



# Metodologías estadísticas para el análisis de la asociación entre dos estructuras de datos de tres vías

Claudia Lorena Montes Mora

Universidad del Valle  
Facultad de Ingeniería  
Escuela de Estadística  
Santiago de Cali  
2024

# Metodologías estadísticas para el análisis de la asociación entre dos estructuras de datos de tres vías

**Claudia Lorena Montes Mora**

Trabajo de grado presentado como requisito parcial para optar al  
título de:

**Magister en Estadística**

Director:

Ph. D. Jaime Mosquera Restrepo

Universidad del Valle  
Facultad de Ingeniería  
Escuela de Estadística  
Santiago de Cali, Colombia  
2024

## Dedicatoria

Con cariño a mis familiares y amigos,  
especialmente a mi madre, Ana, por apoyarme  
incondicionalmente, por el amor y por ser mi  
fuerza para culminar este proceso.



# Agradecimientos

Expreso mi profundo agradecimiento a todos aquellos que han sido parte fundamental de mi travesía académica y la culminación de esta maestría.

A mis respetados profesores de la Escuela de Estadística, agradezco sinceramente su orientación, paciencia y dedicación en compartir sus conocimientos, especialmente al profesor Jaime Mosquera por sus valiosos aportes, los cuales han enriquecido enormemente este trabajo de investigación, guiándome hacia un camino de aprendizaje significativo.

A mis compañeros, agradezco el intercambio de ideas y el apoyo mutuo que hemos compartido a lo largo de este viaje. Juntos hemos enfrentado desafíos y celebrado logros, creando recuerdos que atesoraré siempre.

A mi familia y amigos, les agradezco por su constante aliento y comprensión. Su apoyo incondicional ha sido mi mayor motivación, impulsándome a alcanzar este importante hito en mi vida académica y ser parte de este capítulo significativo de mi vida.

Este logro no solo representa el resultado de mi esfuerzo individual, sino también el resultado del esfuerzo colectivo de todos aquellos que me han rodeado. Estoy profundamente agradecida por la oportunidad de aprender y crecer en este entorno académico enriquecedor.



## Resumen

La necesidad de analizar datos de múltiples dimensiones ha dado lugar al desarrollo de metodologías para el análisis de estas estructuras. Frecuentemente, los datos se organizan en estructuras bidimensionales que implican dos vías de análisis (individuos y variables), y se pueden representar fácilmente mediante matrices. Las estructuras tridimensionales añaden una tercera dimensión (tiempo, lugar, condición, entre otras), generando conjuntos de datos más complejos que pueden organizarse como cubos de datos e incluso se puede extender a arreglos de más de  $N$ -vías.

En la actualidad, con el crecimiento de los datos, es frecuente enfrentarnos a estructuras de tres vías, especialmente en áreas de investigación como las ciencias sociales y económicas, biología, quimiometría, ingeniería, entre otras áreas. Las metodologías para el análisis de estas estructuras ha tenido un crecimiento en los últimos años y brinda gran utilidad para identificar patrones entre los datos y las diferentes vías que compone la estructura. Sin embargo, en algunas ocasiones surge la necesidad de analizar dos estructuras de datos en tres vías y la asociación entre ellas.

Este trabajo se centra en la revisión de metodologías para el análisis de datos organizados en estructuras de tres vías, explorando métodos como Análisis Factorial Múltiple (AFM), Mínimos Cuadrados Parciales (sus siglas en inglés PLS) y Mínimos Cuadrados Parciales Multivía ( $N$ -PLS) para analizar simultáneamente dos estructuras de datos tridimensionales. Para la aplicación de las metodologías toma como caso de estudio datos del Producto Interno Bruto - PIB y el Índice de Pobreza Multidimensional - IPM en los departamentos de Colombia entre los años 2018 y 2021, buscando identificar la técnica más eficaz para estudiar la relación entre estas dos estructuras de datos. Al realizar la comparación se evidencia que las técnicas empleadas para el análisis de las dos estructuras de datos de tres vías ofrecen resultados comparables, a pesar de tener finalidades diferentes.

**Palabras clave:** Análisis multivariante, Análisis de datos en dos vías, Análisis de datos en tres vías, Asociación de dos arreglos de datos, Asociación de dos arreglos de dato de tres vías, AFM, PLS, N-way PLS.

# Abstract

The need to analyze data from multiple dimensions has led to the development of methodologies for analyzing these structures. Frequently, data is organized into two-dimensional structures involving two avenues of analysis (individuals and variables), and can be easily represented using matrices. Three-dimensional structures add a third dimension (time, place, condition, among others), generating more complex data sets that can be organized as data cubes and can even be extended to arrays of more than N-ways.

Currently, with the growth of data, it is common to encounter three-way structures, especially in research areas such as social sciences, biology, chemometrics, engineering, among others. The methodologies for analyzing these structures have seen growth in recent years and provide great utility for identifying patterns among the data and the different avenues that compose the structure. However, sometimes there arises the need to analyze two data structures in three ways and the association between them.

This work focuses on reviewing methodologies for analyzing data organized in three-way structures, exploring methods such as Multiple Factor Analysis (MFA), Partial Least Squares (PLS), and Multivariate Partial Least Squares (N-PLS) to simultaneously analyze two three-dimensional data structures. For the application of the methodologies, data from Gross Domestic Product (GDP) and the Multidimensional Poverty Index (MPI) in the departments of Colombia between 2018 and 2021 are taken as a case study, aiming to identify the most effective technique for studying the relationship between these two data structures. Upon comparison, it is evident that the techniques employed for analyzing the two three-way data structures offer comparable results, despite having different purposes.

**Keywords:** Multivariate analysis, Two-way data analysis, Three-way data analysis, Association of two data arrays, Association of two three-way data arrays, AFM, PLS, N-way PLS.

# Contenido

|                                                                                      |           |
|--------------------------------------------------------------------------------------|-----------|
| <b>1. Introducción</b>                                                               | <b>1</b>  |
| <b>2. Planteamiento del problema</b>                                                 | <b>3</b>  |
| 2.1. Definición del problema . . . . .                                               | 3         |
| 2.2. Justificación . . . . .                                                         | 6         |
| 2.3. Objetivos . . . . .                                                             | 9         |
| 2.3.1. Objetivo General . . . . .                                                    | 9         |
| 2.3.2. Objetivos Específicos . . . . .                                               | 9         |
| <b>3. Revisión de antecedentes</b>                                                   | <b>10</b> |
| 3.1. Análisis bibliométrico . . . . .                                                | 11        |
| 3.1.1. Análisis de impacto de autores e instituciones más relevantes . . . . .       | 12        |
| 3.1.2. Volumen de citas . . . . .                                                    | 14        |
| 3.1.3. Redes de colaboración y Cocitación de autores . . . . .                       | 15        |
| 3.2. Análisis de datos de dos vías . . . . .                                         | 18        |
| 3.3. Análisis de la asociación entre dos tablas de datos de dos vías . . . . .       | 20        |
| 3.4. Análisis de datos de tres vías . . . . .                                        | 23        |
| 3.5. Análisis de la asociación entre dos estructuras de datos de tres vías . . . . . | 28        |
| <b>4. Marco Teórico</b>                                                              | <b>30</b> |
| 4.1. Análisis de estructura de datos en dos vías . . . . .                           | 30        |
| 4.1.1. Análisis de Componentes Principales - ACP . . . . .                           | 31        |
| 4.1.2. Análisis Factorial Múltiple - AFM . . . . .                                   | 36        |
| 4.2. Análisis de la asociación entre estructuras de datos en dos vías . . . . .      | 44        |
| 4.2.1. Mínimos Cuadrados Parciales - PLS . . . . .                                   | 44        |
| 4.2.2. PLS1 . . . . .                                                                | 46        |
| 4.2.3. PLS2 . . . . .                                                                | 50        |
| 4.3. Análisis de estructura de datos en tres vías . . . . .                          | 56        |
| 4.3.1. Multiway - ACP (MPCA) . . . . .                                               | 56        |
| 4.3.2. PARAFAC . . . . .                                                             | 60        |
| 4.3.3. STATIS . . . . .                                                              | 64        |
| 4.4. Análisis de la asociación entre estructuras de datos en tres vías . . . . .     | 67        |
| 4.4.1. Unfold-PLS . . . . .                                                          | 67        |
| 4.4.2. N-PLS . . . . .                                                               | 68        |

|                                                                                                                   |            |
|-------------------------------------------------------------------------------------------------------------------|------------|
| 4.4.3. Otras aproximaciones . . . . .                                                                             | 70         |
| <b>5. Metodología</b>                                                                                             | <b>74</b>  |
| 5.1. Descripción de las estructuras de datos . . . . .                                                            | 75         |
| 5.2. Formulación de alternativas de despliegue para el análisis de dos estructuras de datos de tres vías. . . . . | 76         |
| 5.2.1. Alternativa de despliegue vertical . . . . .                                                               | 77         |
| 5.2.2. Alternativa de despliegue horizontal . . . . .                                                             | 77         |
| 5.2.3. Alternativa de despliegue combinada . . . . .                                                              | 78         |
| 5.2.4. Aplicación de técnicas . . . . .                                                                           | 78         |
| 5.3. Selección de técnicas para el análisis de dos estructuras de datos de tres vías. . . . .                     | 79         |
| 5.3.1. Análisis Factorial Múltiple (AFM) . . . . .                                                                | 79         |
| 5.3.2. Mínimo Cuadrados Parciales (PLS) . . . . .                                                                 | 81         |
| 5.3.3. $N$ - PLS . . . . .                                                                                        | 82         |
| 5.3.4. Análisis de comparación de los resultados y hallazgos . . . . .                                            | 83         |
| 5.4. Aplicación: Caso de estudio generación de riqueza y condiciones de pobreza en Colombia. . . . .              | 84         |
| 5.5. Software . . . . .                                                                                           | 84         |
| <b>6. Resultados</b>                                                                                              | <b>85</b>  |
| 6.1. Análisis Exploratorio de Datos . . . . .                                                                     | 85         |
| 6.1.1. Indicadores del Producto Interno Bruto . . . . .                                                           | 85         |
| 6.1.2. Indicadores de Pobreza . . . . .                                                                           | 90         |
| 6.2. Análisis bajo una estructura de datos en dos vías . . . . .                                                  | 95         |
| 6.2.1. Análisis Factorial Múltiple (AFM) . . . . .                                                                | 95         |
| 6.2.2. Análisis comparativo de las alternativas a través de AFM . . . . .                                         | 114        |
| 6.3. Asociación entre dos estructuras de datos en dos vías (Despliegue) . . . . .                                 | 116        |
| 6.3.1. Mínimos Cuadrados Parciales (PLS) . . . . .                                                                | 116        |
| 6.3.2. Análisis comparativo de las alternativas a través de PLS . . . . .                                         | 127        |
| 6.4. Asociación entre dos estructuras de datos en tres vías . . . . .                                             | 129        |
| 6.4.1. Mínimos Cuadrados Parciales Multiway ( $N$ -PLS) . . . . .                                                 | 129        |
| 6.5. Análisis Comparativo de los resultados . . . . .                                                             | 134        |
| <b>7. Conclusiones y recomendaciones</b>                                                                          | <b>139</b> |
| 7.1. Conclusiones . . . . .                                                                                       | 139        |
| 7.2. Recomendaciones . . . . .                                                                                    | 140        |
| <b>Bibliografía</b>                                                                                               | <b>142</b> |
| <b>A. Anexo A: Diccionario de datos de las matrices IPM y PIB</b>                                                 | <b>150</b> |
| A.1. Diccionario de datos - Producto Interno Bruto . . . . .                                                      | 151        |

|                                                                          |     |
|--------------------------------------------------------------------------|-----|
| A.2. Diccionario de datos - Índice de Pobreza Multidimensional . . . . . | 152 |
|--------------------------------------------------------------------------|-----|

# Lista de Figuras

|                                                                                                                                |    |
|--------------------------------------------------------------------------------------------------------------------------------|----|
| 3-1. Producción científica por año de técnicas para el análisis de datos de tres vías desde el 2002 al 2023 . . . . .          | 12 |
| 3-2. Fuentes con mayor volumen de producción enfocado a técnicas para el análisis de datos de tres vías . . . . .              | 13 |
| 3-3. Mapa de producción Científica por país enfocado a las técnicas para el análisis de datos de tres vías . . . . .           | 14 |
| 3-4. Países con mayor volumen de citación de artículos sobre el análisis de datos de tres vías . . . . .                       | 14 |
| 3-5. Red de colaboración entre los autores sobre el análisis de datos de tres vías .                                           | 15 |
| 3-6. Red de co-citación entre los autores más influyentes para el análisis de datos de tres vías . . . . .                     | 16 |
| 3-7. Red de colaboración por país sobre el análisis de datos de tres vías . . . . .                                            | 17 |
| 3-8. Red de co-ocurrencia de palabras clave sobre el análisis de datos de tres vías                                            | 17 |
| 4-1. Estructura de datos en dos vías . . . . .                                                                                 | 30 |
| 4-2. Representación de la nube de individuos. ( <a href="#">Lebart et al., 1995</a> ) . . . . .                                | 32 |
| 4-3. Circulo de correlaciones, ( <a href="#">Lebart et al., 1995</a> ) . . . . .                                               | 34 |
| 4-4. Individuos y variables representados conjuntamente, ( <a href="#">González, 2014</a> ) . . . .                            | 35 |
| 4-5. Matriz X conformada por la yuxtaposición de tablas, ( <a href="#">Escofier y Pagès, 1992a</a> )                           | 36 |
| 4-6. Ejemplo de representación de los grupos de variables en el AFM . . . . .                                                  | 38 |
| 4-7. Ejemplo de representación de las variables sobre los ejes factoriales del AFM, <a href="#">Garcia A (2021)</a> . . . . .  | 39 |
| 4-8. Ejemplo de representación de los individuos sobre los ejes factoriales del AFM, <a href="#">Garcia A (2021)</a> . . . . . | 40 |
| 4-9. Ejemplo de representación superpuesta de la nube de individuos en AFM, <a href="#">Garcia A (2021)</a> . . . . .          | 42 |
| 4-10. Obtención de las coordenadas de $w_1$ por proyección de $x_j$ sobre $y$ <a href="#">González (2014)</a> . . . . .        | 48 |
| 4-11. Regresión de $y$ sobre las componentes ortogonales $t_1, t_2$ <a href="#">González (2014)</a> . . .                      | 49 |
| 4-12. Los residuos en $x$ y en $y$ son ortogonales a $t_1$ <a href="#">González (2014)</a> . . . . .                           | 50 |
| 4-13. Estructura de las matrices $\mathbf{X}$ y $\mathbf{Y}$ para PLS2 adaptado de <a href="#">Wold et al. (2001)</a>          | 50 |
| 4-14. Representación geométrica del modelo PLS2 <a href="#">Wold et al. (2001)</a> . . . . .                                   | 51 |
| 4-15. Representación de la correlación entre las variables de $X$ y $Y$ . . . . .                                              | 55 |

|                                                                                                                                                                |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4-16. Representación de las componentes T y U . . . . .                                                                                                        | 55 |
| 4-17. Estructura de datos de tres vías . . . . .                                                                                                               | 56 |
| 4-18. Formas de despliegue de una matriz de datos de tres vías, ( <a href="#">Henrion, 1994</a> ) . . . . .                                                    | 57 |
| 4-19. Diferentes tipos de escalado de datos, ( <a href="#">P. Kroonenberg, 1983</a> ) . . . . .                                                                | 59 |
| 4-20. Diferentes tipos de escalado de datos en MPCA según <a href="#">Henrion (1994)</a> . . . . .                                                             | 60 |
| 4-21. Representación del modelo PARAFAC con R componentes., ( <a href="#">Leardi, 2005</a> ) . . . . .                                                         | 61 |
| 4-22. Modelo PARAFAC de dos componentes para una estructura de datos de tres vías $\mathbf{X}$ , ( <a href="#">Bro-Jorgensen, 1998</a> ) . . . . .             | 62 |
| 4-23. Estructura de datos de tres vías y despliegue con STATIS, ( <a href="#">Stanimirova et al., 2004</a> ) . . . . .                                         | 65 |
| 4-24. Representación de las matrices T y U., ( <a href="#">Hervás et al., 2018</a> ) . . . . .                                                                 | 69 |
| 4-25. Representación de los pesos $W^J$ y $W^K$ ., ( <a href="#">Hervás et al., 2018</a> ) . . . . .                                                           | 70 |
| 4-26. Representación de la segunda vía y tercer vía de la estructura de datos con respecto a $\mathbf{X}$ ., ( <a href="#">Hervás et al., 2018</a> ) . . . . . | 70 |
| 4-27. Esquema de análisis de datos para la técnica CO-STATIS., ( <a href="#">Medina-Hernández et al., 2021</a> ) . . . . .                                     | 71 |
| 4-28. Esquema de análisis de datos para la técnica X-STATIS., ( <a href="#">Medina-Hernández et al., 2021</a> ) . . . . .                                      | 72 |
| 4-29. Esquema de análisis de datos para la técnica Coinercia., ( <a href="#">Medina-Hernández et al., 2021</a> ) . . . . .                                     | 72 |
| 5-1. Fases metodológicas del trabajo de investigación . . . . .                                                                                                | 74 |
| 5-2. Datos PIB . . . . .                                                                                                                                       | 75 |
| 5-3. Datos IPM . . . . .                                                                                                                                       | 75 |
| 5-4. Alternativa de despliegue vertical . . . . .                                                                                                              | 77 |
| 5-5. Alternativa de despliegue horizontal . . . . .                                                                                                            | 77 |
| 5-6. Alternativa de despliegue combinado de la vía 2 (variables) y vía 3 (tiempo) . . . . .                                                                    | 78 |
| 5-7. Tabla conjunta de las estructuras del PIB e IPM bajo la alternativa de despliegue vertical . . . . .                                                      | 80 |
| 5-8. Tabla conjunta de las estructuras del PIB e IPM bajo la alternativa de despliegue horizontal . . . . .                                                    | 80 |
| 5-9. Tabla conjunta de las estructuras del PIB e IPM bajo la alternativa de despliegue horizontal combinada . . . . .                                          | 81 |
| 5-10. Pls aplicado a las estructuras del PIB e IPM bajo la alternativa de despliegue vertical . . . . .                                                        | 82 |
| 5-11. Pls aplicado a las estructuras del PIB e IPM bajo la alternativa de despliegue horizontal . . . . .                                                      | 82 |
| 6-1. Distribución de los indicadores del PIB por año . . . . .                                                                                                 | 87 |
| 6-2. Correlación entre las variables del PIB por año . . . . .                                                                                                 | 89 |
| 6-3. Distribución de los indicadores del IPM por año . . . . .                                                                                                 | 92 |

|                                                                                                                        |     |
|------------------------------------------------------------------------------------------------------------------------|-----|
| 6-4. Correlación entre las variables del IPM por año . . . . .                                                         | 93  |
| 6-5. Correlación entre las variables del IPM y PIB por año . . . . .                                                   | 94  |
| 6-6. Gráfico de sedimentación de valores propios del AFM . . . . .                                                     | 96  |
| 6-7. Representación de los grupos en el análisis global . . . . .                                                      | 98  |
| 6-8. Circulo de correlación de las variables en el análisis global . . . . .                                           | 99  |
| 6-9. Contribución de las variables a la inercia a la primera dimensión . . . . .                                       | 100 |
| 6-10. Contribución de las variables a la inercia a la segunda dimensión . . . . .                                      | 100 |
| 6-11. Representación de los territorios en el análisis global . . . . .                                                | 101 |
| 6-12. Contribución de los territorios a la primera dimensión . . . . .                                                 | 102 |
| 6-13. Contribución de los territorios a la segunda dimensión . . . . .                                                 | 102 |
| 6-14. Gráfico de sedimentación de valores propios del AFM alternativa de despliegue<br>horizontal . . . . .            | 103 |
| 6-15. Representación de los grupos en el análisis global . . . . .                                                     | 104 |
| 6-16. Circulo de correlación de las variables en el análisis global . . . . .                                          | 106 |
| 6-17. Contribución de las variables a la inercia del primer eje factorial . . . . .                                    | 107 |
| 6-18. Contribución de las variables a la inercia del segundo eje factorial . . . . .                                   | 107 |
| 6-19. Representación de los territorios en el análisis global . . . . .                                                | 107 |
| 6-20. Contribución de los territorios a la primer componente . . . . .                                                 | 108 |
| 6-21. Contribución de los territorios a la segunda componente . . . . .                                                | 108 |
| 6-22. Gráfico de sedimentación de valores propios del AFM alternativa de despliegue<br>combinada . . . . .             | 109 |
| 6-23. Representación de los grupos en el análisis global . . . . .                                                     | 111 |
| 6-24. Circulo de correlación de las variables en el análisis global alternativa de des-<br>pliegue combinada . . . . . | 112 |
| 6-25. Contribución de las variables a la inercia del primer eje factorials . . . . .                                   | 112 |
| 6-26. Contribución de las variables a la inercia del segundo eje factorial . . . . .                                   | 112 |
| 6-27. Representación de los territorios en el análisis global . . . . .                                                | 113 |
| 6-28. Contribución de los territorios a la primer componente . . . . .                                                 | 114 |
| 6-29. Contribución de los territorios a la segunda componente . . . . .                                                | 114 |
| 6-30. Circulo de correlaciones del método PLS . . . . .                                                                | 120 |
| 6-31. Nube de individuos en T . . . . .                                                                                | 121 |
| 6-32. Nube de individuos en U . . . . .                                                                                | 121 |
| 6-33. Circulo de correlaciones del método PLS . . . . .                                                                | 125 |
| 6-34. Nube de individuos en T . . . . .                                                                                | 126 |
| 6-35. Nube de individuos en U . . . . .                                                                                | 127 |
| 6-36. Gráfico de puntuación de los dos primeros componentes en la matriz T . . . .                                     | 132 |
| 6-37. Gráfico de puntuación de los dos primeros componentes en la matriz U . . . .                                     | 132 |
| 6-38. Gráfico de ponderaciones de la matriz $W_j$ . . . . .                                                            | 133 |
| 6-39. Gráfico de ponderaciones de la matriz $W_k$ . . . . .                                                            | 133 |

# Lista de Tablas

|                                                                                                                                                  |     |
|--------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3-1. Resumen análisis bibliométrico . . . . .                                                                                                    | 11  |
| 3-2. Indicadores bibliométricos de los autores más relevantes sobre el análisis de estructuras de datos de tres vías . . . . .                   | 13  |
| 3-3. Resumen de técnicas identificadas en los antecedentes . . . . .                                                                             | 29  |
| 4-1. Ejemplo de identificación de variables que más aportan a la construcción de los ejes factoriales, <a href="#">Garcia A (2021)</a> . . . . . | 39  |
| 5-1. Conjunto de variables - Ramas de actividad económica del PIB (Ver Anexo A.1) . . . . .                                                      | 75  |
| 5-2. Conjunto de variables - Indicadores para el IPM (Ver Anexo A.2) . . . . .                                                                   | 76  |
| 5-3. Metodologías aplicadas a las alternativas de despliegue realizadas . . . . .                                                                | 79  |
| 6-1. Estadísticas descriptivas de los indicadores del PIB 2018 - 2021 . . . . .                                                                  | 86  |
| 6-2. Estadísticas descriptivas de los indicadores del IPM 2018 - 2021 . . . . .                                                                  | 91  |
| 6-3. Valores propios del AFM . . . . .                                                                                                           | 96  |
| 6-4. Valores propios y ponderaciones de los ACP separados por grupo . . . . .                                                                    | 97  |
| 6-5. Contribuciones del conjunto de datos del PIB y el IPM a la construcción de los ejes . . . . .                                               | 97  |
| 6-6. Coeficientes Lg y RV . . . . .                                                                                                              | 98  |
| 6-7. Valores propios del AFM alternativa de despliegue horizontal . . . . .                                                                      | 103 |
| 6-8. Valores propios, Coordenadas, contribuciones y ponderaciones del conjunto de datos del PIB y el IPM a la construcción de los ejes . . . . . | 104 |
| 6-9. Coeficientes Lg . . . . .                                                                                                                   | 105 |
| 6-10. Coeficientes RV . . . . .                                                                                                                  | 105 |
| 6-11. Valores propios del AFM alternativa de despliegue combinada . . . . .                                                                      | 109 |
| 6-12. Valores propios, ponderaciones, y contribuciones del conjunto de datos del PIB y el IPM a la construcción de los ejes . . . . .            | 110 |
| 6-13. Correlación entre T y U . . . . .                                                                                                          | 117 |
| 6-14. Coeficientes $R^2$ . . . . .                                                                                                               | 118 |
| 6-15. Estadísticos VIP . . . . .                                                                                                                 | 119 |
| 6-16. Correlación entre T y U . . . . .                                                                                                          | 122 |
| 6-17. Coeficientes $R^2$ . . . . .                                                                                                               | 123 |
| 6-18. Estadísticos VIP . . . . .                                                                                                                 | 124 |

|                                                                                                                    |     |
|--------------------------------------------------------------------------------------------------------------------|-----|
| <b>6-19.</b> Validación Cruzada para la selección del número de componentes . . . . .                              | 130 |
| <b>6-20.</b> Contribuciones a las componentes de las actividades económicas . . . . .                              | 131 |
| <b>6-21.</b> Contribuciones a las componentes del modo tiempo . . . . .                                            | 131 |
| <b>6-22.</b> Valores predichos para algunos indicadores del IPM . . . . .                                          | 134 |
| <b>6-23.</b> Resultados comparativos de % de varianza obtenida por AFM y Correlación<br>obtenida por PLS . . . . . | 135 |
| <b>A-1.</b> Diccionario de datos de las variables de la matriz del PIB . . . . .                                   | 151 |
| <b>A-2.</b> Diccionario de datos de las variables de la matriz del IPM . . . . .                                   | 152 |

# 1. Introducción

En la actualidad, la recolección de datos es una tarea cotidiana para muchas áreas de la ciencia, de la industria y del sector oficial. A menudo, estos datos no se limitan a estructuras de dos vías (individuos y variables) sino que involucran múltiples dimensiones. Las estructuras de dos vías implican dos variables o dimensiones y se representan fácilmente en forma de matriz, habitualmente individuos en las filas y variables en las columnas. En contraste, las estructuras de tres vías añaden una tercera dimensión, creando conjuntos de datos más complejos que se pueden organizar como cubos de datos. En este contexto, el análisis de datos de tres vías o multivía emerge como un campo de investigación en crecimiento, dado que las estructuras de datos son cada vez más complejas y se requieren análisis integrales de los datos ([Cuarita Ajno, 2020](#)).

En la práctica, a menudo nos enfrentamos a arreglos de datos que se componen de tres índices o vías, donde la primer vía identifica los individuos que son la unidad del estudio, la segunda vía identifican las variables que son medidas en los individuos y la tercer vía se identifica como variables adicionales, comunmente relacionadas con la temporalidad o variación de las condiciones del fenómeno. Si bien el análisis de un arreglo de datos de tres vías brinda gran utilidad para identificar patrones, con frecuencia la asociación con otra estructura igual (otro arreglo de tres vías) resulta de gran interés y relevancia para comprender mejor el fenómeno de análisis.

De forma más general, el análisis de datos de  $N$  vías se refiere al análisis de datos que se encuentran indexados y organizados en múltiples dimensiones. En otras palabras, son datos que provienen de un formato  $N$ -dimensional de forma  $X_{IxJxKx\dots}$ . Estos arreglos difieren de las estructuras conocidas, que corresponden a matrices bidimensionales  $X_{IxJ}$  ([Hervás et al., 2018](#)).

Actualmente existen metodologías desarrolladas para el análisis de un arreglo de datos de tres vías, tales como el Análisis Factorial Múltiple desarrollado por [Escofier et al. \(1984\)](#), además, cabe resaltar los métodos desarrollados por [P. M. Kroonenberg \(1983\)](#) denominados los modelos Tucker 1, Tucker 2 y Tucker 3, el modelo del Análisis Factor Paralelo, por sus siglas en inglés Parallel Factor Analysis propuesto por [Harshman \(1970\)](#), los métodos de Descomposición Canonica de [Carroll y Chang \(1970\)](#) y la técnica desarrollada por [Lavit et al. \(1994\)](#) denominada Structuration des Tableaux ‘a Trois Indices de la Statistique (STATIS) o también conocida como Analyse Conjointe de Tableaux (ACT), PARAFAC [Westerhuis et](#)

al. (1999) y N-PLS Hervás et al. (2018). Particularmente, los datos de tres vías representan una clase especial de datos que involucran tres conjuntos de variables o factores, es decir, un caso especial del análisis de datos de  $N$  vías. Sin embargo, es importante destacar que recientemente han surgido metodologías que intentan abordar el desafío para el análisis simultáneo e integrado de dos estructuras de datos de tres vías.

Este trabajo de investigación tiene como objetivo explorar y evaluar metodologías para el análisis integrado y simultáneo de dos estructuras de datos de tres vías, a partir de una visión multivariante, proporcionando una comprensión más clara de estas metodologías y destacando su importancia en la investigación actual y futura. Al explorar y abordar las alternativas y metodologías para el análisis de estas estructuras de datos, se avanzará en la comprensión de estas estructuras multidimensionales y su complejidad para extraer conocimientos valiosos, a partir de la evaluación de las diferentes metodologías aplicadas.

Como técnicas de análisis se utilizará Mínimos Cuadrados Parciales (PLS), Mínimos Cuadrados Parciales Multiway ( $N$ -PLS) y se compara con el Análisis Factorial Múltiple (AFM). A partir de estos métodos se describe de forma simultánea la relación existente entre las dos estructuras de datos de tres vías, además, las técnicas PLS y  $N$ -PLS permite describir la relación a partir de la maximización de la correlación mientras que la técnica AFM lo describe a partir de la maximización de la varianza.

Como caso de estudio se realizará la aplicación de las técnicas anteriormente indicadas sobre el conjunto de datos relacionado con el Producto Interno Bruto (PIB) nacional desde el 2018 al 2021, como fuente de información de la generación de riqueza para los departamentos de Colombia y el conjunto de datos del Índice de Pobreza Multidimensional (IPM) desde el 2018 al 2021 como fuente de información de las condiciones de pobreza para los departamento de Colombia. A partir de la aplicación de PLS,  $N$ -PLS y AFM, analizando los individuos (los 32 departamentos de Colombia) mediante el conjunto de variables de cada una de las dos estructuras de datos indicadas. Con el propósito de mejorar la comprensión sobre la relación entre la generación de riqueza y la prevalencia de la pobreza en el país, así como para mostrar aplicaciones prácticas de las técnicas y reconocer tanto sus propiedades como sus limitaciones.

El contenido de este documento se encuentra dividido en los siguientes capítulos: Introducción como Capítulo 1. El planteamiento de la investigación, como Capítulo 2, en el cual se realiza la definición del problema, justificación y los objetivos de la investigación. El Capítulo 3 resume los antecedentes estadísticos más relevantes encontrados en la literatura. El Capítulo 4 contiene el Marco Teórico. La metodología de análisis de datos se presenta en el Capítulo 5, los resultados obtenidos se resumen en el Capítulo 6. Por último, el Capítulo 7 condensa las conclusiones principales y recomendaciones que surgen de esta investigación.

## 2. Planteamiento del problema

Este capítulo incluye la definición del problema y establece los objetivos del proyecto. Además, se proporciona una justificación desde una perspectiva estadística, de la necesidad de conocer y comprender metodologías capaces de abordar de manera efectiva la asociación entre dos estructuras de datos tridimensionales.

### 2.1. Definición del problema

A menudo nos enfrentamos a arreglos de datos que se componen de tres índices o vías. La primer vía identifica los individuos que son la unidad del estudio, la segunda vía identifican las variables que son obtenidas sobre los individuos y la tercera vía se identifica como variables adicionales, que indican la variación de condiciones en la recolección de datos. Si bien el análisis de un arreglo de datos de tres vías brinda gran utilidad para identificar patrones, con frecuencia la asociación con otra estructura es necesaria para comprender el comportamiento del fenómeno de forma más amplia.

En la revisión de literatura se identifican técnicas y métodos que modelan la estructura de tres vías, tales como: los métodos de Descomposición Canonica de [Carroll y Chang \(1970\)](#), el modelo del Análisis Factor Paralelo, por sus siglas en inglés Parallel Factor Analysis propuesto por [Harshman \(1970\)](#), el Análisis de Procrustes desarrollado por [Gower \(1975\)](#), los métodos desarrollados por [P. M. Kroonenberg \(1983\)](#) denominados los modelos Tucker 1, Tucker 2 y Tucker 3, el Análisis Factorial Múltiple desarrollado por [Escofier et al. \(1984\)](#), la técnica desarrollada por [Lavit et al. \(1994\)](#) denominada Structuration des Tableaux ‘a Trois Indices de la Statistique (STATIS) o también conocida como Analyse Conjointe de Tableaux (ACT), el análisis de coinerencia desarrollado por [Dolédéc y Chessel \(1994\)](#), el método X-STATIS, inicialmente propuesto en 1978 por [Jaffrenou \(1978\)](#) e implementado finalmente por [Abdi et al. \(2012\)](#). Estos métodos son útiles para abordar cada una de estas estructuras de manera individual. Originalmente han sido concebidos para estudiar la variabilidad dentro de una estructura de datos de tres vías.

En este contexto, la revisión de metodologías para analizar la asociación de dos estructuras de datos de tres vías es el interés en este trabajo. Aunque existen en la literatura diversas técnicas y métodos que son útiles para abordar estas estructuras de manera individual, originalmente han sido concebidos para estudiar la variabilidad dentro de una estructura y no

necesariamente para investigar en profundidad la asociación entre dos estructuras de datos de tres vías. Por lo tanto, se ha identificado que son escasas las metodologías que se acercan a explorar la asociación entre dos estructuras de tres vías.

A partir de lo anterior, es fundamental distinguir entre la variabilidad dentro de una estructura de datos y la asociación entre dos estructuras de datos distintas. Mientras que las técnicas tradicionales de análisis multivariado pueden ser útiles para entender la variabilidad dentro de una estructura de datos individual, como la relación entre variables dentro de un conjunto de datos específico, estas pueden no ser adecuadas para capturar la asociación entre dos estructuras de datos con múltiples dimensiones o “vías”. En este sentido, la búsqueda de metodologías que puedan abordar de manera efectiva esta asociación entre dos estructuras de datos de tres vías se convierte en un campo de investigación crucial. Este tipo de análisis no solo permite comprender mejor las relaciones complejas entre diferentes conjuntos de datos, sino que también puede proporcionar información valiosa sobre cómo estas asociaciones pueden influir en diversos fenómenos o procesos de interés en distintos campos de estudio.

