring provinces of *j* ($p_{i,-j,t}$), the average after-tax annual income of local residents ($INC_{p,t}$), and the season of the survey ($Season_t$):

$$p_{i,j,t} = \beta_0 + \beta_1 p_{i,-j,t} + \beta_2 INC_{p,t} + \beta_3 Season_t + \varepsilon_{i,j,t}$$

where $\varepsilon_{i,j,t}$ is the error term.

Model 2: rescaled OLS

In Model 1, variables differ greatly in magnitude. For example, the after-tax annual income is typically in the five-digit range, while food prices are usually in the single-digit range.

Such differences may lead to inflated standard errors, multicollinearity, and unstable estimates. Therefore, in Model 2, we rescale the magnitude of $p_{i,j,t}$ to ensure stable coefficient estimates, as the ultimate goal is to predict exact prices.

Model 3: random effects

On the basis of Model 1, we introduce a product-level random effect to account for unobserved heterogeneity across food products, such as consumer preferences for local products, transportation costs, subsidies, and shelf-life differences:

$$p_{i,j,t} = \beta_0 + \beta_1 p_{i,-j,t} + \beta_2 INC_{p,t} + \beta_3 Season_t + \mu_i + \varepsilon_{i,j,t} \tag{3}$$

where μ_i represents the random effect specific to product *i*.

Model 4: random effects with year fixed effects

To account for year-specific influences, we add survey-year fixed effects γ_t:

$$p_{i,j,t} = \beta_0 + \beta_1 p_{i,-j,t} + \beta_2 INC_{p,t} + \beta_3 Season_t + \mu_i + \gamma_t + \varepsilon_{i,j,t} \tag{4}$$

This model improves coefficient accuracy and allows testing the robustness of the random effects by comparing results with Model 3.

Model 5: product-specific estimation

Once Models 3 and 4 both confirm the existence of random effects, we estimate separate regressions for each food product without the random-effect term. This grouped regression approach allows capturing product-specific price determinants more accurately:

$$p_{i,j,t} = \beta_0 + \beta_1 p_{i,-j,t} + \beta_2 INC_{p,t} + \beta_3 Season_t + \gamma_t + \varepsilon_{i,j,t}$$

Appendix 1.3: Tables

Here we report estimated coefficients for models 1 to 4 when we use minimum (Table 3) and maximum prices (Table 4) as outcome.

Table 3 Regression results for data with minimum prices

	OLS regressions		Mixed effect regressions	
	Model 1	Model 2	Model 3	Model 4
Neighbor price	0.993***	3.772***	0.129***	0.129***
	(0.001)	(0.003)	(0.004)	(0.004)
Average income	− 0.000***	− 0.011***	− 0.000***	− 0.000***
	(0.000)	(0.003)	(0.000)	(0.000)
Season	− 0.000	− 0.000	− 0.001	− 0.001
	(0.005)	(0.005)	(0.004)	(0.004)
Year_2022	–	–	–	0.014***
	–	–	–	(0.007)
Year_2023	–	–	–	0.003
	–	–	–	(0.007)
Year_2024	–	–	–	− 0.002
	–	–	–	(0.008)
Constant	0.130***	2.690***	2.625***	2.614***
	(0.021)	(0.004)	(0.258)	(0.258)
N	116,228	116,228	116,228	116,228
R-squared	0.951	0.951	–	–
Adjusted R-squared	0.951	0.951	–	–
REML	–	–	261,162.2	261,178.7
Var. (product)	–	–	11.040	11.044
Var. (residual)	–	–	0.546	0.546
Std. Dev. (product)	–	–	3.323	3.323
Std. Dev. (residual)	–	–	0.739	0.739
Scaled	No	Yes	No	No
Product RE	No	No	Yes	Yes
Year FE	No	No	No	Yes

(1) Standard errors are shown in brackets. (2) Model 5 is not reported as a single regression equation because it involves estimating separate regressions for each food product. After confirming the presence of product-specific effects, we perform product-level regressions and use the resulting coefficients to impute missing prices. For brevity, these individual specifications are not listed here but are fully documented in "Appendix 1.2"

Table 4 Regression results for Data with Maximum Prices

	OLS regressions		Mixed effect regressions	
	Model 1	Model 2	Model 3	Model 4
Neighbor price	0.983***	8.354***	0.428***	0.421***
	(0.001)	(0.008)	(0.004)	(0.004)
Average income	0.000***	0.131***	0.000***	0.000***
	(0.000)	(0.006)	(0.000)	(0.000)
Season	0.002	0.002	0.007	0.007
	(0.012)	(0.012)	(0.011)	(0.011)
Year_2022	–	–	–	− 0.025***
	–	–	–	(0.017)
Year_2023	–	–	–	− 0.305***
	–	–	–	(0.017)
Year_2024	–	–	–	− 0.320***
	–	–	–	(0.019)
Constant	− 10.6***	8.354***	1.862***	1.957***
	(0.052)	(0.008)	(0.413)	(0.419)
N	116,228	116,228	116,228	116,228
R-squared	0.954	0.954	–	–
Adjusted R-squared	0.954	0.954	–	–
REML	–	–	477,499.6	477,138.4
Var. (product)	–	–	28.112	28.843
Var. (residual)	–	–	3.517	3.505
Std. Dev. (product)	–	–	5.302	5.371
Std. Dev. (residual)	–	–	1.875	1.872
Scaled	No	Yes	No	No
Product RE	No	No	Yes	Yes
Year FE	No	No	No	Yes

(1) Standard errors are shown in brackets. (2) Model 5 is not reported as a single regression equation because it involves estimating separate regressions for each food product. After confirming the presence of product-specific effects, we perform product-level regressions and use the resulting coefficients to impute missing prices. For brevity, these individual specifications are not listed here but are fully documented in "Appendix 1.2"

Appendix 1.4: Figures

See Figs. 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22.

