Swiss CPI Microdata for Research: Data Structure, Panel Construction, and Applications*

Barbara Rudolf† Pascal Seiler‡

August 2026

Abstract

This paper describes the microdata underlying the Swiss Consumer Price Index (CPI) and provides a fully reproducible pipeline for their use in empirical research. It documents the features most relevant for research—product hierarchy, collection methods, expenditure weights, temporary sales, product substitutions, and missing observations—and details a complete procedure for transforming the raw files into a research-ready price panel, making all design choices explicit. Two case studies illustrate how different research objectives map into different sample constructions: one reconstructing official CPI inflation, one tailored to the analysis of price-setting behavior. An open-source R repository implements every pipeline step, with key treatment choices available as configurable switches, allowing researchers with authorized data access to reproduce all results and adapt the pipeline to their own questions.

JEL classification: C80, C81, E31

Keywords: CPI microdata; price-setting behavior; reproducible research; longitudinal price panels; inflation measurement

---

*We would like to thank Corinne Becker and Hans-Markus Herren of the Swiss Federal Statistical Office for their valuable help with the data.

†Swiss National Bank, Monetary Policy Analysis, Börsenstrasse 15, CH–8022 Zurich, Switzerland. E-mail: barbara.rudolf@snb.ch.

‡ETH Zurich, KOF Swiss Economic Institute, Leonhardstrasse 21, CH–8092 Zurich, Switzerland. E-mail: seiler@kof.ethz.ch. Web: https://www.pascalseiler.ch.

1 Introduction

The consumer price index is among the most consequential statistics produced by modern economies. It serves as a target variable for monetary policy, as the basis for wage and contract indexation, and as an important measure of changes in the cost of living, to name a few of its roles. The microdata underlying CPIs—individual price quotes collected across products, outlets, and time periods—have become a standard input for research on the very price dynamics the index is meant to capture. Since the early contributions of Bils and Klenow (2004), Klenow and Kryvtsov (2008), and Nakamura and Steinsson (2008) on price rigidity in the United States, and the cross-country work of Dhyne et al. (2006) and the subsequent PRISMA network (Gautier et al., 2024) for the euro area, CPI microdata have been central to research on price-setting behavior, inflation dynamics, monetary transmission, and consumption heterogeneity. Evidence for Switzerland includes Kaufmann (2009), who studies quarterly data from 1993 to 2005, and Rudolf and Seiler (2022), who analyze monthly data from 2008 to 2025.[1]

Despite their widespread use, CPI microdata are difficult to work with, for two distinct reasons. The first is one of transformation. The raw data are organized for the operational task of index construction, but not for the longitudinal panel analyses that most economic research requires. Turning raw price quotes into a research-ready dataset demands a long sequence of non-trivial choices—about how to link observations over time, how to handle product replacements and quality adjustments, whether to retain or remove temporary sales, and how to treat missing observations and imputations. These choices are consequential: different treatments of the same raw data can produce materially different estimates of price-change frequency, price-spell durations, or the distribution of price changes across products and sectors.

The second difficulty is confidentiality. CPI microdata contain sensitive information about prices charged by individual outlets, and access is therefore restricted in virtually every country. This restriction makes direct data sharing impossible, which in turn makes reproducibility challenging. Findings based on CPI microdata are difficult for other researchers to verify or build upon unless the data-processing steps are documented with unusual precision—a standard that,

[1] Klenow and Malin (2010), Nakamura and Steinsson (2013), Luo and Villar (2025), and Dedola et al. (2025) provide comprehensive reviews of this literature.

in practice, is rarely met.

This paper addresses both difficulties using the Swiss CPI microdata as its setting. The Swiss Federal Statistical Office (FSO) provides a rich longitudinal microdata series combining individual price quotes across regions, outlets, and product categories with annual expenditure weights derived from the Household Budget Survey. Over its roughly three-decade history, the Swiss CPI has incorporated scanner data, online price collection, and repeated methodological revisions. The resulting complexity is representative of modern CPI production systems in high-income countries.

The paper makes three contributions covering the period from May 2000 through December 2025, the latest data available at the time of writing. First, it provides a research-oriented description of the Swiss CPI microdata. Rather than replicating the methodological documentation of the FSO (FSO, 2022), which focuses on the construction of official index series, the paper describes the data from the perspective of a researcher building a longitudinal panel—covering the unit of observation, the product hierarchy, collection methods and their heterogeneity, expenditure weights, and the four features that most frequently require explicit treatment decisions: temporary sales, product substitutions and quality adjustments, missing observations, and the periodicity of price collection. Second, the paper documents in full how a longitudinal micro price panel can be constructed from the raw data, making every assumption and design choice at each step of the transformation explicit. Third, it introduces an open-source R repository that implements all pipeline steps in a modular and reproducible way. The repository takes the confidential raw files as input and produces research-ready datasets as output, with key treatment choices—on sales, substitutions, index-based prices, administered prices, and others—implemented as configurable switches that researchers can adjust to suit their own questions. The accompanying open-source repository is available at `https://github.com/pascalseiler/swiss-cpi-microdata`.

To illustrate how different research objectives translate into different data constructions, the paper presents two case studies. The first retains essentially all price quotes used in the official index and uses them to reconstruct official CPI inflation from the microdata. The second applies a set of economically motivated restrictions—excluding index-based prices, administered prices, and other observations that do not correspond to decentralized market-price decisions—to construct a sample suited to the analysis of price-setting behavior. These case studies are intended not as substantive empirical applications but as concrete illustrations of the pipeline in operation, and of the trade-offs inherent in each approach.

The approach taken here is relevant beyond Switzerland. CPI microdata in other countries share the core challenges described in this paper: the absence of explicit longitudinal identifiers, product turnover and quality adjustment, the treatment of temporary sales, and confidentiality constraints that prevent direct data sharing. Transparent pipeline documentation and reproducible code are therefore general-purpose tools for advancing the cumulative use of official price statistics in economic research.

The remainder of the paper is organized as follows. Section 2 discusses data access, confidentiality, and the reproducibility strategy adopted in this paper. Section 3 describes the Swiss CPI microdata from a research perspective. Section 4 documents the construction of a longitudinal price panel. Section 5 presents the two case studies. Section 6 describes the accompanying R repository. Section 7 concludes.

2 Data Access, Confidentiality, and Reproducibility

The microdata underlying the Swiss CPI are subject to strict confidentiality requirements. Because the data contain detailed information about prices charged by individual outlets—and in some cases about contractual terms or supplier relationships—they cannot be made publicly available. Access is granted by the Federal Statistical Office under research agreements that specify permissible uses and prohibit redistribution. Analysis is typically conducted within secure computing environments or under data-use agreements that require outputs to be reviewed before publication.

These constraints are common to CPI microdata in most countries, and they create a tension that is worth stating explicitly. The value of microdata-based research depends on the ability of others to scrutinize, replicate, and extend published findings. Under confidentiality restrictions, such scrutiny is possible in principle—any researcher granted data access can run the same code on the same data—but it is not possible in practice unless the data-processing steps are documented with sufficient precision that a second researcher, starting from the same raw files, would arrive at an identical analysis sample. In practice, this standard is rarely met. Studies using CPI microdata typically describe their data constructions in a few sentences of a data section, leaving many consequential choices—on the handling of missing observations, the classification of substitutions, the treatment of sales, the exclusion of specific product categories—undocumented or documented only partially.

The cumulative cost of this gap is substantial. Differences in undocumented data treatments can produce meaningfully different estimates of the same underlying economic objects, making it difficult to assess whether variation in published results reflects genuine economic heterogeneity or differences in how the data were prepared. It also raises the entry cost for new researchers: without detailed documentation, each study using CPI microdata must independently reconstruct a pipeline that is largely common across applications, a process that is time-consuming and error-prone.

This paper adopts a reproducibility strategy designed to close this gap while fully respecting the legal constraints on data sharing. The strategy rests on a clean separation between code and data. All data-processing and panel-construction steps are implemented in a publicly available R repository that contains no confidential information. The repository accepts the raw files provided by the FSO as input and produces a research-ready longitudinal panel as output. A researcher with authorized access to the microdata can reproduce all results, tables, and figures in this paper by running the published code without modification, provided the input files are placed in the directory structure specified in the repository documentation. The pipeline is fully deterministic: given the same inputs and the same configuration, it always produces the same output.