Por ejemplo, en el desarrollo del caso de estudio se analizará las tablas de datos del Producto Interno Bruto (PIB) y los Índices de Pobreza Multidimensional (IPM). Durante años, el análisis de la relación entre estos dos tensores han sido de interés, dada la importancia de conocer la dinámica económica, el crecimiento y la generación de riqueza de un país, como también su impacto en el mejoramiento de las condiciones de calidad de vida de su población. Siguiendo este objetivo, surge la intención de analizar de forma integrada las componentes de variación multivariante que se presentan en la dinámica de la generación de riqueza (PIB) y las condiciones de pobreza de las subdivisiones administrativas de Colombia durante el período 2018 al 2021.

Los estudios desarrollados para el análisis de estos indicadores han abordado técnicas estadísticas que van desde las series temporales hasta el análisis multivariado de datos. Para el análisis del PIB desde una perspectiva temporal con el objetivo de realizar proyecciones y pronósticos a futuro, se han ajustado modelos de tipo ARIMA ([Herrera y Diaz, 2002](#)), ARIMAX ([Virguez, 2019](#)), VECM ([Campo y Sarmiento, 2011](#)) y VAR ([Quinde et al., 2019](#)), desde una perspectiva econométrica, con el objetivo de explicar o capturar la heterogeneidad no observable en el tiempo, se han ajustado modelos de Datos Panel, realizado por [Vargas \(2018\)](#) y desde una perspectiva multivariada para observar de manera simultánea varias variables, se ha aplicado la técnica Multi Way - PCA y Tucker 3, realizado por [Ruiz y Bonilla \(2018\)](#). Es preciso señalar de este último trabajo que los resultados principales sugirieron que existen algunos factores e inductores que han influido en el crecimiento de la economía colombiana desde el 2000 al 2015, estos fueron la manufactura, los servicios, la explotación de recursos naturales y el sector social.

En cuanto al análisis de la pobreza, se han abordado técnicas multivariantes como el Análisis de Componentes Principales (PCA) ([Medina-Hernández et al., 2021](#)), Análisis Factorial Múltiple (AFM) ([Plaza, 2022](#)), técnicas de georreferenciación y regresión ponderada ([Arias Ramírez et al., 2020](#)), modelo econométrico de Umbral Generalizado ([C. Marrugo-Arnedo et al., 2015](#)), regresión lineal múltiple ([Cardona et al., 2013](#)) y el análisis factorial de mínimos cuadrados ponderados y Kaiser-Meyer-Olkin (KMO) ([Barahona, 2018](#)). Concluyendo a partir de estos, que existen factores que influyen y determinan los niveles de pobreza de una población, y así, generar estrategias que impacten en la reducción de la pobreza de un país.

Las anteriores investigaciones novedosamente desarrollan técnicas ampliamente aplicadas al análisis del PIB y la pobreza. Sin embargo, el ejercicio de análisis se ha realizado de forma individual para cada conjunto de indicadores, sin investigar la relación existente entre ambas estructuras, ya que según [V. Marrugo-Arnedo \(2013\)](#), a través de su estudio para analizar el comportamiento del crecimiento económico y el desarrollo humano en Colombia durante la década 2000-2010, sugiere que para realizar el análisis del crecimiento económico (PIB) del país, no es suficiente hacerlo con el aumento de un período a otro, es necesario relacionar este con variables sociales que influyen en la población y que generan bienestar, como lo son el acceso a la salud, el mercado laboral, entre otros.

De igual manera, otras investigaciones sugieren esta relación. Por ejemplo, [Vargas \(2018\)](#), a partir de su estudio para medir la relación entre el PIB Per Cápita y la pobreza monetaria a nivel departamental en Colombia durante el 2002 hasta el 2014, concluyó que existe una relación inversa entre el PIB per cápita y la pobreza, puesto que a medida que el PIB per cápita aumenta la pobreza disminuye en un 0,207 %. Asimismo, [Arias \(2022\)](#), en su investigación para la caracterización de la relación entre la pobreza y la actividad económica en Colombia para el año 2018, a partir del análisis de tablas múltiples aplicando técnicas de AFM y PLS, sugiere que existe una relación entre las condiciones de vida de la población y el aporte en la productividad de los departamentos, concluyendo que el crecimiento económico del país tiene una focalización en algunos departamentos estratégicos, cercanos entre sí, lo cual demuestra que a mayor productividad hay bajos niveles de pobreza, como lo fue en la Región Andina por ejemplo. Por otro lado, la cercanía entre departamento, evidencia similares niveles de pobreza y actividad económica.

De lo anterior se concluye que, aunque hay aproximaciones para el análisis de la relación entre estos indicadores económicos y sociales, por un lado, integrando la temporalidad y por el otro integrando la estructura multivariante, no se evidencian investigaciones de esta relación integrando la estructura multivariante y temporal, por ello, es importante precisar algunas técnicas para análisis de tablas múltiples o de tres vías.

Finalmente, se evidencia una brecha en la literatura, la cual permite a partir de esta investigación aportar con el planteamiento de una metodología que integre los componentes multivariante y temporal que se presentan en dos estructuras de datos de tres vías, tomando como caso aplicado el conjunto de variables que miden la generación de riqueza y pobreza de las subdivisiones administrativas de Colombia durante el período del 2000 al 2021.

## 2.2. Justificación

Las estructuras de datos de tres vías, tuvieron sus primeras manifestaciones y aplicaciones significativas a problemas específicos del Control de Procesos Químicos, definido como Quimiometría, la cual utiliza métodos matemáticos y estadísticos para proporcionar la máxima información química relevante mediante el análisis de datos químicos ([Deming et al., 1988](#)).

Las estructuras de los datos en esta área de investigación involucran conjuntos de datos con muestras, variables químicas y condiciones experimentales, por lo tanto, surgió la necesidad de desarrollar metodologías para representar y analizar la variabilidad de estas estructuras complejas de manera efectiva que no podía ser completamente capturada por enfoques bidimensionales tradicionales ([Martens, 2001](#)).

Aunque se reconoce que estas estructuras tuvieron sus primeras apariciones en el área de investigación de la Quimiometría, se desatacan algunos estudios relevantes realizados por autores como [Bro \(1996\)](#), [Smilde \(2001\)](#), [Singh et al. \(2006\)](#). En la actualidad las estructuras de datos de tres vías son cada vez más comunes en muchas otras áreas de investigación. Por ejemplo, en las Ciencias Sociales con estudios de autores como [Medina-Hernández et al. \(2021\)](#), [Bautista \(2009\)](#), [Camiz \(2011\)](#), en Biología con investigaciones realizadas por [De-caëns et al. \(2009\)](#) [Feki-Sahnoun et al. \(2018\)](#), e Ingeniería ([Swiercz y Mroczkowska, 2020](#)), entre otras áreas. Realizar una exhaustiva investigación para la revisión de metodologías que permita analizar estas estructuras aborda una necesidad concreta en estos campos, ya que estos datos pueden ser extremadamente complejos de analizar debido a su dimensionalidad, lo que presenta varios desafíos.

Desde una perspectiva estadística, una de las limitaciones actuales son las pocas metodologías para el análisis integrado y simultáneo de dos estructuras de datos de tres vías. Por esta razón, este trabajo propone realizar un aporte en la revisión de metodologías para el análisis integrado de dos estructuras de datos de tres vías, lo cual por el momento, ha sido poco común encontrar enfoques que aborden esta tarea de manera integral y se ha optado por realizar un análisis de estas estructuras de forma individual, lo que se presenta como un riesgo al omitir la correlación que puede existir entre dos bloques de datos con la misma estructura y puede ser aprovechable para el análisis estadístico.

La investigación de metodologías capaces de abordar de manera efectiva la asociación entre dos estructuras de datos de tres vías emerge como un campo de estudio crucial. Este análisis no solo permite una comprensión más profunda de las complejas relaciones entre conjuntos de datos diversos, sino que también puede brindar información valiosa sobre cómo estas asociaciones impactan en una variedad de fenómenos o procesos de interés en distintos ámbitos de estudio.

Considerando lo anterior, aprovechar la estructura multivariante de los bloques de datos y su división en subtablas representa una ventaja significativa para la aplicación de técnicas como AFM, PLS y *N*-PLS, y mediante diversas alternativas de despliegue de los bloques de datos, llevar a cabo un estudio efectivo que permita realizar el análisis simultáneo de ambos bloques de datos. Lo anterior ayudaría a identificar la correlación entre los dos bloques, así como a caracterizar a los individuos y otras variables de interés de manera precisa.

La importancia de analizar datos de tres vías y sus relaciones radica en su capacidad para enfrentar la complejidad inherente a conjuntos de datos multidimensionales. Estas estructuras de datos, brindan una visión más completa y detallada de los fenómenos estudiados que los enfoques bidimensionales convencionales. Al explorar cómo se relacionan dos estructuras de datos de tres vías, es posible descubrir patrones y conexiones que de otra manera podrían pasar desapercibidos.

Como caso de estudio, se aplicará la metodología propuesta para el análisis integrado de las estructuras de datos del PIB (2018 al 2021) e IPM (2018 - 2021). En el cual se podrá estudiar, a partir de las metodologías planteadas, la relación entre la generación de riqueza y la pobreza en los departamentos de Colombia, dado que diversos estudios han demostrado que existe una relación inversa entre ambos aspectos ([Ruiz y Bonilla, 2018](#); [Vargas, 2018](#); [Arias, 2022](#)).

En el caso específico de estudio, los resultados podrían brindar información crucial para la formulación de políticas públicas efectivas. Al comprender mejor la relación entre la generación de riqueza y la pobreza, la toma de decisiones podrían enfocarse en diseñar estrategias que promuevan un crecimiento económico inclusivo que beneficie a todos los segmentos de la sociedad. Además, estos resultados podrían ayudar a identificar áreas prioritarias para la inversión pública y dirigir los recursos para que tengan un mayor impacto en la reducción de la pobreza, en la mejora del bienestar social en los departamentos de Colombia y contribuir al desarrollo socioeconómico sostenible del país.

Los resultados de esta investigación y la revisión de las diferentes metodologías, pueden tener diversas aplicaciones prácticas en otros campos de investigación. Los hallazgos pueden proporcionar una comprensión más profunda de estas estructuras que son cada vez más comunes

en las diferentes áreas de investigación y la asociación entre dos estructuras similares, lo que podría conducir a avances significativos en áreas como Biología, Ciencias Sociales, Ingeniería y otras áreas, que podría ampliar la aplicación y el análisis de estas estructuras.

## **2.3. Objetivos**

### **2.3.1. Objetivo General**

Realizar evaluación metodológica para el análisis integrado de las componentes de variación y asociación multivariante que se presentan en dos estructuras de datos de tres vías.

### **2.3.2. Objetivos Específicos**

- Identificar en la literatura científica las metodologías propuestas para el análisis simultáneo de dos arreglos de datos de tres vías.
- Documentar y examinar diferentes métodos para el análisis simultáneo de dos arreglos de datos de tres vías.
- Utilizar el conjunto de variables de la generación de riqueza y condiciones de pobreza en Colombia para el periodo del 2018 al 2021 como caso de estudio identificando e implementando la metodología más conveniente para el análisis de las estructuras de datos.

### 3. Revisión de antecedentes

En la literatura se identifican numerosos estudios e investigaciones sobre aplicación de metodologías para el análisis de datos de tres vías.

La revisión de antecedentes de este trabajo se realiza partiendo de una búsqueda bibliográfica, identificando inicialmente un ecuación de búsqueda que facilite el análisis de tendencias de producción científica sobre el análisis de estructura de datos de tres vías en la base de datos de Scopus entre las fechas de Agosto a Octubre del 2023.

```
Ecuación de búsqueda = ( "Three-way data analysis" OR "Three-dimensional data analysis" OR "Data analysis with three factors" OR "Multivariate three-way data analysis" OR "Three-way statistics" OR "Three-way factorial analysis" OR "Three dimensions in statistics" OR "Three-way principal component analysis" OR "Three-way statistical models" )
```

El análisis bibliométrico se desarrolló en el aplicativo biblioshiny disponible en la librería de Bibliometrix del software estadístico R ([Aria y Cuccurullo, 2021](#)). Este capítulo responde y da cumplimiento al primer objetivo de este trabajo de investigación.

Una vez finalizado el análisis bibliométrico, se relacionan las fuentes más relevantes identificadas de las diferentes metodologías utilizadas para el análisis de estructuras de datos de tres vías. Los trabajos son presentados en orden cronológico, lo que indica que van desde la publicación más antigua hasta la más reciente.

Las técnicas y métodos útiles que modelan la estructura de datos de tres vías son tales como el modelo del Análisis Factor Paralelo, por sus siglas en inglés Parallel Factor Analysis propuesto por [Harshman \(1970\)](#), los métodos de Descomposición Canonica de [Carroll y Chang \(1970\)](#), los métodos desarrollados por [P. M. Kroonenberg \(1983\)](#) denominados los modelos Tucker 1, Tucker 2 y Tucker 3, el Análisis Factorial Múltiple desarrollado por [Escofier et al. \(1984\)](#) y la técnica desarrollada por [Lavit et al. \(1994\)](#) denominada Structuration des Tableaux 'a Trois Indices de la Statistique (STATIS) o también conocida como Analyse Conjointe de Tableaux (ACT) y finalmente, [Bro \(1996\)](#) desarrolló una metodología denominada N-PLS.

### 3.1. Análisis bibliométrico

A partir de este análisis se identificó que en los últimos años han sido grandes los avances que se han reportado. En el año 2022 se publicaron 6 trabajos con respecto al análisis de datos de tres vías.

Resultado de la búsqueda bibliográfica en la base de datos SCOPUS, se obtuvieron un total de 146 documentos publicados desde 1991 hasta el año 2023. Se observó que este tema de investigación tuvo una tendencia creciente a través de estos años. Se realiza el análisis del promedio de citas por año, autores más relevantes, Número total de apariciones de los autores, número de artículo por autor, Índices de citación.

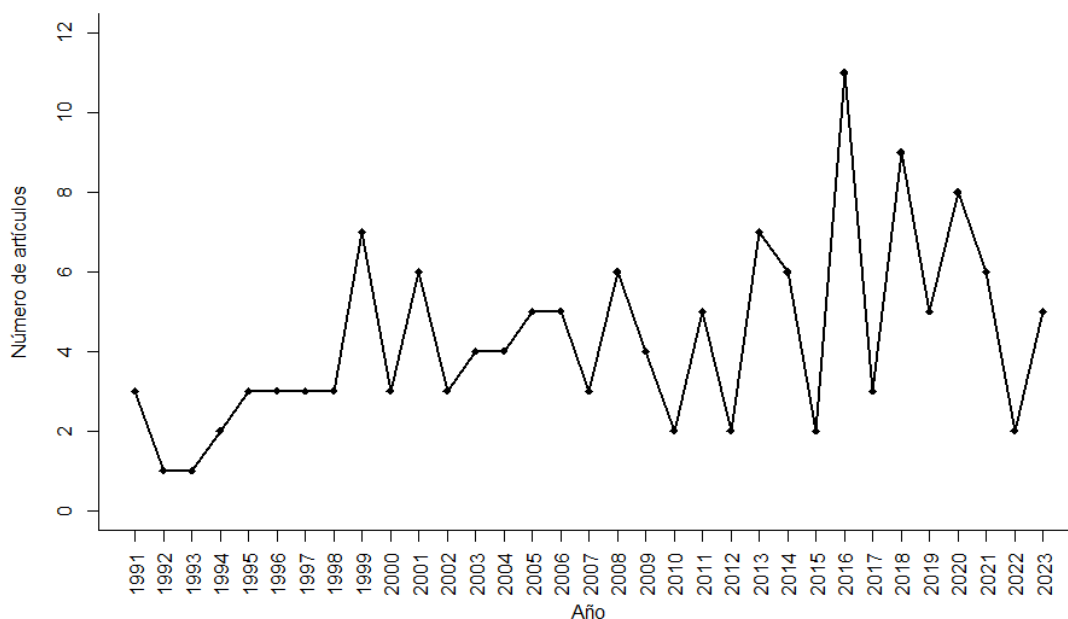
| Descripción                     | Resultados |
|---------------------------------|------------|
| Documentos                      | 146        |
| Fuentes                         | 107        |
| Promedio de citas por documento | 35.38      |
| Autores                         | 434        |
| Co-autores por documento        | 3.53       |
| Artículos                       | 115        |
| Capítulo de libro               | 4          |
| Otro tipo de documento          | 27         |

**Tabla 3-1.:** Resumen análisis bibliométrico

En los resultados consignados en la Tabla **3-1** se evidencia que la ecuación de búsqueda permitió obtener cerca de 146 documentos en los cuales se identifican técnicas utilizadas para el análisis de datos de tres vías, estos fueron publicados en 107 revistas y/o fuentes de divulgación científica. Por otro lado, el promedio de citación por documento es de 35.38, indicando así que estos generan alta divulgación en la comunidad académica. Finalmente, se evidencia una participación de 434 autores con coautorías de 3 personas aproximadamente por documento. Igualmente, cabe resaltar que los artículos son el tipo de documento más prevalente con una frecuencia de 115 artículos publicados sobre el tema de investigación, sin embargo, se evidencian 27 documentos de tipo diversos de materiales como informes técnicos, memorias de conferencias u otros tipos de publicaciones.

En cuanto a la producción científica por año, la Figura **3-1** ilustra el crecimiento desde 1991 en el análisis de datos de tres vías. Se destaca un aumento significativo en la producción científica en este campo a partir de 2016. No obstante, se observa una disminución en la producción para el año 2022, seguida de un repunte en 2023. En 2016, se registró el mayor número de publicaciones, con aproximadamente 11 documentos publicados durante ese año. Además, se puede notar que las publicaciones oscilan entre 3 y 6 por año en los demás

períodos analizados.



**Figura 3-1.:** Producción científica por año de técnicas para el análisis de datos de tres vías desde el 2002 al 2023

Lo anterior evidencia que estas técnicas de análisis de datos de tres vías aún no se utilizan ampliamente debido a su complejidad. Por tanto, es importante impulsar la aplicación de estas técnicas para un uso más frecuente.

### 3.1.1. Análisis de impacto de autores e instituciones más relevantes

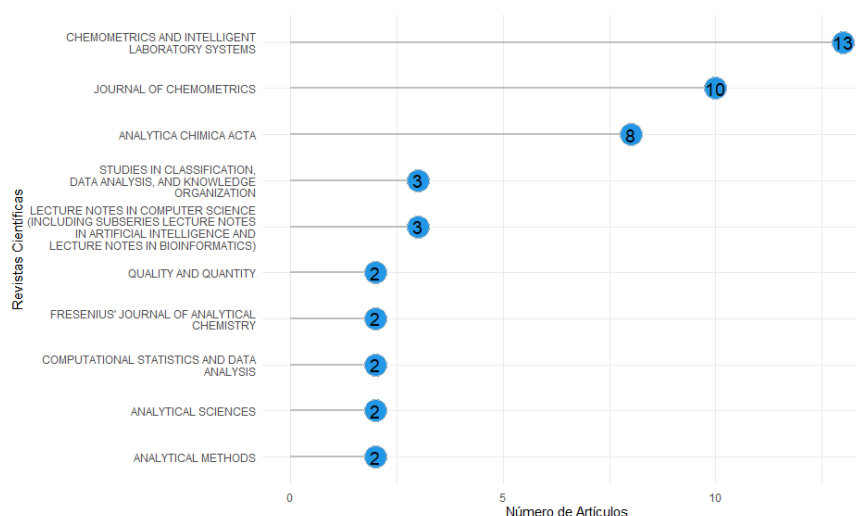
El análisis de autores e instituciones más relevantes permite identificar quiénes son las personas y organizaciones más influyentes en el tema de investigación de este trabajo. Este análisis se basó en la obtención de diferentes métricas como el índice H, el número total de citas, el número total de artículos publicados y el año de publicación.

En la Tabla **3-2** se proporciona información sobre la productividad y el impacto de diferentes autores sobre el análisis de datos de tres vías. El autor *Jiang J-H* es el autor con índice H más alto, lo que indica que al menos dos de sus artículos han sido citados al menos dos veces cada uno. Igualmente, *Smilde A* tiene el mayor número de citas (884), seguido por el autor *Bro R* con 114 citas.

| Autores        | Índice H | Total Citas | Total artículos | Año inicio |
|----------------|----------|-------------|-----------------|------------|
| Smilde A       | 1        | 884         | 1               | 1995       |
| Bro R          | 1        | 114         | 1               | 1998       |
| Kroonenberg PM | 1        | 54          | 1               | 1994       |
| Jiang J-H      | 2        | 51          | 2               | 1999       |
| Singh KP       | 1        | 32          | 1               | 2006       |
| Singh VK       | 1        | 32          | 1               | 2006       |
| Smilde AK      | 1        | 18          | 1               | 1990       |
| Slimani N      | 1        | 13          | 1               | 2017       |
| Hanafi M       | 1        | 3           | 1               | 2020       |

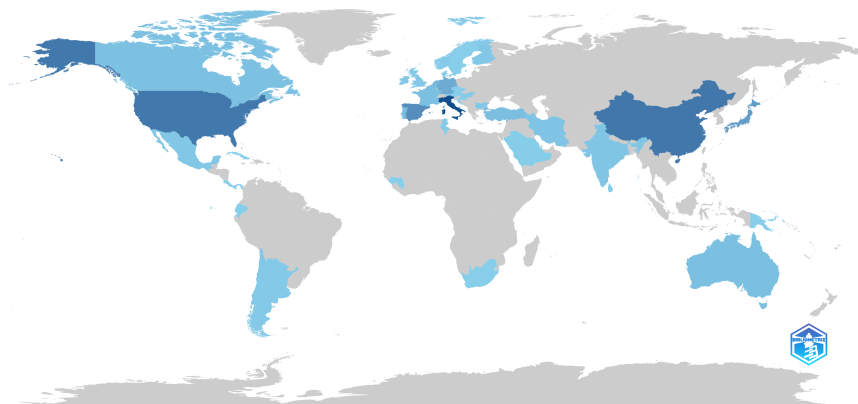
**Tabla 3-2.:** Indicadores bibliométricos de los autores más relevantes sobre el análisis de estructuras de datos de tres vías

Por otro lado, para evaluar el impacto de la producción científica, se analizan las revistas que han publicado la mayor cantidad de artículos. Según se muestra en la Figura 3-2, la revista “**Chemometrics and Intelligent Laboratory Systems**” encabeza la lista con un total de 13 artículos publicados sobre este tema de investigación. Le sigue de cerca “**Journal of Chemometrics**” con 10 artículos. Las demás revistas han publicado 3 y 2 artículos respectivamente, lo que sugiere una difusión limitada de esta temática y evidencia el bajo impacto de las revistas centradas en teoría estadística. Además, revela que el desarrollo de estos métodos ha surgido como respuesta a problemas prácticos en el Control de Procesos Químicos, también conocido como Quimiometría.



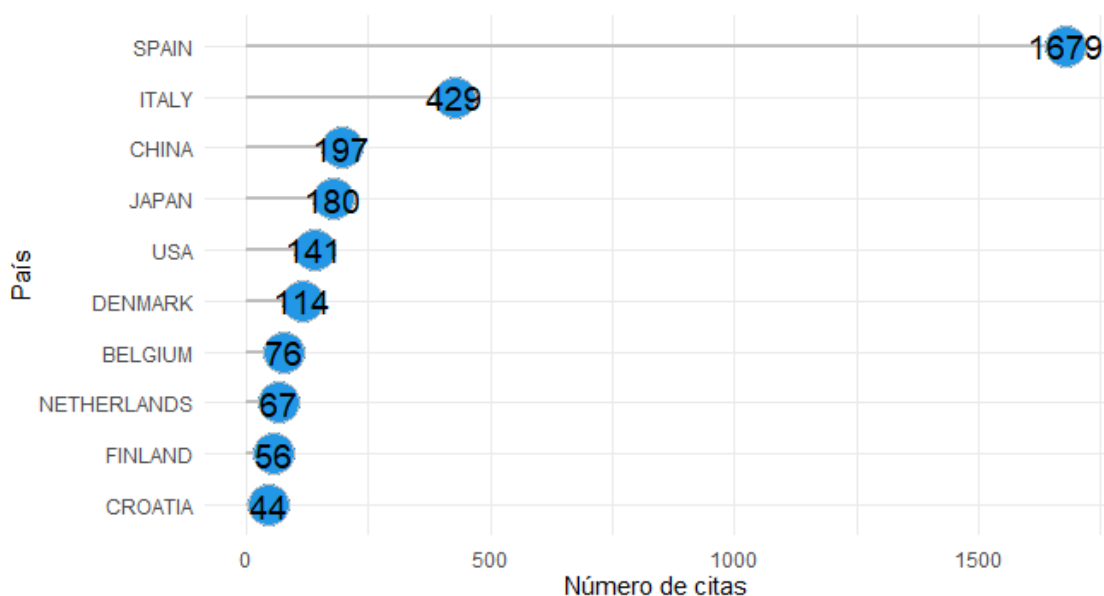
**Figura 3-2.:** Fuentes con mayor volumen de producción enfocado a técnicas para el análisis de datos de tres vías

### 3.1.2. Volumen de citaciones



**Figura 3-3.:** Mapa de producción Científica por país enfocado a las técnicas para el análisis de datos de tres vías

La Figura **3-3** evidencia el rezago en Colombia con respecto a los otros países en cuanto al análisis de datos de tres vías. El país con mayor producción científica en este tema es Italia, seguido por países como China, Estados Unidos y España. Por su parte, Colombia no se ubica como potencial productor científico en tópicos relacionados con el análisis de datos de tres vías, de hecho, no tiene producción. Por tanto, se presenta como una oportunidad para impulsar la investigación en este campo de la estadística en el país.

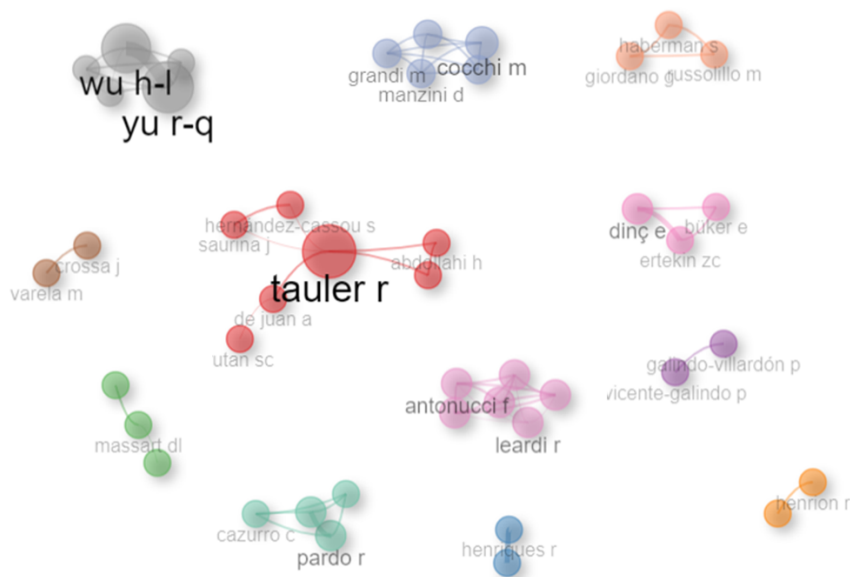


**Figura 3-4.:** Países con mayor volumen de citación de artículos sobre el análisis de datos de tres vías

A pesar de la información presentada en la Figura 3-3, en la Figura 3-4 se observa que España lidera en el índice de citación con un total de 1.679 citas, seguido por Italia con 429 citas. Estos datos subrayan la influencia significativa de estos países en la producción científica y colaborativa relacionada con el análisis de datos de tres vías. En contraste, países como Finlandia y Croacia muestran índices de citación más bajos, con 56 y 44 citas respectivamente.

### 3.1.3. Redes de colaboración y Cocitación de autores

Las redes de colaboración de autores, también conocidas como redes de coautoría, son representaciones gráficas de las interacciones colaborativas entre diferentes autores, en este caso, sobre el análisis de estructuras de datos de tres vías. Las redes, permiten representar a los autores y las conexiones con otros autores que representan colaboraciones en la escritura de artículos, libros u otros tipos de publicaciones científicas.



**Figura 3-5.:** Red de colaboración entre los autores sobre el análisis de datos de tres vías

La Figura 3-5 muestra una representación visual de cómo los autores colaboran entre sí en el campo del análisis de datos de tres vías. En esta red, es posible observar a aquellos investigadores que han trabajado juntos en proyectos y estudios relacionados con esta temática, mayormente en la aplicación. Sin embargo, es notable que algunos autores prominentes como Smilde A, Bro R y otros, mencionados en la Tabla 3-2, no están presentes en esta red de colaboración.

A pesar de la falta de colaboración directa en este tema entre estos autores, es importante destacar que han trabajado juntos en otras temáticas, como el Análisis de Componente Principales [Bro y Smilde \(2014\)](#). Esto sugiere que aunque puedan tener enfoques individuales en ciertos temas, aún existe la posibilidad de colaboración en otros ámbitos de estudio.

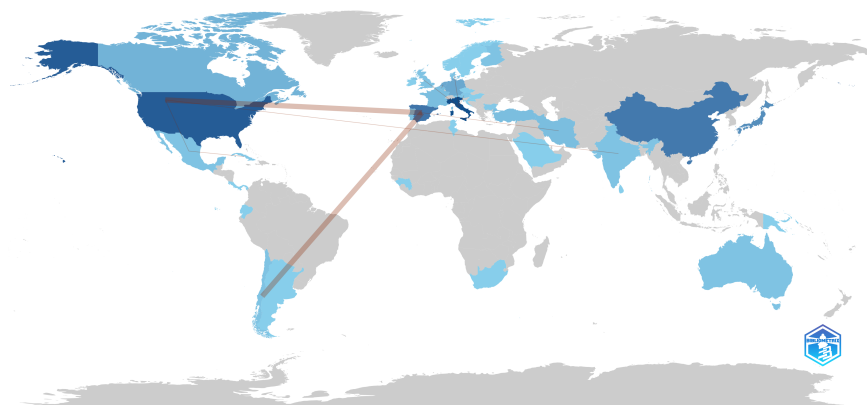
Figure 1 displays six network graphs illustrating the evolution of research clusters in the field of organizational learning. Each graph represents a specific year or range of years, showing the relationships between authors (nodes) and their associated research clusters (edges). The nodes are labeled with the author's name and the year of their contribution.

- 1980:** A central node (2) connected to kroonenberg p.m. 1980 and kroonenberg p.m. 1983-1.
- 1983-1:** A central node (2) connected to kroonenberg p.m. 1980 and kroonenberg p.m. 1983-1.
- 1986-1:** A central node (2) connected to kroonenberg p.m. 1980 and kroonenberg p.m. 1983-1.
- 1986-2:** A central node (2) connected to kroonenberg p.m. 1980 and kroonenberg p.m. 1983-1.
- 1986-3:** A central node (6) connected to andersson c.a. 2000-2 and tucker l.r. 1966-3.
- 1989:** A central node (5) connected to pravdova v. 2002 and henriou r. 1994-1.

The graphs show a growing complexity and interconnectedness of research clusters over time, with the 1989 graph showing the most extensive network of relationships.

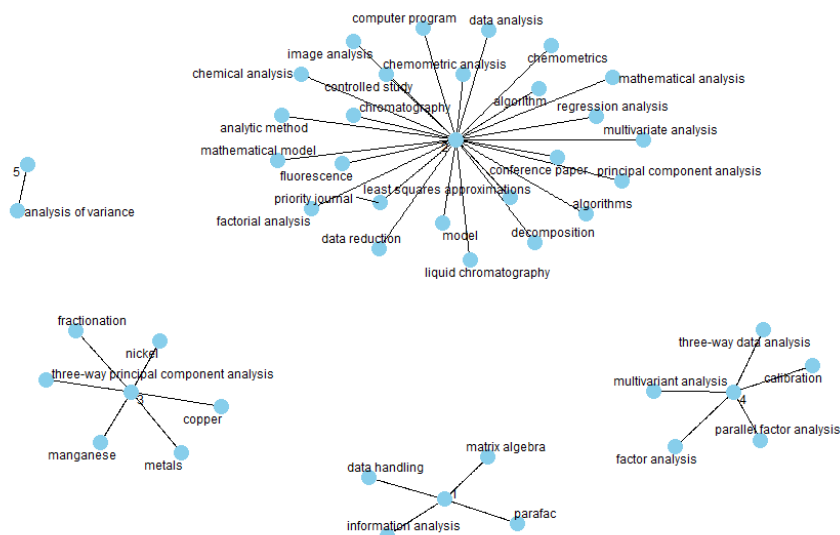
La Figura **3-6** representa la red de co-citación de autores, en este se logra evidenciar 7 clusters relevantes de asociación, sin embargo, tres de estos presentan una mayor co-citación de autores. El cluster 1, 4 y 7, presentan conexiones relevantes entre autores como *Bro R.*, *Tucker I.*, *Kroonenberg P.*, *Smilde A.*, entre otros. Lo anterior implica que estos autores

presentan técnicas estrechamente relacionadas con análisis de estructuras de datos de tres vías, lo cual resulta de insumo y relevante para la relación de autores que se presenta en las subsecciones de *Análisis de datos de tres vías* y *El análisis de la asociación entre dos tablas de datos de tres vías* de este mismo capítulo.



**Figura 3-7.:** Red de colaboración por país sobre el análisis de datos de tres vías

La Figura 3-7 representa la colaboración por país en referencia al análisis de datos de tres vías, donde, España presenta una alta colaboración con Argentina y Chile, por tanto, a nivel suramericana, estos dos países representan avances en este campo estadístico. Por otro lado, Italia tiene una conexión de colaboración con países como Reino Unido y Hungría. Aunque China presenta una alta producción científica no tiene conexiones significativas con otros países.



