## **Crosswalk Methodology**

### **Overview**

We constructed a reproducible, end-to-end pipeline to crosswalk Brazilian CBO occupation codes to O*NET-SOC 2019 codes. This five-stage mapping sequence consists of: CBO → ISCO-08 → SOC-2010 → SOC-2018 → O*NET-SOC 2019, followed by appending job titles and descriptions. Each stage was executed using Python and pandas, ensuring consistent formatting, rigorous logging, validation of join integrity, and preservation of mappings.

### **1. Data Sources**

* **CBO → ISCO-08**: cbo-isco-conc.csv ([official source)](https://econweb.ucsd.edu/muendler/download/brazil/cbo/cbo-isco-conc.csv)
* **ISCO-08 → SOC-2010**: isco\_soc\_crosswalk - ISCO-08 to 2010 SOC.csv ([official source](https://www.bls.gov/soc/isco_soc_crosswalk.xls))
* **SOC-2010 → SOC-2018**: soc\_2010\_to\_2018\_crosswalk.xlsx ([official source](https://www.bls.gov/soc/2018/soc_2010_to_2018_crosswalk.xlsx))
* **SOC-2018 → O\*NET-SOC 2019**: 2019\_to\_SOC\_Crosswalk.csv ([official source](https://www.onetcenter.org/taxonomy/2019/soc/2019_to_SOC_Crosswalk.csv?fmt=csv))
* **Occupation Titles & Descriptions**: Occupation Data.xlsx ([official source](https://www.onetcenter.org/dl_files/database/db_29_3_excel/Occupation%20Data.xlsx))

#### **A. Approach**

This subsection details the raw data utilized and the foundational steps taken to prepare it for accurate mapping. The integrity of the final crosswalk is heavily dependent on the quality and appropriate handling of these initial datasets.

The following raw data files were identified and utilized for the crosswalk construction:

* **CBO to ISCO08:** The cbo-isco-conc.csv file 3 served as the primary source for the initial mapping from Brazilian CBO codes to ISCO-08. This file contains the cboid (CBO identifier), cbotrans (CBO translated title), iscoid (ISCO-08 identifier), and iscodesc (ISCO-08 description) fields, which are essential for this foundational stage.
* **ISCO08 to SOC2010:** The isco\_soc\_crosswalk - ISCO-08 to 2010 SOC.csv file 3 provided the mapping from ISCO-08 to the 2010 version of the U.S. SOC system. Key columns from this file include ISCO-08 Code, ISCO-08 Title EN (English title), 2010 SOC Code, and 2010 SOC Title.
* **SOC2010 to SOC2018:** The soc\_2010\_to\_2018\_crosswalk.xlsx - Sorted by 2010.csv file 1 was used to transition from SOC2010 to the updated SOC2018 classification. This file contains 2010 SOC Code, 2010 SOC Title, 2018 SOC Code, and 2018 SOC Title.
* **SOC2018 to O\*NET-SOC 2019 and Final Descriptions:** The 2019\_to\_SOC\_Crosswalk.csv file 3 provided the direct mapping from SOC2018 to O*NET-SOC 2019 codes and titles. To enrich the final O*NET-SOC 2019 codes with their detailed descriptions, the Occupation Data.xlsx - Occupation Data.csv file 2 was utilized. It is important to note that snippets 2, and 2 refer to the same underlying data, with 2 providing the comprehensive descriptions.

### **2. Cleaning and Standardization**

Across all datasets:

* Code columns were stripped of whitespace, converted to string types, and padded with leading zeros where required (e.g., ISCO codes to 4 digits).
* Missing values were standardized as NaN.
* Only relevant columns were retained (e.g., code mappings), while metadata like comments or extra descriptors were dropped.
* Text fields (titles/descriptions) were cleaned to remove trailing spaces, non-standard characters, and ensure uniform encoding.

### **3. Mapping Pipeline**

#### **Stage 1: CBO → ISCO-08**

* Standardized cboid and ISCO08 columns.
* Filtered out rows with missing ISCO codes (~15% unmatched).
* Verified one-to-one structure and performed a left join.

#### **Stage 2: ISCO-08 → SOC-2010**

* Renamed and standardized ISCO and SOC2010 codes.
* Performed a left join to preserve unmatched rows.
* Accepted one-to-many mappings.
* Logged match rate (~82–85%).

#### **Stage 3: SOC-2010 → SOC-2018**

* Left-joined on cleaned SOC2010 values.
* Managed one-to-many mappings.
* Match rate monitored and flagged if below 90%.

#### **Stage 4: SOC-2018 → O\*NET-SOC 2019**

* Cleaned SOC2018 and ONET codes.
* Left-joined and tracked matches.

#### **Stage 5: Append Title & Description**

* Joined on ONET-SOC 2019 Code.
* Cleaned title and description fields to ensure consistent formatting.

### **4. Validation, Audit & Deduplication**

* Each join was validated using pandas join integrity checks.
* We introduced Boolean flags to indicate whether each mapping stage succeeded.
* For one-to-many joins, all rows were retained. Duplicates introduced during expansion were addressed by removing only exact duplicates, ensuring no valid mappings were lost.
* Intermediate statistics were logged for transparency.
* The final dataset can be filtered to retain only complete crosswalks (i.e., rows with non-null values in all mapping columns).

### **5. Final Output**

The final crosswalk includes:

* cboid: Brazilian CBO code
* ISCO08: 4-digit ISCO-08 code
* SOC2010: 2010 U.S. SOC code
* SOC2018: 2018 U.S. SOC code
* O\*NET-SOC 2019 Code: Latest U.S. O\*NET occupation code
* Title, Description: O\*NET job title and description

Data was exported as CSV and JSON (record-oriented) to support various analytical tools. Character encoding and code formats were preserved.

### **6. Limitations**

* Some CBO codes do not map to ISCO or SOC equivalents.
* One-to-many mappings result in data expansion; use caution in downstream analysis.
* The data sources we used were the most complete and compatible available on the internet at the time of this research.