The repository is organized around a distinction between fixed pipeline steps and configurable treatment choices. Fixed steps—date harmonization, variable renaming, basket merging, weight assignment—are the same regardless of research application and are applied unconditionally. Treatment choices—on whether to retain or remove temporary sales, how to handle product substitutions, which categories to exclude—depend on the research question and are implemented as named switches in a central configuration file. Setting a switch changes the behavior of one or more pipeline modules without altering any other part of the code. This architecture serves two purposes. It makes the pipeline transparent: every deviation from a default construction is visible as a change to a documented configuration parameter. And it makes the pipeline extensible: researchers can adapt it to new questions by adjusting switches or adding modules, without modifying the core code.

The two case studies in Section 5 illustrate this architecture in practice, using different switch configurations to construct two distinct but internally consistent analysis samples from the same raw data. By making both configurations and the code that implements them publicly available, the paper provides a reference point against which future work using the same data can be compared.

3 Swiss CPI Microdata

This section describes the Swiss CPI microdata from the perspective of empirical research. The focus is on features that are essential for constructing longitudinal price panels and for interpreting micro-level price dynamics. It abstracts from the institutional details of CPI production and refers instead to the methodological documentation of the Federal Statistical Office for a comprehensive treatment (e.g., FSO, 2016, 2022).

3.1 Unit of Observation and Panel Structure

The fundamental unit of observation in the CPI microdata is a price quote: the transaction price of a narrowly defined good or service observed in a specific sales outlet during a given reference period. Prices are recorded inclusive of indirect taxes and compulsory charges and are intended to reflect prices actually paid by consumers.

Each price quote is characterized by three dimensions. The time dimension corresponds to the month of price collection. The product dimension is defined at the level of CPI varieties—the most granular level of the basket of goods and services, which is designed to be representative of private household consumption in Switzerland. The outlet dimension identifies the specific sales outlet or data source in which the price was observed; each outlet is associated with one of the price-collection regions used in the Swiss CPI, a variable available from January 2006 onward.

The raw microdata are not organized as a longitudinal panel. Price quotes appear as independent monthly observations, and there is no explicit identifier linking the same product-outlet combination across time periods. Constructing panel series therefore requires an explicit linking step, described in Section 4.

3.2 Product Classification and Hierarchy

The basket of goods and services underlying the Swiss CPI is organized according to the Classification of Individual Consumption by Purpose (COICOP), adopted in May 2000. For research applications, three levels of this hierarchy are particularly important.

The highest level divides the overall CPI into twelve main groups (*Hauptgruppen*, HG), such as food and non-alcoholic beverages, housing and energy, and transport. Below this, product groups provide finer categorizations. At the lowest level for which the FSO publishes official price indices and expenditure weights, the CPI defines expenditure items (*Indexpositionen*, IP), corresponding approximately to five-digit COICOP positions.

Below expenditure items are varieties (*Erhebungspositionen*, EP), which constitute the lowest level of the basket of goods and services and define the concrete goods and services for which individual prices are collected. Varieties are not directly weighted in the official CPI calculation, and the set of varieties changes over time as consumer behavior and market offerings evolve. Most microeconomic research operates at the variety level, making it necessary to link price quotes both across time and, in some cases, across successive variety definitions.

3.3 Temporal Coverage and Price Collection

The Swiss CPI microdata are available from 1993, though the adoption of the COICOP classification in May 2000 introduced a comprehensive basket revision that creates a structural break at that date. The pipeline constructed in this paper therefore begins in May 2000. While earlier observations may in principle be linkable through concordance exercises across basket revisions,[2] such an extension is beyond the scope of the present paper.

Since January 2008, the large majority of goods and services are priced on a

[2] Kaufmann (2009) notes that the transition to the COICOP classification largely involved a redistribution of pre-existing index positions into the new classification structure. Extending the panel back to 1993 may therefore be feasible with additional concordance work. However, other challenges remain, including differences in product definitions, collection frequencies, and data structure across vintages.

monthly basis. Prior to 2008, a substantial share of categories was collected at lower frequencies, which limits the scope for consistent monthly analysis over the earlier period. Some categories, such as rents, continue to be collected quarterly; others are collected seasonally or at irregular intervals when price changes occur.

Prices are collected through several complementary channels. Traditionally, the dominant mode was in-person collection by trained price collectors visiting physical outlets. Over time, this has been supplemented by scanner data from major retailers, centrally collected prices for products with nationally uniform pricing, and an increasing volume of online collection through web scraping and direct digital reporting (FSO, 2021).

The distinction between collection modes matters for research because prices collected online and prices collected in physical outlets may exhibit systematically different adjustment patterns. Changes in the dominant collection mode for specific product categories can also introduce structural breaks in observed price dynamics that are unrelated to changes in firms' underlying pricing behavior. Scanner data pose a further complication in this respect: unlike traditional CPI price quotes, which correspond to a single product-outlet observation, scanner-based prices are typically derived from transaction records and represent average prices over a set of transactions within a given product category and reference period. This aggregation means that scanner-sourced observations are not directly comparable to collector-recorded quotes at the level of an individual product-outlet pair, which has implications for how price changes and adjustment frequencies are measured across collection modes. While the microdata identify the responsible outlet or data source, they do not always contain an explicit indicator for the collection mode. For several expenditure items, the transition to online or scanner-based collection occurred at a known point in time—for example, consumer electronics shifted to online collection in January 2016—enabling researchers to construct collection-mode indicators at the item level.

3.4 Expenditure and Regional Weights

The Swiss CPI uses two types of expenditure weights, which differ in their level of application and update frequency.

Expenditure weights are defined at the item level and reflect the relative importance of each item in private household consumption, as measured by the Household Budget Survey. They are updated annually.

Regional weights reflect the relative importance of the price-collection regions within the official aggregation procedure. Unlike item weights, they are held constant within each basket period and updated only every five years when the basket is revised.

In the official CPI, item-level and regional weights enter a hierarchical aggregation procedure that also accounts for distribution channels. While item-level and regional weights are publicly available and can be incorporated in micro-level analyses, the channel-specific weights used in the official index construction are not. As a result, micro-based aggregations necessarily approximate the official procedure, as they cannot fully replicate the channel dimension of weighting. Researchers should be aware of this limitation when comparing weighted microdata aggregates to official CPI series.

3.5 Temporary Sales

The Swiss CPI aims to measure actual transaction prices. Temporary price reductions—including promotions and short-term discounts—are therefore included in the official index if they meet specific criteria: the reduction must apply to an identical product, be available to all consumers, and not be conditional on bulk purchases or loyalty schemes. The microdata contain a binary flag recording whether each observation was identified as a temporary sale by the price collector.

For microeconomic research, temporary sales pose a well-known challenge (Nakamura and Steinsson, 2008; Klenow and Kryvtsov, 2008). They introduce high-frequency price movements that may not reflect deliberate repricing decisions by firms, and they can mechanically inflate measures of the repricing frequency. At the same time, sales are genuine price reductions paid by consumers and are economically relevant for index construction and consumption analysis.

The appropriate treatment of sales therefore depends on the research question. Studies aiming to replicate or understand official index dynamics will generally retain sales, while studies of price rigidity and regular price adjustment commonly strip them out. The distinction between these two orientations is a key axis along which the two case studies in Section 5 differ.

3.6 Product Substitutions and Quality Adjustment

Products are periodically replaced in the CPI. When an item disappears from the market or ceases to be representative of consumer spending, it may be substituted by a comparable product. Because the replacement product is typically not identical to its predecessor, the price change at the time of substitution conflates a pure price movement with a quality change. The FSO applies various quality-adjustment procedures to isolate the pure price component.

The microdata record product substitutions through a replacement code variable. The coding distinguishes several types of event: direct replacements that are treated as continuous (code 1), replacements with explicit quality adjustment using linked prices for the preceding period (codes 2 and 7), and other substitution events. FSO (2022) documents these codes.