**Figura 3-8.:** Red de co-ocurrencia de palabras clave sobre el análisis de datos de tres vías

La red de Co-ocurrencia de las palabras claves de la Figura **3-8** muestra las palabras claves más frecuentes en los documentos consultados. Se observan 5 clusters de palabras, de los cuales, el Cluster 2 presenta la mayor cantidad de relación entre las palabras, y a partir de este, se identifica que el análisis de datos de tres vías tienen mayor relación con áreas de conocimiento de Matemáticas, Química, Computación, Estadística multivariada, entre otras.

El Análisis Bibliométrico ha permitido encontrar estudios de metodologías utilizadas para el análisis de estructuras de datos de tres vías y la asociación entre dos estructuras. Lo anterior permite identificar las metodologías propuestas para el análisis de los dos arreglos, y por consiguiente, las técnicas adecuadas que se aplicarán en este trabajo. Igualmente, los autores identificados se relacionan en las siguientes subsecciones de este capítulo.

## 3.2. Análisis de datos de dos vías

Esta sección presenta las investigaciones más relevantes identificadas a partir de la revisión bibliográfica relacionada con el análisis de datos de estructuras de dos vías, destacando los diversos métodos de análisis empleados para abordar estas estructuras bidimensionales.

- [Olaya et al. \(2013\)](#) investigaron las asociaciones entre la composición socioeconómica familiar y la generación urbana de residuos sólidos domiciliarios por lados de manzana y por comunas, utilizando datos de la ciudad de Cali, Colombia.

Se caracterizó la composición socioeconómica por lados de manzana y comunas, utilizando variables socioeconómicas de una muestra de familias encuestadas para determinar los perfiles de los generadores de residuos sólidos. Se empleó la metodología de Análisis de Componentes Principales como herramienta estadística multivariada para estudiar posibles asociaciones entre la composición socioeconómica de las familias y la generación de residuos sólidos. Los resultados obtenidos mostraron altos grados de asociación entre niveles educativos y edades, y también indicaron que la generación y composición de residuos sólidos domiciliarios varían según el estrato socioeconómico.

- [Garcia A \(2021\)](#) llevó a cabo un Análisis Factorial Múltiple para examinar las condiciones de salud percibida de la población priorizada en la ciudad de Cali durante el año 2018. Los datos se recolectaron mediante un cuestionario tipo encuesta que exploraba las condiciones de vivienda, individuales y del entorno. Los resultados permitieron caracterizar los diferentes barrios de la ciudad, destacando los principales problemas de salud percibida que los afectan. Además, el AFM facilitó la síntesis de la información recopilada en un índice sintético, el cual generó un ranking de los barrios con mayor impacto en la salud en la ciudad. Este índice resultará útil para la administración municipal, ya que su aplicación permitirá priorizar las intervenciones sociales y de salud pública en los barrios más vulnerables.

- [Hernández et al. \(2021\)](#) estudió la pobreza multidimensional en los hogares colombianos utilizando los últimos datos del Índice de Pobreza Multidimensional (IPM) del DANE. El estudio examinó las privaciones en los hogares y las variables sociodemográficas de sus miembros.

Aplicando la técnica de Análisis de Componentes Principales (ACP), se identificó los factores que condicionan la pobreza en diferentes regiones. Los resultados revelaron que el número de miembros del hogar, el nivel educativo y el acceso al empleo formal son variables que distinguen a los hogares pobres de los no pobres. Adicionalmente, el estudio encontró una mayor incidencia de pobreza en las regiones costeras y la región amazónica.

- [Plaza \(2022\)](#) desarrolló un análisis multivariado de datos para la evaluación de la dinámica de la pobreza monetaria en Colombia por departamentos durante el período 2012-2020.

Para tal caracterización realizó la aplicación de dos métodos estadísticos, Análisis Factorial Múltiple y Análisis de Conglomerados. Los datos consistieron en un arreglo bidimensional que contenía información sobre la pobreza monetaria en 24 departamentos de Colombia.

Los resultados principales mostraron que los departamentos más pobres están ubicados en zonas más alejadas del distrito capital, como los departamentos de la región pacífica y el caribe. Además, se observó que algunos departamentos han mantenido una situación persistente de pobreza monetaria a lo largo del tiempo.

- [Gutiérrez et al. \(2022\)](#) desarrolló un análisis multivariado de los factores socioeconómicos asociados al rendimiento en las pruebas Saber Pro 2016, específicamente en el caso de los estudiantes de Licenciatura en Matemáticas en Colombia. Utilizando un conjunto de datos que incluía información socioeconómica y variables relacionadas con los resultados en competencias genéricas de los estudiantes, sus familias y las instituciones educativas que ofrecen programas de licenciatura en matemáticas para el año 2016.

Se empleó la técnica de análisis conocida como Análisis Factorial Múltiple, la cual reveló que los departamentos mejor posicionados en los resultados fueron Tolima, Boyacá, Norte de Santander, Magdalena, Quindío, Meta, Sucre, Atlántico, Nariño y Caquetá. En estos departamentos, los estudiantes provenían de instituciones de educación superior locales, residían en áreas rurales, habitaban viviendas de estratos 1 y 2, tenían padres jornaleros o trabajaban en pequeñas empresas familiares, eran mayoritariamente de género femenino y habían sido becados durante su carrera universitaria.

Esto indica que los factores determinantes en el desempeño en la prueba incluyen el género del estudiante, su área de residencia, el título de bachiller obtenido, la ocupación de la madre y el nivel educativo del padre.

- [Mejia \(2023\)](#) utilizó el Análisis Factorial Múltiple (AFM) para explorar los patrones de mortalidad y el impacto del Índice de Pobreza Multidimensional (IPM) en los departamentos de Colombia. Su objetivo era clasificar los departamentos que necesitan una mayor atención de políticas públicas para abordar la pobreza y las causas de mortalidad asociadas.

Para llevar a cabo su investigación, Mejía recopiló datos de mortalidad entre los años 2018 y 2021, utilizando la lista corta de tabulación 6/67 propuesta por la OPS, que clasifica 67 causas de muerte en 6 grupos. También se incluyeron las tasas brutas de mortalidad y los indicadores del IPM, que abarcan 15 dimensiones de pobreza. Se optó por trabajar con la medida resumen IPM. El análisis inicial consistió en una correlación lineal de Pearson y un AFM. Posteriormente, se realizó un análisis de agrupamiento o cluster utilizando las coordenadas globales resultantes del AFM, junto con el análisis Silhouette para validar la clasificación de los departamentos.

Los resultados revelaron que algunas áreas del país están experimentando múltiples fases epidemiológicas simultáneamente. Por ejemplo, se observó un aumento en las enfermedades no transmisibles, como las relacionadas con el sistema circulatorio y las neoplasias, así como enfermedades perinatales y signos y síntomas mal definidos en los departamentos más pobres multidimensionalmente. Además, se encontró una relación inversa entre el IPM y las tasas de mortalidad, lo que sugiere una conexión con la longevidad y un aumento en enfermedades propias de la vejez. El AFM y la clasificación por clusters proporcionaron una representación visual de la diversidad entre los departamentos, destacando asociaciones directas con causas externas y el eje de accidentalidad y violencia en departamentos como Putumayo y Cauca.

### 3.3. Análisis de la asociación entre dos tablas de datos de dos vías

En esta sección se exponen algunas investigaciones encontradas durante la revisión bibliográfica sobre el análisis de la asociación entre dos estructuras de datos de dos vías. Se resaltan los distintos métodos de análisis utilizados para abordar la asociación entre estas estructuras bidimensionales.

- [Wold et al. \(2001\)](#) llevaron a cabo una revisión exhaustiva de la metodología de regresión PLS y su aplicación significativa en los campos de la química y la ingeniería. Se examinaron el modelo, los supuestos y los diagnósticos necesarios para una interpretación adecuada de dicho modelo. Se profundizó en un tipo específico de análisis multivariado, donde la regresión PLS utiliza el modelo PLS predictivo de dos bloques

de dos vías para modelar la relación entre dos matrices,  $\mathbf{X}$  e  $\mathbf{Y}$ . Además, la regresión PLS modela la “estructura” de ambas matrices,  $\mathbf{X}$  e  $\mathbf{Y}$ .

Presentó dos ejemplos ilustrativos: primero, se utilizó un conjunto de datos relacionados con péptidos para mostrar cómo desarrollar, interpretar y refinar un modelo PLSR en el contexto de QSAR/QSPR. Segundo, se analizó un conjunto de datos relacionados con la fabricación de papel reciclado para demostrar el modelado de series temporales de datos de proceso mediante regresión PLS, incluyendo variables  $\mathbf{X}$  con retardo temporal.

- [Alciaturi et al. \(2003\)](#) realizó una investigación sobre la regresión de mínimos cuadrados parciales (PLS) y su aplicación al análisis del carbón mineral explicando los fundamentos del método PLS, mostrando una aplicación a la predicción de propiedades de carbones minerales a partir de datos del infrarrojo medio, con el objetivo de desarrollar métodos de análisis rápidos y no destructivos para estos materiales.

Inicialmente, desarrolló un análisis teórico del método, clasificándolo como “caballo de batalla” de la quimiometría, identificando los algoritmos utilizados para su aplicación. Luego, se presentó un ejemplo de aplicación en la predicción de propiedades de carbones minerales. La espectroscopía infrarroja permite estudiar la estructura de materiales proporcionando información cualitativa (sobre los tipos de grupos funcionales presentes en la muestra) y cuantitativa.

Para la aplicación práctica, se consideró un conjunto de 49 muestras de carbón de Venezuela y Colombia. Se determinaron propiedades como materia volátil, carbono fijo, cenizas, poder calórico, carbono elemental, hidrógeno, nitrógeno, azufre total y reflectancia máxima de vitrinita ( $R_o \max$ ) mediante métodos convencionales según normas, y se correlacionaron con el espectro de infrarrojo. En este caso, la matriz  $\mathbf{X}$  estaba representada por los espectros de infrarrojo (en unidades de absorbancia), y el vector  $\mathbf{y}$  por los datos de referencia. Se obtuvieron buenas correlaciones para carbono fijo, cenizas, reflectancia de vitrinita y concentración de carbono e hidrógeno, demostrando el potencial de la espectroscopía infrarroja en el análisis rápido, económico y no destructivo del carbón mineral.

- [Preda et al. \(2007\)](#) realizó la clasificación de datos funcionales, a partir de un enfoque de mínimos cuadrados parciales (PLS) para estimar las funciones del coeficiente discriminante aplicado para el análisis discriminante lineal (LDA) ya que los predictores son datos de tipo funcional (curvas). Desarrolló un estudio de simulación, así como una aplicación a datos de amasado, comparando los resultados del modelo PLS con los dados por otros métodos.

Los datos reales correspondían a la resistencia de la masa en el proceso de amasado, para lo cual se aplicó la regresión PLS para predecir la calidad de las galletas a partir de curvas que representan la resistencia (densidad) de la masa observada durante el proceso de amasado. Para una harina dada, la resistencia de la masa se registra

durante los 480 segundos del proceso de amasado. Para 115 tipos diferentes de harina, se obtuvo 115 curvas que pueden considerarse como trayectorias de muestra de un proceso estocástico continuo, por tanto se puede representar la matriz  $\mathbf{X}$  como  $\mathbf{X} = \{X_t, t \in [0, 480]\}$ . Cada curva se observa en 240 puntos de tiempo equiespaciados dentro del intervalo de tiempo  $[0, 480]$ . Así, una curva de amasado estaba representada por el conjunto de 241 puntos  $\{(t_i, X_{ti}), i = 0, \dots, 240\}$ .

Después del amasado, la masa fue procesada para obtener galletas y observar su calidad, por tanto, para cada tipo de harina, se obtuvo la calidad ( $\mathbf{Y}$ ) de las galletas, que se clasificaban en Buena, Ajustable y Mala. La muestra contenía 115 observaciones que se distribuían en 50 observaciones para  $Y = \text{Buena}$ , 25 para  $Y = \text{Ajustable}$  y 40 para  $Y = \text{Mala}$ .

Los resultados concluyeron que el enfoque PLS es una alternativa simple e interesante a los métodos lineales clásicos basados en componentes principales, y en general ofrece mejores resultados en sus coeficientes para ser utilizados en un LDA.

- [Arias \(2022\)](#), desarrolló una investigación para la caracterización de la relación entre las condiciones de pobreza y la actividad económica en Colombia para el año 2018.

Para realizar el análisis [Arias \(2022\)](#) realizó la aplicación de técnicas multivariadas para el análisis de múltiples tablas, como el Análisis Factorial Múltiple (AFM) y Mínimos Cuadrados Parciales (PLS). Según los resultados estas técnicas se complementan, a pesar que sus objetivos de análisis son diferentes. Adicionalmente realiza un análisis de clusters para identificar las similitudes de los departamentos en cuanto a niveles de pobreza y actividad económica.

Finalmente, los resultados principales lograron identificar que el aporte en productividad de las regiones se ve ciertamente relacionado con las condiciones de vida de su población. Asimismo, se observó que el crecimiento económico del país tiene una focalización en algunos departamentos estratégicos, cercanos entre sí, lo cual demuestra que a mayor productividad hay bajos niveles de pobreza, como lo fue en la Región Andina por ejemplo. Por otro lado, la cercanía entre departamento, evidencia similares niveles de pobreza y actividad económica.

- [Vaquer Rivera et al. \(2022\)](#) realizó un trabajo de investigación enfocado al análisis del método PLS y sus aplicaciones en quimiometría. El método de Partial Least Squares (PLS) es un método de análisis multivariante muy aplicado al analizar conjuntos de datos en dimensión alta. La Quimiometría es uno de los campos donde esta metodología es claramente de gran aplicabilidad debido a la gran abundancia de datos generada por la instrumentalización analítica.

Este trabajo desarrolla la revisión tanto a nivel metodológico como de software, de la aplicación del método PLS en este entorno comprobando su aplicabilidad al comparar

los diferentes algoritmos desarrollados para este método. Se realizó la parte práctica mediante el software R, con la librería “chemometrics”, en la cual se incluyen todos los conjuntos de datos analizados y las funciones utilizadas para realizar la regresión PLS y PCR. Se tomaron cuatro conjuntos de datos, como el conjunto de datos de Compuestos aromáticos policíclicos, datos de cenizas, datos de cereales y datos de radiación infrarroja cercana.

### 3.4. Análisis de datos de tres vías

Esta sección presenta las investigaciones más relevantes identificadas a partir de la revisión bibliográfica relacionada con el análisis de datos de estructuras de tres vías, destacando los diversos métodos de análisis empleados para abordar estas estructuras tridimensionales.

- [Jiang et al. \(1999\)](#) desarrolló un método denominado resolución de vectores acoplados alternos (ACOVER), para el análisis trilineal de una matriz de datos de tres vías. El ACOVER se basa en la resolución de cuatro errores (errores asociados con la estimación de las componentes de los vectores como error de medida, modelo, identificación y estimación) de resolución de vectores acoplados (COVER) propuestos en primer lugar, utilizando el principio de mínimos cuadrados. En segundo lugar, ideó un procedimiento que resuelve sucesivamente los perfiles de cada componente. Esta característica de resolución sucesiva del ACOVER proporciona una manera natural de evitar la degeneración de dos factores.

Simuló un conjunto de datos medido utilizando un sistema de cromatografía líquida de alto rendimiento (HPLC) con detección de matriz de diodos (DAD) en diez muestras. Al conjunto de datos se le aplicó el método propuesto, así como el algoritmo PARAFAC para fines de comparación.

Los resultados experimentales indican que el método ACOVER tiene una ventaja significativa, mostrando que los perfiles de interés analítico resueltos son extremadamente estables en relación con el número de componentes estimado. Esto es particularmente efectivo cuando el número de componentes se elige para que sea igual o mayor que la dimensionalidad real del modelo. Este enfoque evita el dilema de determinar un número adecuado de componentes para el modelo, un desafío difícil de manejar para el algoritmo PARAFAC. Además, el método ACOVER exhibe una tasa de convergencia mucho más rápida en comparación con el algoritmo PARAFAC

- [Westerhuis et al. \(1999\)](#) proporcionó una discusión crítica respecto al tratamiento metodológico de datos multidireccionales o que se encuentran en una matriz de tres vías. Inicialmente, realiza una presentación de los métodos de Análisis de Componentes Principales de múltiples vías (MPCA), que también es llamado Tucker1, el Análisis Factorial Paralelo (PARAFAC) y el método Tucker 3. Estos métodos han sido ampliamente utilizados para descomponer las estructuras de datos de tres vías, especialmente

bajo el contexto de análisis de datos para el monitoreo de procesos multivariantes con observaciones por lotes.

Mostró una comparación de estos métodos y discute las diferentes formas de desplegar los datos, centrado (consiste en restar la media de cada variable de la estructura de datos y a centrarlos alrededor de cero) y escalar (se escalan los datos dividiendo cada variable por su desviación estándar, lo que estandariza la escala de las variables). Aunque [Nomikos y MacGregor \(1994\)](#) proponen una forma apropiada para centrar y escalar los datos para aplicar la técnica de MPCA, recientemente se han utilizado otras formas de desplegar, centrar y escalar. En este artículo, se examinó el impacto que este tipo de preprocesamiento sobre el análisis de datos por lotes.

- [Smilde \(2001\)](#) analizó algunos modelos usados para el análisis de datos de tres vías, discutiendo algunos aspectos teóricos del uso de diferentes modelos para estas estructuras de datos en procesos por lotes y además, sugiere tener presente algunas consideraciones y consecuencias prácticas. Se aplica a su vez validación cruzada, para modelos PARAFAC, que permite establecer las formas de establecer el número de componentes significativas en un PCA y procesamiento de datos, a partir de diferentes formas de desplegar el arreglo de tres vías (despliegue matricial donde si el arreglo tridimensional es de dimensiones  $(I, J, K)$ , se puede desplegar en una matriz de dimensiones  $(I \times JK)$  o  $(IJ \times K)$ ), lo cual puede influir en favorecer un método sobre otro.
- [Singh et al. \(2006\)](#) realizó un análisis de datos multidireccionales de suelos regados con aguas residuales, utilizando modelos PCA de dos y tres vías para evaluar la contaminación del suelo/subsuelo y la movilidad de los contaminantes en los perfiles de suelo, la estructura de los datos es de 19 variables, 23 sitios medidos y 5 profundidades. Aunque el ACP bidireccional proporcionó información sobre la asociación entre los contaminantes relacionados con las aguas residuales y su acumulación en las capas superiores del suelo, debido a la estructura de datos de tres vías, no pudo extraer información sobre los patrones generales de comportamiento de los sitios de muestreo. El análisis de datos de tres vías se realizó utilizando modelos PARAFAC y Tucker3 validados para la estabilidad y bondad del ajuste. Un modelo PARAFAC de dos componentes, que explica el 36,23 % de la varianza de los datos, lo que permitió la interpretación de los datos. El primer componente se relacionó con las características generales de los suelos con poca variación vertical, mientras que el segundo componente está relacionado con los sitios y la profundidad contaminada con metales. La interpretación de los elementos centrales (Tucker3) que revelaron interacciones entre componentes de diferentes modos permitió inferir información más realista sobre el patrón de contaminación de los suelos tanto a lo largo de las coordenadas horizontales como verticales en comparación con las técnicas tradicionales de análisis de datos. Concluyó que los suelos en sitios más cercanos al punto de salida de aguas residuales están más contaminados con metales. El estudio concluyó además que los metales pesados relacionados con las aguas residuales

confinados a los perfiles superiores del suelo son más accesibles para la absorción de plantas de cultivo/vegetación cultivadas en la región y su hiperacumulación en partes comestibles de cultivos y vegetación forrajera puede representar un riesgo para los consumidores.

- [Bautista \(2009\)](#) presentó una comparación entre los métodos multivariantes de Tucker 3 y Análisis Factorial Múltiple (AFM) para el análisis de datos de tres vías, mediante la expresión del modelo y la función de pérdida, la inercia, las contribuciones y ayudas a la interpretación y las representaciones de individuos, variables, condiciones y trayectorias.

Para el desarrollo de la comparación, se tomó una base de datos que describía las características sociodemográficas y de salud de algunos departamentos de Colombia en el periodo 2004-2007. Los datos corresponden a un arreglo de dimensión  $(24 \times 12 \times 4)$ , en el cual  $I = 24$  representa los departamentos de Colombia,  $J = 12$  las variables sociodemográficas, como tasa bruta de natalidad (TBN), esperanza de vida al nacer (EVN), tasa bruta de mortalidad (TBM), entre otras, y  $K = 4$ , representa las condiciones correspondientes a los años 2004 a 2007. Aplicó a este conjunto de datos las dos metodologías (Tucker 3 y AFM) y se comparó sus resultados.

Del trabajo, las principales conclusiones que se obtienen respecto a las metodologías aplicadas sugieren que, a pesar de las diferencias metodológicas que se evidenciaron en los dos métodos evaluados, coincidieron en algunos elementos como las representaciones de individuos, variables, condiciones y trayectorias de individuos a través de condiciones.

Respecto a conclusiones de la base de datos por ambas metodologías, sugirieron que ambas metodologías se resumieron en tres factores de análisis, el primer factor fue la pobreza y el desarrollo, donde Bogotá presentó las mejores condiciones en comparación con Chocó, con niveles extremos de pobreza. El segundo factor estuvo asociado con la natalidad y mortalidad, la costa atlántica presentó mayores niveles de natalidad mientras que en la zona central del país son bajos, y por último, el tercer factor correspondía con maternidad materna y neonatal donde Chocó presentó los niveles más altos, Sucre y Córdoba los más bajos del país.

- [Camiz \(2011\)](#) realizó análisis de las elecciones en la provincia de Brescia - Italia, con el uso de técnicas de análisis de datos de tres vías. Para el estudio se tomó la información de la evolución política en Italia, desde 1948 hasta 1994. Los datos recopilados corresponden a los resultados electorales para cada municipio de la provincia de Brescia en las elecciones a la Cámara de Diputados desde 1948 hasta 1994, por tanto, la estructura de los datos puede considerarse como una matriz de 12 capas (una para cada año de elección), con 206 filas (el número de municipios en la provincia) y la cantidad de variables que corresponde al número de listas desafiantes en cada elección. Estos datos

fueron sometidos a las tres técnicas para el análisis de datos de tres vías, los cuales son Análisis Canónico Generalizado, Análisis Factorial Múltiple y STATIS, y finalmente se compararon con la técnica de Análisis de Componentes Principales (sin especificar el despliegue utilizado). Los resultados obtenidos mostraron que las diferencias de los análisis dependen de las diferentes aproximaciones de cada método. Bajo el caso de estudio, se evidenció una visión clara del contexto político y su estabilidad a lo largo del tiempo.

- [Feki-Sahnoun et al. \(2018\)](#) desarrolló la aplicación de un análisis con enfoque multi-tabla para evaluar la biogeografía del fitoplancton. El estudio se basó en el análisis del conjunto de datos recopilados durante el período 2004-2007 a lo largo de las zonas costeras del Golfo de Gabès, con una estructura de  $X_{(n,p,k)}$  donde n es el número de muestreos por sitio (n=48), p el número de unidades taxonómicas (p=57) y k representa el número del sitio (k=1...15). utilizando el método de tablas múltiples de STATIS. Este método permitió medir la similitud entre las 15 zonas de muestreo, la visualización de los conjuntos de especies que muestran variaciones espaciales similares y la reproducibilidad de este patrón para la asociación de cada zona. El alto gradiente espacial fue evidente para la mayoría de las especies y se obtuvieron tres grandes regiones. Luego se detectaron trece conjuntos de taxones diferentes que realizaban un patrón común de composición de fitoplancton. Para cada ensamblaje, se identificó la región más representativa que abarca grupos de zonas con propiedades similares. Se incrimina que los procesos hidrodinámicos y las propiedades geomorfológicas de las diferentes áreas, así como las actividades antropogénicas, desempeñan un papel importante en la configuración de los patrones estructurales del fitoplancton a lo largo de las zonas costeras del Golfo de Gabès. Este estudio sugirió considerar las características biogeográficas en la evaluación del conjunto espacial del fitoplancton.
- [Ruiz y Bonilla \(2018\)](#) realizaron un trabajo de investigación, con el objetivo de evaluar las diferencias en la estructura económica de las regiones de Colombia. Se realizó un análisis del PIB generado para los 32 departamentos que componen el país, teniendo en cuenta 10 variables importantes para este indicador nacional, en un período de tiempo desde el 2000 hasta el 2015, tomando valores del PIB por actividades productivas cada tres años (2000 - 2003 - 2006 - 2009 - 2012 - 2015).

Para el cumplimiento del objetivo planteado en este trabajo se realizó un análisis de datos de tres vías, debido a que se cuenta con una base de datos con variables (ramas de actividad económica en Colombia), individuos (Departamentos) y el tiempo (Los años tomados). Estos datos conforman una estructura de tres vías, cuyo análisis se realizó desplegando por filas y columnas el arreglo para obtener una perspectiva bidimensional de los datos. Este despliegue permitió observar y comparar las trayectorias de los departamentos (despliegue por filas) y la ubicación relativa global de los departamentos en los periodos de análisis (despliegue por columnas). Finalmente, se

aplican técnicas multivariantes como el Análisis de Componentes Principales Múltiple (Multi Way - PCA) y el método Tucker 3.

Los resultados principales, sugieron que existen 3 inductores en la economía nacional (Manufactura y Servicios, Explotación de Recursos Naturales, Sector Social), cuya evaluación temporal brindó una referencia de la evolución económica de los departamentos, además, permitió identificar las fortalezas productivas y caracterizar las unidades con altos niveles de desarrollo económico y aquellas unidades con desarrollo económico rezagado. Por otro lado, las técnicas mostraron niveles de inercia retenidos diferentes, siendo para MPCA 65.52 % y 64.44 % en ambos casos de despliegue, y para Tucker de 46.43 % en los dos despliegues utilizados. A pesar de lo anterior, para este estudio en particular se obtuvo resultados y conclusiones muy similares bajo las dos metodologías implementadas.

- [Swiercz y Mroczkowska \(2020\)](#) presentó la utilidad del método Multiway PCA (MPCA) para la detección temprana de fugas en el sistema de tuberías de una caldera de vapor en una planta de energía termoeléctrica. Para la aplicación del método MPCA para la detección temprana de fugas en un sistema de tuberías de una caldera de vapor en Elektrociepłownia Białystok. La estructura de los datos correspondía a registros diarios de variables de proceso medidas con fines de control en el período de 2011 al 2016. Los datos fueron almacenados en una base de datos de procesos.

Un largo segmento de mediciones de variables de proceso seleccionadas se dividió en una serie de “lotes” (que representan registros diarios del comportamiento normal de la planta) y se utilizó para crear el modelo MPCA de un sistema “saludable” en un espacio reducido de tres componentes principales (PC).

Utilizó el modelo MPCA actualizado periódicamente para establecer el elipsoide de confianza para el sistema “sano” en las coordenadas del PC. La decisión del personal de la probable detección de fugas está respaldada por la comparación de la ubicación actual del punto de operación (en la “trayectoria de falla”) con los límites del elipsoide de confianza. Cabe destacar que debido a los cambios diarios y estacionales de las demandas de calor/electricidad, las variables de proceso tienen una variabilidad sustancialmente mayor que en los ejemplos de procesos por lotes estudiados en la literatura. A pesar de esos desafíos reales para el método MPCA, los ejemplos numéricos confirmaron que el enfoque presentado pudo prever las fugas antes que el operador, generalmente de 3 a 5 días antes del apagado de la caldera. La metodología presentada puede ser útil en la implementación de un sistema en línea, desarrollado para mejorar la seguridad y el mantenimiento de las calderas en una planta de energía termoeléctrica.

- [Ajno \(2020\)](#) realizó la aplicación de la descomposición tensorial Tucker3 aplicado a tablas de contingencias de tres vías, utilizando datos de la Encuesta de Hogares del año 2011. La muestra total fue de 28644 personas. La tabla de contingencia tiene como

primera vía la variable Identidad, con cinco niveles (Quechua, Aymara, Otros Nativos, Ninguno y NS/NR). La segunda vía está asociada a la variable Idioma Materno, que tiene cuatro niveles (Aymara, Castellano, Otros y Quechua), y la tercera vía corresponde a la variable Departamento, que tiene nueve niveles. En este artículo se observan las bondades y utilidades del modelo Tucker 3, siendo una alternativa para el análisis de datos de tres vías en el campo multivariante, además, son técnicas que van en escala de crecimiento dado el que las estructuras de datos son cada vez más complejas y permiten el análisis integral de los datos.

### 3.5. Análisis de la asociación entre dos estructuras de datos de tres vías

En esta sección se exponen las investigaciones más relevantes encontradas durante la revisión bibliográfica sobre el análisis de la asociación entre dos estructuras de datos de tres vías, los cuales son el objeto de análisis de este proyecto. Se resaltan los distintos métodos de análisis utilizados para abordar la asociación entre estas estructuras tridimensionales.

- [Bro \(1996\)](#) presentó el método denominado mínimos cuadrados parciales de N-way (N-PLS). El algoritmo desarrollado es superior a los métodos de despliegue, principalmente debido a una estabilización de la descomposición. Esta estabilización proporciona potencialmente una mayor interpretabilidad y mejores predicciones.

El algoritmo es rápido en comparación con otras técnicas como PARAFAC, porque consiste en resolver problemas de valores propios. Realiza la aplicación del método con datos de la industria azucarera y se compara con la técnica unfold-PLS. Utiliza las matrices de emisión (EEM) para predecir el contenido de basura o cenizas en el azúcar. Los resultados mostraron que ambos métodos son comparables. Sin embargo, presentan una diferencia notable en su interpretación.

- [Slimani et al. \(2017\)](#) presentó la aplicación y uso de métodos denominados STATICO y COSTATIS. Técnicas altamente utilizadas en el análisis de datos ecológicos, donde se presentan estructuras de datos de tres vías. Se desarrollaron aplicaciones en el programa estadístico R ([R Core Team, 2020](#)) utilizando los paquetes ade4 ([Chessel et al., 2009](#)) y adegraphics ([Siberchicot et al., 2017](#)).

Se aplican al contexto de la ecología, (1) una serie de tablas de datos de especies y (2) una serie de tablas de parámetros ambientales. Los resultados obtenidos mostraron como los dos métodos se pueden utilizar para interpretar las relaciones entre las dos tablas de datos de tres vías.

- [Hervás et al. \(2018\)](#) presentan un paquete estadístico desarrollado bajo el entorno del programa estadístico R ([R Core Team, 2020](#)), para desarrollar mínimos cuadrados

parciales de N vías (N-PLS). Según el autor este método es una regresión superior a otros métodos para el análisis de datos de tres vías basados en el despliegue de matrices, que realiza una mejor descomposición de las estructuras de datos de tres vías, además proporciona una mejor interpretabilidad y mejores predicciones.

Este trabajo mostró las funciones del paquete sNPLS en el entorno de R, para realizar el análisis de datos de tres vías.

- [Hanafi et al. \(2020\)](#) realizó un análisis de datos multibloque y tridireccional. En este estudio describe una revisión de los métodos populares en quimiometría para analizar datos multibloque y de tres vías. Estos métodos abarcan estrategias de análisis para el aprendizaje no supervisado cuyo objetivo fue investigar la estructura de los datos, y el aprendizaje supervisado para relacionar datos univariados o multivariados con datos multibloque o tridireccionales. Los métodos no supervisados incluyen estrategias de análisis como el análisis de componentes principales multibloque y el modelo PARAFAC.

Para el aprendizaje supervisado, se esbozan PLS multibloque y PLS N-way. Se discuten ilustraciones basadas en estudios de casos reales.

En resumen de este capítulo, se concluye que, a pesar de contar con más de 20 años de desarrollo en la investigación, son escasas las metodologías que exploran la asociación entre dos estructuras de tres vías como se observa en la Tabla **3-3**.

| Análisis individual |                                                         | Asociación de estructuras |                                            |
|---------------------|---------------------------------------------------------|---------------------------|--------------------------------------------|
| 2 vías              | 3 vías                                                  | 2 vías                    | 3 vías                                     |
| PCA                 | PARAFAC<br>TUCKER 3<br>MPCA<br>AFM<br>ANCOVER<br>STATIS | PLS<br>AFM                | Unfold-PLS<br>N.PLS<br>STATICO<br>COSTATIS |

**Tabla 3-3.:** Resumen de técnicas identificadas en los antecedentes

Por tanto, la búsqueda de metodologías capaces de abordar eficazmente la asociación entre dos estructuras de datos de tres vías permite comprender mejor las complejas relaciones entre estos conjuntos de datos. Además, proporciona información valiosa sobre cómo estas asociaciones pueden influir en diversos fenómenos o procesos de interés en diferentes campos de estudio.

## 4. Marco Teórico

En este capítulo presenta el sustento teórico de los temas principales para el desarrollo de esta investigación. A partir del capítulo de antecedentes y el desarrollo de un análisis bibliométrico se obtuvo un marco estadístico con los fundamentos teóricos que sustentan los métodos de análisis multivariantes empleados en el análisis de datos de tres vías y son insumos para el planteamiento de la metodología propuesta para la aplicación en el caso de estudio. Se define inicialmente los métodos utilizados para el análisis de estructuras de datos de dos vías, como lo son el Análisis de Componentes Principales (ACP), que es la base fundamental para los métodos multivariados y pueden ser extendidos a otras técnicas multivariantes como lo es el Multiway ACP, además, el Análisis Factorial Múltiple (AFM), los métodos utilizados para el análisis de datos de tres vías, para este caso se identifican las técnicas Multiway PCA, PARAFAC y STATIS. Finalmente, se describe teóricamente las técnicas para la asociación entre estructuras de datos de dos y tres vías como lo son los Mínimos Cuadrados Parciales (PLS) y Mínimos Cuadrados Parciales Multiway (N-PLS). Algunas de estas serán las técnicas seleccionadas para el desarrollo del caso de estudio que se realiza en este trabajo. Una descripción más detallada de estos métodos puede ser consultada en [Lebart et al. \(1995\)](#) y [Escofier y Pagès \(1992a\)](#) y otros trabajos citados en el desarrollo de este capítulo.

Asimismo, este capítulo responde y da cumplimiento al segundo objetivo de este trabajo.

