**Metodologia.**

**Dados.**

Para analisar os padrões de trabalhadores por conta própria, coletamos dados da Pesquisa Nacional por Amostragem de Domicílio Contínua (PnadC) dos anos de 2023 e 2024. A pesquisa é realizada trimestralmente e acompanha indivíduos ao longo de 5 trimestres consecutivos, sendo que todos em todos os períodos tem um grupo amostral iniciando e finalizando a pesquisa. Dessa forma, ao longo de 8 trimestres, 4 grupos amostrais completam a pesquisa.

Para o desenho deste estudo, selecionamos os indivíduos que exerceram ocupação de trabalho por conta própria em alguma das entrevistas e que tivessem idade entre 14 e 64 anos, isso resultou em uma amostra de 47.449 pessoas. A variável no qual a ocupação é analisada é V4012, cujo tem sete categorias de ocupações e a não se aplica caso o indivíduo esteja desocupado ou fora da força de trabalho. Para simplificação da análise, algumas transformações foram aplicadas, sendo elas: agregação do trabalhador doméstico como empregado do setor privado e exclusão das ocupações trabalhador familiar não remunerado, empregado do setor público e militares (será que precisa justificar melhor?). O quadro abaixo sintetiza as transformações mencionadas.

|  |  |  |
| --- | --- | --- |
| **Antes do tratamento** | **Após o tratamento** | **Sigla** |
| Empregado do setor privado | Empregado do setor privado | ESP |
| Trabalhador doméstico |
| Trabalhador por conta própria | Trabalhador por conta própria | TCP |
| Empregador | Empregador | EMP |
| Fora da força de trabalho | Pessoa fora de trabalho | PFT |
| Desocupado |
| Trabalhador familiar não remunerado | Excluídos | |
| Empregado do setor público (inclusive empresas de economia mista) |
| Militar do exército, da marinha, da aeronáutica, da polícia militar ou do corpo de bombeiros militar |

Fonte: elaborado pelos autores.

Os tratamentos implicaram em uma redução da amostra para 43.527. Os indivíduos foram identificados através da definição da variável id resultado da junção das seguintes variáveis: Unidade Primária de Amostragem (UPA), número de seleção do domicílio, grupo de amostra, estrato, dia, mês e ano de nascimento e sexo.

**Análise de Sequência**

A Análise de Sequência é realizada seguindo as três etapas definidas pelos autores Abbott e Tsay (2000): codificação de narrativas ou processos como sequências, medição de dissimilaridades em pares entre sequências e redução de dados. Para a primeira etapa são necessários três elementos, primeiro as unidades observacionais, que são representadas pelos indivíduos através da identificação id, segundo os pontos no tempo, identificado pelo número da entrevista do total de 5 e os estados, representada pelas ocupações definidas no quadro anterior (Liao et al., 2025).

Na segunda etapa, medição de dissimilaridades, é realizada para cada par distinto de indivíduos. O objetivo é quantificar o quanto duas sequências ou narrativas são diferentes. É aplicada a técnica Optimal Matching com valor 2 para substituições e 1,5 para inserções e exclusões de estados, o cálculo é realizado minimizando o custo de tornar uma sequência igual a outra. Para ilustrar esse processo, considere as seguintes trajetórias:

* Sequência A: TCP, TCP, ESPR, TCP, TCP;
* Sequência B: PFT, TCP, ESPR, ESPR, ESP;
* Sequência C: TCP, TCP, TCP, TCP, TCP.

Os únicos pontos no tempo em que não houve correspondência entre as sequências A e B são a primeira, quarta e quinta entrevista, totalizando um custo igual a 6. Enquanto, para tornar sequências A e C iguais, tem um custo igual a 2, e portanto são mais similares. O resultado desse processo é uma matriz de dissimilaridades de dimensões n x n, em que o n é o tamanho da amostra. Tanto a codificação, quanto a medição de dissimilaridades são executadas usando a biblioteca TraMineR em linguagem R.

A medição da dissimilaridade é especialmente importante para a etapa posterior denominada redução de dados. Empregamos o método de clusterização em que o objetivo é identificar padrões de trajetórias classificando-os em grupos de forma de forma que elas sejam o mais semelhante possível dentro do cluster e dissimilar entre os grupos (Liao et al., 2025). Primeiro é aplicada a técnica de redução de dimensionalidade com o Escalonamento Multidimensional Clássico (MDS) sobre a matriz de dissimilaridades, reduzindo-a para duas dimensões. E por fim, a clusterização é realizada através do K-means, definindo três clusters. Esta última etapa é executada usando funções nativas do R Studio.