For research purposes, the critical question is not which adjustment method the FSO applied, but whether price changes at substitution events should be treated as part of the same price series or as a discontinuity. The CPI methodology implicitly treats many substitutions as continuous observations, accepting the quality-adjusted price change as a valid price movement. In studies of price rigidity, however, it may be preferable to treat each substitution as the termination of one price spell and the beginning of another, thereby excluding the price change at the moment of replacement from the analysis. Both approaches are valid, but they yield different measures of the frequency of price changes, and their consequences are large enough to matter empirically. This paper implements both as a configurable option.

3.7 Missing Prices and Imputations

Price quotes may be absent for a given product-outlet-month combination for several reasons: seasonal unavailability, temporary stockouts or outlet closures. In the official CPI, missing prices are typically handled by carrying forward the last observed price, subject to a maximum imputation horizon, or by applying higher-level imputation rules.

At the micro level, imputed prices appear as repeated values without an explicit flag identifying them as such. This creates two research-relevant problems. First, imputed observations artificially extend price spells and increase measured price rigidity if not accounted for. Second, the imputation horizon and rules differ across product categories and over time, making it difficult to construct a consistent treatment rule.

During the COVID-19 pandemic, a larger-than-usual share of price observations was imputed, particularly for products that became temporarily unavailable or unobservable due to shop closures. The FSO published information at the expenditure-item level about which items were affected and during which months, enabling researchers to flag likely imputations even where no explicit observation-level indicator exists. This external information is incorporated in the construction pipeline described in Section 4.

4 Constructing a Longitudinal Micro Price Panel

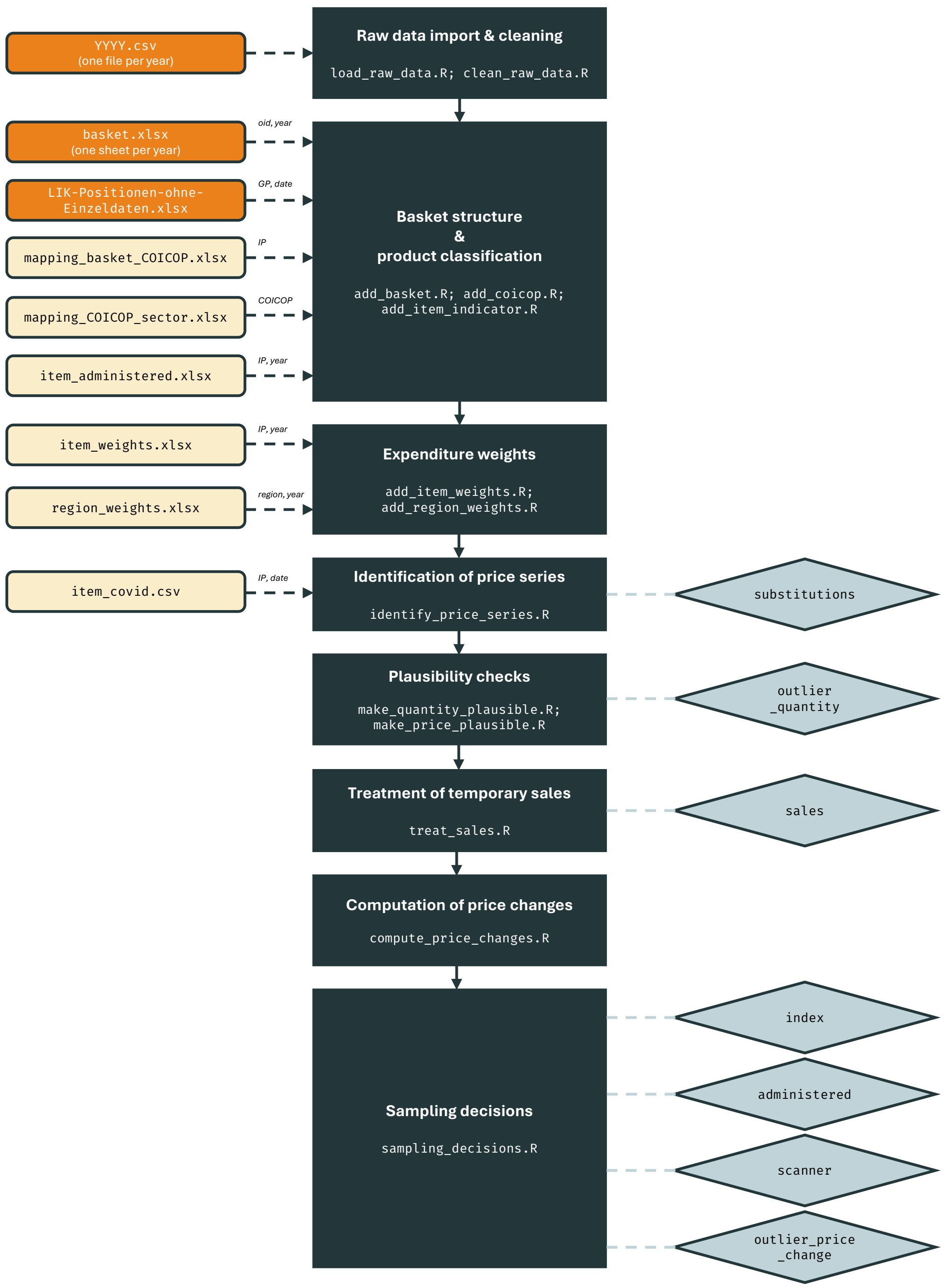
This section documents the construction of a longitudinal panel of micro prices from the raw Swiss CPI microdata. The objective is to transform the raw files into a research-ready dataset in a transparent and reproducible way, making all design choices explicit.

The resulting panel covers the period from May 2000 to December 2025. May 2000 marks the first month following the adoption of the COICOP classification and the associated basket revision; prior observations are not comparable in product definition and are therefore excluded. December 2025 corresponds to the most recently available data at the time of writing.

Figure 1 provides an overview of the pipeline. Each step corresponds to one or more dedicated scripts in the accompanying R repository and is described in a subsection below. The pipeline runs sequentially: each step takes the output of the previous step as its input.

All configurable treatment choices are collected in a single configuration file (`R/00_config.R`), which defines sample years, file paths, and a set of named treatment switches that control key analytical decisions. These switches are described in context below and summarized in Table 1.

Figure 1. Data pipeline for constructing the longitudinal CPI micro price panel



Notes: The figure illustrates the eight sequential processing steps of the data pipeline (in the center), together with their input data sources (left) and configurable treatment switches (right). Confidential FSO inputs are shown in orange; publicly available inputs included in the repository are shown in yellow. Italic labels along arrows indicate merge keys. The final output is a longitudinal micro price panel. Treatment switches are documented in Table 1.

Table 1. Treatment switches in R/00_config.R

Switch	Type (default)	Description
substitutions	Logical (TRUE)	If TRUE, product substitutions are treated as continuous price series and do not trigger spell breaks. If FALSE, any observation carrying a substitution-related replacement code (1, 2, 3, or 7) starts a new spell; COVID-19-related imputations (item_covid.csv) receive a synthetic code (8) and are treated analogously.
sales	Logical (TRUE)	If TRUE, temporary sale prices are retained as observed. If FALSE, each sale price is replaced by the last non-sale price within the spell (carry-forward), eliminating mechanical price changes at the start and end of promotions.
outlier_quantity	Numeric (10)	Factor threshold for detecting implausible quantity jumps within a price spell. Consecutive log-quantity ratios exceeding this factor in either direction are flagged and replaced by the last plausible value carried forward. Set to 0 to disable.
index	Logical (TRUE)	If FALSE, expenditure items for which the microdata record index values rather than true micro prices are excluded, using the time-varying index_indicator and a fixed list of item codes.
administered	Logical (TRUE)	If FALSE, administered and semi-administered expenditure items (e.g., health insurance, public transport) are excluded, using the annually updated item_administered.xlsx list merged by expenditure item and year.
scanner	Logical (TRUE)	If FALSE, expenditure items collected via scanner data are excluded, using the time-varying scanner_indicator.
outlier_price_change	Numeric (0.01)	Percentile threshold $p \in [0, 1)$ for trimming extreme log price changes. Within each variety, observations whose absolute log price change falls below the p-th or above the $(1 - p)$-th percentile of the variety-level distribution are dropped. Set to 0 to disable.

Notes: All switches are set in R/00_config.R and propagate automatically to the relevant pipeline module without requiring changes to any other script. Logical switches default to the value shown; setting the opposite value activates the alternative treatment.