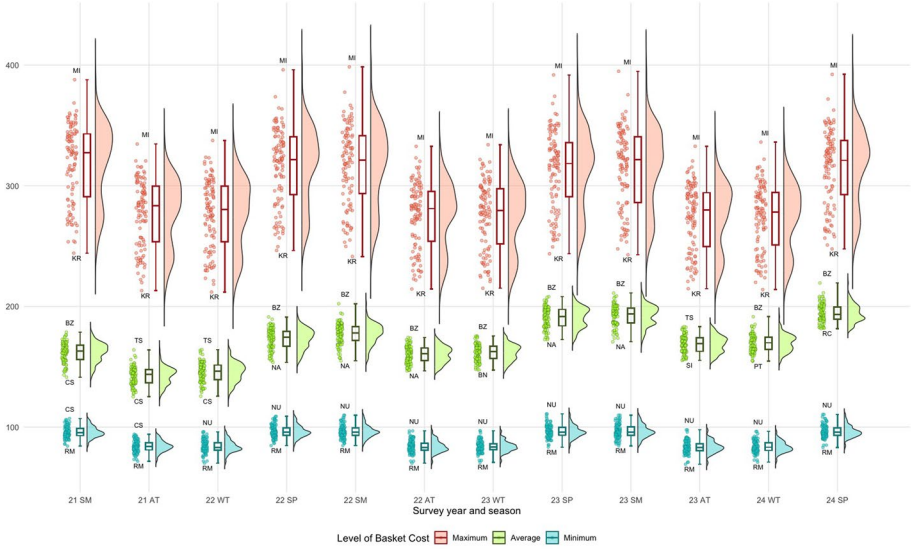

Fig. 9 Quantity and distribution of basket costs for adult males in each survey season at minimum, average, and maximum levels. The two-character notations above and below each scatter plot follow the standard abbreviations of Italian provinces, which include BN (for Benevento), BZ (for Bolzano), CS (for Cosenza), KR (for Crotone), MI (for Milan), NA (for Naples), NU (for Nuoro), PT (for Pistoia), RC (for Reggio Calabria), RM (for Rome), SI (for Siena), and TS (for Trieste)

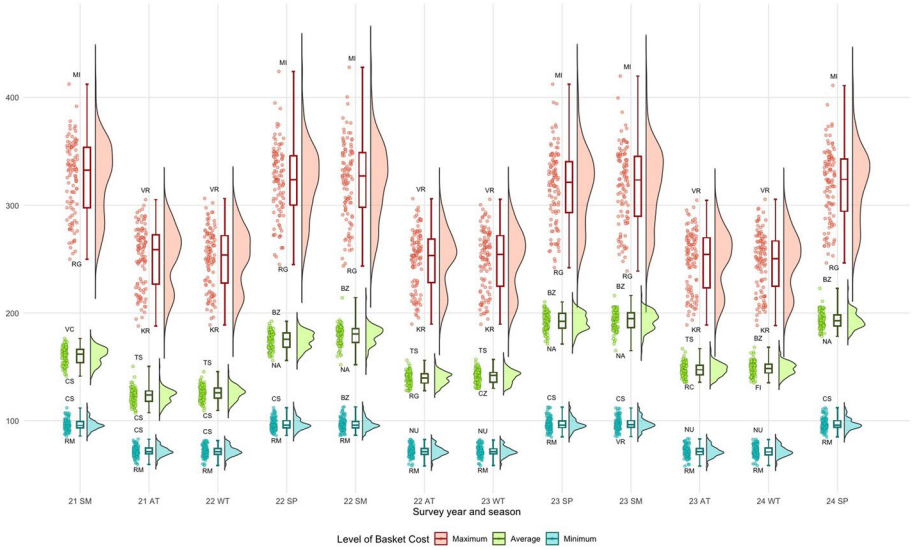

Fig. 10 Quantity and distribution of basket costs for elders in each survey season at minimum, average, and maximum levels. The two-character notations above and below each scatter plot follow the standard abbreviations of Italian provinces, which include BZ (for Bolzano), CS (for Cosenza), CZ (for Catanzaro), FI (for Firenze), KR (for Crotone), MI (for Milan), NA (for Naples), NU (for Nuoro), RC (for Reggio Calabria), RG (for Ragusa), RM (for Rome), TS (for Trieste), and VC (for Vercelli)