### 4.1. Análisis de estructura de datos en dos vías

La estructura de las tablas a las cuales comunmente nos enfrentamos corresponde a una tabla bidimensional, en las cuales se registran variables e individuos. La representación de esta estructura de datos se observa en la Figura 4-1.

p variables cuantitativas

|  |          |          |          |          |
|--|----------|----------|----------|----------|
|  | $x_{11}$ | $x_{12}$ | ...      | $x_{1p}$ |
|  | $x_{21}$ | $x_{22}$ | ...      | $x_{2p}$ |
|  | $x_{31}$ | $x_{32}$ | ...      | $x_{3p}$ |
|  | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
|  | $x_{n1}$ | $x_{n2}$ | ...      | $x_{np}$ |

n individuos

$n \times p$

Figura 4-1.: Estructura de datos en dos vías

### 4.1.1. Análisis de Componentes Principales - ACP

El Análisis de Componentes Principales (ACP) es una de las técnicas con mayor importancia y aplicación del análisis multivariado de una matriz de datos (2 vías), presentado por Pearson en 1901, sin embargo fue hasta 1933 donde Hotelling logró su integración como una técnica estadística matemática (Pearson, 1901; Hotelling, 1933).

El objetivo de esta técnica es obtener una representación óptima de un espacio de dimensiones reducidas a partir de los datos originales, además de transformar los datos a nuevas variables no correlacionadas llamadas factores o componentes principales (Pearson, 1901), lo que permite comprender las relaciones existentes entre individuos y variables que componen la hoja de datos. En otras palabras, con esta técnica es posible la explicación de la varianza de un conjunto de datos a partir de combinaciones lineales de las variables originales (Nomikos y MacGregor, 1994).

El ACP se aplica a arreglos estructurados con  $n$  filas y  $p$  columnas ( $n \times p$ ), cuyas columnas contienen  $p$  variables continuas y  $n$  filas que representan los individuos a quién se les miden las variables (Lebart et al., 1995), como se muestra en la Figura 4-1.

En la matriz  $X$  de la Figura 4-1, cada elemento  $x_{ij}$  representa la medida de la variable  $j$  en el individuo  $i$ . Inicialmente, antes de aplicar la técnica de ACP, debido a las diferentes unidades de medida de cada variables (columnas), se suele realizar un proceso de estandarización por columnas de la siguiente manera:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Donde,  $\bar{x}_j$  y  $s_j$  son el promedio y desviación estándar de la variable  $j$ , calculados de la siguiente manera:

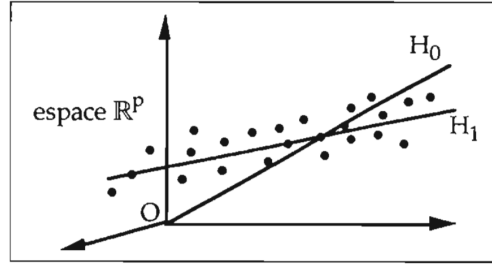
$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2; \quad s_j = \sqrt{s_j^2}$$

Con este proceso de estandarización se garantiza que todas las variables tengan la misma escala de medida y por consiguiente, evitar la excesiva influencia de una variable sobre el resultado. Además, según Ochoa (2018) para aplicar esta técnica se deben cumplir los siguientes requerimientos, los cuales garantizan resultados adecuados.

- Las variables son continuas.
- La matriz  $X$  no puede tener datos faltantes.
- Supuesto de linealidad en las variables, la cual establece que las variables están relacionadas de manera lineal entre sí.

### Análisis de la nube de individuos

La nube de individuos en ACP es el ajuste de  $n$  individuos en el espacio de las variables medidas en dichos individuos, ajustando los  $n$  individuos a un subespacio de dos dimensiones y así obtener una representación visual (ver Figura 4-2) en el que se identifique las proximidades de los individuos con respecto a las  $p$  variables.



**Figura 4-2.:** Representación de la nube de individuos. (Lebart et al., 1995)

Para el análisis e interpretación de la nube de individuos que se encuentran en el espacio  $R^p$ , es necesario observar su cercanía, lo cual se logra a partir de la distancia euclidiana:

$$d^2(i, i') = \sum_{j=1}^p m_j (x_{ij} - x_{i'j})^2$$

Donde,  $m_j$  es la métrica que pondera la influencia de cada variable  $j$ . Por tanto, para analizar la proximidad de los individuos es necesario, según Lebart et al. (1995), maximizar la suma de cuadrados de las distancias entre los pares de individuos  $\text{Max}_H \{ \sum_{i, i'}^p d_H^2(i, i') \}$ , lo cual se logra al obtener un vector unitario  $H$  que realiza la mejor representación en el eje  $\alpha(H_\alpha)$ , que a su vez equivale a maximizar la suma de los cuadrados de las distancias entre los puntos y el centro de gravedad de la nube  $G \text{Max}_H \{ \sum_i^p d_H^2(i, G) \}$ .

Por consiguiente, dada la estandarización inicial a los datos, los individuos se trasladan alrededor del origen  $O$ , por tanto se desarrolla una diagonalización de la matriz de correlaciones  $ZNZ$ , donde la matriz  $N$  es una matriz diagonal que representa los pesos que se asocian a cada individuo  $N = \text{diag}(1/n)$  y  $Z$  es la matriz de datos estandarizados. Las coordenadas de los individuos quedarán determinadas por  $\psi_\alpha = Zu_\alpha$  ( $u_\alpha$  son los ejes ortogonales), que a su vez son los componentes de dicho vector y donde las  $\alpha$  componentes se obtienen por combinación lineal de las variables originales. Por consiguiente, cada componente se obtiene de la siguiente manera:

$$\psi_1 = u_{11}Z_1 + u_{12}Z_2 + \dots + u_{1p}Z_p$$

$$\psi_2 = u_{21}Z_1 + u_{22}Z_2 + \dots + u_{2p}Z_p$$

$$\vdots$$

$$\psi_p = u_{p1}Z_1 + u_{p2}Z_2 + \dots + u_{pp}Z_p$$

El orden de las componentes se encuentra en función del porcentaje de varianza de la matriz  $z$  que dicho componente puede explicar. Por tanto, el primer componente será el que mayor porcentaje de variabilidad explica de los datos.

### Análisis de la nube de variables

Con respecto al análisis de la nube de variables, las cuales se asocian a un espacio  $R^n$ . La distancia entre estas variables se obtiene aplicando la formula usual de la distancia euclídea de la siguiente manera:

$$d^2(j, j') = \sum_{i=1}^n (z_{ij} - z_{ij'})^2$$

De esta ecuación, según [Lebart et al. \(1995\)](#), se obtiene que:

$$\sum_{i=1}^p z_{ij}^2 = \sum_{i=1}^p z_{ij'}^2 = 1$$

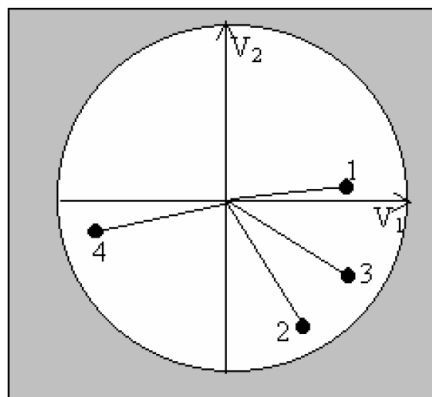
$$\sum_{i=1}^p z_{ij} z_{ij'} = c_{jj'}$$

Por tanto, dado que  $c_{jj'}$  es el coeficiente de correlación entre las variables  $j$  y  $j'$ , entonces se puede expresar la distancia como:

$$d^2(j, j') = 2(1 - c_{jj'})$$

Y esto implica que la distancia estará entre 0 y 4 ( $0 \leq d^2(j, j') \leq 4$ ).

De lo anterior, según [Lebart et al. \(1995\)](#) la interpretación de la nube de variables se realiza a partir del círculo de correlaciones. Se considera que dos variables están muy próximas cuando se encuentran fuertemente correlacionadas, esto implica que  $c_{jj'} = 1$ , son ortogonales cuando  $c_{jj'} = 0$  y en caso contrario se encuentran muy alejadas cuando  $c_{jj'} = -1$ .



**Figura 4-3.:** Círculo de correlaciones, (Lebart et al., 1995)

Para el cálculo de las coordenadas se hace uso del principio de relación de transición, a partir del cual es posible relacionar el espacio de los individuos con el de las variables y obtener las coordenadas o componentes principales asociadas a las variables. Por tanto, estas se pueden calcular de la siguiente manera:

$$\varphi_{\alpha} = \sqrt{\lambda_{\alpha}} u_{\alpha}$$

### Selección de componentes

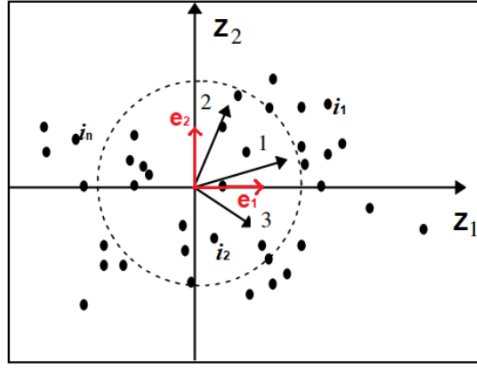
El objetivo del ACP es la simplificación de una estructura de datos de  $j$  variables en un espacio de dimensión  $q$ , lo cual implica la sustitución de  $j$  variables originales en  $q$  variables nuevas que resuman la información de manera óptima. El número adecuado de componentes se seleccionan a partir de algunos criterios que pretenden condensar de mejor manera la estructura de los datos. A continuación se destacan las siguientes:

- **Criterio de Kaiser:** Consiste en eliminar las componentes cuyos valores propios sean menores a uno (Braeken y van Assen, 2017).
- **El scree test:** Consiste en realizar un gráfico descendiente de los valores propios. Los componentes seleccionados son aquellos en los que se identifica un aporte significativo y diferencial a la varianza explicada por los componentes (Cattell, 1966).
- **Porcentaje de Inercia acumulada:** Se busca el mejor subespacio de dimensión ( $q \leq p$ ) que recoja gradual y decrecientemente las proporciones de inercia proyectadas (González, 2014).

### Representación Simultánea

En el ACP las nubes de individuos y variables se encuentran definidas en espacios, orígenes y bases distintos. El origen para la nube de individuos es dado por el centro de gravedad

de los individuos de dimensión  $p$  mientras que para la nube de variables es de dimensión  $n$ , los ejes factoriales para la nube de individuos se representa con  $u_\alpha$  para la nube de variables se representa con  $v_\alpha$ . Lo anterior nos indica que los puntos fila y columna se encuentran en espacios distintos y es imposible visualizar las dos nubes en un solo espacio [González \(2014\)](#). No obstante, la posibilidad de una representación simultánea consiste en tomar las direcciones de los puntos variables del gráfico de correlaciones y representarlas en el gráfico de los individuos. Es decir, reside en la proyección en  $R^p$ .



**Figura 4-4.:** Individuos y variables representados conjuntamente, ([González, 2014](#))

La dirección de una variable define zonas para los individuos: de un lado, aquellos que toman valores fuertes para esta variable y, al lado opuesto, aquellos que toman valores bajos. Nos interesará el alargamiento de los individuos en la dirección de la variable ([González, 2014](#); [Lebart et al., 1995](#)).

### Interpretaciones del ACP

Los ejes factoriales permiten obtener la mejor visualización aproximada de las distancias entre los individuos de una parte y de otra parte las variables que a su vez califican y caracterizan tanto a los ejes como a los individuos ([Lebart et al., 1995](#); [González, 2014](#)).

**Contribuciones de los individuos:** Se define la contribución de un individuo en un eje de inercia como la importancia de dicho individuo a lo largo del eje:

$$CTR(i, \alpha) = \frac{\psi_{(i, \alpha)}^2}{n\lambda_\alpha}, \text{ donde } \sum_{i=1}^n CTR(i, \alpha) = 1$$

**Contribuciones de las variables:** La contribución de una variable en un determinado eje de inercia es una medida de la importancia de la variable en dicho eje y, en concreto, determina qué parte de la inercia total que existe a lo largo del eje se debe a la variable.:

$$CTR(j, \alpha) = \frac{u_{(j, \alpha)}^2}{n\lambda_\alpha}, \text{ donde } \sum_{j=1}^n CTR(j, \alpha) = 1$$

La calidad de la representación presenta una idea acerca de que tan cercano se encuentra un punto a un componente principal. Su cálculo se realiza mediante los cosenos cuadrados entre cada punto y el componente principal de la siguiente forma:

Individuos:

$$\cos^2(i, \alpha) = \frac{\psi_{(i, \alpha)}^2}{d^2(i, G)}$$

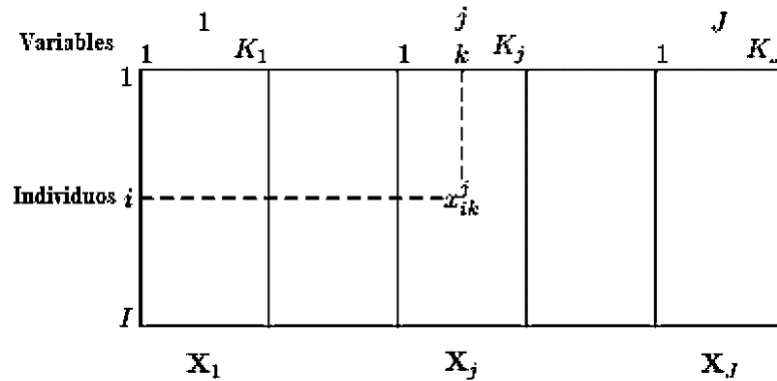
Variabes:

$$\cos^2(j, \alpha) = \varphi_{(j, \alpha)}^2 = \text{cor}^2(j, \alpha)$$

#### 4.1.2. Análisis Factorial Múltiple - AFM

El Análisis Factorial Múltiple (AFM) hace parte de los métodos para el análisis multivariado, desarrollado por [Escofier y Pagès \(1992a\)](#). El AFM permite representar la nube de puntos en un espacio de menor dimensión para observar las distancias entre individuos, en general, se busca disminuir la dispersión o variabilidad del espacio debido a los grupos.

En general, el AFM se aplica a tablas de datos en las que los mismos individuos son medidos por diferentes grupos de variables según su naturaleza (cualitativas o cuantitativas), donde  $J_c$  representa las tablas de variables cuantitativas,  $J_q$  las tablas de variables cualitativas (nominales) y  $J_o$  las tablas de variables cualitativas (ordinales) como se muestra en la figura 4-5.



**Figura 4-5.:** Matriz X conformada por la yuxtaposición de tablas, ([Escofier y Pagès, 1992a](#))

El objetivo principal al realizar un AFM es identificar la existencia de estructuras comunes a todas, o a parte, de las nubes parciales. Lo anterior indica que el AFM es un ACP ponderado, donde se busca representar la nube de puntos en un espacio de menor dimensión, conservando la distancia entre individuos.

El método AFM se divide en dos etapas de desarrollo

- **Primer etapa:** Se realiza un análisis separado para obtener los factores parciales y los valores propios mayores, lo que indica que se analiza el grupo de variables de naturaleza cuantitativa ( $Jq$ ) mediante la aplicación de un ACP y ACM para el grupo de variables de naturaleza cualitativa ( $Jq$  o  $Jo$ ). De lo anterior, de cada grupo de variable analizado por separado se obtiene el primer valor propio denotado por  $\lambda_j^1$  (Escofier y Pagès, 1992a).
- **Segunda etapa:** Se realiza un análisis global, ACP ponderado donde cada tabla es ponderada por el inverso del primer valor propio denotado como  $1/\lambda_j^1$ , conformando la matriz diagonal  $M$ . La ponderación garantiza la estructura de las tablas ya que cada variable que pertenece al mismo grupo recibe el mismo peso, asimismo bajo lo anterior se logra equilibrar la inercia de los grupos, debido a que la inercia máxima que toma cada una de las nubes de individuos definida para los diferentes grupos, vale 1 en cualquier dirección (Abascal y Landaluce, 2002).

El AFM considera tres espacios de análisis de los datos, el espacio de la nube de individuos  $R^k$ , el espacio de la nube de variables  $R^I$ , el espacio de la nube de grupos de variables  $R^{I^2}$ .

### Ponderación de los grupos de variables

Inicialmente se halla la ponderación de los grupos de variables para equilibrar el papel de estos en todos los aspectos del análisis, para lo cual se debe obtener el peso de cada variable del grupo  $j$ , donde según (Escofier y Pagès, 1992a) se divide por  $\lambda_j^1$ , donde  $\lambda_j^1$  es el primer propio del ACP o ACM (según la naturaleza de la variable).

Posteriormente, se realiza un ACP para la tabla general (estandarizada y ponderada) y se obtienen los valores y vectores para el análisis general en el que la inercia de cada factor está conformada proporcionalmente por las variables que se encuentran activas en cada grupo para ser usadas como coordenadas para la gráfica.

### La representación de los grupos de variables

Uno de los objetivos del estudio de varios grupos de variables es comparar globalmente los grupos. Para ello, se introduce el espacio  $R^{I^2}$  base para la representación como ayuda para la interpretación del ACP de la tabla completa  $X$  pero que posee su propia optimización.

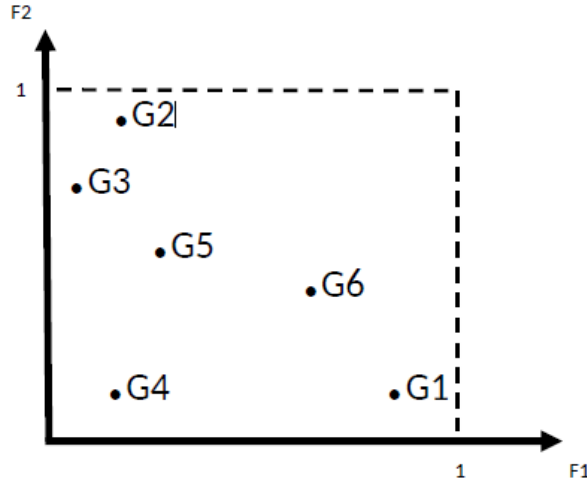
La nube  $N_j$  se conforma por las distancias entre los grupos de variables, las cuales se derivan del producto escalar  $L_g = \langle W_j D, W_l D \rangle = \text{traza}(W_j D W_l D)$ , donde  $W_j D = X_j M_j X_j' D$ , siendo  $X_j$  la matriz de datos asociada al grupo  $j$  y  $D$  es la matriz diagonal de los pesos de los individuos.

El producto escalar se interpreta como una medida de relación entre los grupos y a partir de este se construye el coeficiente de correlación vectorial  $RV$ . Su valor está comprendido

entre 0 y 1, donde 0 indica la no existencia de relación entre las variables y 1 las nubes que representan a los grupos son homotéticas, lo cual se refiere a relaciones o transformaciones que conservan la forma y la proporción.

$$RV = \frac{\langle W_j D, W_l D \rangle}{\|W_j D\| \|W_l D\|} = \frac{\text{traza}(W_j D W_l D)}{\|W_j D\| \|W_l D\|}$$

Donde  $W_j D = X_j X_j' D$ , siendo  $X_j$  la matriz de datos asociada al grupo  $j$  y  $D$  es la matriz diagonal de los pesos de los individuos. La norma de  $W_j D$  se expresa como  $\|W_j D\| = \sqrt{\sum_s (\lambda_s^j)^2}$  donde  $\lambda_s^j$  es el  $j$ -ésimo valor propio del análisis de la tabla  $j$ . El numerador representa los productos escalares entre los representantes de los grupos.



**Figura 4-6.:** Ejemplo de representación de los grupos de variables en el AFM

En la Figura 4-6 se observa un ejemplo de representación gráfica de los grupos de variables en el AFM, según esta gráfica se muestra que el primera factor se debe principalmente a los grupos 1 y 6, el segundo factor se encuentra más relacionado con los grupos 2 y 3 y por otro lado, el grupo 4 no se encuentra relacionado con ninguno de los dos ejes factoriales.

En resumen, valores altos en las coordenadas de los grupos sugieren una fuerte asociación del factor con las variables de esos grupos, mientras que valores bajos indican una relación más débil del factor con esas variables.

### Factores comunes

Según [Escofier y Pagès \(1992b\)](#), si la correlación entre los grupos (ya sea todos o algunos de ellos) y el eje factorial es fuerte, implica la presencia de un “factor común”. Este factor se calcula a partir del concepto de inercia acumulada de las variables representadas en el grupo con respecto a dicho factor, lo que puede denominarse como el “peso” del factor en el grupo.

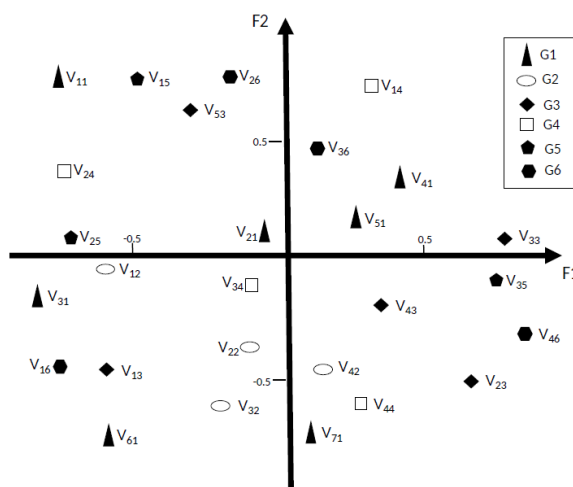
### Representación de las variables

La representación se obtiene del ACP de la tabla completa  $X$ , previamente ponderada y se utilizan las coordenadas de las proyecciones sobre los ejes factoriales, igualmente, estas se pueden calcular usando las relaciones de transición del ACP.

$$\varphi_s = \sqrt{\lambda_s} u_s; \quad s = 1, 2, \dots, K$$

La representación de las variables puede considerarse a la vez:

- Como una ayuda para la interpretación de la representación de la nube de individuos.
- Como una representación óptima de las correlaciones entre variables



**Figura 4-7.:** Ejemplo de representación de las variables sobre los ejes factoriales del AFM, [Garcia A \(2021\)](#)

La Figura 4-7 muestra la proyección de las variables en el primer plano factorial. Con base en esta representación, en la Tabla 4-1 se identifican las variables de cada grupo que están más relacionadas con los factores.

| Factores | G1               | G2                       | G3       | G4       | G5               | G6               |
|----------|------------------|--------------------------|----------|----------|------------------|------------------|
| Factor 1 | $V_{31}, V_{11}$ | $V_{12}$                 | $V_{33}$ | $V_{24}$ | $V_{25}, V_{35}$ | $V_{46}, V_{16}$ |
| Factor 2 | $V_{71}$         | $V_{42}, V_{22}, V_{32}$ | $V_{53}$ | $V_{14}$ | $V_{15}$         | $V_{26}$         |

**Tabla 4-1.:** Ejemplo de identificación de variables que más aportan a la construcción de los ejes factoriales, [Garcia A \(2021\)](#)

Los componentes principales maximizan la inercia de las proyecciones para todas las variables. La inercia proyectada por cada nube  $N_j^k$  puede ser interpretada como la contribución

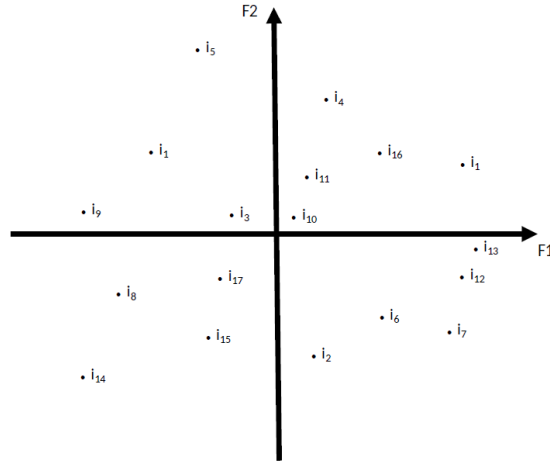
de un grupo. La ponderación de los grupos  $(\frac{1}{\lambda_j})$  equilibra la influencia en el sentido de que la contribución de un grupo a la construcción de un eje está limitada a 1. Por tanto, ningún grupo por si solo puede determinar el primer eje y que un grupo influye sobre tantos ejes cuanto mayor sea su dimensionalidad.

### Representación de los individuos

Sea  $u_s$  el eje de inercia de orden  $s$  de la nube de individuos  $N_I$  asociada a la tabla  $X$  en  $R^K$ . Este se deduce de la componente principal  $F$  por la relación:  $u_s = (1/\sqrt{\lambda_s})X'DF_s$  donde  $\lambda_s$  es el valor propio de  $WD$  asociado a  $F_s(W = \sum_j W_j)$ . La proyección de  $N_I^j$  sobre  $u_s$  se escribe:

$$F_s^j = \tilde{X}_j M u_s = (1/\sqrt{\lambda_s}) \tilde{X}_j M X' D F_s = (1/\sqrt{\lambda_s}) W_j D F_s$$

Las  $J$  nubes de individuos, llamadas “nubes parciales”, se proyectan como ilustrativas sobre los ejes.



**Figura 4-8.:** Ejemplo de representación de los individuos sobre los ejes factoriales del AFM, [Garcia A \(2021\)](#)

La Figura 4-8 es una representación gráfica ficticia que ilustra la disposición de los individuos en el primer plano factorial. Se observa que el individuo  $i_{14}$  es el más excéntrico en ambos ejes, mientras que los individuos  $i_{12}$  e  $i_{13}$  se alinean con el primer factor (a diferencia del individuo  $i_9$ ). Por otro lado, el individuo  $i_5$  tiene una contribución significativa al segundo factor. Por último, los individuos  $i_3$ ,  $i_{10}$ ,  $i_{11}$  e  $i_{17}$  se consideran como “individuos promedio” debido a su débil influencia en la construcción del primer plano factorial.

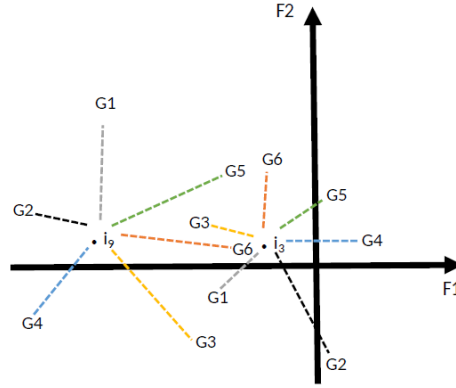
## Representación de los individuos

La representación simultánea en AFM se refiere a la visualización de múltiples nubes de individuos en un mismo espacio, lo que permite comparar y analizar las relaciones entre ellas de manera conjunta. En el Análisis Factorial Múltiple (AFM), estas nubes de individuos representan diferentes grupos o categorías de datos, y su representación simultánea proporciona información sobre cómo se relacionan entre sí en el espacio de variables latentes. Según [Escofier y Pagès \(1998\)](#), la representación simultánea de las  $J$  nubes  $N_I^j$  en  $\mathbb{R}^K$  se logra a partir de las siguientes consideraciones:

- La nube  $N_I^j$  está en  $\mathbb{R}_j^K$ , subespacio de  $\mathbb{R}^K$ .
- $\mathbb{R}^K = \bigoplus_j \mathbb{R}_j^K$  (suma directa de los subespacios ortogonales dos a dos).
- $M_j$  es la métrica en  $\mathbb{R}_j^K$ , submatriz con las  $K_j$  filas y  $K_j$  columnas de  $M$ .
- Las coordenadas de los puntos de  $N_I^j$  son las filas de  $X_j$ , entonces las coordenadas de estos puntos en  $\mathbb{R}^K$  son las filas de la matriz  $\tilde{X}_j$  definida como:

$$\tilde{X}_j = [0 \cdots 0 X_j 0 \cdots 0]$$

- $X = \sum_j \tilde{X}_j$ , entonces con el fin de buscar coincidir  $F_s(i)$  con el punto medio de los  $J$  puntos parciales  $F_s^j(i)$ , las filas de  $\tilde{X}_j$ , denominadas filas parciales, se proyectan como puntos suplementarios amplificados por  $J$ . En efecto  $X = \frac{1}{J}(J \sum_j \tilde{X}_j)$ .
- La unión de las  $J$  nubes parciales forma la nube  $N_I^J$  de  $IJ$  puntos, que también se puede considerar como la unión de  $I$  nubes  $N_i^J$  cada una con los  $J$  puntos parciales  $i^j$  de la misma fila  $i$ .
- La inercia de la mnube  $N_I^J$  se puede expresar como  $InerciaIntra + InerciaEntre$  subnubes  $N_i^J$ . La inercia de la nube media  $N_I$  del AFM corresponde a la  $InerciaEntre$ , cuando las coordenadas parciales se amplifican por  $J$  ([Escofier y Pagès, 1998](#)).



**Figura 4-9.:** Ejemplo de representación superpuesta de la nube de individuos en AFM, [Garcia A \(2021\)](#)

La Figura 4-9 presenta un ejemplo de la representación superpuesta, que para efectos de mayor claridad en la interpretación se realiza para dos individuos, por tanto, el individuo  $i_9$  está mejor representado por los grupos G5 y G6, mientras que para estos mismos grupos de variables el individuo  $i_3$  tiene un aporte más bajo.

### Ayudas para la interpretación

Al igual que en el ACP, para identificar el concepto subyacente que representan los ejes factoriales del AFM y comprender las relaciones entre individuos, variables y grupos, se calcula la contribución de cada uno de estos objetivos de análisis a los ejes factoriales.

### Contribución de los individuos a una dimensión

$$ctr_{i,s} = \frac{m_i(F_s^j)^2}{\lambda_s}$$

Donde  $m_i$  y  $F_s^j$  son el peso del  $i$ -ésimo individuo (usualmente  $\frac{1}{n}$ ) y la coordenada del  $i$ -ésimo individuo en el  $s$ -ésimo eje factorial respectivamente.

### Contribución de las variables a una dimensión

Al igual que con los individuos, se puede medir la importancia de las variables para una dimensión dada. Formalmente, la contribución de la  $k$ -ésima variable a la  $s$ -ésima dimensión está dada por

$$ctr_{k,s} = \frac{\psi_s^2}{\lambda_j^1}$$

Todas las variables de un mismo grupo comparten el mismo peso dado por  $\frac{1}{\lambda_j^1}$ .

### Contribución de un grupo de variables a una dimensión

Está definida como la sumatoria de las contribuciones del  $j$ -ésimo grupo de variables en el  $s$ -ésimo eje factorial.

$$ctr_{j,s} = \sum_{k=1}^{K_j} ctr_{k,s}$$

Las contribuciones toma valores entre 0 y 1. Para un eje factorial dado, la suma de las contribuciones de todos los individuos (o variables) es igual a 1. Individuos (o variables) con altas contribuciones y con coordenadas factoriales con diferentes signos pueden entonces contrastarse para ayudar a interpretar el componente ([García A, 2021](#)).

### Elementos suplementarios

Al igual que en el Análisis de Componentes Principales (ACP), en el Análisis Factorial Múltiple (AFM) también es posible proyectar elementos suplementarios. Estos elementos suplementarios brindan la oportunidad de explorar asociaciones con los elementos activos o enriquecer el análisis. Se denominan activos a los elementos que participan directamente en el análisis.

### Individuos suplementarios

A menudo, es necesario sustituir el estudio de cada individuo por el de las clases de individuos que comparten una característica en común. Para esto, se introducen como elementos suplementarios los centros de gravedad de estas clases. Estos individuos no influyen en las representaciones de los individuos activos: se calcula su proyección en la nube  $N_I^*$  y en las diferentes nubes  $N_I^j$ .

### Grupos de variables suplementarios

Si hay un grupo adicional que pueda servir como referencia, resulta interesante compararlo con otros grupos utilizando todos los medios disponibles para estos últimos, sin que el grupo de referencia influya en la nube media ni en los resultados del análisis. La mayoría de los cálculos (aunque no todos) realizados sobre los grupos activos se aplican también a un grupo suplementario ([Escofier y Pagès, 1998](#)).

- Normalización de la nube  $N_K^j$ : es necesario normalizarla de la misma forma sobreponderando las variables del grupo para comparar la nube asociada al grupo suplementario con las otras nubes.

- Proyección de las componentes principales del grupo: a través de ella es posible comparar la forma general de la nube  $N_K^j$  con la de la nube media  $N_K$  y la de las nubes asociadas a otros grupos de variables.
- Proyección de los  $W_jD$ : La coordenada de un elemento suplementario  $W_jD$  sobre el eje de rango  $s$  concuerda con la medida de la relación entre el  $s$ -ésimo factor y el grupo  $j$ .

### Selección del número de ejes factoriales

Como se ha mencionado anteriormente, el AFM puede considerarse una generalización del ACP, por lo que la selección del número de ejes factoriales a retener puede seguir los mismos criterios de varianza explicada o valores propios propios superiores a 1 utilizados en el ACP.

## 4.2. Análisis de la asociación entre estructuras de datos en dos vías

En esta sección, se abordan técnicas estadísticas empleadas para examinar y comprender la relación entre dos estructuras de datos bidimensionales. En el marco de los antecedentes, se detallan métodos como PLS y PLS2, los cuales resultan útiles cuando se dispone de múltiples conjuntos de datos bidimensionales y se busca comprender su asociación.

### 4.2.1. Mínimos Cuadrados Parciales - PLS

El método de Mínimos Cuadrados Parciales (Partial Least Squares - PLS por sus siglas en inglés) fue introducido inicialmente por el estadístico Herman Wold en 1975 para un tratamiento de cadenas de matrices y aplicaciones en econometría. Sin embargo, el método PLS fue por mucho tiempo desconocido y misterioso, hasta que Svante Wold y otros introducen el método en otros campos de la ciencia donde toma gran fuerza ([Gaviria, 2016](#)).

El método PLS cuenta con características muy versátiles para realizar el análisis de múltiples variables independientes  $X$  con múltiples variables dependientes  $Y$ , con el objetivo de reducir la dimensión de la regresión múltiple con la garantía de que las primeras componentes cumplen la condición de ortogonalidad y además, mejoran la explicación de  $Y$ .