4.1 Raw Data Import and Cleaning

The raw CPI micro price data are provided by the Federal Statistical Office as one CSV file per calendar year (YYYY.csv). All files covering the selected sample period are read and stacked into a single dataset.

A series of mechanical cleaning and harmonization steps is then applied. In particular, the date variable is standardized, as raw files employ different string formats across vintages, including abbreviated month names and two-digit year suffixes. Duplicate December observations arising from CPI chain-linking procedures are removed, retaining only the observation corresponding to the current year. Price quotes belonging to the Harmonized Index of Consumer Prices (HICP), identified via the product description, are excluded, as they are not part of the Swiss CPI.

Core variables are subsequently renamed to a consistent naming scheme and cast to appropriate data types. Numeric price and quantity fields are cleaned of formatting characters before conversion. Regional identifiers are initialized for observations prior to January 2006, when regional information was not yet systematically recorded. Table 2 documents the variables contained in the confidential raw files, reporting both the original identifiers supplied by the FSO and the standardized variable names used throughout the data pipeline.

Table 2. Raw CPI microdata: variable documentation

ID (ID in raw data)	**Description**	**Data type**	**Example**
`date` (`Datum`)	Month of price collection.	Date	2025-01-01
`outlet` (`PMS_Nummer`)	Numerical identifier of the outlet in which the price is collected.	Integer	454898
`region` (`Region`)	Price-collection region. Available from January 2006 onward.	Character	Bâle
`oid` (`Schemapos_OID`)	Variety-level identifier; serves as the merge key linking price observations to the basket.	Integer	16310
`price_series` (`Price_series_NB`)	Price-series number within a variety–outlet combination.	Integer	2
`description` (`Description_g_TX`)	Text description of the variety.	Character	Aspargus, green
`price` (`Price_NB`)	Observed transaction price.	Numeric	7.57
`quantity` (`Quantity_NB`)	Quantity associated with the observed price (e.g. weight, volume, or number of units).	Numeric	0.75
`replacement_code` (`Processing_CD`)	Processing code indicating product replacements and the associated quality-adjustment method. For a detailed description of individual codes, see FSO (2016).	Integer	99
`price_pre` (`Pre_period_corr_price_NB`)	Quality-adjusted price of the preceding period.	Numeric	7.77
`quantity_pre` (`Pre_period_corr_quantity_NB`)	Quality-adjusted quantity of the preceding period.	Numeric	0.50
`sales_indicator` (`Reduction_FLG`)	Binary flag for temporary price reductions (sales), as recorded by price collectors.	Binary indicator	0

Notes: The table documents the structure of the confidential CPI microdata. The raw data are provided by the Federal Statistical Office as one CSV file per calendar year (`YYYY.csv`), each containing price quotes for narrowly defined goods and services observed in specific outlets and reference periods. The table maps original variable identifiers to the harmonized naming convention used in the data pipeline and summarizes the key fields at the variety-outlet-month level.

4.2 Basket Structure and Product Classification

The basket of goods and services is provided by the Federal Statistical Office as a confidential Excel file containing one sheet per calendar year (`basket.xlsx`).[3] Each sheet lists all basket positions for that year and records, for each position, its hierarchical level (ranging from the aggregate CPI down to individual varieties), its product codes, and its scheduled price-collection frequency. Table 3 summarizes and describes the variables contained in the basket data.

The basket data are processed year by year and stacked into a single dataset. In the raw file, each level of the hierarchy is recorded as a separate row, while higher-level identifiers are not repeated on subsequent lower-level rows. To recover the complete hierarchy, parent-level identifiers are carried forward until the next entry at the same hierarchical level appears. This procedure assigns to each variety its unique expenditure item, product group, and main group as defined in the CPI basket for that year.

The reconstructed basket hierarchy is merged into the CPI micro price data using a left join on the variety identifier (`oid`) and calendar year. As a result, each price quote inherits the basket classification and collection design that are applicable at the time of observation. In particular, the basket data record the periodicity of price collection through two variables: a periodicity code and a twelve-character periodicity string indicating the months in which prices are scheduled to be collected. These variables play a central role in the computation of price changes, as they allow scheduled and unscheduled observations to be distinguished in the price-change measures described in Section 4.7.

[3] While related basket information is publicly available, the research version provided by the FSO contains additional information, including mappings from varieties to basket positions and collection-frequency metadata.

Table 3. Basket of goods and services: variable documentation

ID (ID in raw data)	Description	Data type	Example
GP (GP.Nr.)	Unique product identifier assigned at each of the six hierarchical levels of the CPI basket (levels 1–6). Present for every row in the basket file; used to reconstruct the product hierarchy.	Integer	1410
oid (Schemapos_oid)	Variety-level identifier; serves as the merge key linking price observations to the basket.	Integer	16310
position (Pos.Typ)	Hierarchical level within the CPI basket, taking values 1 (aggregate CPI) through 6 (variety). Used to recover the full product hierarchy by carrying higher-level GP codes forward to child rows.	Integer	6
periodicity_code (Periodizitaet.Code)	Code indicating the collection frequency of prices for the variety. Common values include M (monthly), V (quarterly), H (semi-annual), J (annual), and A (aperiodic).	Character	M
periodicity_string (Periodizitaetsmonat)	Twelve-character string indicating the scheduled collection months. Each character position corresponds to a calendar month (January through December); X marks a scheduled collection month and - marks a non-collection month. Used to construct the collection indicator in the price-change computation step.	Character	-XXXXX——

Notes: The table documents the structure of the basket of goods and services. The raw data are provided by the Federal Statistical Office as a confidential Excel file (basket.xlsx) with one sheet per calendar year, each listing basket positions across all hierarchy levels. The table maps original variable identifiers to the harmonized naming convention used in the data pipeline and summarizes the key fields describing the product hierarchy and collection design at the variety level.

Two additional classification layers are added at this stage. First, each expenditure item is mapped to its corresponding COICOP position using a publicly available concordance table (mapping_basket_COICOP.xlsx). Second, broader sectoral aggregates—food, energy, non-energy industrial goods, and services—are derived from the COICOP classification (mapping_COICOP_sector.xlsx).

Beyond these structural classifications, a set of item-level indicators provided by the FSO as part of the confidential microdata is merged into the dataset (LIK-Positionen-ohne-Einzeldaten.xlsx). These indicators identify expenditure items for which prices are collected via scanner data, items whose prices are derived from index-based methods rather than direct price observations, and items classified as services. Because the introduction of scanner-based and index-based collection varies across item groups and over time, the indicators are activated conditionally on the observation date, following the periods documented in the metadata.

Finally, administered and semi-administered items are identified by constructing an `admin_price_indicator`. This indicator is obtained by merging the publicly available file `item_administered.xlsx` into the data using expenditure item and year, allowing these categories to be retained or excluded depending on the research objective.

4.3 Expenditure Weights

Item-level and regional expenditure weights are added in two separate steps, each drawing on a publicly available input file included in the repository.

Item-level weights are sourced from `item_weights.xlsx`, which contains one sheet per basket revision period: `LIK2000` (2000–2005), `LIK2005` (2006–2010), `LIK2010` (2011–2015), `LIK2015` (2016–2020), and `LIK2020` (2021–2025). Each sheet provides annual weights for all expenditure items in that revision period. The five sheets are read, stacked into a long-format table of item–year pairs, and merged into the microdata by expenditure item and calendar year, so that each price observation inherits the weight applicable in its reference year.

Regional weights are sourced from `region_weights.xlsx`, organized by the same five basket revision periods. Unlike item weights, regional weights do not vary within a basket period: a single weight is defined per region per revision, and all years within a period are assigned the same value. The weights are merged by region name and calendar year using the same period-to-year mapping as for item weights.

4.4 Identification of Price Series

A central challenge in working with CPI microdata is the absence of explicit longitudinal identifiers. The raw data record each month's prices as independent observations; constructing a panel requires linking price quotes that refer to the same product being sold at the same outlet over consecutive periods.