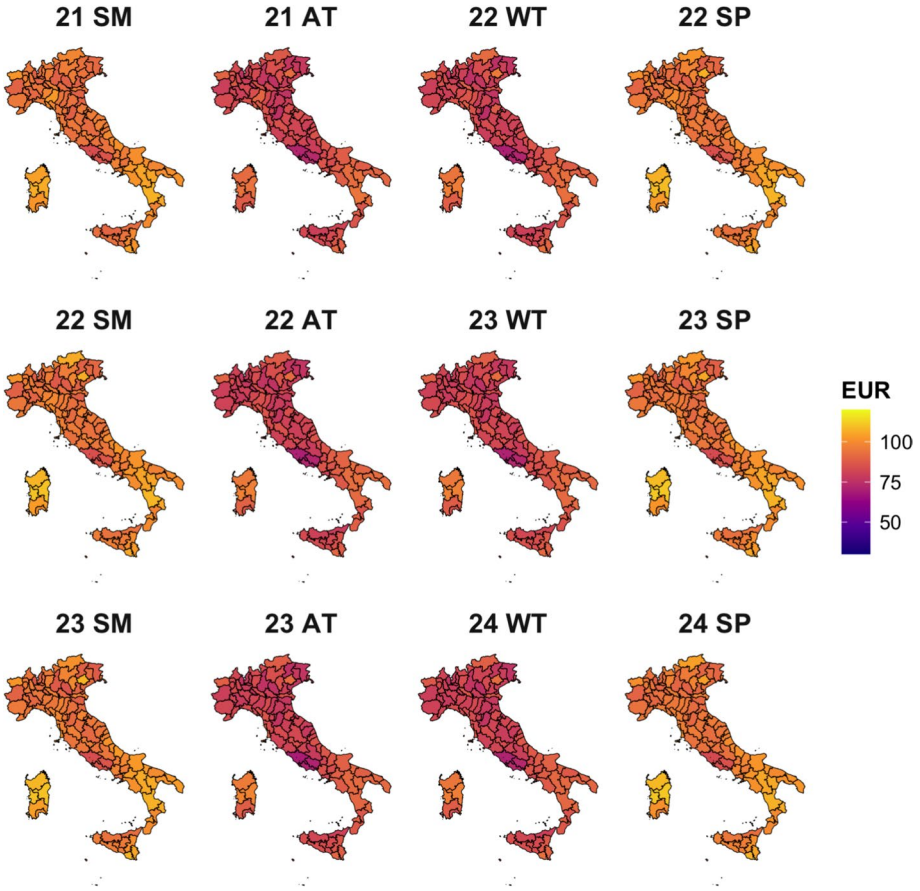

Fig. 11 Minimum basket cost for adult males in each survey period

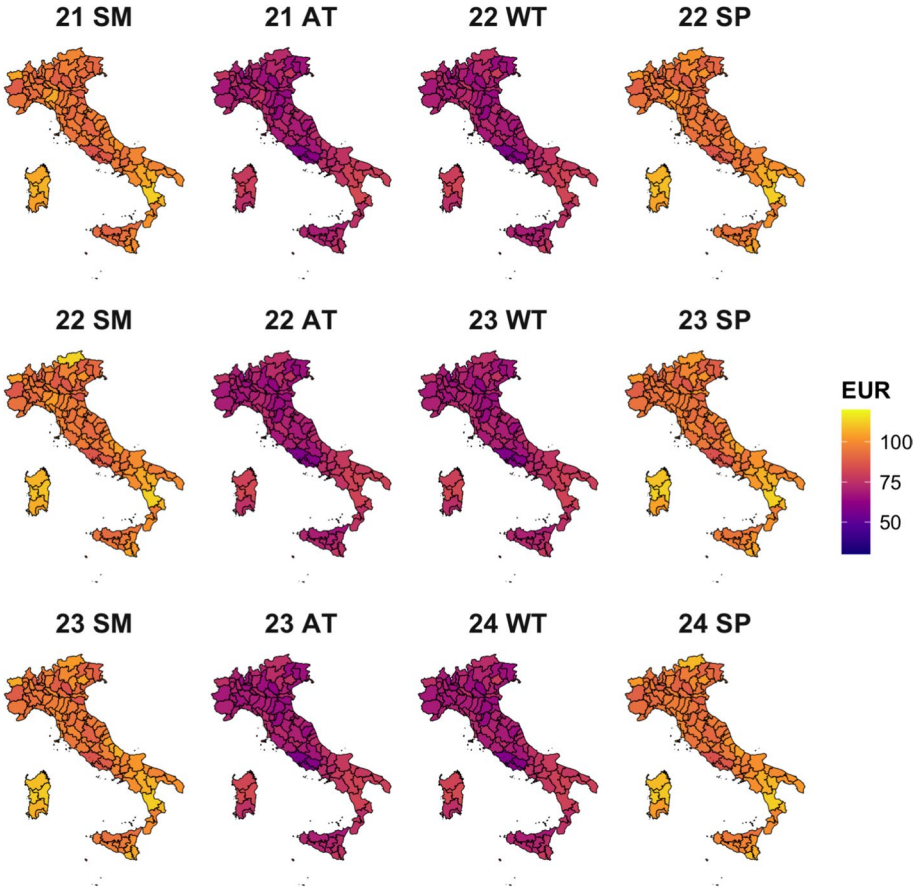

Fig. 12 Minimum basket cost for elders in each survey period

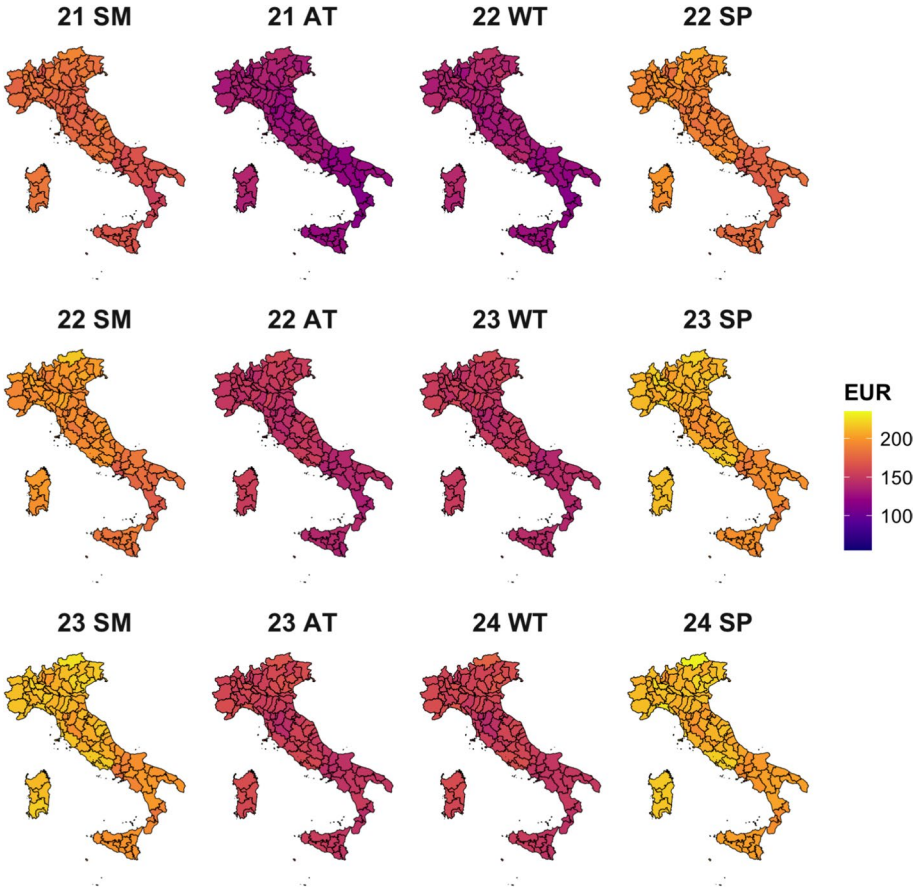

Fig. 13 Average basket cost for adult females in each survey period

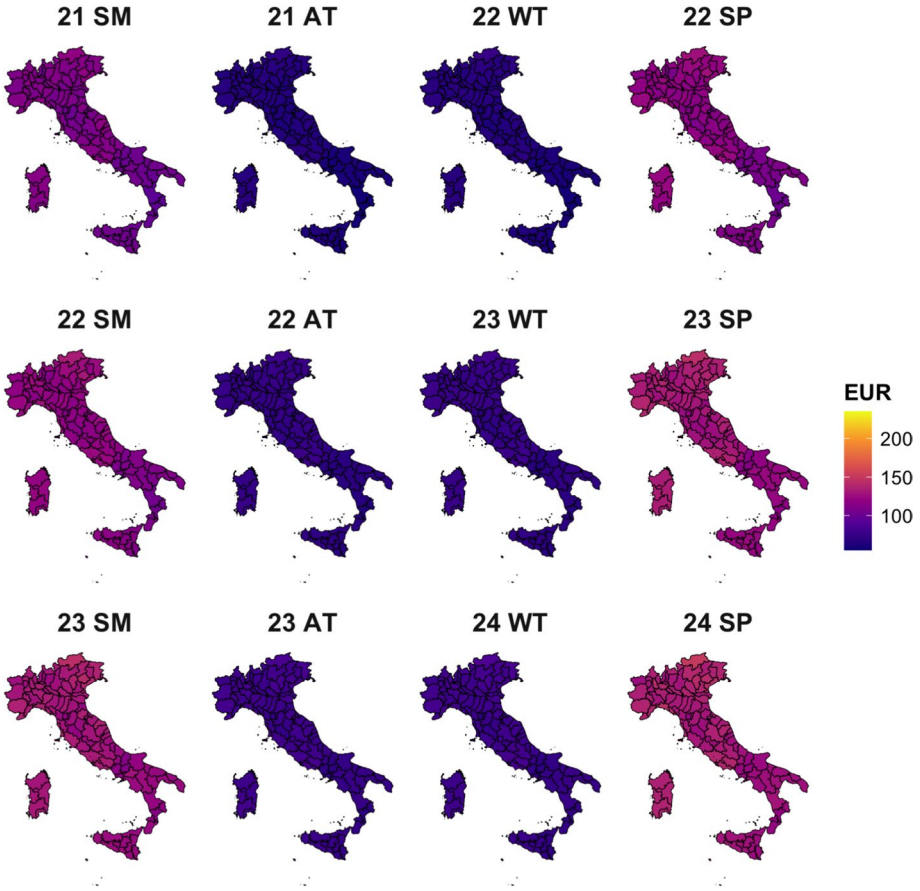

Fig. 14 Average basket cost for adolescents in each survey period

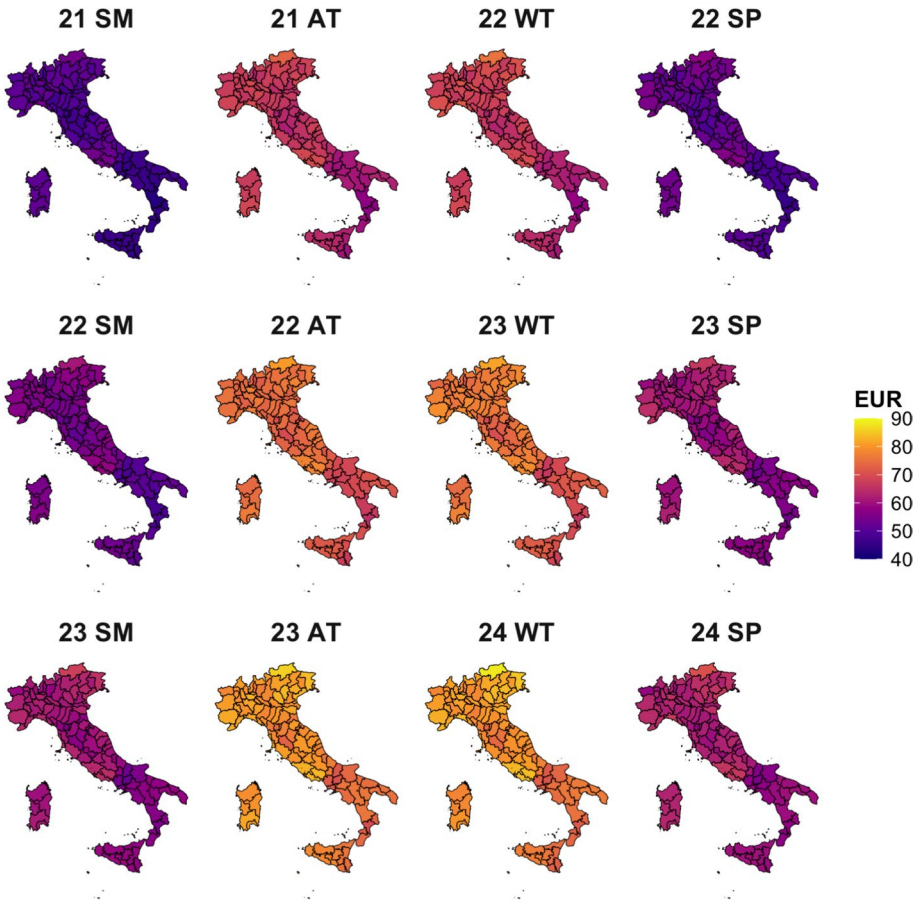

Fig. 15 Average basket cost for babies in each survey period

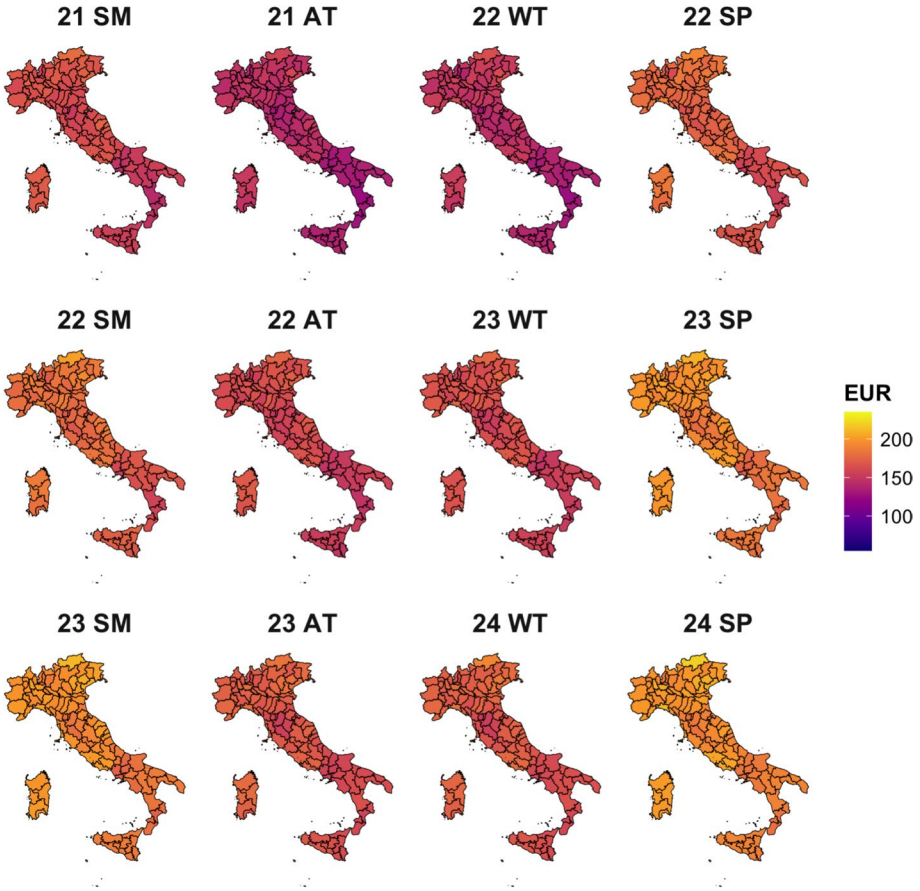

Fig. 16 Average basket cost for adult males in each survey period