Según [Gaviria \(2016\)](#), los métodos PLS y PCR (Regresión por Componentes Principales) tienen estructuras similares, ya que ambas metodologías realizan una transformación de las variables explicativas en un conjunto de variables latentes linealmente independientes, su diferencia radica en que la regresión PCR no considera la variable de respuesta para determinar las variables latentes, mientras que PLS determina las variables latentes a partir

de las variables independientes y dependientes. Por lo anterior, PLS es una metodología que combina la metodología multivariante PCA para la maximización de la varianza en las variables explicativas y los MCO para maximizar la correlación entre las respuestas.

Asimismo, [Gaviria \(2016\)](#) indica que existen unos factores que representan unas ventajas del uso de esta metodología que se mencionan a continuación:

- La regresión PLS es un método potente, ya que es flexible con el número de variables cuando es muy grande y considera la multicolinealidad.
- El modelo predice la(s) respuesta(s) a partir de un conjunto de variables linealmente dependientes
- Para obtener el modelo final, durante el proceso se usa un número pequeño de componentes PLS calculados.
- El modelo PLS puede ser optimizado para un alto rendimiento en la predicción a partir del número de componentes PLS y estas a su vez determinan la complejidad del modelo.

La regresión PLS puede ser aplicado cuando se tiene una sola variable respuesta y recibe el nombre de PLS1. Cuando hay más de una variable respuesta, la regresión PLS recibe el nombre de regresión PLS2.

Existen algoritmos para el desarrollo de ambas regresiones, donde para el caso de PLS1 se procede de la siguiente manera:

1. Inicialmente, se calcula el primer componente PLS que es la variable latente que contiene la máxima covarianza entre los scores y la variable de respuesta y.
2. Luego, se elimina la información de esta componente de las variables explicativas, generalmente el proceso es llamado deflación que indica que lo anterior se puede ver como una proyección del espacio de las variables de respuesta sobre un hiperplano ortogonal. Se obtiene una matriz residual con el mismo número de variables que la matriz original la diferencia radica en que esta tiene una dimensión reducida a uno.
3. Con la matriz residual se calcula la siguiente componente.
4. Se repite el proceso hasta no evidenciar un cambio significativo en el modelamiento de la variable de respuesta.

El algoritmo PLS2 es una extensión del algoritmo PLS1 anteriormente descrito. Sin embargo, este realiza el modelamiento para más de una variable explicativa o en caso contrario, el algoritmo PLS1 es un caso particular para el algoritmo extendido PLS2.

### 4.2.2. PLS1

El algoritmo PLS1 se realiza cuando se tiene solo una variable de respuesta, es un caso particular del PLS2 ([González, 2014](#)).

Se quiere estimar via PLS la regresión multiple:

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Que corresponde a una sola variable a explicar  $Y$  sobre todas las variables explicativas  $X_1, \dots, X_p$  que pueden estar altamente correlacionadas, lo que indica que presentan multicolinealidad, y ser mayor en número que las observaciones, esto es  $p > n$ . Tanto las  $X_j$  como  $Y$  pertenecen al espacio  $R^n$  y a su vez, son centradas y escaladas.

Sea  $T$  la matriz formada por estas componentes, se tiene el modelo PLS dado por:

$$Y = Tc^T + \epsilon$$

Donde  $\epsilon$  es el vector de residuos y  $c^T$  es el vector de coeficientes de los componentes. Al postular  $T = XW^*$ , donde  $W^*$  es la matriz de los coeficientes de las variables  $x_j$  en cada componente  $t_h$ , tenemos la expresión directa para la variable de respuesta  $Y$  usando los predictores  $x_j$ :

$$Y = XW^*c^T + \epsilon$$

Expandiendo la ecuación anterior se tiene que:

$$Y_i = \sum_{h=1}^H (c_h w_{1h}^* x_{i1} + \cdots + c_h w_{ph}^* x_{ip}) + \epsilon_i$$

Donde  $H$  es el número de componentes retenidos en el modelo final con  $H \leq \text{rango}(X)$ . Además,  $p$  es el número de variables de la matriz  $\mathbf{X}$ . Los coeficientes  $c_h w_{jh}^*$  muestran la relación entre el vector  $\mathbf{Y}$  y las variables  $x_j$  a través de las componentes  $t_h$ .

### Normalización de los datos

La normalización de los datos se puede desarrollar de dos formas que a continuación se relacionan:

- Forma 1 La primera forma de normalización consiste en restar para cada una de las variables su media y dividir por la raíz cuadrada de la suma de los cuadrados de las desviaciones de sus media:

$$1. y_i^{[1]} = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \text{ para } i = 1, 2, \dots, n.$$

$$2. x_{ij}^{[1]} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}} \text{ para } i = 1, 2, \dots, n \text{ } j = 1, 2, \dots, p.$$

- Forma 2 La segunda forma consiste en restar para cada una de las variables su media y dividir por la raíz cuadrada de la suma de los cuadrados de las desviaciones de su media dividido por  $n - 1$ :

$$1. y_i^{[2]} = \frac{y_i - \bar{y}}{\sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}} \text{ para } i = 1, 2, \dots, n.$$

$$2. x_{ij}^{[2]} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n - 1}}} \text{ para } i = 1, 2, \dots, n \text{ } j = 1, 2, \dots, p.$$

Las operaciones intermedias que hay que realizar para obtener los coeficientes de regresión difieren del tipo de normalización de los datos. Sin embargo, para los coeficientes de regresión que son asociados a las variables  $x_1^{[1]}, x_2^{[1]}, \dots, x_p^{[1]}$  como resultado de la primera normalización y las variables  $x_1^{[2]}, x_2^{[2]}, \dots, x_p^{[2]}$  como resultado de la segunda normalización son los mismos, lo que quiere decir que  $\beta^{[1]} = \beta^{[2]}$  (Gaviria, 2016).

### Algoritmo PLS1

Este algoritmo es tomado y adaptado de Garthwaite (1994); Vega-Vilca y Guzmán (2011). La entrada de datos corresponde a la matriz de predictoras  $\mathbf{X}_{(n \times p)}$  y el vector respuesta  $Y_{(n \times 1)}$ , los cuales han sido centrados o estandarizados por columnas.

1. Entrada  $X_0, Y_0$
2. Para  $h = 1, \dots, H$ 
  - $a = COV(X, Y)$ , luego,  $w = \frac{a}{||a||}$ , [Normar  $w$  a 1]
  - $T_h = Xw$ , [Componente  $t_h$  de  $X$ ]
  - $p = \frac{T_h' X}{T_h' T_h}$ , [Coeficientes de regresión  $x_j$  sobre  $t_h$ ]
  - $c = \frac{T_h' Y}{T_h' T_h}$ , [Coeficientes de  $t$  sobre  $y$ ]
  - $X_h = X_{h-1} - T_h P$ , [Residuo  $X$  y ortogonalidad]
  - $Y_h = Y_{h-1} - T_h C$ , [Residuos en  $y$ ]
3. Próximo  $h$

Donde,

- $a$ : es un vector de covarianzas entre  $Y$  e  $X$ .

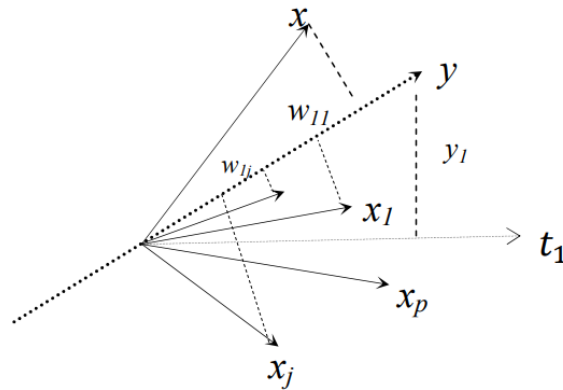
- $w$ : es un vector propio que se utiliza para construir las componentes PLS  $t_h$ .
- $T$ : son las componentes PLS (scores).
- $p$ : coeficientes que relacionan las componentes PLS con las variables explicativas  $X$ . (loadings)
- $c$ : coeficientes que relacionan las componentes PLS con la variable respuesta  $Y$ . (loadings)
- $h = 1, \dots, H$ : son los ejes factoriales

La matriz  $X_h$  es el residuo de la regresión de  $X_{h-1}$  sobre  $t_h$ . La variable  $y_h$  es el vector de residuos en la regresión de  $y_{h-1}$  sobre  $t_h$ .

A partir del algoritmo, se puede escribir la fórmula de regresión de  $Y$  sobre las componentes  $t_1, \dots, t_H$  en función de las variables  $X$ , entonces:

$$\begin{aligned}\hat{Y} &= T_h C_h = T_1 C_1 + \dots + T_H C_H \\ &= X W_h^* C_h \\ &= X W_h (p_h' W_h)^{-1} C_h = X \hat{\beta}_h\end{aligned}$$

Donde  $\hat{\beta}_h = W_h (p_h' W_h)^{-1} C_h$  es el vector estimado de coeficientes de regresión PLS de  $Y$  sobre  $X$  utilizando  $h$  componentes.



**Figura 4-10.:** Obtención de las coordenadas de  $w_1$  por proyección de  $x_j$  sobre  $y$  [González \(2014\)](#)

El número de componentes  $t_1, \dots, t_H$  a retener en la regresión es determinado por validación cruzada. Para cada valor  $h$  se calcula las predicciones  $\hat{y}_{hi}$  e  $\hat{y}_{h(-i)}$  de  $y_i$  con la ayuda del modelo con  $h$  componentes, calculadas utilizando todas las observaciones, y luego sin utilizar

la observación  $i$  (González, 2014).

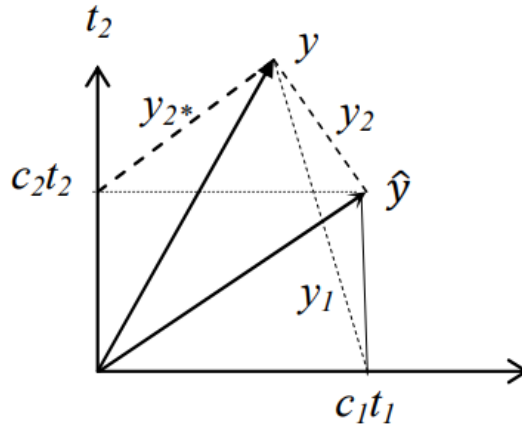
Se calcula enseguida los criterios  $RSS_h$  (Residual Sum of Squares) y  $PRESS_h$  (Prediction Error Sum of Squares) definidos por:

$$RSS_h = \sum (y_i - \hat{y}_{hi})^2, \quad PRESS_h = \sum (y_i - \hat{y}_{h(-i)})^2$$

Así, la componente  $t_h$  es retenida si:

$$\sqrt{PRESS_h} \leq 0.95 \sqrt{RSS_{h-1}}$$

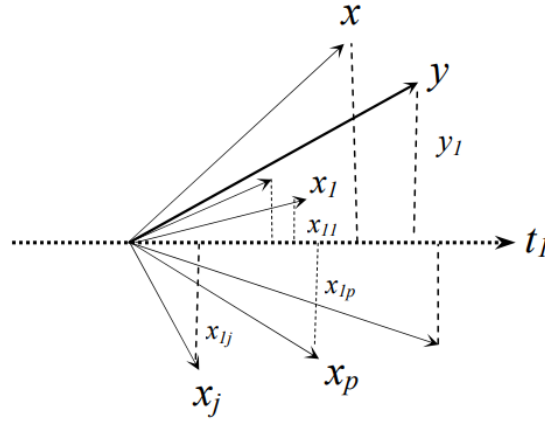
González (2014) indica que en lugar de eliminar las observaciones una por una, se puede también eliminar bloque por bloque.



**Figura 4-11.:** Regresión de  $y$  sobre las componentes ortogonales  $t_1, t_2$  González (2014)

De estas regresiones se puede deducir las propiedades de ortogonalidad e igualmente se describen un conjunto de propiedades cuando no hay datos faltantes, las cuales también aplican para PLS2:

- Las componentes obtenidas son ortogonales a los residuos de cualquiera de las variables explicativas  $T_h'X = 0$
- Las componentes obtenidas son ortogonales a los residuos de cualquiera de las variables  $Y$   $T_h'Y = 0$
- Las componentes PLS, obtenidas a partir de las variables explicativas y sus valores residuales, son ortogonales entre sí, es decir  $T_h'T_l = 0, l > h$
- Las componentes obtenidas en una etapa cualquiera son ortogonales a los residuos de cualquiera de las variables explicativas de la misma etapa o de etapas superiores  $T_h'X_l = 0, l \geq h$

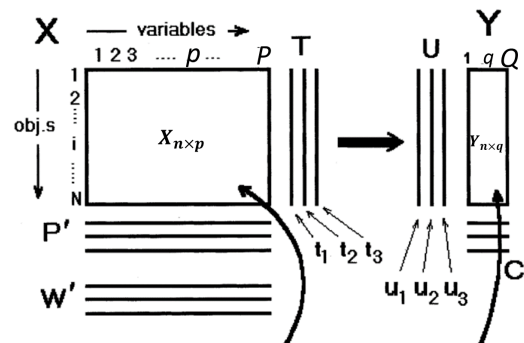


**Figura 4-12.:** Los residuos en  $x$  y en  $y$  son ortogonales a  $t_1$  [González \(2014\)](#)

### 4.2.3. PLS2

El algoritmo PLS2 busca explicar un conjunto de variables  $y_1, y_2, \dots, y_q$  mediante un conjunto de variables explicativas  $x_1, x_2, \dots, x_p$ , a través de un modelo predictivo basado en estructuras latentes de  $\mathbf{X}$  y de  $\mathbf{Y}$ . Este algoritmo aplica el método de PCA a partir de un conjunto de variables  $x_1, x_2, \dots, x_p$  bajo la condición de que estas componentes expliquen a su vez el conjunto de variables  $y_1, y_2, \dots, y_q$  ([Gaviria, 2016](#)).

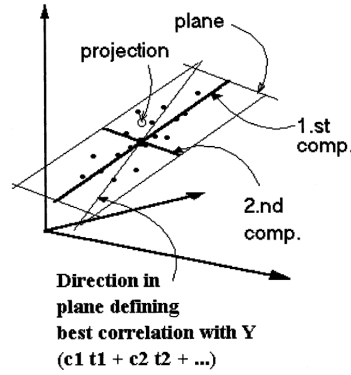
El desarrollo del algoritmo consiste en la construcción de las matrices  $\mathbf{X}$  y  $\mathbf{Y}$ , entendiendo que la matriz  $X$  contiene los vectores  $x_1, x_2, \dots, x_p$  y la matriz  $Y$  los vectores  $y_1, y_2, \dots, y_q$ . Cada vector perteneciente a cada matriz, pertenece al espacio vectorial  $R^n$  y la matriz  $\mathbf{X}$  tiene espacio vectorial  $R_{n \times p}$  y la matriz  $\mathbf{Y}$  espacio vectorial  $R_{n \times q}$ . A partir de este algoritmo se realizan proyecciones en ambos espacios sobre hiperplanos de dimensiones reducidas, las coordenadas de los puntos en esos hiperplanos forman los elementos de las matrices  $T$  y  $U$  ([Gaviria, 2016](#)).



**Figura 4-13.:** Estructura de las matrices  $\mathbf{X}$  y  $\mathbf{Y}$  para PLS2 adaptado de [Wold et al. \(2001\)](#)

El algoritmo PLS2 pretende lograr dos objetivos básicos:

1. Maximizar la correlación entre los conjuntos de variables  $x_1, x_2, \dots, x_p$  y  $y_1, y_2, \dots, y_q$
2. Aproximar a través del hiperplano lo mejor que se pueda a los espacios vectoriales generados por los conjuntos de variables  $x_1, x_2, \dots, x_p$  y  $y_1, y_2, \dots, y_q$ , es decir, la información que estos poseen.



**Figura 4-14.:** Representación geométrica del modelo PLS2 [Wold et al. \(2001\)](#)

### Regresión PLS2

La metodología de regresión por mínimos cuadrados parciales PLS puede ser visto como un algoritmo que busca maximizar una función objetivo que está sujeta a ciertas restricciones. En este contexto, la función objetivo es la covarianza entre  $\mathbf{X}$  y  $\mathbf{Y}$  y la restricción usualmente es la ortogonalidad entre los scores.

Por tanto, según [González \(2014\)](#) se define PLS2 como:

Sea  $\mathbf{Y}$  la matriz de variables dependientes  $Y_1, Y_2, \dots, Y_r$  y  $\mathbf{X}$  la matriz de variables independientes  $X_1, X_2, \dots, X_p$  de rango  $H$  medidas sobre  $n$  individuos. Todas las variables estandarizadas. Puede existir multicolinealidad al interior de cada bloque, presencia de datos faltantes y más variables que  $n$ .

Por tanto,  $\mathbf{Y}$  y  $\mathbf{X}$  son dos conjuntos de variables para los cuales se asume que existe una relación subyacente entre los dos bloques explicada por  $H$  componentes latentes ortogonales  $t_h (h = 1, 2, \dots, H)$  calculadas como una combinación lineal de las variables del conjunto predictor  $\mathbf{X}$ , y guardando alta relación con  $\mathbf{Y}$  a través de su combinación lineal  $u_h = \mathbf{Y}c_h$ . Entonces, las matrices predictor y respuesta son descompuestas de la siguiente forma:

$$\begin{aligned} X &= T_H P'_H + X_H \\ Y &= T_H C'_H + Y_H \end{aligned}$$

Donde  $P_H$  y  $C_H$  son las matrices de cargas conteniendo los parámetros del modelo, y  $X_H$  y  $Y_H$  las matrices de residuos representando la parte de la variabilidad de los datos no explicada por los modelos de parámetros (González, 2014).

### Transformación, escalado y centramiento de las matrices $\mathbf{X}$ y $\mathbf{Y}$

Antes de realizar el análisis por PLS o en general una regresión, las variables de  $\mathbf{X}$  e  $\mathbf{Y}$  son transformadas con el objetivo de que las distribuciones alcancen un comportamiento simétrico. Por ejemplo, las variables que tengan un rango de más de una magnitud de 10 suelen ser transformadas logarítmicamente.

Igualmente, Wold et al. (2001) indica que los resultados de métodos de proyección como PLSR dependen de la escala de los datos. Con una escala adecuada, se puede enfocar el modelo en variables  $\mathbf{Y}$  más importantes y utilizar la experiencia para aumentar los pesos de las variables de la matriz  $\mathbf{X}$  y que estas sean más informativas. En ausencia de conocimiento sobre la importancia relativa de las variables, el enfoque multivariado estándar consiste en: **i)** escalar cada variable a varianza unitaria dividiéndolas por sus desviaciones estándar, y **ii)** centrarlas restando sus promedios, conocido como auto-escalado. Esto corresponde a darle a cada columna de variable el mismo peso, la misma importancia en el análisis. Sin embargo, a menudo el auto-escalado no es la mejor forma de escalar las variables de  $\mathbf{X}$ .

Para facilitar la interpretación y para estabilidad numérica, se recomienda centrar los datos antes del análisis. Esto se hace, ya sea antes o después del escalado, restando los promedios de todas las variables tanto en  $\mathbf{X}$  como en  $\mathbf{Y}$ . El centrado no altera la interpretación de los resultados.

### Algoritmo PLS2

El siguiente algoritmo es adaptado de González (2014). La entrada de datos corresponde a la matriz de predictoras  $X_{(n \times p)}$  y a la matriz de respuestas  $Y_{(n \times r)}$ .

Entrada  $X_0, Y_0$ .

Para  $h = 1, \dots, H$

1. Inicializo  $u_h$  ( $u_1$ : primera columna de  $Y_{h-1}, \dots$ )
2. Repetir hasta convergencia de  $w_h$ 
  - $w_h = X'_{h-1}u_h/||u_h||$  ;  $w_h/||w_h||$  [Coeficientes de regresión simple  $x_j$  sobre  $u_h$ ]
  - $t_h = X_{h-1}w_h/(w'_h w_h)$  [*Componente<sub>h</sub>* de  $\mathbf{X}$ ]
  - $c_h = (Y'_{h-1}t_h)/(t'_h t_h)$  [Vector de  $R^r$ ]
  - $u_h = Y_{h-1}c_h/(c'_h c_h)$  [*Componente<sub>h</sub>* de  $\mathbf{Y}$ ]

3.  $p_h = \frac{X'_{h-1}t_h}{t'_h t_h}$ , [Vector P de coeficientes sobre  $\mathbf{X}$  en  $R^p$ ]
4.  $X_h = X_{h-1} - t_h p'_h$ , [Residuo  $\mathbf{X}$  y ortogonalidad]
5.  $Y_h = Y_{h-1} - t_h c'_h$ , [Residuos en  $\mathbf{Y}$ ]
6. Próximo h

Donde,

- w: es un vector propio que se utiliza para construir las componentes PLS  $t_h$ .
- T: son las componentes PLS (scores).
- p: coeficientes que relacionan las componentes PLS con las variables explicativas  $\mathbf{X}$  (loadings).
- c: coeficientes que relacionan las componentes PLS con la variable respuesta  $\mathbf{Y}$  (loadings).
- $h = 1, \dots, H$ : son los ejes factoriales

A partir del algoritmo, se puede escribir la fórmula de regresión de  $\mathbf{Y}$  sobre las componentes  $t_1, \dots, t_H$  en función de las variables contenidas en la tabla  $\mathbf{X}$ . Como:

$$Y = T_h C_h = T_1 C'_1 + \dots + T_H C'_H + Y_H$$

Donde  $c = [c_1, c_2, \dots, c_H]$  Entonces la ecuación de regresión se puede expresar matricialmente como:

$$\begin{aligned} Y &= T_H C'_H + Y_H \\ &= X W_H^* c'_H + Y_H \\ &= X W_H (p'_H W_H)^{-1} C_H = X \hat{\beta}_H + Y_H \end{aligned}$$

Donde  $\hat{\beta}_H = W_H (p'_H W_H)^{-1} C'_H$  es el vector estimado de coeficientes de regresión PLS de  $\mathbf{Y}$  sobre  $\mathbf{X}$  utilizando H componentes. El número de componentes  $t_1, \dots, t_H$  a retener en la regresión es determinado por validación cruzada.

Un indicador importante en el modelo PLS2 es el VIP el cuál, indica el poder explicativo de cada variable  $x_j$  en las variables respuestas  $y_j$  sobre el espacio de proyección de la nube de puntos.

La medida VIP viene definida por:

$$VIP_{H_j} = \sqrt{\frac{p}{Rd(Y : t_1, \dots, t_h) \sum_{h=1}^H Rd(Y : t_h)}} w_{hj}^2$$

Donde  $\sum_{j=1}^p VIP_{H_j}^2 = p \forall h$ .  $H$  representa el número de componentes que se retienen en el modelo.

La contribución de una variable  $x_j$  a la construcción de la componente  $t_h$  se mide por el coeficiente  $w_{hj}^2$ . Para cada  $h$  la suma de esos valores al cuadrado sobre el conjunto de las  $p$  variables explicativas  $x_j$  asciende a 1, lo que quiere decir que  $\sum_{j=1}^p w_{hj}^2 = 1$ .

Para medir la contribución de la variable  $x_j$  a la construcción  $\mathbf{Y}$  a partir de las componentes  $t_h$  es necesario tener en cuenta el poder explicativo de dicha componente, el cual se mide a través de la redundancia  $Rd(Y; t_h)$ . Donde  $Rd(Y : t_h)$  se entiende como:

$$Rd(Y; t_h) = \frac{1}{r} \sum_{k=1}^r corr^2(y_r, t_h) = \frac{1}{r} \sum_{k=1}^r \lambda_h r_{hk}^2$$

No es otra cosa que una medida de las correlaciones al cuadrado entre la componente y las variables que forman el conjunto  $\mathbf{Y}$  (variable a explicar). Los valores (VIP) permiten clasificar las variables  $x_j$  en función del poder explicativo que tiene sobre  $\mathbf{Y}$  y sobre el espacio de proyección de la nube de puntos. Las variables con un valor grande  $VIP(> 1)$  son las más importantes en la construcción de  $\mathbf{Y}$  (Márquez Ruiz, 2017).

Si una variable explicativa posee un coeficiente de regresión relativamente pequeño en valor absoluto y el VIP es también pequeño (en este caso Wold considera como pequeño cualquier valor VIP inferior a 0.8), entonces se convierte en una candidata a ser eliminada del análisis, aunque en ningún caso éste sea un objetivo perseguido por la regresión PLS (Márquez Ruiz, 2017).

Antes de eliminar una variable, se pueden representar gráficamente las variables en el plano  $(w_i, w_h)$  para  $i, h$  componentes retenidas (sobre todo las principales 1, 2), si las variables candidatas a eliminar se encuentran cerca del origen del citado plano entonces se tendría argumento para eliminarlas. Por tanto, se pueden desarrollar los siguientes argumentos según (Márquez Ruiz, 2017):

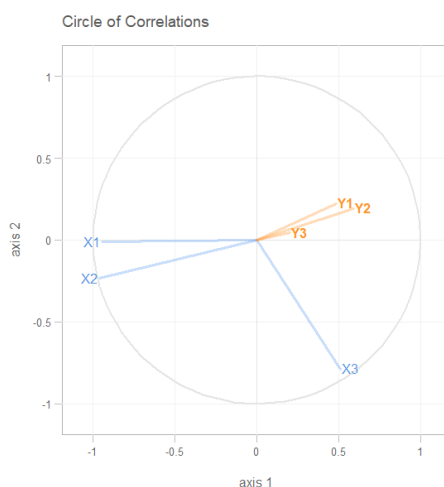
- $VIP_{H_j} < 0.8$ .
- $\beta_j$  pequeño de  $\mathbf{Y}$ .
- Vector  $w$  próximo al origen.

Igualmente, Márquez Ruiz (2017) indica que estas condiciones podrían servir para eliminar variables y comenzar nuevamente el algoritmo con las variables restantes. No obstante, como

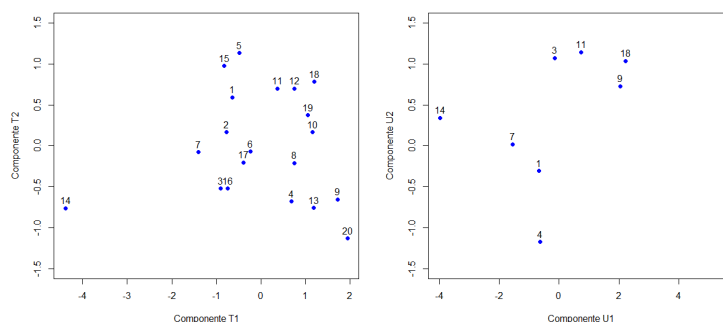
el objetivo principal que persigue la regresión PLS es la predicción de las variables y no la descripción de las relaciones, muchos autores aconsejan no eliminar variables a no ser que se den todas las circunstancias anteriormente descritas y que los valores previstos no varíen excesivamente tras su eliminación.

### Representación gráfica e interpretación de resultados

La representación gráfica e interpretación de los resultados de la técnica PLS2 se presentan a continuación.



**Figura 4-15.:** Representación de la correlación entre las variables de X y Y



**Figura 4-16.:** Representación de las componentes T y U

La Figura 4-15, representa la correlación entre las variables de la matriz X con la matriz Y, mientras que la Figura 4-16 ilustra la distribución de los individuos de la matriz X en los dos primeros componentes T y la distribución de los individuos de la matriz Y en los dos primeros componentes U,

### 4.3. Análisis de estructura de datos en tres vías

La estructura de las tablas corresponde a un cubo de datos que contemplan tres modos identificados como variables, individuos y una tercera condición que puede ser el tiempo. La representación de las dos estructuras de datos se observa en la Figura 4-17.

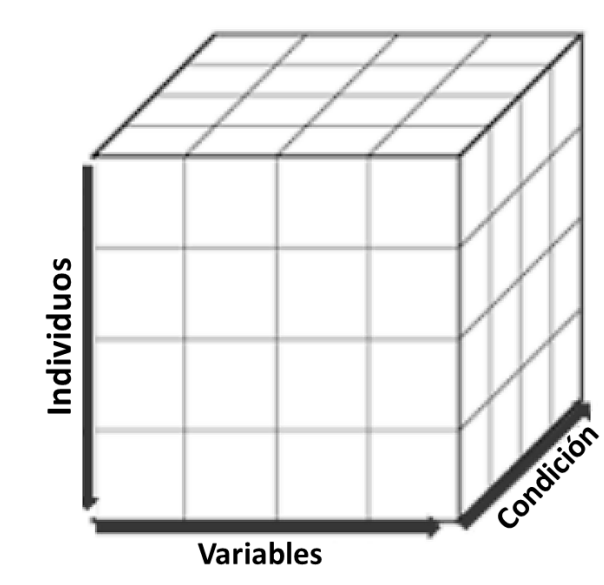


Figura 4-17.: Estructura de datos de tres vías

La estructura representa un cubo de la forma  $X_{I \times J \times K}$ , donde la primera vía  $I$  representa a los individuos, la segunda vía  $J$  representa las variables y la tercer vía  $K$  representa una condición o tiempo.

#### 4.3.1. Multiway - ACP (MPCA)

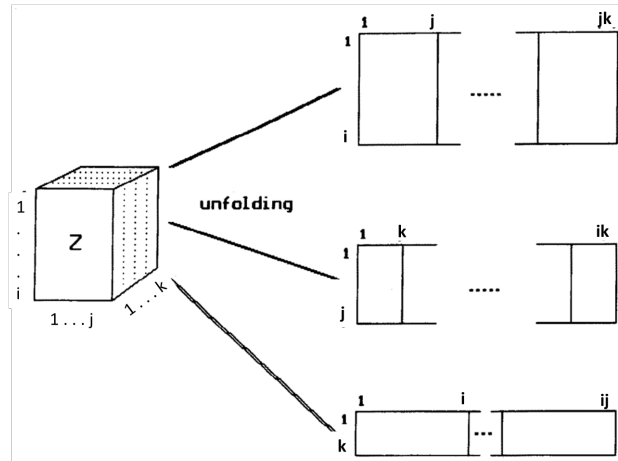
A parte de las ventajas propias de la técnica ACP, siendo un algoritmo simple, fácil y efectivo para analizar un arreglo bidimensional, este también se puede extender para cuando se presentan tablas o arreglos de más de dos vías, como es el caso en campos como la química, análisis de imágenes, entre otros, donde la estructura de los datos adopta una forma de tres vías, que significa una matriz de la forma  $X_{IJK}$  (Nomikos y MacGregor, 1994).

El análisis de componentes principales multidireccionales (Multiway ACP) es una extensión del análisis de componentes principales para datos de tres vías. Se lleva a cabo desplegando inicialmente la matriz de datos tridimensionales en una matriz bidimensional y luego aplicando el análisis de componentes principales a la matriz bidimensional resultante. Este enfoque es particularmente útil para procesar datos de proceso por lotes, ya que se recopila un segmento de datos para un número de lotes o individuos  $i$  ( $i = 1, 2, \dots, I$ ), con  $j$

( $j = 1, 2, \dots, J$ ) variables de medición en intervalos de tiempo  $k$  ( $k = 1, 2, \dots, K$ ), por lo que la información sobre el proceso se puede organizar en una estructura de datos de tres vías,  $\mathbf{X}$  (Swiercz y Mroczkowska, 2020).

Ante esta estructura, una técnica que puede manejar este tipo de datos es MPCA, considerado un proceso algorítmicamente consistente al ACP, cumpliendo los mismos objetivos de interpretación y representación de una estructura de tres vías.

Según Nomikos y MacGregor (1994), la relación que existe entre MPCA y ACP es que el MPCA se puede aplicar de forma equivalente al ACP cuando se forma una matriz bidimensional al desplegar la matriz de tres vías de varias maneras posibles. La Figura 4-18 muestra diferentes despliegues horizontales, realizando una combinación de los índices o vías de la estructura dejando fijo una de las vías en las filas y variando las otras vías en los diferentes despliegues. Por tanto, en el primer caso, permanece en las filas la vía  $I$  y desplegando de forma horizontal las vías  $K$  y  $J$  obteniendo una matriz  $I \times JK$ , la segunda opción presenta las filas de  $J$  desplegando de forma horizontal las vías  $I$  y  $K$  y obtener una matriz bidimensional  $J \times IK$  y como tercera opción permanece en las filas la vía  $K$  y se realiza el despliegue de las vías  $I$  y  $J$  resultando una matriz de tipo  $(K \times IJ)$ .



**Figura 4-18.:** Formas de despliegue de una matriz de datos de tres vías, (Henrion, 1994)

Las representaciones en la Figura 4-18 muestran udespliegues observados desde diferentes ángulos y perspectivas de cada vía. Sin embargo, al transponer estas matrices, se convierten en despliegues verticales, lo que significa que aumentan las filas y se mantienen las columnas en el análisis. Por tanto, estas matrices pueden ser vistas a su vez en arreglos de la siguiente forma  $X(JK \times I)$ ,  $X(IK \times J)$ ,  $X(IJ \times K)$ .

MPCA descompone  $\mathbf{X}$  en una suma del producto de  $t$  vectores de puntuación ( $t$ ) y  $p$  matrices

de carga ( $\mathbf{P}$ ), más una matriz residual ( $\mathbf{E}$ ) que se minimiza en un sentido de mínimos cuadrados.

$$\mathbf{X}_{I \times J \times K} = \sum_{r=1}^R \mathbf{t}_r \otimes \mathbf{P}_r + \mathbf{E}_{I \times JK} \quad (4-1)$$

donde  $R$  es el número de componentes principales usado en MPCA.

Un vector de puntuación  $t_r$  representa la relación entre  $I$  lotes (cada elemento del vector de puntuación expresa la proyección del  $i$ -ésimo lote sobre el espacio reducido), mientras que la matriz de carga  $\mathbf{P}_r$  está relacionado con las variables ( $j$ ) y sus variaciones de tiempo ( $k$ ). Las matrices de carga almacenan toda la información sobre cómo las lecturas del sensor se desvían de sus valores medios en cada muestra (Swiercz y Mroczkowska, 2020).

Según Hoo et al. (1994), esta descomposición condensa y reduce los datos en espacios de puntuaciones de baja dimensionalidad, abarcando tanto las variables como el tiempo. Estos espacios capturan la variabilidad principal en los lotes a lo largo de todo el periodo de tiempo. Cada matriz de carga  $\mathbf{P}$  resume las principales variaciones temporales de las variables en relación con sus trayectorias promedio en todos los lotes. Al hacerlo, MPCA considera no solo la magnitud de la desviación de cada variable de su trayectoria promedio, sino también las correlaciones entre ellas. Igualmente, el número adecuado de componentes principales puede determinarse mediante validación cruzada (Hoo et al., 1994).