The pipeline defines a price series as a sequence of price quotes sharing the same variety identifier, outlet, and FSO price-series number, observed in consecutive months without a gap exceeding 31 days. Observations are sorted by variety, outlet, price-series number, and date, and a new-spell flag is set whenever the time gap to the preceding observation within a group exceeds one month. Each resulting spell is assigned a globally unique numeric identifier (`series_counter_number`) used as the key grouping variable in all downstream computations.

The treatment of product substitutions represents the main discretionary choice at this stage of the panel construction. When a product replacement occurs, the CPI microdata record this event through dedicated processing codes and, in some cases, changes in auxiliary identifiers associated with the price quote. However, the raw data do not provide a unique longitudinal identifier that mechanically determines whether a replacement should be interpreted as a continuation of the same price series or as the start of a new one.

The key question for research purposes is therefore whether observations following a product replacement should be treated as belonging to the same underlying price series—thereby accepting the CPI's implicit or explicit quality adjustment—or whether they should be interpreted as terminating the previous series and initiating a new one. Both approaches are common in the literature (Klenow and Kryvtsov, 2008; Klenow and Malin, 2010) and depend on the research objective.

The pipeline implements both approaches via a treatment switch (`substitutions`). When substitutions are treated as continuous (`TRUE`), only temporal gaps trigger new spells; when they are excluded (`FALSE`), any observation carrying a substitution-related replacement code (1, 2, 3, or 7) also starts a new spell. In the latter case, COVID-19-related imputations are additionally flagged at the expenditure-item level using FSO press-release data (`item_covid.csv`), and observations in affected items and months receive a synthetic replacement code (8) that triggers a spell break. This prevents imputed price spells from being treated as genuine price observations.

4.5 Plausibility Checks

Before computing price changes, the pipeline applies a small number of deterministic corrections to address known data quality issues.

For quantities, zero and negative values are replaced by one, reflecting the assumption that the recorded price corresponds to a single unit. Large implausible quantity jumps within a price series—which typically reflect unit changes, such as a shift from per-kilogram to per-100-gram reporting, rather than actual changes in the product sold—are detected by computing the ratio of consecutive log-quantities within each spell. Ratios exceeding a configurable threshold (default: a factor of ten in either direction, controlled by `outlier_quantity`) are flagged, and the implausible value is replaced by the last plausible observation carried forward.

For prices, zero and negative values are set to missing and then filled by carrying the last valid price forward within the spell. This situation is rare in the Swiss CPI data but can arise due to data entry errors or specific collection conventions for certain product categories.

4.6 Treatment of Temporary Sales

Temporary sale prices, identified by the binary sales indicator recorded by price collectors, are handled according to a configurable treatment switch (`sales`). When sales are retained (`TRUE`), no modification is made and the recorded transaction price is used as-is. When sales are excluded (`FALSE`), each sale price is replaced by the last observed non-sale price within the same spell, implementing a carry-forward of the regular price. This approach eliminates the mechanical price changes associated with the start and end of temporary promotions—which would otherwise appear as price decreases at the start of the promotion followed by symmetric price increases at its end—while preserving the underlying regular price trajectory and the full observation count of the panel.

The choice between these options has a large effect on the frequency of price changes, particularly in product categories with high promotional intensity such as food and personal care products. The two case studies in Section 5 differ on this dimension, illustrating its empirical relevance.

4.7 Computation of Price Changes

The final pipeline step computes price-level and price-change variables for each observation and introduces an indicator distinguishing scheduled from unscheduled collection months.

Log prices are computed for both the nominal price and the unit price. Month-to-month log price changes are computed within each price spell. A zero log-change is re-coded to missing to distinguish observations where the price was unchanged from missing price-change observations; the binary price-change indicator is set to one when a non-zero log-change is observed.

A crucial adjustment accounts for the fact that not all items are collected every month. Using the periodicity string from the basket data—a twelve-character sequence in which each character corresponds to a calendar month, with `X` marking a scheduled collection month—the pipeline constructs a binary collection indicator for each observation. Price changes and the price-change indicator are set to missing in months that fall outside the scheduled collection window. This prevents spurious frequencies of price changes from arising when prices are carried forward for CPI aggregation purposes into months in which no actual price collection took place. The correction is particularly important for quarterly-collection items such as rents and for seasonal products. Items recorded with a fully empty periodicity string are treated as collected every month.

The final dataset is arranged in a consistent column order grouping date and time variables, hierarchical product identifiers from broad to narrow, outlet and region identifiers, spell identifiers, all price and price-change variables, auxiliary flags, and weight variables. This constitutes the complete micro price panel that serves as the input for the sampling and analysis stages described in Section 5 and Section 6.

4.8 Sampling Decisions

The final pipeline step translates the full micro price panel into an analysis-ready sample by applying a set of item- and observation-level filters. Unlike the preceding steps, which are applied unconditionally to all data, the filters at this stage are discretionary: each reflects a research-oriented judgment about which observations carry economically meaningful price information. All filters are controlled by named switches in `R/00_config.R` and can be activated or deactivated independently. These defaults reflect one reasonable configuration among several equally defensible options; researchers pursuing different questions—for instance studying regulated-price pass-through, for which administered prices would be retained, or comparing scanner and non-scanner subsamples—may find that different combinations of switches better suit their purposes, and the modular architecture is designed to make such alternative constructions straightforward.

Index-based Prices. When `index = FALSE`, observations belonging to expenditure items that do not report individual price quotes in the microdata but instead enter as pre-aggregated price indices (e.g., rents or books) are removed. For these items, the FSO does not collect or publish underlying transaction-level prices; only the resulting index series is available.

Administered Prices. When `administered = FALSE`, expenditure items classified as administered or semi-administered are excluded. The list covers items whose prices are set or strongly influenced by regulatory or contractual mechanisms, including health insurance premiums or public transport fares, and is updated annually to reflect changes in administrative pricing arrangements.

Scanner Data. When `scanner = FALSE`, observations belonging to expenditure items collected via retailer scanner systems are removed, using the time-varying `scanner_indicator`. This option is available for researchers who wish to focus on manually collected outlet prices, for example to study price-setting behavior in segments where scanner coverage may introduce selection.

Price-change Outliers. When `outlier_price_change` > 0, extreme log price changes are trimmed at both ends of the distribution. Trimming is performed separately within each variety: the upper and lower quantile thresholds are computed on the distribution of absolute log price changes within the item, at the $(1 - p)$ and p percentiles respectively, where p is the value of the switch. Observations whose absolute log price change falls outside these bounds are dropped. The default value of $p = 0.01$ removes the bottom and top one percent of absolute price changes within each item, eliminating a small number of observations that likely reflect data entry errors, unit-change corrections not captured by the quantity plausibility step, or other recording artifacts. Trimming is applied at the item level rather than globally to account for the substantial heterogeneity in the typical magnitude of price changes across product categories.

5 Case Studies

This section presents two illustrative sample constructions that bracket the range of choices researchers are likely to face. The first retains nearly all price quotes to reconstruct official CPI inflation from the microdata. The second applies a set of economically motivated restrictions to construct a sample suited to the analysis of price-setting behavior. The two constructions are not competing approaches to the same question—they are answers to different questions—and the gap between them illustrates the trade-offs inherent in working with CPI microdata.

5.1 Case Study I: Reconstructing the CPI

Sample Definition. The objective of the first case study is to reconstruct official CPI inflation from the microdata. To this end, the full CPI price universe is organized as a longitudinal panel while deviating as little as possible from the official data. Nearly all price quotes are retained, including prices collected via scanner data, index-based prices, and administered prices. Temporary sales, product substitutions, and imputed observations are preserved as they appear in the raw data. Product substitutions are treated as continuous—spell breaks are not introduced at substitution events—consistent with the FSO's implicit treatment of quality-adjusted replacements as ongoing price series. The sample spans the full available period from May 2000 to December 2025.