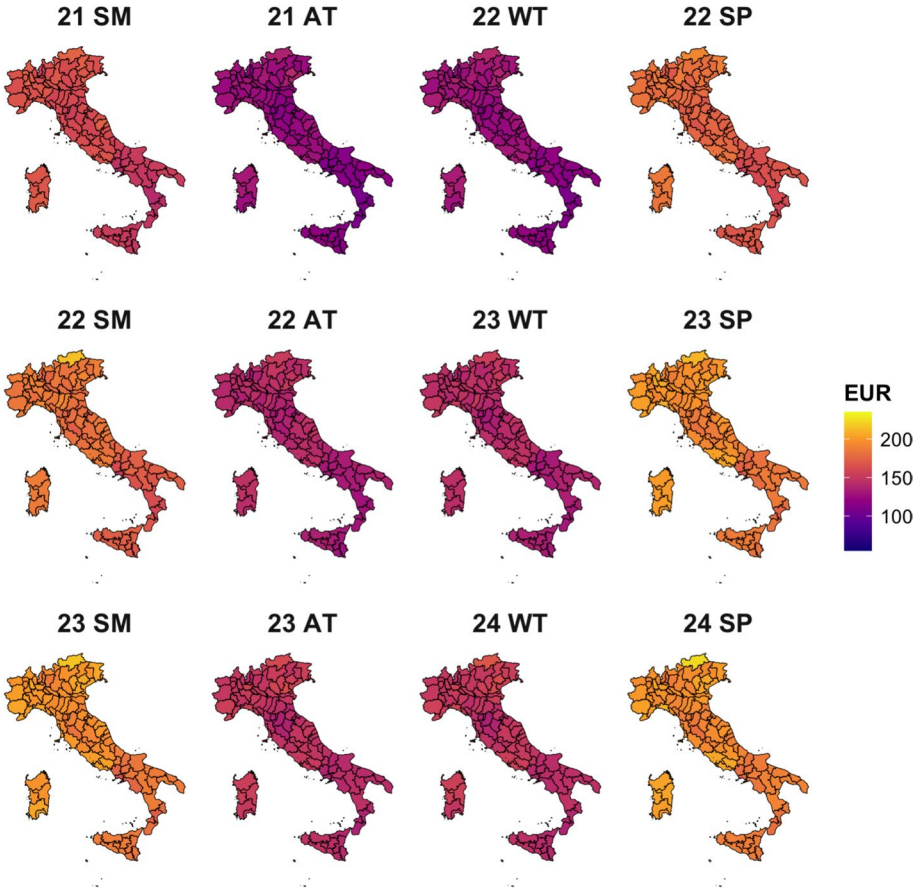

Fig. 17 Average basket cost for elders in each survey period

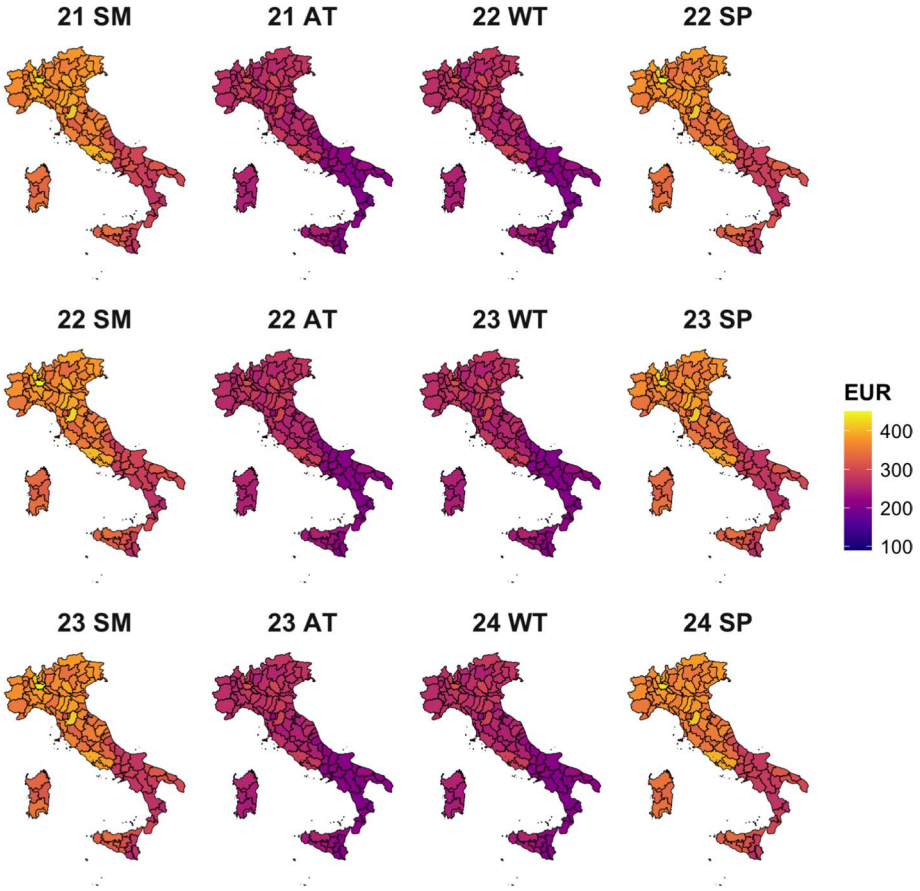

Fig. 18 Maximum basket cost for adult females in each survey period

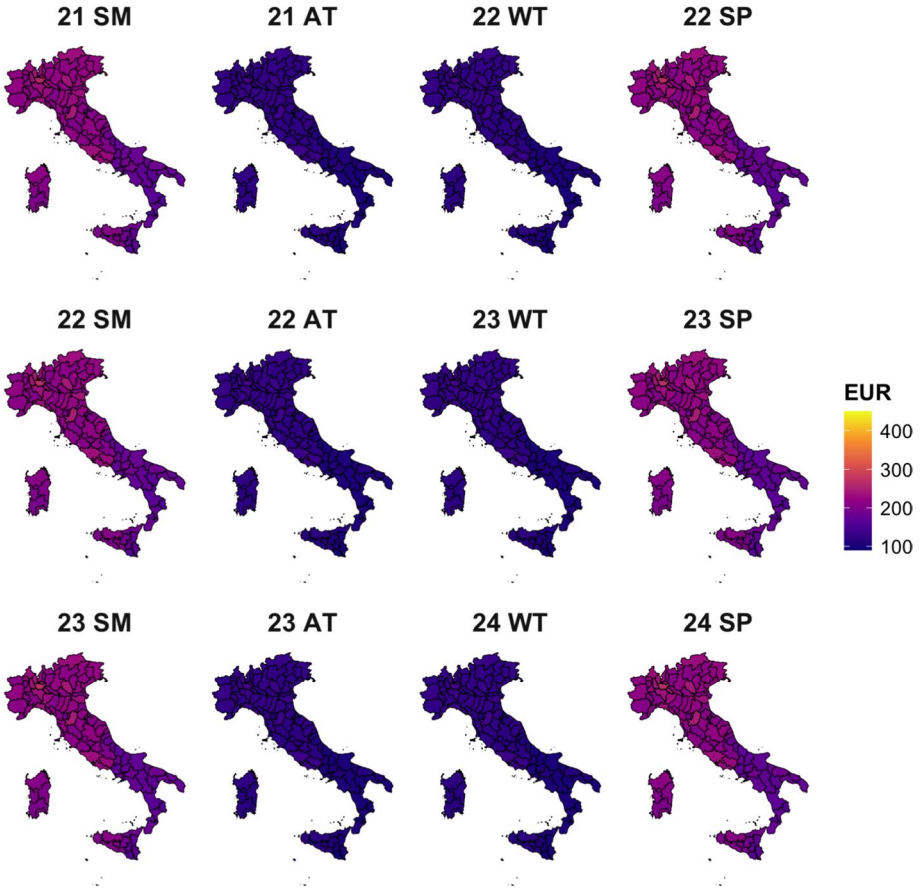

Fig. 19 Maximum basket cost for adolescents in each survey period

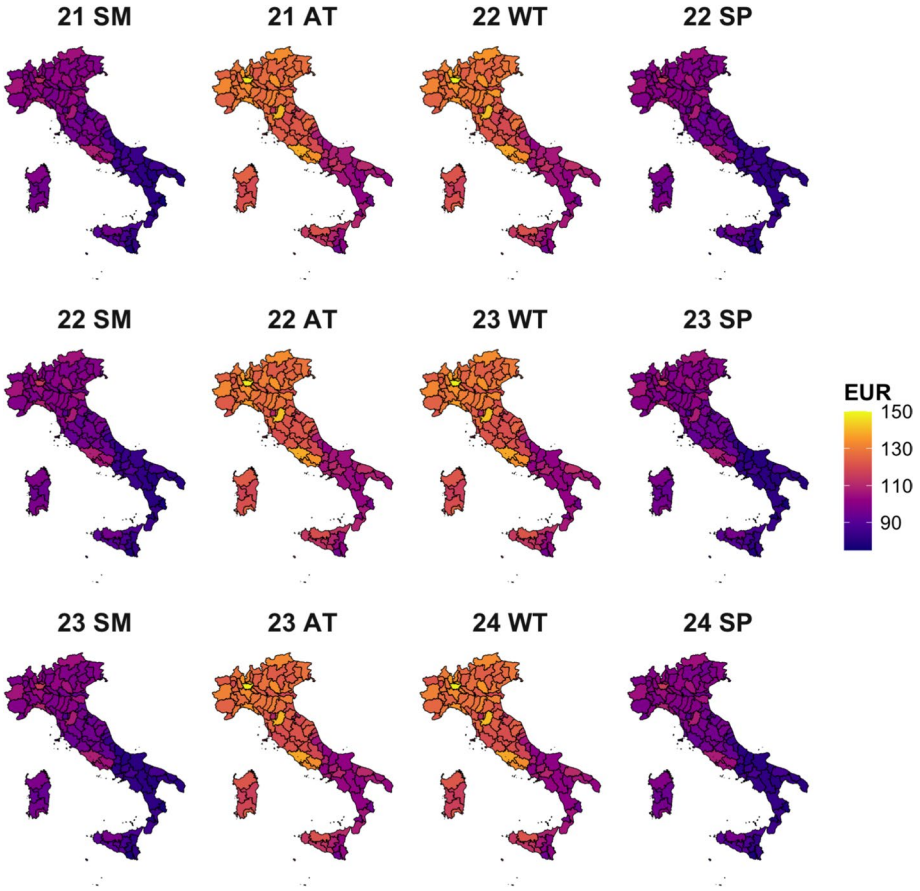

Fig. 20 Maximum basket cost for babies in each survey period

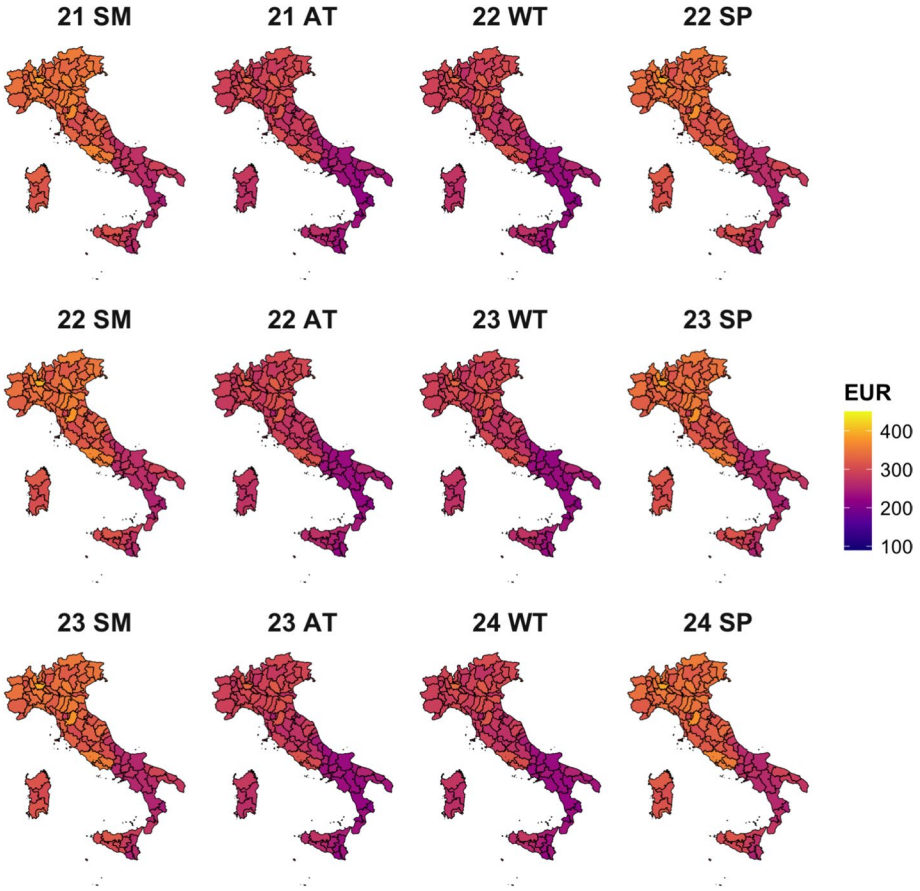

Fig. 21 Maximum basket cost for adult males in each survey period

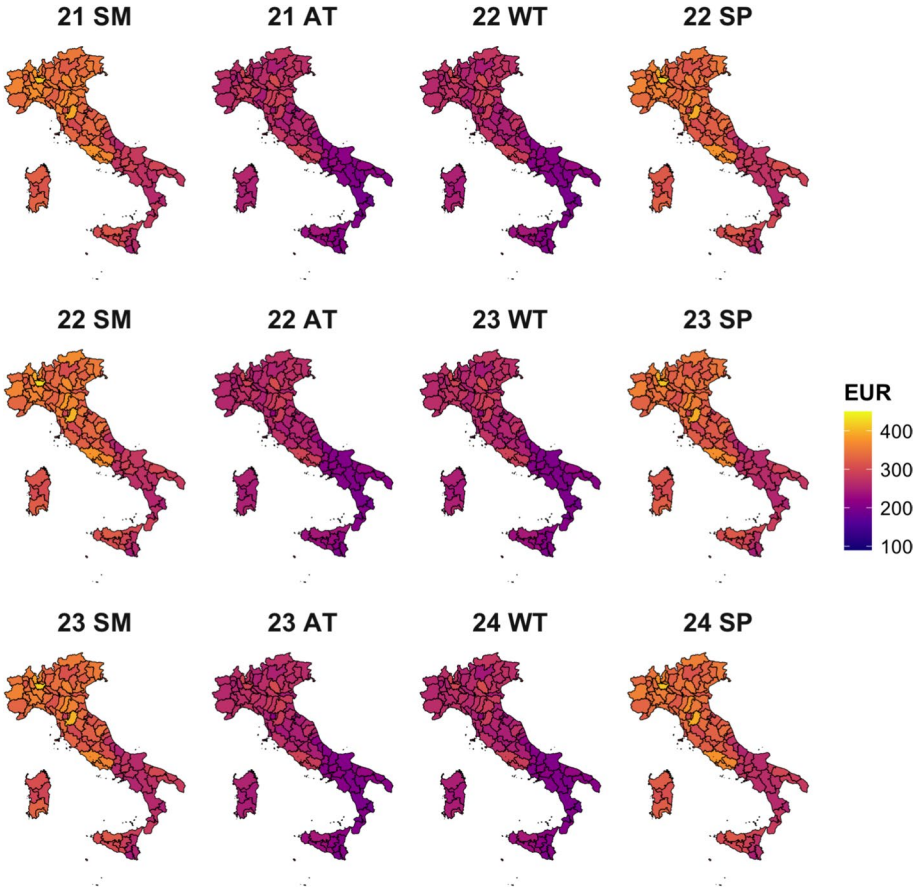