Asimismo, la varianza explicada se calcula comparando los datos reales del proceso con los datos reconstruidos generados por el modelo MPCA. Los datos reconstruidos se generan a partir de las puntuaciones para un lote particular y las cargas correspondientes.

$$\hat{\mathbf{X}} = \sum_{r=1}^R \mathbf{t}_r \mathbf{P} \quad (4-2)$$

La suma de la varianza explicada para cada variable puede usarse para identificar variables cuyas variaciones no están relacionadas con otras variables, como en el caso de variables que están bien controladas (Hoo et al., 1994).

## Escalado

Según Henrion (1994), el escalado adecuado de los datos es importante para el MPCA. Las posibilidades de escalado pueden ser no escalado, escalado, escalado global, escalado

de modos individuales y escalado de modos de combinación, donde el escalado se refiere al centrado y/o estandarización adicional de datos.

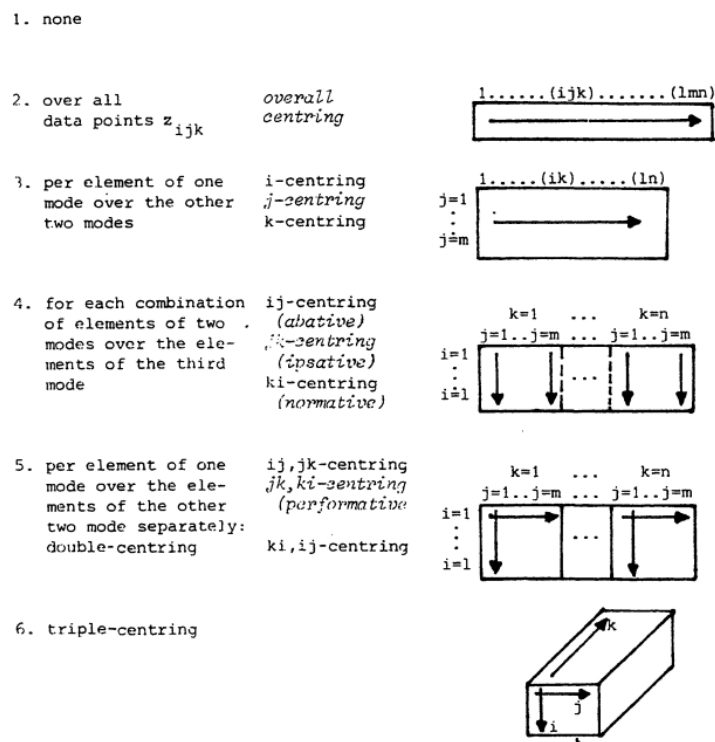


Figura 4-19.: Diferentes tipos de escalado de datos, (P. Kroonenberg, 1983)

En la Figura 4-19 se presentan los diferentes tipos de escalado de datos para MPCA, donde las flechas indican las columnas de la matriz desplegada que tienen media cero y desviación estándar unitaria.

Henrion (1994) indica que es importante tener cuidado al elegir apropiadamente el tipo de escalado, ya que este influye considerablemente en las soluciones obtenidas y sus interpretaciones posteriores.

Una forma bastante natural de pretratamiento de datos es realizar el escalado de variables sobre toda la matriz de datos, es decir, a lo largo del modo de combinación de objetos y condiciones. También se habla de “ $j$  escalado” haciendo referencia al índice  $j$  habitual para el modo de variables. Después de este tipo de escalado, las diferencias de medias y dispersiones entre condiciones y entre objetos se conservan en la estructura de datos. Por lo tanto, el escalado  $j$  se prefiere si se van a investigar las diferencias entre objetos y condiciones.

En el caso de tener interés principalmente en las diferencias entre objetos mientras que las diferencias absolutas entre condiciones deben ser eliminadas, se debe realizar la estandariza-

ción solo a lo largo de los objetos. Esto se llama escalado *jk* con referencia a los índices de los modos sujetos a escalado. De manera simétrica, también se pueden eliminar las diferencias absolutas entre objetos y en su lugar centrarse en las diferencias entre condiciones (escalado *ji*).

Estas tres formas típicas de escalado de datos se ilustran en la Figura 4-20, donde las flechas indican las columnas de la matriz desplegada que tienen media cero y desviación estándar unitaria.

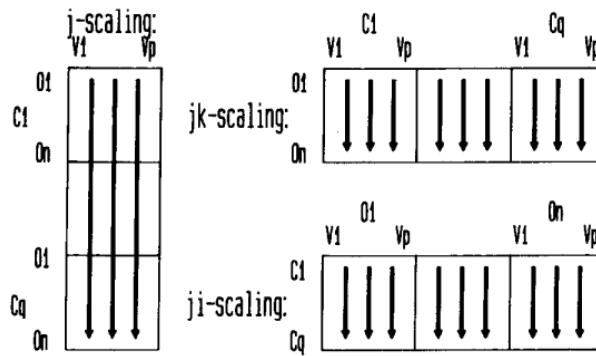


Figura 4-20.: Diferentes tipos de escalado de datos en MPCA según [Henrion \(1994\)](#)

### 4.3.2. PARAFAC

La técnica PARAFAC (Análisis Factorial Paralelo) fue propuesta por Harshman (1970) y por Carroll y Chang (1970). En esta técnica se realiza una descomposición de los datos de tres vías. PARAFAC puede considerarse como una extensión del Análisis de Componentes Principales. Sin embargo, la diferencia entre PCA y PARAFAC es que PARAFAC no exige el cumplimiento de ortogonalidad de las componentes ([Arce Antón et al., 2021](#)).

#### Estructura del Modelo

El modelo estructural detrás del Análisis de Componentes Principales (ACP) para estructuras de datos de dos vías es un modelo bilineal de la siguiente forma:

$$x_{ij} = \sum_{r=1}^R a_{ir} b_{jr} + e_{ij}; \quad i = 1, \dots, I, \quad j = 1, \dots, J$$

Mientras que para el modelo PARAFAC se aplica a un cubo de datos de la forma  $\mathbf{X}(I \times J \times K)$  compuesto por  $x_{ijk}$ , representado por la siguiente ecuación:

$$x_{ijk} = \sum_{r=1}^R a_{ir} b_{jr} c_{kr} + e_{ijk}; \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad k = 1, \dots, K \quad (4-3)$$

Donde  $R$  es el número de componentes usado en el modelo PARAFAC y  $e_{ijk}$  es el termino residual que contiene toda la variación no explicada.

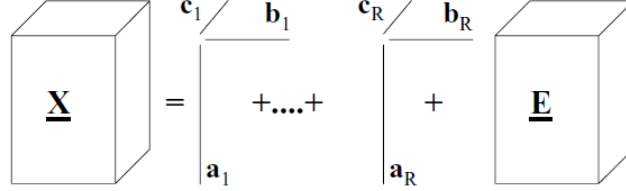
Usando la multiplicación de Kronecker, la estructura del modelo se puede escribir de la siguiente forma:

$$X_{I \times JK} = \sum_{r=1}^R \mathbf{a}_r (\mathbf{c}_r^T \otimes \mathbf{b}_r^T) + \mathbf{E}_{I \times JK} \quad (4-4)$$

Donde  $X_{I \times JK}$  es el modelo de tres vías desplegado a una matriz  $I \times JK$ , lo cual es bastante cercano a la forma en que se especifica el modelo de ACP:

$$X_{I \times J} = \sum_{r=1}^R \mathbf{a}_r \mathbf{b}_r^T + \mathbf{E}_{I \times J}$$

Donde  $\mathbf{a}_r$ ,  $\mathbf{b}_r$  y  $\mathbf{c}_r$  son las  $r$ -ésimas columnas de las matrices de carga  $\mathbf{A}$ ,  $\mathbf{B}$  y  $\mathbf{C}$ .



**Figura 4-21.:** Representación del modelo PARAFAC con  $R$  componentes., ([Leardi, 2005](#))

De esta forma, el modelo PARAFAC de un cubo de tres vías está formado por tres matrices de carga,  $\mathbf{A}$ ,  $\mathbf{B}$  y  $\mathbf{C}$ , cuyas columnas se corresponden con los vectores de carga anteriormente mencionados y las dimensiones de estas matrices son  $(I \times R)$ ,  $(J \times R)$  y  $(K \times R)$  respectivamente. Los vectores de carga se determinan por mínimos cuadrados. Con estas matrices, el modelo PARAFAC puede representarse a partir de la siguiente ecuación:

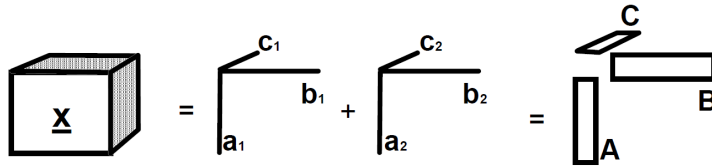
$$\mathbf{X}_K = \mathbf{A} \mathbf{D}_K \mathbf{B}^T + \mathbf{E}_K \quad K = 1, \dots, K \quad (4-5)$$

Donde  $\mathbf{D}_k$  es la matriz diagonal con la forma  $(R \times R)$  donde sus elementos no nulos son la  $k$ -ésima fila de la matriz  $\mathbf{C}$  en su diagonal y  $\mathbf{E}_k$  es la matriz de residuos  $(I \times J)$ . El superíndice T se refiere a la transposición de una matriz.

La ecuación 4-5 del modelo PARAFAC parece indicar que hay una diferencia en cómo se tratan los diferentes modos en el modelo. Esto no es el caso. El modelo PARAFAC es simétrico

en el sentido de que si el arreglo se reordena de modo que, por ejemplo, el primer modo se convierta en el tercero y viceversa, entonces los modelos PARAFAC de los dos conjuntos de datos serán idénticos, excepto que las cargas de los modos primero y tercero se han intercambiado (Bro-Jorgensen, 1998).

La estructura  $\mathbf{A}\mathbf{D}_K\mathbf{B}^T$  arroja algo de luz sobre cómo se pueden interpretar las cargas del tercer modo, específicamente sobre cómo se relaciona el modelo PARAFAC con el principio de perfiles paralelos proporcionales. Sea  $\mathbf{S}_k = \mathbf{A}\mathbf{D}_K$ , entonces  $\mathbf{X}_k = \mathbf{S}_k\mathbf{B}^T + \mathbf{E}_k$ , es decir, para cada variable  $k$  se obtiene un modelo bilineal de  $\mathbf{X}_k$  y para diferentes  $k$ s los modelos serán idénticos excepto que los componentes individuales están ponderados de manera diferente a través de la matriz  $\mathbf{D}_K$ . En la Figura 4-22 se ilustra gráficamente un modelo de dos componentes de un arreglo de tres vías  $\mathbf{X}$ .



**Figura 4-22.:** Modelo PARAFAC de dos componentes para una estructura de datos de tres vías  $\mathbf{X}$ , (Bro-Jorgensen, 1998)

Igualmente, el modelo de la ecuación 4-5, se puede simplificar mediante la introducción del producto *Khatri-Rao*. A partir de este, el modelo PARAFAC puede formularse en términos de arreglos desplegados como:

$$X_{I \times JK} = \mathbf{A}(\mathbf{C} \otimes \mathbf{B})^T + \mathbf{E}_{I \times JK} \quad (4-6)$$

Donde  $X_{I \times JK}$  es el modelo de tres vías desplegado a una matriz  $I \times JK$ . De esta manera, el modelo completo puede expresarse en una notación matricial simple, muy en línea con un modelo bilineal como

$$X = \mathbf{A}\mathbf{B}^T + \mathbf{E} \quad (4-7)$$

El producto de Khatri-Rao también facilita la comparación del modelo PARAFAC con la formulación matricial del modelo Tucker3, así como también permite formular modelos PARAFAC de orden superior utilizando notación matricial (Bro-Jorgensen, 1998).

### Determinación del número de componentes

Lo más importante en esta técnica es la elección de los factores en los cuales se realizará la descomposición de PARAFAC. El número de factores adecuado será aquel que describa el

comportamiento del tercer modo ([Arce Antón et al., 2021](#)).

Para la elección y determinación de los factores se utilizan técnicas de validación cruzada ([Gourvénec et al., 2005](#)), que consiste en la partición del cubo de datos que se denomina “split-half” ([DeSarbo, 1984](#)). Este proceso consiste en la división del cubo en dos mitades, y estas son analizadas de manera independiente. Si al realizar el “split-half” se observa estabilidad en el modelo esto indica que se está obteniendo la variación esencial del modelo.

Generalmente, el criterio utilizado es el índice CONCORDIA (Core Consistency Diagnostic), que fue desarrollado para PARAFAC ([Bro y Kiers, 2003](#); [Kamstrup-Nielsen et al., 2013](#)) y entre más cercano se encuentre a 100 % indicará que el modelo es apropiado para el cubo de datos. Se recomiendan como regla aproximada que valores en torno al 90 % deben ser interpretados como trilineales. Cuando existe una falta de trilinealidad ([Arce Antón et al., 2021](#)).

A pesar de lo anterior, [Bro \(1997\)](#) indica que es difícil decidir el mejor número de componentes para un modelo PARAFAC. Esta área aún no está muy bien establecida y definitivamente se necesita investigación. En general, la validación cruzada en el ACP bilineal, desinfla la matriz  $X$  después del cálculo de cada componente, y por lo tanto, eventualmente los componentes describen ruido en lugar de variación sistemática. Esto se observa como un aumento en los residuos del modelado de muestras independientes. Esta situación puede ser aún peor en PARAFAC. En PARAFAC, uno no deflacta el arreglo, porque se puede demostrar que el modelo trilineal calculado simultáneamente para todos los componentes se ajusta mejor al arreglo que si los componentes se calcularan sucesivamente, como es posible en PCA. Como consecuencia, extraer demasiados componentes no solo significa que el ruido se está modelando cada vez más, sino también que los factores reales están siendo modelados por más componentes (correlacionados) ([Bro, 1997](#)).

### CONCORDIA (Core Consistency Diagnostic)

Este es un diagnóstico comúnmente utilizado para determinar el número óptimo de componentes y evaluar la “adecuación” del modelo PARAFAC.

El principio del índice de CONCORDIA es evaluar el grado de superdiagonalidad o, de hecho, la similitud entre el conjunto principal implícitamente impuesto  $T$  de unos y el ajustado por mínimos cuadrados  $G$  para una serie de modelos con un número gradualmente creciente de componentes, comenzando desde el modelo de un solo componente ([Bro y Kiers, 2003](#)).

Bajo lo anterior, [Martí y Baldrich \(2015\)](#) indica que se asume que los elementos diagonales en el arreglo  $D$ , asociados a una solución PARAFAC para un número dado de componen-

tes serán mucho más grandes que el resto de las soluciones, siempre y cuando el número de componentes sea el correcto. El modelo PARAFAC puede escribirse como una versión restringida de un modelo Tucker-3,

$$X = AD(C| \otimes |B)^T$$

Una vez ajustado el modelo PARAFAC y calculadas las matrices A, B y C, la matriz en núcleo D puede calcularse mediante un procedimiento de mínimos cuadrados:

$$\operatorname{argmin}_D ||X - AD(C| \otimes |B)^T||^2$$

Se puede construir un modelo de regresión para resolver el problema de tal manera que:

$$\min_D ||\operatorname{vec}X - (C \otimes B \otimes A)\operatorname{vec}D||^2$$

Para un modelo PARAFAC válido, la matriz principal G debería idealmente tener valores cercanos a uno en la superdiagonal y valores cercanos a cero fuera de la superdiagonal. Cuando esto no ocurre, los datos no están modelados adecuadamente por un modelo PARAFAC trilineal, se han utilizado demasiados componentes para ajustar el modelo o los datos tienen valores atípicos. Una forma sencilla de evaluar si la estructura del modelo es razonable es, por lo tanto, monitorear la distribución de los elementos en la superdiagonal y fuera de la superdiagonal de D. El índice de CORCONDIA, calculado como un único parámetro, se define como:

$$\text{Core consistency} = 100. \left( \frac{1 - \sum_{d=1}^R \sum_{e=1}^R \sum_{f=1}^R (d_{def} - t_{def})^2}{\sum_{d=1}^R \sum_{e=1}^R \sum_{f=1}^R t_{def}^2} \right)$$

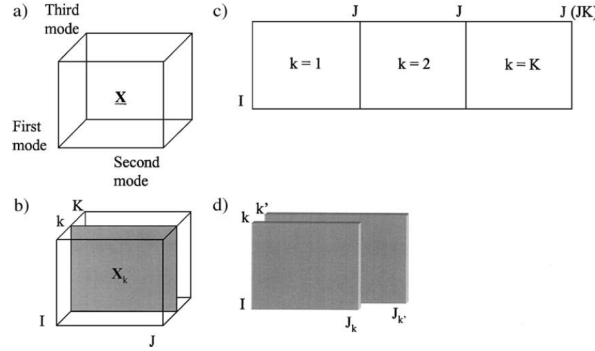
Donde la suma involucra todos los valores de la matriz principal D. Los elementos de  $t_{def}$  son los valores teóricos de la matriz central, es decir, unos cuando  $d = e = f$  y ceros en caso contrario. La consistencia del núcleo toma valores cercanos al 100 % para modelos con un bajo número de componentes y comienza a disminuir cuando se seleccionan demasiados componentes. El número óptimo de componentes corresponde al modelo más grande que aún mantiene una alta consistencia del núcleo ([Martí y Baldrich, 2015](#)).

### 4.3.3. STATIS

STATIS (Structuration des Tableaux ATrois Indices de la Statistique) es una técnica conocida para el análisis de estructuras y conjuntos de datos de tres vías en estadística ([Lavit et al., 1988](#)). Esta consiste en realizar la comparación de diferentes matrices de datos que contienen el mismo número de filas y/o columnas. Sin embargo, son obtenidas bajo diferentes condiciones experimentales. El conjunto de datos de tres vías se denota como  $\mathbf{X}$  con dimensiones  $I, J, K$ , donde  $I$  corresponde al número de filas,  $J$  al número de columnas y

$K$  al número de tablas (Stanimirova et al., 2004).

Cada dirección se llama modo y el número de niveles en el modo se llama dimensión, como se muestra en la siguiente figura.



**Figura 4-23.**: Estructura de datos de tres vías y despliegue con STATIS, (Stanimirova et al., 2004)

La Figura 4-23 indica los modos de las estructuras de datos de tres vías (a), para el análisis y aplicación de la técnica STATIS se realiza la fijación del índice del primer o segundo modo, es una alternativa muy habitual para los métodos de análisis de datos de  $N$  vías.

Para el desarrollo del método STATIS se considera inicialmente el despliegue de la matriz como se observa en la Figura 4-23(c), aunque con esta es posible desarrollarse un ACP desplegado. Sin embargo, no sería posible observar las similitudes de las tablas de datos, por tanto STATIS considera la varianza de cada tabla, a la cual se le realiza una ponderación según las similitudes entre las tablas, lo que resulta en un método con resultados más sólidos e interpretables. Por tanto, las matrices de varianzas-covarianzas  $\mathbf{W}$  se verán así:

$$\mathbf{W} = \sum_{k=1}^K a_k \mathbf{W}_k$$

Donde el vector  $a_k$  es un vector de pesos. Finalmente, la técnica STATIS se puede considerar como un PCA desplegado con una ponderación de  $a_k$  diferente a 1.

Para cada tabla de datos, se calcula la matriz de varianza-covarianza,  $\mathbf{W}_{I \times I}$ , que refleja las similitudes entre los  $I$  objetos dentro de esta tabla de datos:

$$\mathbf{W}_k = \mathbf{X}_k \mathbf{Q}_k \mathbf{X}_k^T$$

Donde  $\mathbf{X}_k$  es la matriz de dimensiones  $I \times J_k$  y  $\mathbf{X}_k^T$  es la matriz traspuesta. La matriz  $\mathbf{Q}_k$  suele ser la matriz identidad con dimensiones  $J \times J$ , por tanto, las similitudes entre dos matrices de varianza-covarianza se puede calcula de la siguiente forma:

$$\langle \mathbf{W}_k, \mathbf{W}_{k'} \rangle = \text{trace}(\mathbf{W}_k \mathbf{D} \mathbf{W}_{k'} \mathbf{D})$$

donde  $\mathbf{D}$  es una matriz de dimensión  $I \times I$ , cuyos elementos diagonales son iguales a  $1/I$ . La medida de cercanía más utilizada entre las matrices de varianza-covarianza es el llamado coeficiente RV, introducido por Robert y Escoufier [Robert y Escoufier \(1976\)](#). Para las tablas de datos  $k$ -ésimo y  $k'$ -ésimo, se define de la siguiente manera:

$$RV\langle \mathbf{W}_k, \mathbf{W}_{k'} \rangle = \frac{\langle \mathbf{W}_k, \mathbf{W}_{k'} \rangle}{\sqrt{\langle \mathbf{W}_k, \mathbf{W}_k \rangle \langle \mathbf{W}_{k'}, \mathbf{W}_{k'} \rangle}}$$

Los coeficientes RV tienen una escala entre 0 y 1 y es una matriz  $K \times K$ . Cuanto más cerca sea el RV a 1, más similares serán las dos matrices de varianza-covarianza. Las similitudes entre tablas se pueden visualizar en el espacio de componentes principales, después de realizar ACP en la matriz RV, lo que se denomina análisis de interestructura. El primer vector propio obtenido después del ACP de la matriz RV no centrada es una variable de tamaño global que representa el **acuerdo entre tablas**. Sus elementos se normalizan de tal manera que su suma es igual a 1 y se usan como pesos ( $a_k$ ) para definir lo que en STATIS se llama un **Compromiso** entre las  $K$  tablas como una suma ponderada de las matrices de varianza-covarianza.  $Wk$ , dando como resultado  $\mathbf{W} = \sum_{k=1}^K a_k \mathbf{W}_k$ .

Por tanto, en STATIS,  $\mathbf{W}$  difiere en términos del cálculo de la matriz de covarianzas del ACP desplegado, debido a la inclusión de ponderaciones proporcionales. Esta característica fortalece la robustez de esta técnica en comparación con el ACP desplegado. La matriz  $\mathbf{W}_{I \times I}$  proporciona información sobre la similitud de los objetos en el primer modo. Su distribución se puede visualizar en el espacio de los componentes principales y la representación se denomina diagrama de compromiso. Asimismo, es factible proyectar matrices de covarianza individuales. Este gráfico revela la posición de cada objeto en el diagrama de compromiso, presentándolo como un centro ponderado de  $K$  ubicaciones individuales para dicho objeto. A partir de las puntuaciones de la matriz de compromiso,  $\mathbf{W}$ , se pueden obtener las cargas de las matrices de varianza-covarianza individuales,  $\mathbf{W}_k$ , y esto es importante para visualizar los modos “ocultos”. Las coordenadas (cargas) de los objetos de la  $k$ -ésima tabla para  $f$  componentes principales en el gráfico de compromiso están determinadas por la siguiente ecuación:

$$\mathbf{C}_k = \mathbf{W}_k \mathbf{L} \mathbf{E}$$

donde  $\mathbf{E}_{f \times f}$  es una matriz diagonal, cuyos elementos diagonales son la inversa de las raíces cuadradas de los valores propios de compromiso. La matriz  $\mathbf{L}$  contiene las puntuaciones del PCA del compromiso y tiene dimensión  $I \times f$ . STATIS ofrece otra posibilidad, lo que hace que este enfoque sea bastante atractivo para el análisis exploratorio.

Finalmente, según [Stanimirova et al. \(2004\)](#), el algotirmo de STATIS se puede resumir en los siguientes pasos:

1. Calcule las matrices  $\mathbf{K}$  de varianza-covarianza como  $\mathbf{W}_k = \mathbf{X}_k \mathbf{Q}_k \mathbf{X}_k^T$ .
2. Calcular la matriz de coeficientes RV ( $K \times K$ ):  $RV\langle \mathbf{W}_k, \mathbf{W}_{k'} \rangle = \frac{\langle \mathbf{W}_k, \mathbf{W}_{k'} \rangle}{\sqrt{\langle \mathbf{W}_k, \mathbf{W}_k \rangle \langle \mathbf{W}_{k'}, \mathbf{W}_{k'} \rangle}}$
3. Realice PCA de la matriz RV  $[\mathbf{S}, \mathbf{V}, \mathbf{S}] = \text{SVD}(\text{RV})$ , SVD es la versión de descomposición en valor singular de PCA  $a_k = s_1 / \sum_{k=1}^K \mathbf{S}_{1k}$ , donde  $s_1(K \times 1)$  es la primer columna de la matriz  $\mathbf{S}$ .
4. Calcule el compromiso entre tablas como:  $\mathbf{W} = \sum_{k=1}^K a_k \mathbf{W}_k$
5. Realizar PCA de  $\mathbf{W}$   $[\mathbf{L}, \mathbf{V}, \mathbf{L}] = \text{SVD}(\mathbf{W})$
6. Crear el gráfico de puntuación de compromiso.

## 4.4. Análisis de la asociación entre estructuras de datos en tres vías

En esta subsección se centra en técnicas estadísticas utilizadas para explorar y comprender la relación entre dos estructuras de datos de tres vías. Bajo el capítulo de antecedentes identificados, se obtiene que técnicas como Unfold-PLS, N-way PLS y COSTATIS son útiles cuando se tienen múltiples conjuntos de datos tres vías y se busca entender cómo están relacionados entre sí.

### 4.4.1. Unfold-PLS

El despliegue o unfold es un concepto importante en el análisis multi-dimensional ([Bro, 1996](#)). Es finalmente, una forma de reorganizar un conjunto de datos con una estructura de tres vías o más en una matriz.

La Figura 4-18 muestra gráficamente las diversas formas de desplegar un conjunto de datos. El despliegue se logra concatenando matrices para diferentes niveles, por ejemplo, del tercer modo uno al lado del otro. En ese sentido, se observa que la dimensión de las columnas de la matriz generada se vuelve bastante grande en el modo que consiste en dos modos anteriores. Esto se debe a que las variables de los modos originales se combinan. No hay una nueva variable que se refiera a una variable original, sino más bien un conjunto de variables ([Bro, 1996](#)).

El método unfold-PLS es aplicable para datos de múltiples vías, aunque no es un método verdaderamente multilíneo porque pierde la representación multivía de los datos, y a su vez, [Camacho et al. \(2007\)](#) indica que para el caso de un despliegue horizontal de la forma  $X(I \times$

$JK$ ) la matriz de datos tiene un gran número de variables altamente correlacionadas. Por lo tanto, algunos investigadores han intentado proponer algoritmos de regresión multilínea, entre ellos N-PLS (Luo et al., 2016).

#### 4.4.2. $N$ -PLS

La regresión de mínimos cuadrados parciales (PLS, por sus siglas en inglés) es un método para construir modelos de regresión entre variables independientes (llamadas  $x$ ) y variables dependientes (llamadas  $y$ ). Para la calibración multivía, el PLS multivía o simplemente N-PLS. Según Bro (1996)  $N$ -PLS es una extensión del modelo de regresión ordinario PLS, donde los datos independientes se modelan de manera que enfatiza la variación especialmente relevante para predecir las variables dependientes (kumar Maji et al., 2014).

El análisis  $N$ -PLS es un método que estudia la relación entre estructuras de datos de tres vías o  $N$  vías ( $\mathbf{X}$ ) y cualquier otra estructura de datos  $\mathbf{Y}$ . Basicamente, es una extensión del PLS a estructuras de más de dos vías ( $N$ -way) cumpliendo el mismo objetivo maximizar la correlación entre los arreglos de  $N$  vías  $\mathbf{X}$  y  $\mathbf{Y}$  (Hervás et al., 2018).

Igualmente, kumar Maji et al. (2014) expresa que terminología general de  $N$ -PLS depende del orden de los datos: un prefijo griego indica el orden de  $x$  y un sufijo árabe después de PLS indica el orden de  $y$ . En tri-PLS1, cada muestra de calibración está caracterizada por una matriz  $J \times K$  de  $X_i$ , para construir un cubo de datos  $\mathbf{X}$  de dimensión  $I \times J \times K$ , y para cada muestra hay una medición conocida que se debe predecir mediante las variables independientes, cuyos valores se recopilan en una matriz  $\mathbf{Y}$  de dimensiones  $I \times L \times M$ . Cuando las mediciones se recopilan en un vector  $I \times 1$  llamado  $y$ . Si se buscan más de una variable se recopilan en una matriz  $\mathbf{Y}$  de dimensiones  $I \times L$ , donde  $L$  es el número de variables diferentes (tri-PLS2, tri-PLS3, etc.).

El método  $N$ -PLS encuentra espacios latentes denominados  $\mathbf{W}_j$  y  $\mathbf{W}_k$  que maximizan la correlación entre las dos estructuras de datos, realizando un despliegue de la matrix  $\mathbf{X}_{IXJK}$  a una estructura  $\mathbf{X}_{IXJK}$  Hervás et al. (2018).

#### Algoritmo $N$ -PLS

El algoritmo es tomado de Bro (1996). La entrada de datos corresponde a la estructura de datos de las variables predictoras  $X_{I \times J \times k}$  y a la estructura de datos de variables de respuesta  $Y_{I \times L \times M}$

Centrar  $\mathbf{X}$  e  $\mathbf{Y}$

Sea  $u$  igual a una columna en  $\mathbf{Y}$

$h = 1$

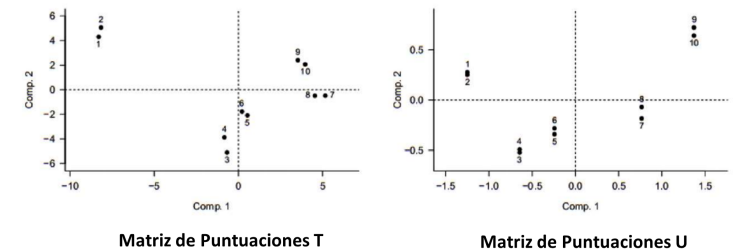
1. Calcular la matriz  $\mathbf{Z}$  (usar  $u$  en lugar de  $y$ , por tanto se determina  $\mathbf{Z}$  mediante  $u^T X_{I \times J \times k}$ )
2. Determinar  $w_J$  y  $w_K$  mediante SVD
3. Calcular  $t$
4.  $q = \frac{Y^T t}{\|Y^T t\|}$
5.  $u = Yq$
6. Si hay convergencia, continuar; de lo contrario, volver al paso 1.
7.  $b = (T^T T)^{-1} T^T u$
8.  $X_i = X - t(w_J)(w_K)^T$  y  $Y = Y - Tbq^T$
9.  $h = h + 1$ . Continuar desde el paso 1 hasta obtener una descripción adecuada de  $Y$ .

El Índice  $h$  es omitido en  $w_J$ ,  $w_K$ ,  $t$ ,  $u$ ,  $q$  y  $b$  por brevedad.

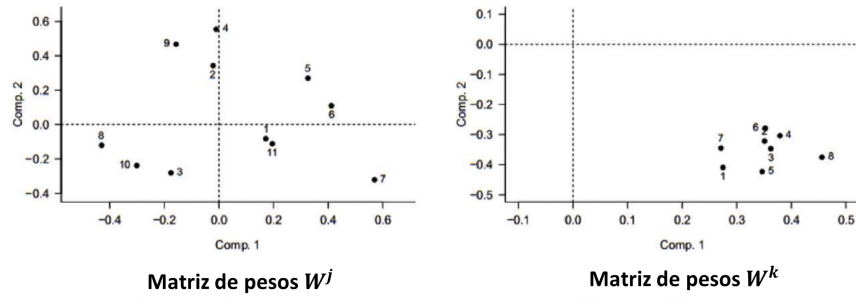
Donde  $\text{SVD}(Z)$  está dada por  $\text{SVD}(Z) = (w^j, w^k) = \max_{(w^j, w^k)} [(w^j)^T Z w^k]$ , lo que significa usar el primer conjunto de vectores normalizados de una descomposición de valores singulares (SVD) en  $Z$  (Bro, 1996). El número de componentes es determinado por validación cruzada.

### Representación gráfica e interpretación de resultados

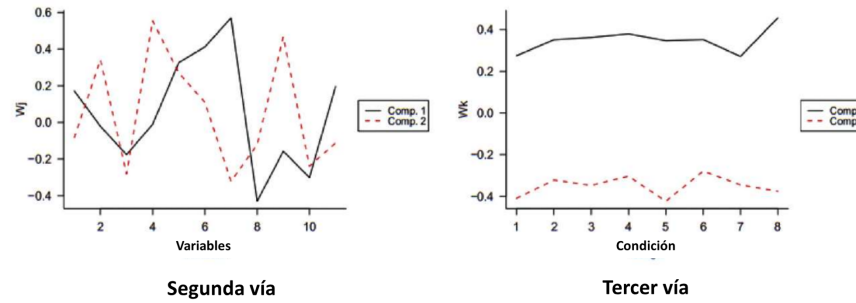
La representación gráfica e interpretación de los resultados de la técnica  $N$ -PLS es tomada a partir del trabajo de investigación desarrollado por Hervás et al. (2018).



**Figura 4-24.:** Representación de las matrices T y U., (Hervás et al., 2018)



**Figura 4-25.:** Representación de los pesos  $W^J$  y  $W^K$ ., (Hervás et al., 2018)



**Figura 4-26.:** Representación de la segunda vía y tercer vía de la estructura de datos con respecto a X)., (Hervás et al., 2018)

La Figura 4-24 ilustra la distribución de los individuos del arreglo X en los dos primeros componentes (puntuaciones matriz T) y la distribución de los individuos del arreglo Y en los dos primeros componentes (puntuaciones matriz U), la Figura 4-25, representa las matrices de los pesos  $W_J$  y  $W_K$ , lo que significa la contribución de las variables y la tercer condición del arreglo X sobre Y. Finalmente, en la Figura 4-26 se ilustra los pesos y contribuciones de las variables del arreglo de X sobre el arreglo Y y de la tercer condición del arreglo X sobre el arreglo Y.