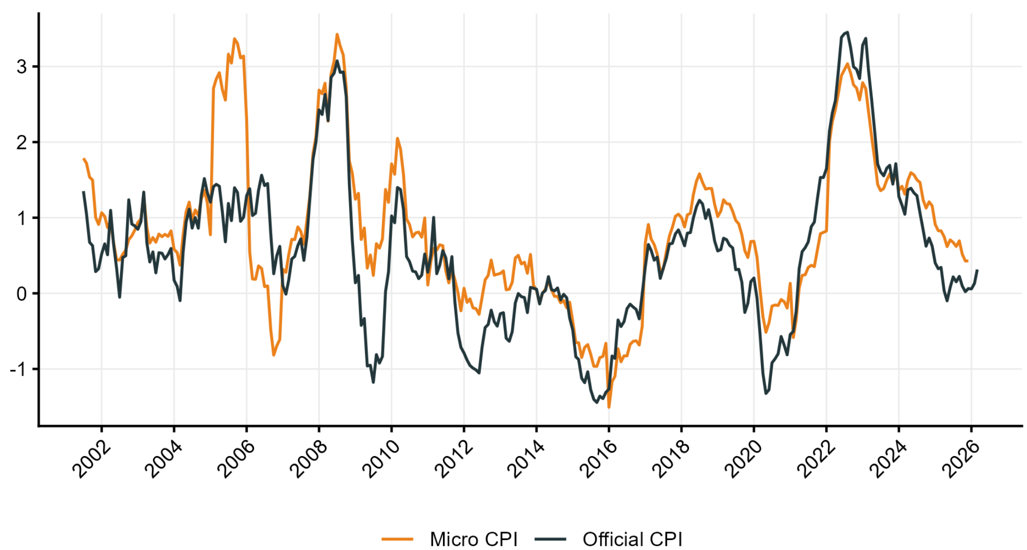
Descriptive Characteristics. Year-on-year inflation is reconstructed by computing monthly price relatives within each product–outlet series, averaging these to the expenditure-item level using official item weights, chaining the resulting weighted index, and computing year-on-year changes. Regional weights are not applied in this aggregation.

Figure 2 compares the reconstructed year-on-year inflation series with the official CPI series published by the FSO over the period May 2001 to December 2025. The two series track each other closely, with a correlation of 0.84, a mean absolute deviation of 0.49 percentage points, and a root mean squared deviation of 0.64 percentage points. The reconstruction captures the main inflation episodes visible in the official series: the low and briefly negative inflation rates of the 2010s, the sharp disinflation following the 2015 exchange rate shock, the deflationary period during COVID-19, and the inflation surge of 2022–2023.

The remaining gap between the two series reflects three distinct sources of approximation. The most important is structural: the official CPI applies a hierarchical weighting procedure that incorporates distribution-channel weights, which are confidential and not released to researchers, so this dimension of the official aggregation cannot be reproduced under any configuration. Second, other differences reflect a deliberate simplification made for consistency across the sample period rather than a data limitation. Regional weights are available in the microdata, but only from January 2006 onward, whereas the reconstruction spans the full period from May 2000; applying them would therefore require switching between regionally weighted and unweighted aggregation partway through the sample, introducing a break in methodology unrelated to genuine price dynamics. Regional weights are therefore omitted throughout, so as to present a single internally consistent series across the full sample period. A regionally weighted variant restricted to the post-2006 period is nonetheless straightforward to construct from the panel for researchers who wish to explore it. A third source lies in the treatment of product substitutions: the reconstruction treats them as continuous using the FSO-provided quality-adjusted linking prices wherever available, which approximates but does not exactly reproduce the FSO's six distinct quality-adjustment procedures (direct replacement, chaining, bridged overlap, option pricing, hedonic adjustment, and non-replacement), several of which rely on information—such as estimated hedonic quality differentials—that is not contained in the confidential microdata. Despite these approximations, the close co-movement of the two series confirms that the panel construction is internally consistent: systematic divergence would indicate errors in the basket merging, weight assignment, or price-change computation steps rather than the sources discussed here.

The gap is somewhat larger in the earlier part of the sample, consistent with several of these factors compounding rather than a single cause. The share of prices collected via scanner data and, later, online collection expanded steadily after 2008, and these more automated channels tend to be more consistently coded in the microdata than earlier collector-recorded quotes, reducing measurement noise in later periods. The FSO's own quality-adjustment and imputation methods were also substantially revised over the sample period—for example, the introduction of hedonic quality adjustment for rents and personal computers around 2010/2011, and of neutral imputation from 2020—and the metadata documenting these treatments is generally more complete for recent vintages. These factors together suggest that the reconstruction should track the official series most closely toward the end of the sample, consistent with the pattern in Figure 2.

Figure 2. Reconstructed versus official CPI inflation



Notes: Year-on-year inflation rates from the reconstructed micro price panel (orange) and the official Swiss CPI (black), May 2001 to December 2025. The reconstructed series chains monthly price relatives within product–outlet series, averages to the expenditure-item level using official item weights, and computes year-on-year changes from the resulting index. The sample retains all price quotes (Case Study I configuration), including administered prices, index-based items, scanner data, and temporary sales. Correlation: 0.84. Mean absolute deviation: 0.49 pp. Root mean squared deviation: 0.64 pp.

Use Cases. This construction is suited to research questions that concern the official CPI itself rather than the behavior of individual price setters. Examples include the study of alternative aggregation methods and their effect on measured inflation, the analysis of index revisions, the comparison of price dynamics across collection modes, and the assessment of how methodological choices—such as the treatment of substitutions—propagate to aggregate price indices. Because this sample imposes minimal restrictions, it is also a natural starting point for researchers who wish to understand the full scope of the data before applying their own exclusions.

5.2 Case Study II: A Sample for Price-Setting Analysis

Sample Restrictions. The second case study constructs a sample suited to studying the price-setting behavior of firms, corresponding closely to the sample used in Rudolf and Seiler (2022), albeit only one of several reasonable configurations for this purpose. The guiding principle is that the sample should contain observations that correspond to decentralized pricing decisions—prices set by individual firms in response to costs, demand, and competitive conditions—and exclude prices that move for reasons unrelated to such decisions.

Several categories of prices are excluded on these grounds. Index-based prices are removed: these are expenditure items for which the microdata record index values rather than true transaction prices (for example rents or books). Because they do not correspond to individual pricing decisions, they contain no information about firm-level pricing behavior. Administered and semi-administered prices—such as health insurance premiums and public transport fares—are excluded for the same reason: their movements are determined by regulatory or contractual processes rather than by market pricing. These two categories together account for a substantial share of the CPI basket by expenditure weight, which highlights the gap between what is measured by the official index and what can be studied using micro price data.

Temporary sales are excluded. Each sale price is replaced by the last observed non-sale price within the spell, so that the measured distribution of price changes reflects deliberate repricing decisions rather than transient promotional discounts.

Price changes at substitution events are treated as spell breaks: when a product replacement occurs, the price series is terminated and a new series is initiated, excluding the quality-adjustment-contaminated price change at the moment of replacement from the analysis. COVID-19-related imputations, identified at the expenditure-item level using FSO press-release information, are treated analogously, preventing the artificial stability of imputed price series from distorting measured price-change frequencies during 2020 and 2021.

Finally, observations with extreme price changes are trimmed at the one-percentile level of the absolute log price change distribution, computed separately within each variety. This removes a small number of observations that likely reflect data entry errors or unit-change corrections not caught by the quantity plausibility step.