Fig. 22 Maximum basket cost for elders in each survey period

Funding The work has been supported by the project "Food MeaSure: Poverty, Vulnerable individuals and Sustainable Diets—New perspectives on Official Statistical data"—PRIN: Progetti di Ricerca di Rilevante Interesse Nazionale-–-Bando 2022 2022PX2RAR - CUP I53D23004820006 (Principal Investigator: Ilaria Benedetti).

Data availability Data from Osservatorio Prezzi e Tariffe are available on this website https://osservaprezzi.mise.gov.it. Full imputed price dataset and basket costs are available from the corresponding author upon request.

Declarations

Conflict of interest The authors have no Conflict of interest to declare.

References

Abay, K.A., Breisinger, C., Glauber, J., Kurdi, S., Laborde, D., Siddig, K.: The Russia–Ukraine war: implications for global and regional food security and potential policy responses. Glob. Food Sec. **36**, 100675 (2023)

Adam, K., Gautier, E., Santoro, S., Weber, H.: The case for a positive Euro area inflation target: evidence from France, Germany and Italy. J. Monet. Econ. **132**, 140–153 (2022)

Alvarez, S., Cavallo, A., MacKay, A., Mengano, P.: Markups and cost pass-through along the supply chain. Available at SSRN (2024)

Argente, D., Lee, M.: Cost of living inequality during the great recession. J. Eur. Econ. Assoc. **19**(2), 913–952 (2021)

Beacom, E., Furey, S., Hollywood, L., Humphreys, P.: Investigating food insecurity measurement globally to inform practice locally: a rapid evidence review. Crit. Rev. Food Sci. Nutr. **61**(20), 3319–3339 (2021)

Bekkers, E., Brockmeier, M., Francois, J., Yang, F.: Local food prices and international price transmission. World Dev. **96**, 216–230 (2017)

Bertsimas, D., Pawlowski, C., Zhuo, Y.D.: From predictive methods to missing data imputation: an optimization approach. J. Mach. Learn. Res. JMLR **18**(196), 1–39 (2017)