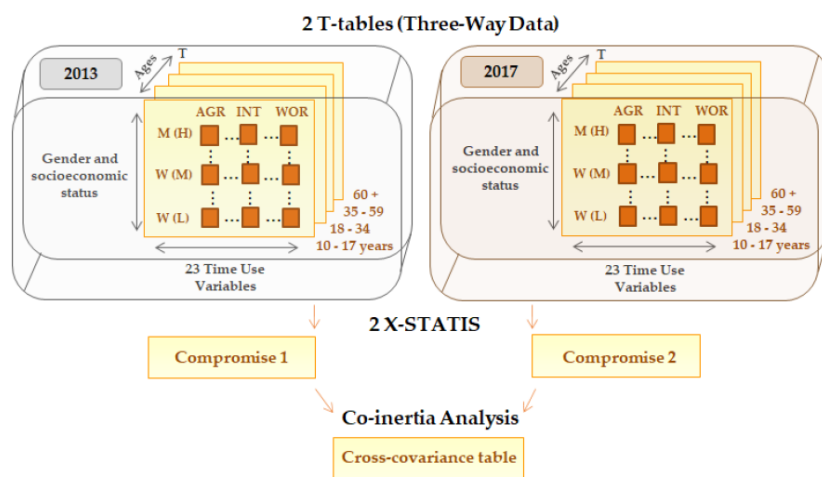
#### 4.4.3. Otras aproximaciones

STATICO o COSTATIS se basa en dos métodos de análisis de datos que ya existían, los cuales son el Análisis de coinerencia desarrollado por COIA, Dolédec y Chessel 1994 y Análisis Triádico Parcial (PTA) desarrollado por Thioulouse y Chessel (1987), los cuales han tenido como objetivo explorar las relaciones entre dos tablas de datos de tres vías.

#### COSTATIS

El método CO-STATIS fue desarrollado por Thioulouse (2011) como un proceso para resumir el análisis de datos de tres vías cuando se utilizan consecutivamente dos análisis X-STATIS,

inicialmente propuesto por [Jaffrenou \(1978\)](#), que tomó el nombre de Análisis triádico por [Thioulouse et al. \(2004\)](#), Análisis Triádico Parcial (PTA) por [P. Kroonenberg \(1989\)](#) y finalmente denominado X-STATIS por [Abdi et al. \(2012\)](#) y un análisis de coinercia desarrollado por [Dolédec y Chessel \(1994\)](#). Esta técnica ha sido utilizada especialmente en estudios de ecología ambiental y para el análisis socioeconómico.

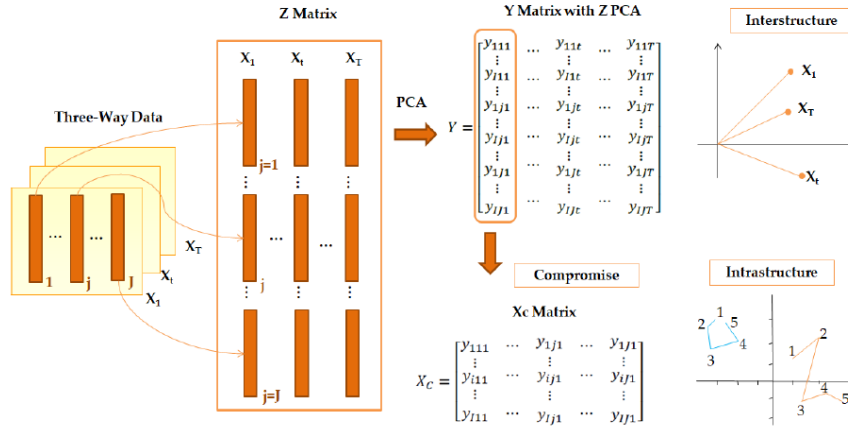


**Figura 4-27.:** Esquema de análisis de datos para la técnica CO-STATIS., ([Medina-Hernández et al., 2021](#))

CO-STATIS se desarrolla en dos etapas:

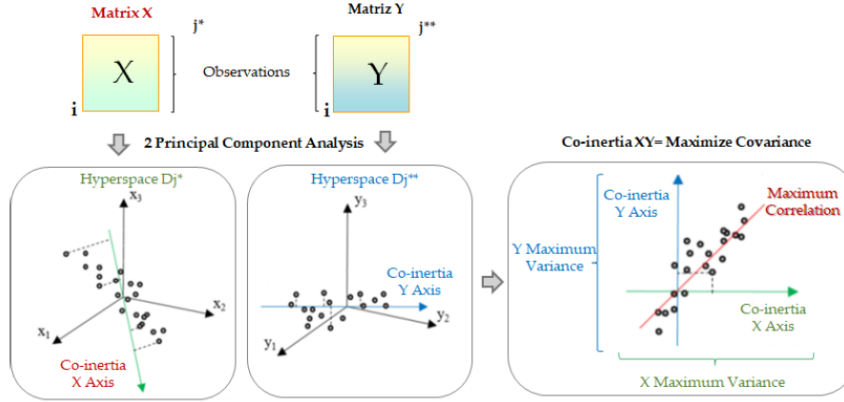
- **Etapas 1: Análisis Triádico Parcial (PTA) o X-STATIS**

Esta etapa se desarrolla en tres subetapas, la primera de las cuales es el análisis de interestructuras, cuyo objetivo es estudiar las similitudes globales entre las tablas. La segunda etapa es el análisis del compromiso, cuyo objetivo es resumir la información de las matrices iniciales para proporcionar una imagen de las estructuras comunes a todas las tablas. Por último se realiza el análisis de intraestructura, que consiste en analizar la reproducibilidad del compromiso. En esta etapa se presentan las posiciones o trayectorias de cada individuo y/o variable incluida en las tablas originales, y se analizan sus posiciones relativas en términos de sus posiciones entre sí y con la posición del compromiso, para comprender este paso ver [4-28](#).



**Figura 4-28.:** Esquema de análisis de datos para la técnica X-STATIS., (Medina-Hernández et al., 2021)

#### ■ Etapa 2: Análisis de coinerencia



**Figura 4-29.:** Esquema de análisis de datos para la técnica Coinercia., (Medina-Hernández et al., 2021)

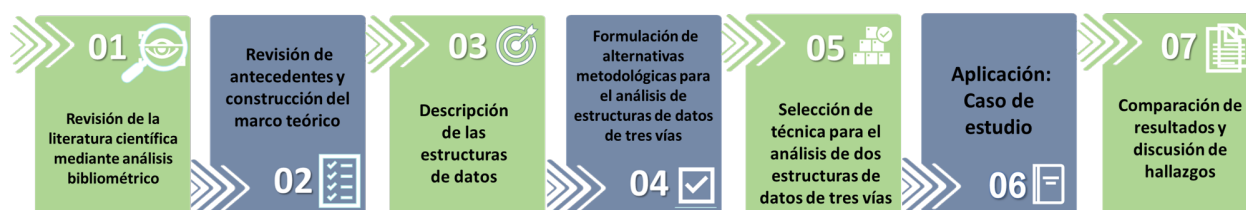
Este análisis permite encontrar estructuras comunes entre dos conjuntos de variables; es decir, analizar simultáneamente dos matrices de datos que contienen los mismos individuos (filas) y variables diferentes o similares (columnas), para describir su coestructura maximizando la covarianza entre las coordenadas de las filas de las dos tablas. Este método pertenece a la familia de técnicas que estudian matrices por pares, como el análisis canónico y el análisis de correspondencia canónica (CCA), pero a diferencia de estos, esta técnica no maximiza las correlaciones entre las coordenadas, sino que maximiza la covarianza entre ellas, como se observa en la Figura 4-29.

Este método es una generalización del análisis de correlación canónica (CANCOR) y del análisis de redundancia (RDA), que presentan más restricciones que el análisis de

coinerencia ([Medina-Hernández et al., 2021](#)).

## 5. Metodología

En este capítulo, se presenta la metodología planteada para el desarrollo de este trabajo, en la Figura 5-1, se resumen las fases metodológicas. Inicialmente, se desarrolla una revisión en la literatura de las técnicas y metodologías utilizadas para el análisis de estructuras de datos de tres vías (presentada en los Capítulos 3 y 4 de este documento), se procede a describir las estructuras de datos que serán insumo para el caso de estudio, y a las cuales se le aplicarán las Fases 3 y 4, de formulación y selección de las técnicas para el análisis y aplicación en el caso de estudio.



**Figura 5-1.:** Fases metodológicas del trabajo de investigación

Bajo lo anterior, es importante resaltar que desde el análisis bibliométrico que se desarrolló en el aplicativo biblioshiny disponible en la librería de Bibliometrix del software estadístico R y sus resultados se encuentran descritos en el Capítulo 3 de este trabajo, en el cual se seleccionan los antecedentes relevantes para este trabajo de grado. A partir de la revisión de antecedentes se realizó la construcción del marco teórico desarrollado en el Capítulo 4. La fase 7 se desarrolla en la sección 6.5 del Capítulo 6.

Las tablas de datos del Producto Interno Bruto (PIB) y el Índice de Pobreza Multidimensional (IPM), se utilizarán para probar las alternativas metodológicas. Estas bases de datos son de carácter público y se pueden encontrar en la página oficial del Departamento Administrativo Nacional de Estadísticas (DANE, 2022b,a) en los siguientes enlaces:

- PIB: <https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-departamentales>
- IPM del 2018 al 2021: <https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-multidimensional/pobreza-multidimensional-2020>

- Proyecciones poblacionales: <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion>

## 5.1. Descripción de las estructuras de datos

En esta sección se describen los conjuntos de variables de cada una de las hojas de datos que se usarán para el caso de estudio, mediante las cuales se logre la implementación de las alternativas de despliegue propuestas e identificación de la metodología más conveniente para las dos estructuras de datos.



**Figura 5-2.: Datos PIB**



**Figura 5-3.: Datos IPM**

Como se observa en las Figuras 5-2 y 5-3, los individuos corresponden a las 33 subdivisiones administrativas, en otras palabras, los departamentos que componen al país.

Las variables se encuentran agrupadas en los dos conjuntos de datos, el primero que corresponde a datos del PIB comprende 12 ramas de actividades económicas y el segundo, corresponde a los datos del IPM que comprende 15 indicadores. La descripción detallada del conjunto de variables se presentan en las Tablas 5-1 y 5-2.

| No. | Abreviatura | Actividad Económica                                                                                    |
|-----|-------------|--------------------------------------------------------------------------------------------------------|
| 1   | Agri        | Agricultura, ganadería, caza, silvicultura y pesca                                                     |
| 2   | Min         | Explotación de minas y canteras                                                                        |
| 3   | Manu        | Industrias manufactureras                                                                              |
| 4   | Elec        | Suministro de electricidad, gas y agua                                                                 |
| 5   | Cons        | Construcción                                                                                           |
| 6   | Come        | Comercio, reparación, restaurantes y alojamiento                                                       |
| 7   | Comu        | Información y comunicaciones                                                                           |
| 8   | Fina        | Actividades financieras y de seguros                                                                   |
| 9   | Inmo        | Actividades inmobiliarias                                                                              |
| 10  | Prof        | Actividades profesionales, científicas y técnicas; actividades de servicios administrativos y de apoyo |
| 11  | Serv.Soc    | Actividades de servicios sociales, comunales y personales                                              |
| 12  | Recre       | Actividades de entretenimiento y recreación                                                            |

**Tabla 5-1.: Conjunto de variables - Ramas de actividad económica del PIB (Ver Anexo A.1)**

Es importante aclarar que se realiza un ajuste a los datos del PIB, para obtener el PIB percapita que se obtiene a partir del cociente entre el PIB por departamento y el número de habitantes proyectados por departamento calculadas con base en los resultados del Censo Nacional de Población y Vivienda - CNPV- 2018 (DANE, 2022c). Lo anterior permite comparar los departamentos que debido a sus diferentes tamaños poblacionales podría presentar sesgo en el análisis. Se obtiene el PIB percapita para cada año que se evalúa.

Por otra parte, los conjuntos de datos comparten las mismas divisiones administrativas y periodicidad anual de evaluación durante el período 2018 a 2021.

| No. | Abreviación | Variables IPM                                            |
|-----|-------------|----------------------------------------------------------|
| 1   | Analfab     | Analfabetismo                                            |
| 2   | B.Ledu      | Bajo logro educativo                                     |
| 3   | B.Servcpi   | Barreras a servicios para cuidado de la primera infancia |
| 4   | B.Access    | Barreras de acceso a servicios de salud                  |
| 5   | DS.Ld       | Desempleo de larga duración                              |
| 6   | HC          | Hacinamiento crítico                                     |
| 7   | Inad.Ee     | Inadecuada eliminación de excretas                       |
| 8   | Ina.Esc     | Inasistencia escolar                                     |
| 9   | MT.Pext     | Material inadecuado de paredes exteriores                |
| 10  | MT.Pisos    | Material inadecuado de pisos                             |
| 11  | Rez.Esc     | Rezago escolar                                           |
| 12  | Sin.Fagua   | Sin acceso a fuente de agua mejorada                     |
| 13  | Sin.Asalud  | Sin aseguramiento en salud                               |
| 14  | Trab.Infan  | Trabajo infantil                                         |
| 15  | Trab.Infor  | Trabajo informal                                         |

**Tabla 5-2.:** Conjunto de variables - Indicadores para el IPM (Ver Anexo A.2)

A partir de la organización de los datos anteriormente descritos, como se presentan en las figuras 5-2 y 5-3, se realizará el análisis simultáneo de estos datos siguiendo las etapas siguientes como estrategia metodológica.

## 5.2. Formulación de alternativas de despliegue para el análisis de dos estructuras de datos de tres vías.

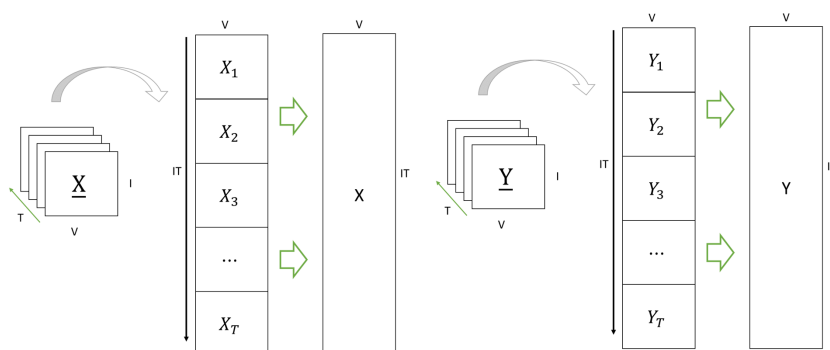
Se realizó una revisión en la literatura científica mediante análisis bibliométrico (Ver Capítulo 3) que permitió conocer el panorama actual y avances científicos de las metodologías utilizadas para el análisis simultáneo de estructuras de datos de tres vías.

Se realizó un análisis por métodos ajustados a 2 vías, partir de diferentes alternativas de despliegue que propone en su investigación Nomikos y MacGregor (1994) y Swiercz y Mroczkowska (2020).

A continuación, se presentan las alternativas de despliegue y metodologías identificadas que se aplican en esta investigación.

### 5.2.1. Alternativa de despliegue vertical

Esta opción de despliegue implicó organizar cada arreglo verticalmente, donde la tercera vía, el tiempo, se dispone uno debajo del otro, formando así una matriz de dos dimensiones ( $IT \times V$ ), tal como se muestra en la Figura 5-4.



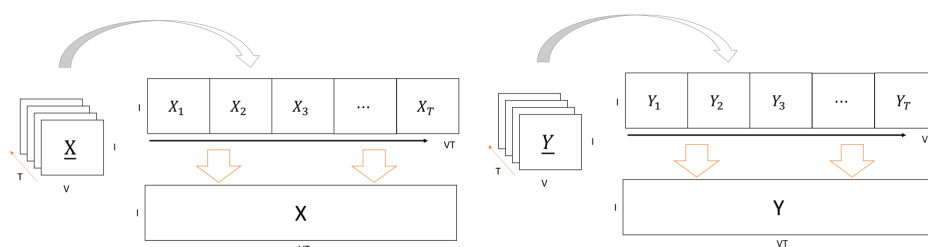
**Figura 5-4.:** Alternativa de despliegue vertical

Donde:

- I: Individuos
- V: Variables
- T: Tiempo

### 5.2.2. Alternativa de despliegue horizontal

Como segunda opción, se llevó a cabo un procedimiento similar al anterior. Sin embargo, en este caso, se desplegó cada arreglo horizontalmente, donde la tercera vía, el tiempo, se dispuso uno al lado del otro, creando así dos matrices con dimensiones ( $I \times VT$ ), como se ilustra en la Figura 5-5.



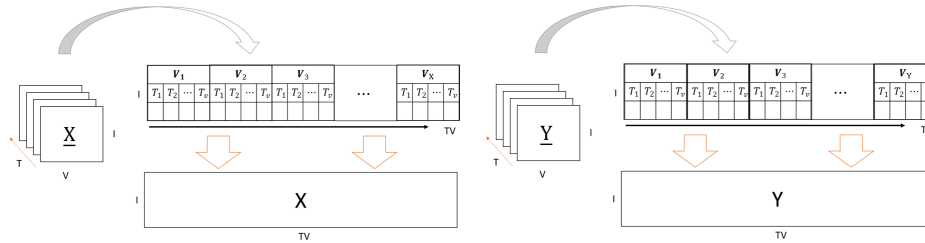
**Figura 5-5.:** Alternativa de despliegue horizontal

Donde:

- I: Individuos
- V: Variables
- T: Tiempo

### 5.2.3. Alternativa de despliegue combinada

Como tercer alternativa se realizó un despliegue horizontal de la matriz, realizando una combinación de las vías 2 y 3, que corresponden a las variables y el tiempo respectivamente. En ese sentido, el despliegue de cada conjunto de datos, se ajustó a un despliegue horizontal, donde cada variable estaba subdividida por sus valores temporales, conformando un arreglo con dimensiones  $(I \times TV)$  como se muestra en la 5-6.



**Figura 5-6.:** Alternativa de despliegue combinado de la vía 2 (variables) y vía 3 (tiempo)

Donde:

- I: Individuos
- V: Variables
- T: Tiempo

### 5.2.4. Aplicación de técnicas

Una vez ajustados los conjuntos de datos a una estructura de dos vías, según las alternativas de las secciones anteriores, se aplicaron las metodologías de Análisis Factorial Múltiple (AFM), Mínimos Cuadrados Parciales (PLS) y Mínimos Cuadrados Parciales N vías (N-way PLS) para el análisis integrado y simultáneo de las dos bases de datos.

| Alternativas                    | AFM | PLS | N-way PLS |
|---------------------------------|-----|-----|-----------|
| Despliegue vertical             | X   | X   |           |
| Despliegue horizontal           | X   | X   |           |
| Despliegue horizontal combinado | X   |     |           |
| Sin despliegue                  |     |     | X         |

**Tabla 5-3.:** Metodologías aplicadas a las alternativas de despliegue realizadas

En la Tabla **6-23**, se puede ver cómo se aplican las diferentes metodologías según las distintas formas de despliegue utilizadas. La metodología AFM se emplea en las tres formas de despliegue propuestas. En cuanto a la metodología PLS, se utiliza tanto en el despliegue vertical como en el horizontal, ya que al aplicarla en el despliegue horizontal combinado se obtienen resultados similares al despliegue horizontal. Por último, la metodología N-Way PLS se aplicó directamente a las estructuras de datos de tres vías (cubos), por lo que no fue necesario realizar ninguna forma de despliegue adicional.

### 5.3. Selección de técnicas para el análisis de dos estructuras de datos de tres vías.

Una vez formuladas las alternativas de despliegue y realizado los respectivos análisis exploratorios de los datos para cada conjunto de datos y un análisis de correlación, se procede a realizar el análisis simultáneo de los dos conjuntos de datos para estudiar el nivel de asociación entre la generación de riqueza y las condiciones de pobreza en Colombia. Para ello, a continuación se indica el proceso utilizado con cada metodología utilizada.

#### 5.3.1. Análisis Factorial Múltiple (AFM)

Este método permite realizar el análisis simultáneo de varios grupos de variables sobre un mismo individuo.

A partir de las alternativas de despliegue anteriormente indicadas, se realiza la fusión de los dos conjuntos de datos en una sola tabla conjunta.

Por tanto, se desarrollaron tres AFM que se detallan a continuación:

- **AFM 1:** Inicialmente, se realiza un AFM con despliegue vertical, donde se aumentarán los individuos por filas, por ende se tendrá un total de 132 individuos que corresponden a 33 departamentos por 4 años medidos (2018 a 2021), los grupos de variables serán, un grupo para las 12 variables del conjunto de datos del PIB y un segundo grupo para el conjunto de variables del IPM, como muestra la Figura **5-7**.

| Ramas de actividad PIB |      |     |     |       | Indicadores IPM |     |     |        |
|------------------------|------|-----|-----|-------|-----------------|-----|-----|--------|
| Departamentos          | Agri | Min | ... | Recre | Anfal           | BLE | ... | T_INFO |
| Antioquía              |      |     |     |       |                 |     |     |        |
| Valle del Cauca        |      |     |     |       |                 |     |     |        |
| ⋮                      |      |     |     |       |                 |     |     |        |
| Bogotá                 |      |     |     |       |                 |     |     |        |
| Departamentos          | Agri | Min | ... | Recre | Anfal           | BLE | ... | T_INFO |
| Antioquía              |      |     |     |       |                 |     |     |        |
| Valle del Cauca        |      |     |     |       |                 |     |     |        |
| ⋮                      |      |     |     |       |                 |     |     |        |
| Bogotá                 |      |     |     |       |                 |     |     |        |
| Departamentos          | Agri | Min | ... | Recre | Anfal           | BLE | ... | T_INFO |
| Antioquía              |      |     |     |       |                 |     |     |        |
| Valle del Cauca        |      |     |     |       |                 |     |     |        |
| ⋮                      |      |     |     |       |                 |     |     |        |
| Bogotá                 |      |     |     |       |                 |     |     |        |
| Departamentos          | Agri | Min | ... | Recre | Anfal           | BLE | ... | T_INFO |
| Antioquía              |      |     |     |       |                 |     |     |        |
| Valle del Cauca        |      |     |     |       |                 |     |     |        |
| ⋮                      |      |     |     |       |                 |     |     |        |
| Bogotá                 |      |     |     |       |                 |     |     |        |

**Figura 5-7.:** Tabla conjunta de las estructuras del PIB e IPM bajo la alternativa de despliegue vertical

- **AFM 2:** El segundo AFM consiste en aplicar el método cuando se realiza un despliegue horizontal de los años, por tanto, se tendrán 33 individuos que corresponde a los departamentos y 4 grupos que corresponden a los años, donde cada grupo contiene las 27 variables de la fusión de los dos conjuntos de datos del PIB y del IPM, como muestra la Figura 5-8.

| Ramas de actividad PIB |      |     |       |  |      |     |       | Indicadores IPM |     |        |  |       |     |        |  |
|------------------------|------|-----|-------|--|------|-----|-------|-----------------|-----|--------|--|-------|-----|--------|--|
| Departamentos          | Agri | ... | Recre |  | Agri | ... | Recre | Anfal           | ... | T_INFO |  | Anfal | ... | T_INFO |  |
| Antioquía              |      |     |       |  |      |     |       |                 |     |        |  |       |     |        |  |
| Valle del Cauca        |      |     |       |  |      |     |       |                 |     |        |  |       |     |        |  |
| ⋮                      |      |     |       |  |      |     |       |                 |     |        |  |       |     |        |  |
| Bogotá                 |      |     |       |  |      |     |       |                 |     |        |  |       |     |        |  |

**Figura 5-8.:** Tabla conjunta de las estructuras del PIB e IPM bajo la alternativa de despliegue horizontal

- **AFM 3:** Por último el tercer AFM consiste en el despliegue horizontal de las variables, por tanto, se tendrán 33 individuos que corresponde a los departamentos y 28 grupos que corresponden a las variables, donde cada grupo contiene los 4 años medidos a cada variable de la fusión de los dos conjuntos de datos del PIB y del IPM, como muestra la Figura 5-9.

| Ramas de actividad PIB |                      |     |                      |       |                       |     |                       | Indicadores IPM       |     |                       |       |                        |     |                        |  |
|------------------------|----------------------|-----|----------------------|-------|-----------------------|-----|-----------------------|-----------------------|-----|-----------------------|-------|------------------------|-----|------------------------|--|
| Departamentos          | Agri <sub>2018</sub> | ... | Agri <sub>2021</sub> | • • • | Recre <sub>2018</sub> | ... | Recre <sub>2021</sub> | Anfal <sub>2018</sub> | ... | Anfal <sub>2021</sub> | • • • | T_INFO <sub>2018</sub> | ... | T_INFO <sub>2021</sub> |  |
| Antioquía              |                      |     |                      |       |                       |     |                       |                       |     |                       |       |                        |     |                        |  |
| Valle del Cauca        |                      |     |                      |       |                       |     |                       |                       |     |                       |       |                        |     |                        |  |
| ⋮                      |                      |     |                      |       |                       |     |                       |                       |     |                       |       |                        |     |                        |  |
| Bogotá                 |                      |     |                      |       |                       |     |                       |                       |     |                       |       |                        |     |                        |  |

**Figura 5-9.:** Tabla conjunta de las estructuras del PIB e IPM bajo la alternativa de despliegue horizontal combinada

La aplicación de este método permitirá observar los principales factores de variabilidad de los departamentos, las relaciones entre los grupos, el grado de semejanza entre los departamentos, relaciones entre las variables de los conjuntos de datos.

Se hizo uso del paquete *FactoMineR* (Husson et al., 2013) *factoextra* (Kassambara, 2016) del software estadístico RStudio.

### 5.3.2. Mínimo Cuadrados Parciales (PLS)

Se emplea la metodología PLS, la cual combina métodos que son Unfold PLS, Análisis de Componentes Principales y Análisis de regresión, permitiendo desarrollar un modelo con múltiples variables predictoras (conjunto de datos del PIB) y una o múltiples variables de respuestas (conjunto de datos del IPM).

A partir de los conjuntos de datos, se toma una matriz como las variables predictoras, para este caso se seleccionó el conjunto de datos que corresponden a la generación de riqueza (PIB) y la matriz de variables de respuesta se tomó el conjunto de datos del IPM.

El desarrollo de este método se realizará a partir de las alternativas de despliegue de los conjuntos de datos, por tanto:

- **PLS 1:** Inicialmente, se realiza un PLS para la alternativa de despliegue vertical de cada conjunto de datos, por tanto, se aumentaron los individuos por filas contando con un total de 132 individuos que corresponden a 33 departamentos por 4 años medidos (2018 a 2021) para cada conjunto de datos. Lo anterior indica que el conjunto de datos de PIB tiene una estructura de  $(132 \times 15)$  y el IMP  $(132 \times 15)$ , como muestra la Figura 5-10.



**Figura 5-10.:** Pls aplicado a las estructuras del PIB e IPM bajo la alternativa de despliegue vertical

- **PLS 2:** El segundo PLS consiste en aplicar el método cuando se realiza un despliegue horizontal de cada conjunto de datos según los años, por tanto, el conjunto de datos del PIB tiene una estructura de  $(33 \times 48)$  que corresponde a 33 departamentos y 48 variables en los 4 años medidos (12 variables por año) y para el conjunto de datos del IPM es de  $(33 \times 60)$  que corresponde a 33 departamentos y 60 variables en los 4 años medidos (15 variables por año), como se muestra en la Figura 5-11.



**Figura 5-11.:** Pls aplicado a las estructuras del PIB e IPM bajo la alternativa de despliegue horizontal

Se hace uso del paquete *plsdepot* (Sánchez, 2016) del software estadístico RStudio.

### 5.3.3. N - PLS

Se emplea el método *N*-PLS, el cual desarrolla un PLS para cada una de las estructuras de datos, variables predictorias **X** (conjunto de datos del PIB) y para variables de respuestas de datos **Y** (conjunto de datos IPM). Finalmente, realiza un modelo PLS donde relaciona

ambos modelos desarrollados de forma individual y así obtener el modelo de regresión  $N$ -PLS.

A partir de los conjuntos de datos, se toman las estructuras de datos de tres vías sin realizar despliegue, por tanto, como variables predictoras se toma el cubo de datos del PIB que corresponden a la productividad de los departamentos de Colombia y como variables de respuesta el cubo de datos del IPM que corresponde a las condiciones de pobreza de los departamentos de Colombia.

Se hace uso del paquete **sNPLS** desarrollado por [Hervás et al. \(2018\)](#) e implementado en el software estadístico RStudio.

#### 5.3.4. Análisis de comparación de los resultados y hallazgos

Para dar cumplimiento a los objetivos se realiza una comparación cualitativa de los resultados y hallazgos de las metodologías aplicadas. Para ello, el análisis consiste en la comprensión de la aplicación de cada una de las metodologías, las fortalezas de cada una según su las alternativas de despliegue propuestas y las limitaciones identificadas en el análisis de datos.

Esta comparación permite identificar la metodología más apropiada según las necesidades, estructura de los datos y despliegue utilizado. A partir de los fundamentos teóricos de cada método relacionado en el capítulo 4, se compara la aplicabilidad y flexibilidad de cada método, para ello, se evalúa la capacidad para manejar diferentes tipos de datos y las diferentes estructuras de los datos.

Asimismo, es fundamental reconocer las limitaciones y restricciones asociadas con cada método, incluyendo consideraciones sobre la complejidad computacional, los supuestos subyacentes y las condiciones de aplicabilidad. Esto permite una evaluación más completa de las fortalezas y debilidades de cada método. La interpretación de los resultados generados por cada método es otro aspecto importante a considerar en el análisis comparativo. La calidad de los resultados obtenidos son consideraciones importantes para evaluar la robustez de los métodos frente a diferentes escenarios de despliegue en este caso específico.

De lo anterior, realizar comparaciones cualitativas entre AFM, PLS y  $N$ -PLS lo que permite una comprensión más profunda de sus características, aplicaciones y limitaciones. Para el caso de AFM se analizó el valor de inercia (varianza explicada). Para las metodologías de PLS y  $N$ PLS se obtienen los valores de bondad de ajuste como el RMSE, que permitió observar la varianza explicada de la matriz de variables de respuesta a partir de la matriz de variables independientes, por tanto, entre mayor sea este mejor es el modelo.

## 5.4. Aplicación: Caso de estudio generación de riqueza y condiciones de pobreza en Colombia.

Se desarrolló el caso de estudio identificando e implementando las metodologías para las estructuras de datos. El caso de estudio propuesto es el conjunto de variables utilizadas para medir la generación de riqueza (PIB) y condiciones de pobreza en Colombia desde el 2018 al 2021.

El Departamento Administrativo Nacional de Estadística (DANE) provee la información con periodicidad anual de los indicadores del PIB, así como las actividades económicas que componen el indicador y aportan a su construcción por subdivisión administrativa (departamentos), a su vez, la misma entidad proporciona los indicadores de condiciones de pobreza (IPM o Pobreza monetaria).

Se realizó un análisis de similitud y similaridad de las subdivisiones administrativas (33 departamentos) según las condiciones de pobreza y generación de riqueza desde el 2018 al 2021. Igualmente, con cada método se analiza la trayectoria de los individuos y de las variables en el tiempo, identificando la influencia del tiempo en la generación de riqueza y las condiciones de pobreza de los departamentos de Colombia. A modo general, se realiza un análisis detallado de cada método según las alternativas de despliegue utilizadas, interpretando los componentes, trayectorias, representaciones de los grupos, variables e individuos, entre otras medidas de interés.

Por tanto, se compara cada alternativa de despliegue utilizada y cada método, que permitió comparar e identificar coincidencias entre las técnicas y sus diferencias de forma cualitativa.

## 5.5. Software

Se utiliza el software estadístico RStudio en su versión 2022.02.3+492 ([R Core Team, 2020](#)). Las principales librerías utilizadas en el desarrollo del código son:

- FactoMinerR ([Husson et al., 2013](#))
- Factoextra ([Kassambara, 2016](#))
- Plsdepot ([Sánchez, 2016](#))
- sNPLS ([Hervás et al., 2018](#))

## 6. Resultados

En este capítulo se describen los resultados obtenidos aplicando la metodología propuesta y descrita en el Capítulo 5 para el desarrollo de este trabajo. Inicialmente se realiza un análisis exploratorio sobre el conjunto de datos del PIB y del IPM, a partir de indicadores de resumen y herramientas gráficas que permitan observar el comportamiento individual de cada una de las estructuras de datos. Posteriormente, se realiza el análisis simultáneo de los dos bloques de datos, aplicando las alternativas y las técnicas propuestas en la metodología, con el fin de comprender la estructura de correlación entre dos arreglos de datos. Finalmente, se realiza la comparación y contraste de los hallazgos obtenidos con la aplicación de las diferentes alternativas implementadas.

### 6.1. Análisis Exploratorio de Datos

Inicialmente, se realizará un análisis descriptivo de cada una de las estructuras de datos. A partir de este análisis se comprenderá la situación actual, una imagen clara de la situación económica y social de los departamentos de Colombia durante el periodo de tiempo del 2018 al 2021. Los indicadores de resumen y los gráficos comparativos, ayudarán a entender la dinámica de la pobreza y la productividad generada en Colombia desde el 2018 al 2021.

#### 6.1.1. Indicadores del Producto Interno Bruto

La Tabla **6-1** y la Figura **6-1** se resume el comportamiento de las variables que compone la estructura de datos del PIB en sus respectivos años. Se observa una notable variabilidad en la producción tanto entre regiones como entre las actividades económicas, dado que algunas de ellas muestran una alta productividad en áreas específicas. Esta variabilidad puede atribuirse a factores como la ubicación geográfica, la riqueza de las regiones, el clima, el nivel de desarrollo, entre otros, que influyen en la generación de altos niveles de producción o en la preferencia por una actividad económica particular.