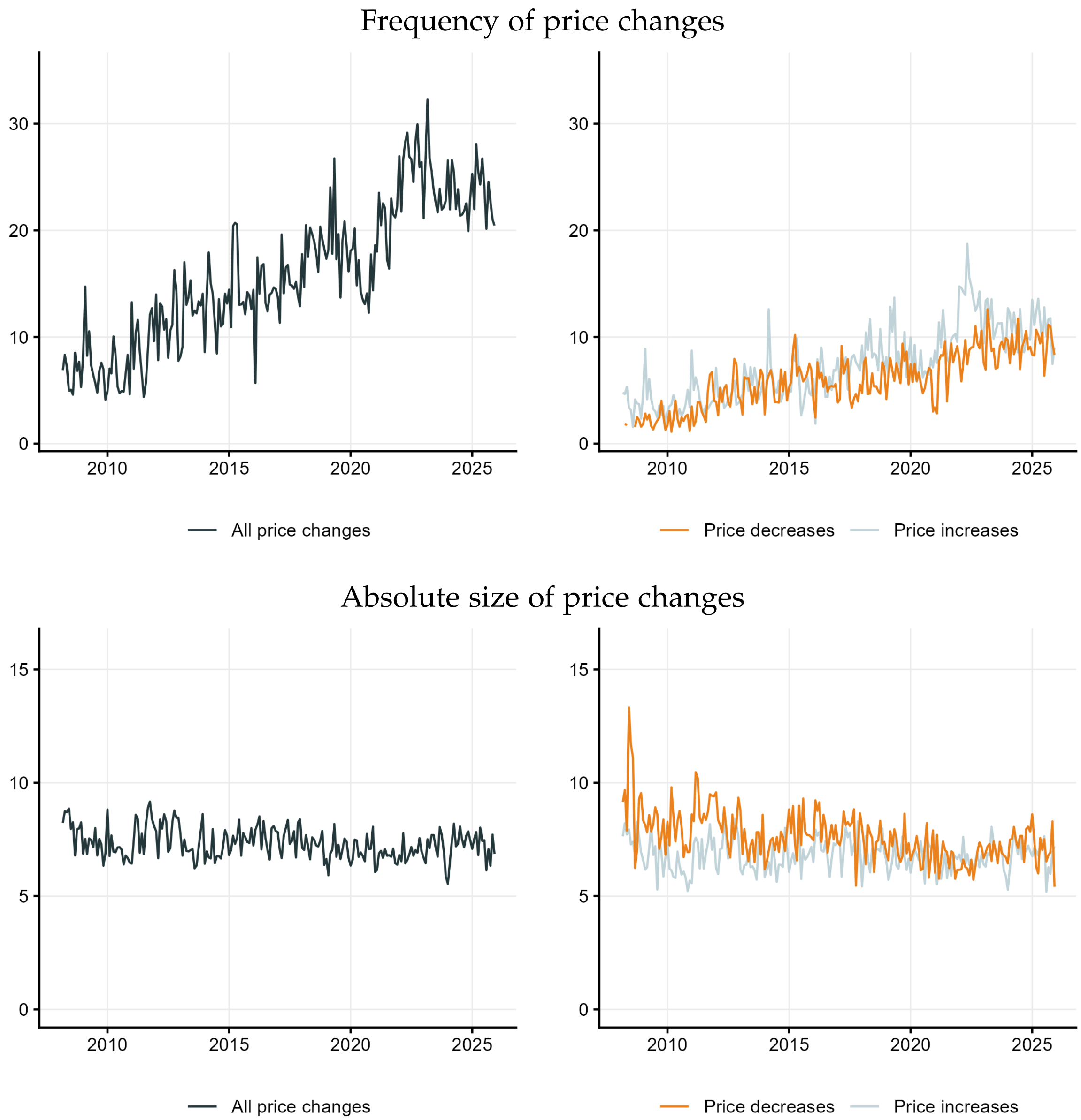
The sample is restricted to the period January 2008 to December 2025. This starting point is motivated by the collection design of the Swiss CPI: prior to 2008, a substantial share of expenditure items was collected at quarterly or lower frequencies, whereas monthly collection became the dominant practice from 2008 onward. Restricting the sample to the post-2008 period therefore improves comparability across products and over time when analyzing monthly price-setting behavior.

Descriptive Characteristics. The sample covers a substantial but incomplete share of the CPI basket. The excluded categories—index-based and administered prices—account in aggregate for roughly half of the total expenditure weight, leaving a sample that represents the market-determined price component of the Swiss consumption basket.

Figure 3 summarizes the main price-setting moments from this sample over the period January 2008 to December 2025. The top panels show the frequency of price changes—the fraction of item-month observations in which a non-zero price change is recorded—both overall and decomposed into price increases and decreases. The bottom panels show the absolute size of price changes, again overall and decomposed by direction. All statistics are computed as expenditure-weighted medians across expenditure items and are expressed as percentages per month.

Several features of the Swiss price-setting environment are visible in these series; a more extensive analysis is provided in Rudolf and Seiler (2022). The overall frequency of price changes increased slightly but steadily over the sample period,

Figure 3. Price-setting moments from CPI microdata



Notes: Frequency (top panels) and absolute size (bottom panels) of price changes, January 2008 to December 2025. Left panels show overall moments; right panels decompose into price increases and decreases. All statistics are expenditure-weighted medians across expenditure items, expressed in percent per month. The sample corresponds to the Case Study II configuration: index-based prices, administered prices, and temporary sales are excluded, and price changes at product substitution events are treated as spell breaks.

rising from below 10 percent in 2008 to more than 20 percent in 2025. Both the frequency of price increases and decreases contribute to the upward trend, though the frequency of price increases exhibits larger high-frequency movements. In addition to this trend, there is visible cyclical variation, most prominently around the inflation episodes of 2022 and 2023, consistent with findings in the broader literature that the frequency of price changes rises during periods of high aggregate inflation (Alvarez et al., 2019; Montag and Villar, 2022; Seiler, 2025). The absolute size of price changes is remarkably stable over time, fluctuating around 7 percent per month throughout the sample period. Price decreases are larger in absolute terms than price increases, a pattern that is persistent across the full sample horizon.

Use Cases. This sample is suited to a wide range of research questions in the price-rigidity and price-setting literature: estimating the frequency and size of price changes at various levels of aggregation, studying heterogeneity in price flexibility across sectors and product types, analyzing the asymmetry between price increases and decreases, and examining how price-setting behavior responds to aggregate shocks. The configurable treatment switches allow researchers to assess sensitivity to key choices—in particular, the treatment of sales and substitutions—without reconstructing the pipeline from scratch.

6 Reproducible Workflow and R Repository

All data-processing steps described in this paper are implemented in an open-source R repository that accompanies the publication and is available at `https://github.com/pascalseiler/swiss-cpi-microdata`. The repository is designed around three objectives: reproducibility of the CPI micro price panel and all results reported in this paper, transparency of every design choice, and extensibility to new research questions.

6.1 Structure

The repository mirrors the logical sequence of the pipeline described in Section 4 and depicted in Figure 1. It is organized into eight directories corresponding to the main processing stages: raw data import and cleaning (`R/10_raw_import`), basket and classification merging (`R/20_classifications`), weight assignment (`R/30_weights`), identification of price series (`R/40_identification`), plausibility checks (`R/50_plausibility`), treatment of temporary sales (`R/60_sales`), computation of price changes (`R/70_prices`), and sampling decisions (`R/80_sampling`). A master script (`main.R`) calls each module in sequence. A single configuration file (`R/00_config.R`) defines the sample period, file paths, and all treatment switches; this is the only file a researcher needs to modify to change the sample definition or switch between case study constructions. An additional directory (`R/90_case_studies`) contains the scripts that produce the figures and statistics for the two case studies presented in Section 5.

Input files are organized into two subdirectories. `input/confidential` holds the raw FSO microdata files and basket documentation that cannot be redistributed and must be obtained from the FSO under a data-use agreement. `input/public` holds the publicly available concordance tables, weight files, and classification mappings that are included directly in the repository. All output files are written to a designated output directory and are not version-controlled.

6.2 Reproducibility

Running `main.R` with the raw FSO files placed in the correct input directory constructs the full CPI micro price panel and reproduces all tables and figures reported in this paper. The pipeline is fully deterministic: no random seeds or stochastic steps are involved. A separate configuration file is provided for each case study (`R/90_case_studies/config_case_I.R` and `config_case_II.R`), allowing users to switch between constructions by sourcing the relevant configuration before running the master script. The R package versions used in the analysis are documented in a `renv` lockfile included in the repository, ensuring that the computational environment can be reconstructed exactly. Approximate running time and storage requirements are documented in the repository `README`.

6.3 Extensibility

The modular structure makes it straightforward to adapt the pipeline to new research questions. Treatment switches can be adjusted to include or exclude specific price categories, change the substitution treatment, or alter outlier thresholds, without modifying any pipeline script. New classification layers—such as additional concordances to external product taxonomies or firm-level identifiers—can be added as new modules in the classification stage. The sample period is controlled by two parameters in `R/00_config.R`. Researchers who wish to construct samples that differ substantially from the two case studies presented here—for instance, a regional price panel or a sample restricted to specific COICOP divisions—can do so by adding a sampling script to `R/80_sampling`, following the structure of the existing module.