Biasini, B., Rosi, A., Scazzina, F., Menozzi, D.: Predicting the adoption of a sustainable diet in adults: a cross-sectional study in Italy. Nutrients **15**(12), 2784 (2023)

Castellini, G., Savarese, M., Castiglioni, C., Graffigna, G.: Organic food consumption in Italy: the role of subjective relevance of food as mediator between organic food choice motivation and frequency of organic food consumption. Sustainability **12**(13), 5367 (2020)

Costantini, E., Lang, K.M., Sijtsma, K., Reeskens, T.: Solving the many-variables problem in mice with principal component regression. Behav. Res. Methods **56**(3), 1715–1737 (2024)

CREA: Inee guida per una sana alimentazione. Centro di Ricerca Alimenti e Nutrizione. CREA (Ed.) (2018)

Dowler, E., Turner, S., Dobson, B.: Poverty bites: food, health and poor families. (No Title) (2001)

Faber, B., Fally, T.: Firm heterogeneity in consumption baskets: evidence from home and store scanner data. Rev. Econ. Stud. **89**(3), 1420–1459 (2022)

French, S.A., Wall, M., Mitchell, N.R.: Household income differences in food sources and food items purchased. Int. J. Behav. Nutr. Phys. Act. **7**(1), 77 (2010)

Giosuè, A., Calabrese, I., Vitale, M., Riccardi, G., Vaccaro, O.: Consumption of dairy foods and cardiovascular disease: a systematic review. Nutrients **14**(4), 831 (2022)

Grimaccia, E., Naccarato, A.: Food insecurity in Europe: a gender perspective. Soc. Indic. Res. **161**(2), 649–667 (2022)

Herforth, A., Bai, Y., Venkat, A., Mahrt, K., Ebel, A., Masters, W.A.: Cost and Affordability of HealthyDiets Across and Within Countries: Background Paper for the State of Food Security and Nutrition in the World 2020. FAO Agricultural Development Economics Technical Study No. 9, vol. 9. Food and Agriculture Organization, Rome (2020)

ISTAT: Le spese per i consumi delle famiglie—ANNO 2023. ISTAT (Ed.) (2023)

Josse, J., Pagès, J., Husson, F.: Multiple imputation in principal component analysis. Adv. Data Anal. Classif. **5**, 231–246 (2011)

Leung, C.W., Stewart, A.L., Portela-Parra, E.T., Adler, N.E., Laraia, B.A., Epel, E.S.: Understanding the psychological distress of food insecurity: a qualitative study of children's experiences and related coping strategies. J. Acad. Nutr. Diet. **120**(3), 395–403 (2020)

Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002). Retrieved from https://CRAN.R-project.org/doc/Rnews/

Mahrt, K., Mather, D., Herforth, A., Headey, D.D.: Household Dietary Patterns and the Cost of a Nutritious Diet in Myanmar, vol. 1854. International Food Policy Research Institute, Washington (2019)

Marchetti, S., Secondi, L.: The economic perspective of food poverty and (in) security: an analytical approach to measuring and estimation in Italy. Soc. Indic. Res. **162**(3), 995–1020 (2022)

Martens, H., Martens, M.: Multivariate Analysis of Quality: An Introduction. Wiley, Hoboken (2001)

Nicholls, J., Drewnowski, A.: Toward sociocultural indicators of sustainable healthy diets. Sustainability **2021**(13), 7226 (2021)

Palladino, M., Cafiero, C., Sensi, R.: Understanding adolescents' lived experience of food poverty. A multimethod study among food aid recipient families in Italy. Glob. Food Secur. **41**, 100762 (2024)

Penne, T., Goedemé, T.: Can low-income households afford a healthy diet? Insufficient income as a driver of food insecurity in Europe. Food Policy **99**, 101978 (2021)
Potsi, A., D'Agostino, A., Giusti, C., Porciani, L.: Childhood and capability deprivation in Italy: a multidimensional and fuzzy set approach. Qual. Quantity **50**, 2571–2590 (2016)
Principato, L., Secondi, L., Cicatiello, C., Mattia, G.: Caring more about food: the unexpected positive effect of the covid-19 lockdown on household food management and waste. Socioecon. Plann. Sci. **82**, 100953 (2022)
Principato, L., Pice, G., Pezzi, A.: Understanding food choices in sustainable healthy diets-a systematic literature review on behavioral drivers and barriers. Environ. Sci. Policy **163**, 103975 (2025)
Riccardi, G., Giosuè, A., Calabrese, I., Vaccaro, O.: Dietary recommendations for prevention of atherosclerosis. Cardiovasc. Res. **118**(5), 1188–1204 (2022)
Sanderson Bellamy, A., Furness, E., Nicol, P., Pitt, H., Taherzadeh, A.: Shaping more resilient and just food systems: lessons from the COVID-19 pandemic. Ambio **50**, 782–793 (2021)
Schneider, K.R., Bellows, A., Downs, S., Bell, W., Ambikapathi, R., Nordhagen, S., Fanzo, J.C.: Inequity in Access to Healthy Foods (2023)
Shabnam, N., Aurangzeb, N., Riaz, S.: Rising food prices and poverty in Pakistan. PLoS ONE **18**(11), e0292071 (2023)
Sheremirov, V., et al.: The drivers of inflation dynamics during the pandemic:(early) evidence from disaggregated consumption data. In: Current Policy Perspectives. Federal Reserve Bank of Boston (2021)
Tang, F., Ishwaran, H.: Random forest missing data algorithms. Stat. Anal. Data Min. ASA Data Sci. J. **10**(6), 363–377 (2017)
Tiberti, L., Tiberti, M.: Food price changes and household welfare: What do we learn from two different approaches? J. Dev. Stud. **54**(1), 72–92 (2018)
Wold, H.: Estimation of principal components and related models by iterative least squares. Multivar. Anal. **23**, 391–420 (1966)
Wright, K.: NIPALS: Principal Components Analysis Using NIPALS or Weighted EMPCA, with Gram-Schmidt orthogonalization [Computer software manual] (2024). Retrieved from https://kwstat.github.io/nipals/ (R package version 1.0)
Zhu, W., Chen, Y., Han, X., Wen, J., Li, G., Yang, Y., Liu, Z.: How does income heterogeneity affect future perspectives on food consumption? Empirical evidence from urban China. Foods **11**(17), 2597 (2022)