Extending the pipeline involves two qualitatively different kinds of updates. The first is the routine (intra-)annual addition of a new data vintage within an existing basket revision period—for example, incorporating the raw price files for a newly released year. Provided the raw data structure, variable naming, and basket documentation format remain unchanged from the preceding year, this requires no changes to the pipeline code: the researcher only needs to update the sample-period parameters in `R/00_config.R` and place the new year's files in the confidential input directories, and the existing modules process the addition automatically.

The second, more involved kind of update follows a full basket revision, which the FSO undertakes roughly every five years (e.g., 2000, 2005, 2010, 2015, 2020, and 2025). Basket revisions typically introduce new varieties and expenditure items, discontinue existing ones, remap items across expenditure items or product groups, and update expenditure weights, regional weights, and occasionally collection frequencies (FSO, 2022). Because such revisions can change the codes, structure, or granularity of the classification system, they generally require researcher-supervised updates to the classification and weight files in `input/public/`—for example, adding a new revision sheet to `item_weights.xlsx` and `region_weights.xlsx`, and updating `mapping_basket_COICOP.xlsx`—together with manual review of how discontinued or newly introduced varieties should be treated at the transition, analogous to the treatment already implemented for ordinary product substitutions (Section 4.4). Where the FSO also changes the structure or variable definitions of the raw confidential files themselves, corresponding adjustments to the raw-import module (`R/10_raw_import`) would additionally be required. The pipeline's separation of fixed processing steps from configurable treatment choices and classification inputs is designed with this in mind: a basket revision changes the input files that feed the pipeline, not the core processing logic.

The pipeline and the results reported here are based on microdata access through the December 2025 vintage; extending the panel to later vintages or future basket revisions requires corresponding data access under a data-use agreement with the FSO, which the modular architecture described above is designed to accommodate for any researcher who undertakes it.

7 Conclusion

This paper documents the microdata underlying the Swiss Consumer Price Index and the procedures required to transform them into research-ready longitudinal price panels. The goal has been to make explicit what is typically left implicit: the structure of the raw data as delivered by the statistical office, the sequence of decisions required to construct a usable panel, and the sensitivity of analytical samples to those decisions. The three contributions are complementary. The research-oriented data description in Section 3 provides a reference that is independent of any particular application, covering the features of the data that any researcher using them will need to understand. The pipeline documentation in Section 4 translates that description into an operational procedure, making each step and each choice traceable. The R repository makes the procedure executable and therefore genuinely reproducible—not in the weaker sense that a sufficiently determined researcher could reconstruct it from written documentation, but in the stronger sense that it can be run directly, inspected in detail, and modified with full visibility of the consequences.

The two case studies serve a dual purpose. They validate the pipeline by showing that the reconstructed CPI universe tracks the official index closely. And they illustrate the practical stakes of the design choices documented in the paper: the sample used for price-setting analysis covers a substantially different share of the consumption basket than the CPI universe, and the measured moments of price adjustment depend meaningfully on whether sales and substitutions are retained or removed. These are not merely technical details—they bear directly on the economic interpretation of findings and on comparability across studies.

The approach developed here is transferable. CPI microdata in other countries share the core challenges described in this paper: the absence of longitudinal identifiers in the raw files, product turnover and quality adjustment, heterogeneous collection methods, and confidentiality constraints that rule out direct data sharing. A pipeline architecture that separates fixed processing steps from configurable treatment choices, and that makes the configuration explicit and version-controlled, is a general solution to these challenges. The present paper is intended as a contribution to that emerging methodological standard—one that treats the construction of research datasets from official statistics not as a preliminary chore to be summarized in a footnote, but as a scientific activity in its own right, subject to the same norms of transparency and reproducibility that apply to the analyses conducted on those datasets.

8 Declarations

8.1 Data availability

The R code and all public input files required to run the pipeline are openly available in the accompanying repository at https://github.com/pascalseiler/swiss-cpi-microdata. The confidential microdata underlying the Swiss Consumer Price Index are not included in the repository and cannot be redistributed. Researchers wishing to replicate the results must obtain the microdata directly from the Swiss Federal Statistical Office under a data-use agreement.

8.2 Competing interests

The authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

8.3 Funding

This research did not receive funding.

8.4 Ethics declaration

Not applicable.

8.5 Author contributions

Pascal Seiler conceived and implemented the data pipeline, conducted the analysis, and drafted the manuscript. Barbara Rudolf contributed to the conceptual design of the data-processing approach and provided guidance on the underlying microdata, drawing on the shared research pipeline developed across several joint projects using the Swiss CPI microdata. Both authors read and approved the final manuscript.

8.6 Acknowledgments

We would like to thank Corinne Becker and Hans-Markus Herren of the Swiss Federal Statistical Office for their valuable help with the data. The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the authors. They do not necessarily reflect the views of the Swiss National Bank (SNB). The SNB takes no responsibility for any errors or omissions in, or for the correctness of, the information contained in this paper.

References

Alvarez, Fernando, Martin Beraja, Martin Gonzalez-Rozada, and Pablo Andrés Neumeyer (2019) "From hyperinflation to stable prices: Argentina's evidence on menu cost models," *The Quarterly Journal of Economics*, Vol. 134, pp. 451–505.

Bils, Mark and Peter J Klenow (2004) "Some evidence on the importance of sticky prices," *Journal of Political Economy*, Vol. 112, pp. 947–985.

Dedola, Luca, Erwan Gautier, Chiara Osbat, and Sergio Santoro (2025) "Price stickiness in the euro area," in *Research Handbook on Inflation*: Edward Elgar Publishing, pp. 265 – 292.

Dhyne, Emmanuel, Luis J Alvarez, Hervé Le Bihan, Giovanni Veronese, Daniel Dias, Johannes Hoffmann, Nicole Jonker, Patrick Lunnemann, Fabio Rumler, and Jouko Vilmunen (2006) "Price changes in the euro area and the United States: Some facts from individual consumer price data," *Journal of Economic Perspectives*, Vol. 20, pp. 171–192.

FSO (2016) *Swiss Consumer Price Index (December 2015=100). Methodological foundations*, No. 855-1500, Neuchâtel: Swiss Federal Statistical Office (FSO).

——— (2021) "The 2020 revisions of price statistics and introduction of an official property price index," Press conference, 26 November 2020.

——— (2022) *Swiss Consumer Price Index (December 2020=100). Methodological foundations*, No. 855-2000, Neuchâtel: Swiss Federal Statistical Office (FSO).

Gautier, Erwan, Cristina Conflitti, Riemer P Faber, Brian Fabo, Ludmila Fadejeva, Valentin Jouvanceau, Jan-Oliver Menz, Teresa Messner, Pavlos Petroulas, Pau Roldan-Blanco et al. (2024) "New facts on consumer price rigidity in the euro area," *American Economic Journal: Macroeconomics*, Vol. 16, pp. 386–431.

Kaufmann, Daniel (2009) "Price-setting behaviour in Switzerland: Evidence from CPI micro data," *Swiss Journal of Economics and Statistics*, Vol. 145, pp. 293–349.

Klenow, Peter J and Oleksiy Kryvtsov (2008) "State-dependent or time-dependent pricing: Does it matter for recent US inflation?" *The Quarterly Journal of Economics*, Vol. 123, pp. 863–904.

Klenow, Peter J and Benjamin A Malin (2010) "Microeconomic evidence on price-setting," in *Handbook of Monetary Economics*, Vol. 3: Elsevier, pp. 231–284.

Luo, Shaowen and Daniel Villar (2025) "Micro evidence: United States," in *Research Handbook on Inflation*: Edward Elgar Publishing, p. 240–264.

Montag, Hugh and Daniel Villar (2022) "Price-setting during the Covid era," Working Paper 547, U.S. Bureau of Labor Statistics.

Nakamura, Emi and Jón Steinsson (2008) "Five facts about prices: A reevaluation of menu cost models," *The Quarterly Journal of Economics*, Vol. 123, pp. 1415–1464.

——— (2013) "Price rigidity: Microeconomic evidence and macroeconomic implications," *Annual Review of Economics*, Vol. 5, pp. 133–163.

Rudolf, Barbara and Pascal Seiler (2022) "Price setting before and during the pandemic: Evidence from Swiss consumer prices," Working Paper 2748, European Central Bank.

Seiler, Pascal (2025) "Nonlinearities of Monetary Policy across States of Price Rigidity," Mimeo, ETH Zurich.