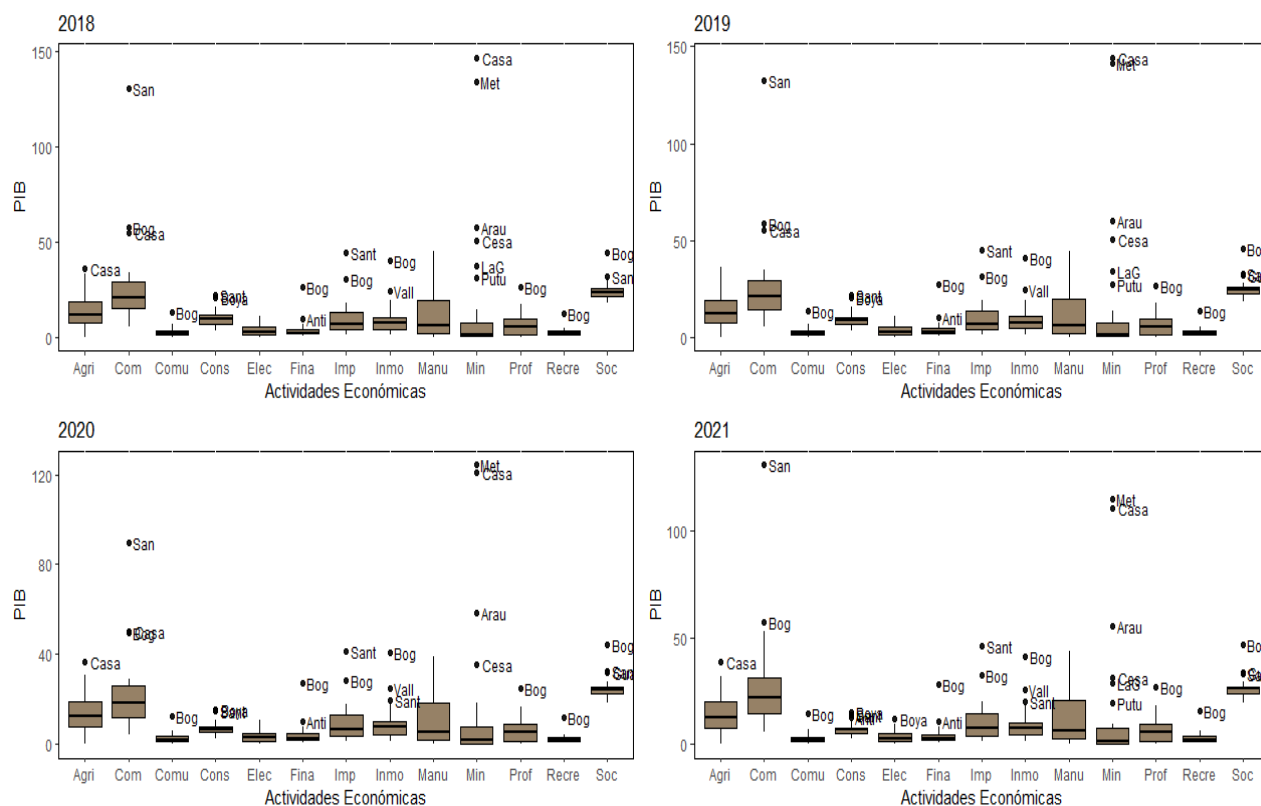
En relación a las estadísticas descriptivas presentadas en la Tabla **6-1**, se observa que las actividades económicas con la mayor producción promedio durante los cuatro años evaluados son el comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas; transporte y almacenamiento; alojamiento y servicios de comida (**Com**), así

como la administración pública y defensa; planes de seguridad social de afiliación obligatoria; educación; actividades de atención de la salud humana y de servicios sociales (**Soc**).

| Año  | Variable | Mínimo | Máximo  | Cuartil 1 | Cuartil 3 | Promedio      | Mediana | DS     |
|------|----------|--------|---------|-----------|-----------|---------------|---------|--------|
| 2018 | Agri     | 0.017  | 36.337  | 7.669     | 19.028    | 13.374        | 11.913  | 8.758  |
|      | Min      | 0.147  | 146.886 | 0.414     | 7.584     | 15.960        | 1.532   | 35.359 |
|      | Manu     | 0.098  | 45.084  | 2.215     | 19.724    | 11.675        | 6.361   | 12.345 |
|      | Elec     | 0.143  | 10.796  | 1.261     | 5.282     | <b>3.581</b>  | 2.928   | 2.934  |
|      | Cons     | 3.425  | 22.069  | 6.783     | 11.489    | 9.823         | 9.435   | 4.364  |
|      | Com      | 5.404  | 130.978 | 14.984    | 29.131    | <b>25.932</b> | 20.988  | 22.124 |
|      | Comu     | 0.054  | 13.485  | 1.188     | 3.625     | <b>2.703</b>  | 2.067   | 2.535  |
|      | Fina     | 0.649  | 26.177  | 1.844     | 4.334     | 4.002         | 2.390   | 4.478  |
|      | Inmo     | 1.415  | 40.658  | 4.381     | 10.630    | 8.951         | 7.362   | 7.651  |
|      | Prof     | 0.066  | 26.468  | 1.383     | 9.766     | 6.271         | 5.339   | 6.056  |
|      | Soc      | 17.955 | 44.837  | 21.574    | 25.459    | <b>24.373</b> | 23.358  | 4.705  |
|      | Recre    | 0.429  | 12.185  | 1.209     | 3.114     | <b>2.374</b>  | 1.882   | 2.119  |
|      | Imp      | 1.373  | 44.462  | 3.963     | 13.170    | 9.858         | 6.911   | 9.075  |
| 2019 | Agri     | 0.017  | 36.288  | 7.553     | 19.278    | 13.390        | 12.367  | 8.682  |
|      | Min      | 0.140  | 144.133 | 0.427     | 7.409     | 15.883        | 1.633   | 35.812 |
|      | Manu     | 0.095  | 44.449  | 2.247     | 19.845    | 11.527        | 6.175   | 12.169 |
|      | Elec     | 0.146  | 10.951  | 1.206     | 5.275     | <b>3.584</b>  | 2.992   | 2.941  |
|      | Cons     | 3.175  | 21.962  | 6.549     | 10.452    | 9.318         | 8.511   | 4.260  |
|      | Com      | 5.382  | 132.578 | 14.582    | 29.120    | <b>26.208</b> | 21.392  | 22.446 |
|      | Comu     | 0.052  | 13.310  | 1.097     | 3.618     | <b>2.660</b>  | 2.002   | 2.507  |
|      | Fina     | 0.658  | 27.130  | 1.933     | 4.515     | 4.155         | 2.501   | 4.651  |
|      | Inmo     | 1.442  | 40.890  | 4.407     | 10.723    | 9.031         | 7.452   | 7.731  |
|      | Prof     | 0.068  | 26.810  | 1.388     | 9.637     | 6.311         | 5.438   | 6.142  |
|      | Soc      | 18.572 | 45.534  | 22.503    | 25.727    | <b>25.066</b> | 24.321  | 4.744  |
|      | Recre    | 0.450  | 13.600  | 1.290     | 3.368     | <b>2.599</b>  | 1.996   | 2.374  |
|      | Imp      | 1.369  | 45.307  | 4.119     | 13.560    | 10.022        | 7.086   | 9.267  |
| 2020 | Agri     | 0.017  | 36.365  | 7.684     | 19.038    | 13.421        | 12.324  | 8.612  |
|      | Min      | 0.102  | 124.368 | 0.284     | 7.869     | 13.075        | 1.816   | 30.646 |
|      | Manu     | 0.086  | 38.702  | 2.006     | 18.616    | 10.328        | 5.391   | 10.722 |
|      | Elec     | 0.139  | 10.758  | 1.135     | 4.996     | <b>3.427</b>  | 2.864   | 2.818  |
|      | Cons     | 2.351  | 15.387  | 5.414     | 7.757     | 6.915         | 6.370   | 2.895  |
|      | Com      | 4.129  | 89.853  | 11.678    | 26.024    | <b>21.306</b> | 18.577  | 16.221 |
|      | Comu     | 0.049  | 12.731  | 0.993     | 3.592     | <b>2.525</b>  | 1.952   | 2.394  |
|      | Fina     | 0.645  | 27.208  | 1.958     | 4.577     | 4.168         | 2.560   | 4.668  |
|      | Inmo     | 1.431  | 40.651  | 4.365     | 9.945     | 8.976         | 7.486   | 7.716  |
|      | Prof     | 0.064  | 24.614  | 1.265     | 8.830     | 5.899         | 5.274   | 5.674  |
|      | Soc      | 18.095 | 44.578  | 22.196    | 25.459    | <b>24.703</b> | 24.292  | 4.692  |
|      | Recre    | 0.413  | 12.072  | 1.145     | 2.856     | <b>2.247</b>  | 1.633   | 2.094  |
|      | Imp      | 1.262  | 41.488  | 3.771     | 12.902    | 9.185         | 6.648   | 8.470  |
| 2021 | Agri     | 0.017  | 38.386  | 7.814     | 19.760    | 13.658        | 12.368  | 9.025  |
|      | Min      | 0.101  | 115.415 | 0.282     | 7.716     | 12.678        | 1.200   | 28.434 |
|      | Manu     | 0.085  | 43.478  | 2.357     | 20.699    | 11.738        | 6.380   | 12.207 |
|      | Elec     | 0.141  | 11.598  | 1.132     | 5.170     | <b>3.549</b>  | 2.870   | 2.963  |
|      | Cons     | 2.385  | 14.974  | 5.212     | 7.798     | 7.108         | 6.938   | 2.898  |
|      | Com      | 5.729  | 131.179 | 14.349    | 31.358    | <b>26.249</b> | 21.987  | 22.142 |
|      | Comu     | 0.051  | 14.320  | 1.044     | 3.465     | <b>2.729</b>  | 2.206   | 2.684  |
|      | Fina     | 0.650  | 27.819  | 2.001     | 4.695     | 4.248         | 2.635   | 4.777  |
|      | Inmo     | 1.460  | 40.926  | 4.386     | 9.936     | 9.085         | 7.589   | 7.819  |
|      | Prof     | 0.068  | 26.682  | 1.369     | 9.480     | 6.367         | 5.718   | 6.167  |
|      | Soc      | 19.155 | 46.802  | 23.725    | 27.051    | <b>26.217</b> | 26.005  | 4.787  |
|      | Recre    | 0.533  | 15.796  | 1.535     | 3.901     | <b>2.984</b>  | 2.090   | 2.771  |
|      | Imp      | 1.350  | 46.300  | 4.012     | 14.435    | 10.188        | 7.483   | 9.529  |

**Tabla 6-1.:** Estadísticas descriptivas de los indicadores del PIB 2018 - 2021

Estas actividades muestran un comportamiento constante en términos de producción a lo largo de los diferentes años. Por otro lado, se observa un rezago en la producción de las actividades económicas de información y comunicaciones (**Comu**), así como actividades artísticas, de entretenimiento y recreación y otras actividades de servicios; actividades de los hogares individuales en calidad de empleadores; actividades no diferenciadas de los hogares individuales como productores de bienes y servicios para uso propio (**Rec**).



**Figura 6-1.:** Distribución de los indicadores del PIB por año

En contraste, la actividad económica de Explotación de minas y canteras (Min) se destaca según su comportamiento en la Figura 6-1. Esta actividad muestra un comportamiento asimétrico con valores atípicos que representan aproximadamente el 50 % de la producción. Estos valores atípicos se concentran en departamentos como Casanare, Meta, Arauca, Cesar, La Guajira y Putumayo, los cuales principalmente pertenecen a la región Orinoquía, reconocida por su actividad minera intensiva y su economía basada en la extracción de petróleo y otros recursos naturales como coltán, plata, oro Naquen, cobre, carbón, aluminio, arcilla, arena y gravas de sol.

De manera distinta a los departamentos anteriores, los departamentos de Cesar y La Guajira que también presentan una alta productividad en esta actividad económica, posiblemente, debido a la presencia de grandes centros de explotación de carbón, siendo este el segundo producto de mayor exportación en el país (ACM, 2021). Según ACM (2021) la participación en la industria minera de La Guajira y Cesar representan el 44 % y 37 % de la economía respectivamente y en exportaciones de carbón aportan el 99 % y 98 % respectivamente. En el caso de Putumayo, su producción minera se centra principalmente en materiales como oro y plata, siendo estas las actividades que más contribuyen a la economía del departamento.

La actividad económica denominada *Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas; transporte y almacenamiento; alojamiento y servicios de comida* (Com) destaca el aporte del departamento de San Andrés, departamento caracterizado por su actividad turística. Por otro lado, departamentos como Bogotá, Antioquia y Valle del Cauca se distinguen por sus actividades asociadas al sector secundario, como *Actividades inmobiliarias* (Inmo), *Actividades financieras y de seguros* (Fina). Es importante señalar que se observa una disminución en la productividad de estas actividades económicas de un año a otro, específicamente entre 2019 y 2020. Esto puede atribuirse a los efectos de la pandemia de la Covid-19, la cual tuvo un impacto significativo a nivel mundial y afectó diversos sectores económicos. En el caso del departamento de San Andrés, se evidencia una disminución particularmente en el ámbito turístico debido a las restricciones y la situación sanitaria durante el año 2020.

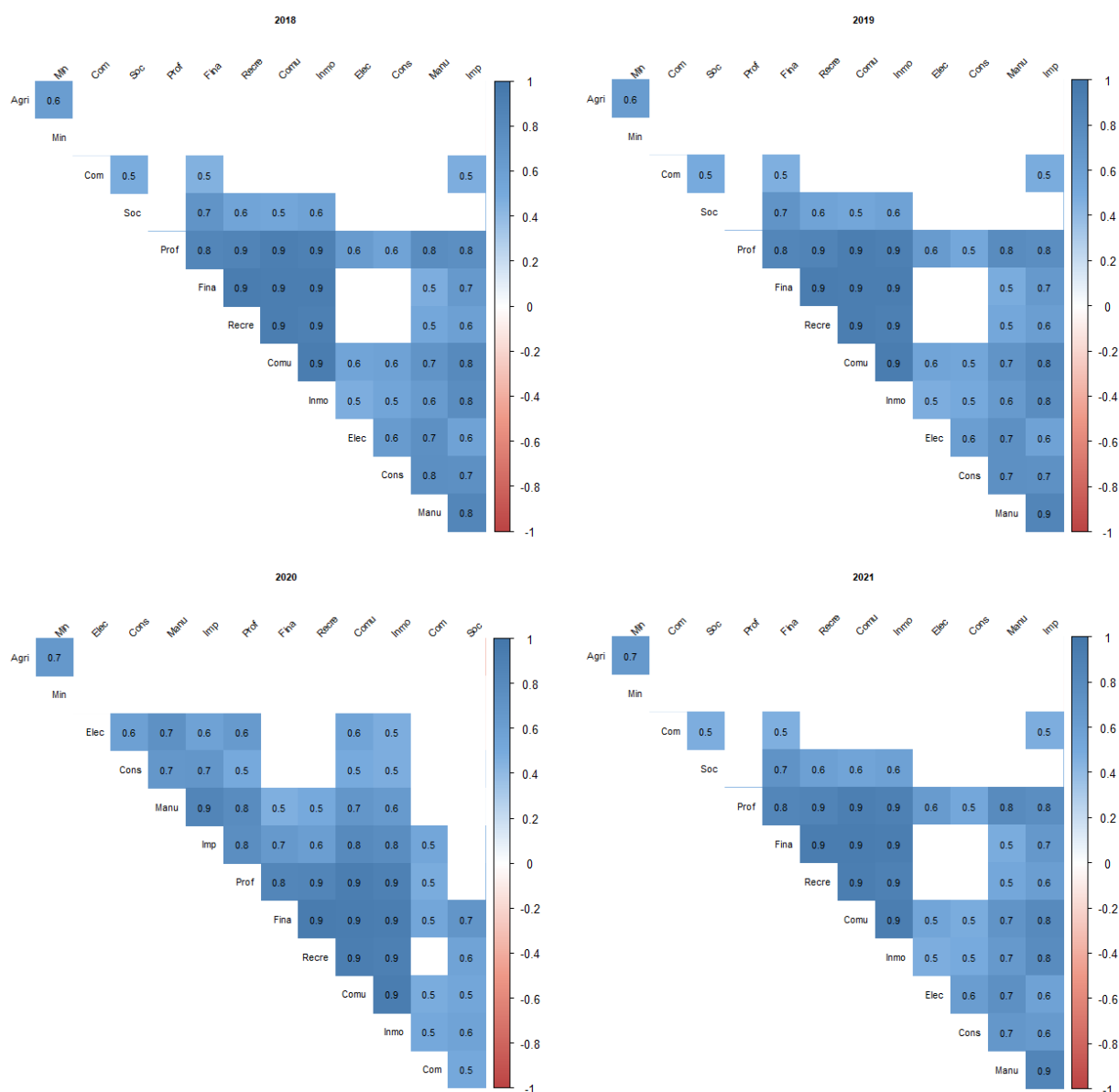
La Figura 6-2 muestra la estructura de correlación entre las diferentes actividades económicas que contribuyen al Producto Interno Bruto (PIB), con una comparación realizada a lo largo de los cuatro años evaluados (2018 al 2021). Esta representación visualiza las correlaciones tanto fuertes como débiles, destacando tanto el coeficiente de correlación como su intensidad de color, siendo azul cuando hay una correlación positiva fuerte y naranja cuando existe una correlación negativa fuerte. Para facilitar el análisis y para efectos de simplificar la representación, aquellas correlaciones no significativas ( $\text{Valor-}p > 0.05$ ) no presentan un color asignado.

Al examinar los resultados, se observa que en el año 2018 existen correlaciones significativas entre las actividades de cada sector productivo. Dentro del sector primario, se destaca una correlación moderada positiva del 60 % entre las actividades de *Agricultura, ganadería, caza, silvicultura y pesca* (Agri) y *Explotación de minas y canteras* (Min). Por otro lado, los sectores secundario y terciario presentan una alta correlación entre sí. *Actividades como Actividades inmobiliarias* (Inmo), *Información y comunicaciones* (Comu), *Actividades profesionales, científicas y técnicas; actividades de servicios administrativos y de apoyo* (Prof) muestran una correlación positiva alta del 90 % entre ellas. Estas actividades también se encuentran altamente correlacionadas con la variable de *Impuestos* (Imp), con una correlación del 80 %.

Además, se observa que la actividad de *Suministro de electricidad, gas, vapor y aire acondicionado; distribución de agua; evacuación y tratamiento de aguas residuales, gestión de desechos y actividades de saneamiento ambiental* (Elec) está estrechamente relacionada con las actividades del sector secundario.

Por otro lado, se destaca la baja correlación existente entre la variable de Impuesto (Imp) con

actividades sociales, la cual es del 30 %, y la baja correlación con actividades de Agricultura y Minería. Las actividades de Agricultura y Minería presentan baja relación con las otras actividades que componen el PIB, siendo no significativa la correlación entre la agricultura y el sector de la manufactura.



**Figura 6-2.:** Correlación entre las variables del PIB por año

En relación a los años 2019 y 2021, se observa un comportamiento de correlación similar al año 2018. Se mantiene la tendencia de una alta correlación dentro de las variables que componen los sectores primario, secundario y terciario, pero una correlación baja entre estos sectores. Los patrones se repiten, mostrando una alta correlación entre variables del sector

financiero y de recreación, así como del sector inmobiliario, comunicación y actividades profesionales.

Por otro lado, se continúa evidenciando una alta correlación entre la variable de impuestos y las variables del sector secundario, mientras que no se observa correlación con las actividades del sector primario. Lo anterior indica que las actividades del sector secundario tienen mayor contribución al recaudo de impuestos en las regiones.

No obstante, el año 2020 representa un cambio en los patrones de correlación. En este año, se observa una correlación más robusta entre las variables del sector secundario, específicamente en los sectores de recreación, financiero, inmobiliario, comunicaciones y actividades profesionales, presentando una correlación del 90 % entre ellos. Por otro lado, la correlación entre el sector agrícola y minero aumenta al 70 %, mientras que la correlación entre el sector manufacturero y la variable impuesto se eleva al 90 %. En línea con los patrones anteriores, se mantiene la ausencia de correlación (0 %) entre los impuestos y las actividades primarias. Es importante resaltar que el 2020 estuvo directamente influenciados por la situación global de la pandemia de la Covid-19.

### 6.1.2. Indicadores de Pobreza

A partir de la Tabla **6-2** y la Figura **6-3** se resume el comportamiento de las variables que compone la estructura de datos del IPM en sus respectivos años. Se observa una notable variabilidad en la pobreza por departamentos, dado que en algunas variables se muestran rezagados con respecto a otros departamento. Esta variabilidad puede atribuirse a factores como oportunidades laborales, políticas públicas, entre otros, que influyen en las condiciones de pobreza.

En relación a las estadísticas descriptivas presentadas en la Tabla **6-2**, se observa que las variables que presentan mayores índices promedio durante los cuatro años evaluados son *Trabajo informal* (**Trab-Infor**), así como *Bajo logro educativo* (**B-Ledu**). Estas variables muestran índices promedios entre 80 y 81, 49 y 54 respectivamente a lo largo de los diferentes años. Por otro lado, se observa un bajo índice en la variable *Trabajo infantil* (**Trab-Infan**).

En la Figura **6-3** permite identificar las dimensiones específicas en las que la población de los departamentos están experimentando privaciones, lo que proporciona información valiosa para la formulación de políticas públicas más efectivas.

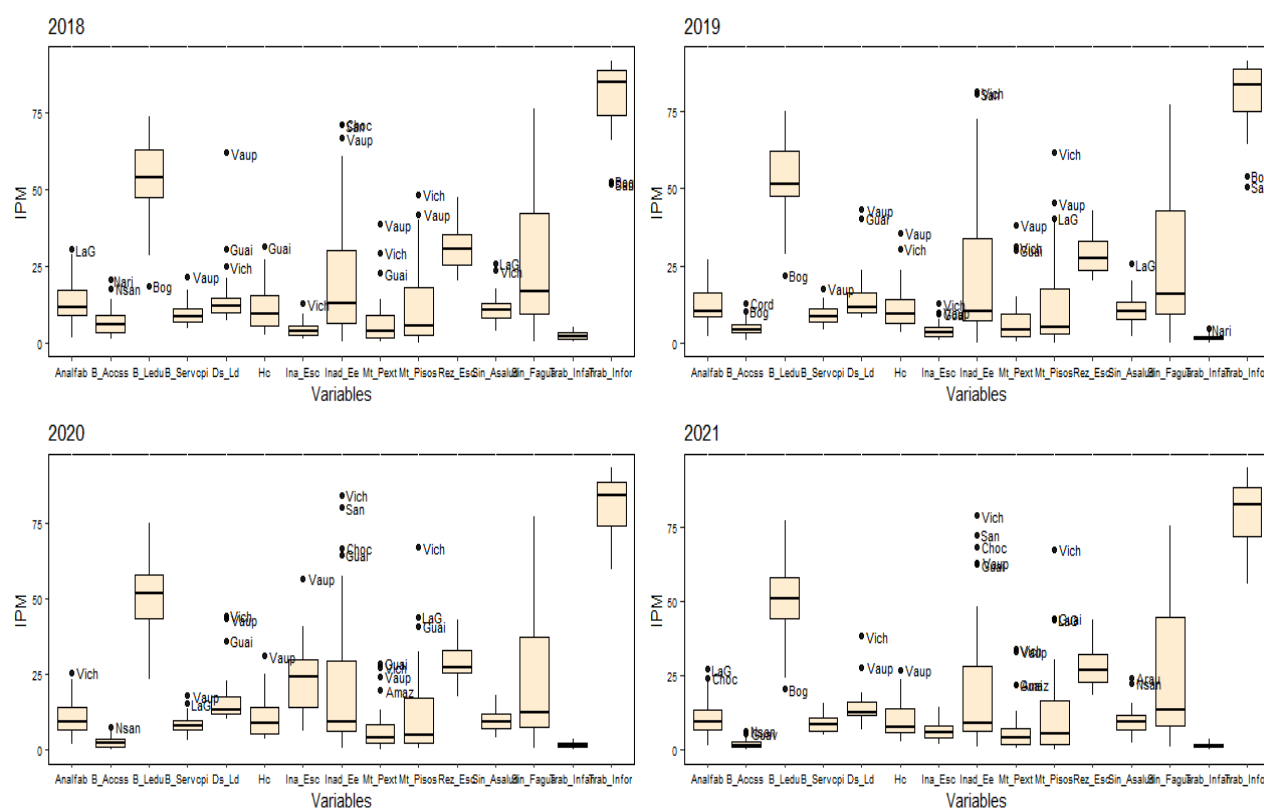
| Año  | Variable   | Mínimo | Máximo | Cuartil 1 | Cuartil 3 | Promedio    | Mediana | DS   |
|------|------------|--------|--------|-----------|-----------|-------------|---------|------|
| 2018 | Analfab    | 1.5    | 30.6   | 8.8       | 17.3      | 13.2        | 11.4    | 7.2  |
|      | B.Ledu     | 18.6   | 73.9   | 47.5      | 62.9      | <b>54.0</b> | 53.8    | 12.3 |
|      | B.Servcpi  | 4.8    | 21.7   | 6.7       | 11.1      | 9.5         | 8.7     | 3.8  |
|      | B.Accss    | 1.1    | 20.5   | 3.6       | 9.1       | 6.8         | 6.0     | 4.7  |
|      | Ds.Ld      | 7.1    | 62.1   | 9.7       | 14.7      | 14.4        | 12.0    | 9.9  |
|      | Hc         | 2.7    | 31.4   | 5.6       | 15.4      | 11.6        | 9.3     | 7.6  |
|      | Inad.Ee    | 0.4    | 71.3   | 6.3       | 30.3      | 22.5        | 12.7    | 21.8 |
|      | Ina.Esc    | 1.2    | 12.7   | 2.7       | 5.5       | 4.5         | 4.0     | 2.5  |
|      | Mt.Pext    | 0.2    | 38.7   | 1.8       | 8.8       | 7.1         | 4.0     | 8.7  |
|      | Mt.Pisos   | 0.0    | 48.5   | 2.7       | 18.0      | 11.6        | 5.4     | 13.7 |
|      | Rez.Esc    | 20.1   | 47.5   | 25.6      | 35.2      | <b>31.0</b> | 30.5    | 6.6  |
|      | Sin.Fagua  | 0.4    | 76.3   | 9.4       | 42.4      | <b>24.8</b> | 16.8    | 21.9 |
|      | Sin.Asalud | 4.0    | 26.0   | 8.3       | 12.8      | 11.2        | 10.6    | 4.1  |
|      | Trab.Infan | 0.4    | 5.2    | 1.4       | 3.2       | <b>2.2</b>  | 2.2     | 1.2  |
|      | Trab.Infor | 52.0   | 92.0   | 74.5      | 88.8      | <b>80.5</b> | 84.9    | 10.8 |
| 2019 | Analfab    | 2.0    | 26.9   | 8.4       | 16.1      | 12.3        | 10.4    | 6.7  |
|      | B.Ledu     | 21.9   | 75.0   | 47.7      | 62.3      | <b>52.8</b> | 51.3    | 11.9 |
|      | B.Servcpi  | 4.2    | 17.7   | 6.6       | 11.0      | 8.9         | 8.3     | 3.2  |
|      | B.Accss    | 0.9    | 12.8   | 3.2       | 5.9       | 4.8         | 4.3     | 2.7  |
|      | Ds.Ld      | 8.0    | 43.2   | 9.7       | 16.1      | 14.4        | 11.5    | 8.0  |
|      | Hc         | 3.4    | 35.6   | 6.5       | 14.1      | 11.7        | 9.3     | 7.8  |
|      | Inad.Ee    | 0.0    | 81.3   | 7.4       | 33.9      | 23.7        | 10.4    | 24.4 |
|      | Ina.Esc    | 0.8    | 12.6   | 2.2       | 5.0       | 4.1         | 3.5     | 2.6  |
|      | Mt.Pext    | 0.3    | 38.3   | 1.9       | 9.3       | 7.5         | 4.1     | 9.3  |
|      | Mt.Pisos   | 0.1    | 61.6   | 2.8       | 17.7      | 12.2        | 5.1     | 15.2 |
|      | Rez.Esc    | 20.1   | 42.7   | 23.7      | 33.0      | 28.7        | 27.6    | 6.3  |
|      | Sin.Fagua  | 0.1    | 77.0   | 9.5       | 42.9      | 26.0        | 15.7    | 23.8 |
|      | Sin.Asalud | 1.9    | 25.6   | 7.5       | 13.3      | 10.7        | 10.1    | 4.7  |
|      | Trab.Infan | 0.1    | 4.6    | 1.2       | 2.2       | <b>1.8</b>  | 1.7     | 0.9  |
|      | Trab.Infor | 50.7   | 91.3   | 75.1      | 88.8      | <b>80.3</b> | 83.7    | 10.6 |
| 2020 | Analfab    | 1.6    | 25.6   | 6.7       | 14.1      | 10.8        | 9.1     | 6.1  |
|      | B.Ledu     | 23.3   | 74.9   | 43.5      | 57.8      | <b>49.8</b> | 51.6    | 11.3 |
|      | B.Servcpi  | 3.2    | 17.8   | 6.5       | 9.6       | 8.5         | 7.8     | 3.2  |
|      | B.Accss    | 0.0    | 7.6    | 0.8       | 3.5       | 2.6         | 2.2     | 2.1  |
|      | Ds.Ld      | 10.2   | 44.3   | 12.0      | 17.6      | 16.6        | 13.2    | 8.6  |
|      | Hc         | 3.4    | 31.1   | 5.1       | 14.2      | 10.4        | 8.6     | 6.7  |
|      | Inad.Ee    | 0.5    | 84.3   | 6.0       | 29.2      | 22.2        | 9.3     | 23.9 |
|      | Ina.Esc    | 6.0    | 56.6   | 14.2      | 29.7      | 23.8        | 24.1    | 11.5 |
|      | Mt.Pext    | 0.2    | 28.6   | 2.2       | 8.5       | 7.0         | 3.9     | 7.7  |
|      | Mt.Pisos   | 0.3    | 66.9   | 2.0       | 17.1      | 12.1        | 5.0     | 15.5 |
|      | Rez.Esc    | 17.4   | 43.0   | 25.4      | 32.7      | 28.2        | 27.3    | 5.5  |
|      | Sin.Fagua  | 0.6    | 77.2   | 7.5       | 37.2      | 23.7        | 12.3    | 23.5 |
|      | Sin.Asalud | 3.8    | 18.0   | 7.2       | 11.8      | 9.7         | 9.4     | 3.7  |
|      | Trab.Infan | 0.1    | 3.4    | 0.8       | 2.2       | <b>1.4</b>  | 1.2     | 0.9  |
|      | Trab.Infor | 59.8   | 93.3   | 74.1      | 88.5      | <b>81.1</b> | 84.4    | 9.6  |
| 2021 | Analfab    | 1.4    | 27.2   | 6.9       | 13.6      | 10.9        | 9.5     | 6.3  |
|      | B.Ledu     | 20.6   | 77.3   | 44.3      | 58.4      | <b>49.6</b> | 51.1    | 11.7 |
|      | B.Servcpi  | 5.1    | 16.0   | 6.6       | 10.8      | 8.9         | 8.6     | 2.7  |
|      | B.Accss    | 0.2    | 6.5    | 0.9       | 2.8       | 2.0         | 1.7     | 1.5  |
|      | Ds.Ld      | 6.8    | 38.4   | 11.6      | 16.2      | 14.5        | 12.9    | 5.8  |
|      | Hc         | 2.9    | 27.0   | 6.0       | 13.9      | 10.2        | 7.9     | 6.0  |
|      | Inad.Ee    | 0.9    | 79.1   | 6.3       | 28.3      | 22.1        | 9.0     | 23.3 |
|      | Ina.Esc    | 2.1    | 14.3   | 4.4       | 8.4       | 6.6         | 6.0     | 3.1  |
|      | Mt.Pext    | 0.7    | 34.0   | 1.8       | 7.2       | 7.1         | 4.1     | 8.8  |
|      | Mt.Pisos   | 0.4    | 67.7   | 2.1       | 16.6      | 11.8        | 5.4     | 15.7 |
|      | Rez.Esc    | 18.7   | 44.1   | 22.8      | 32.3      | 27.9        | 26.8    | 6.5  |
|      | Sin.Fagua  | 0.9    | 75.4   | 8.4       | 44.8      | 24.9        | 13.5    | 23.8 |
|      | Sin.Asalud | 2.4    | 24.3   | 7.0       | 11.6      | 9.8         | 9.5     | 4.9  |
|      | Trab.Infan | 0.2    | 3.6    | 1.0       | 2.2       | <b>1.5</b>  | 1.3     | 0.9  |
|      | Trab.Infor | 55.8   | 95.1   | 72.0      | 88.7      | <b>81.0</b> | 82.7    | 10.5 |

**Tabla 6-2.:** Estadísticas descriptivas de los indicadores del IPM 2018 - 2021

En departamentos como La Guajira, su población presenta privaciones y problemas de analfabetismo y sin aseguramiento en salud. Por otro lado, en el caso del departamento de Vaupés, presentan mayores barreras a servicios para cuidado de la primera infancia, Desempleo de larga duración, Inadecuada eliminación de excretas, Material inadecuado de paredes exte-

riores y Material inadecuado de pisos.

El departamento de Guainía exhibe los índices más altos de Hacinamiento crítico. Esta tendencia se observa de manera destacada en los departamentos de la región de Orinoquía y la región del Amazonas a lo largo de los cuatro años evaluados, en comparación con los departamentos de otras regiones. Estas regiones presentan niveles significativamente más altos en la mayoría de las variables que hacen parte del IPM, lo que indica una situación preocupante que afecta gravemente a su población.



**Figura 6-3.:** Distribución de los indicadores del IPM por año

La Figura 6-4 muestra la estructura de correlación entre las variables que contribuyen al índice de Pobreza Multidimensional (IPM), con una comparación realizada a lo largo de los cuatro años evaluados (2018 al 2021).

Se evidencia una fuerte asociación lineal entre la mayoría de los pares de variables, en algunos casos esta asociación es entre el 40 % y 80 %. Lo que indica que si se presenta el incremento de alguna variable implica el aumento en otra que por consiguiente aumenta las condiciones de pobreza en la población.

Las variables de Inasistencia escolar y Material inadecuado de pisos muestran una correlación del 70 % y 80 % respectivamente con la variable de Hacinamiento crítico. Por otro lado, se observa una correlación no significativa entre variables como barreras a servicios para cuidado de la primera infancia, desempleo de larga duración, inadecuada eliminación de excretas, material inadecuado de paredes exteriores y la falta de acceso a una fuente de agua mejorada, con variables como barreras de acceso a servicios de salud, falta de aseguramiento en salud y trabajo infantil.

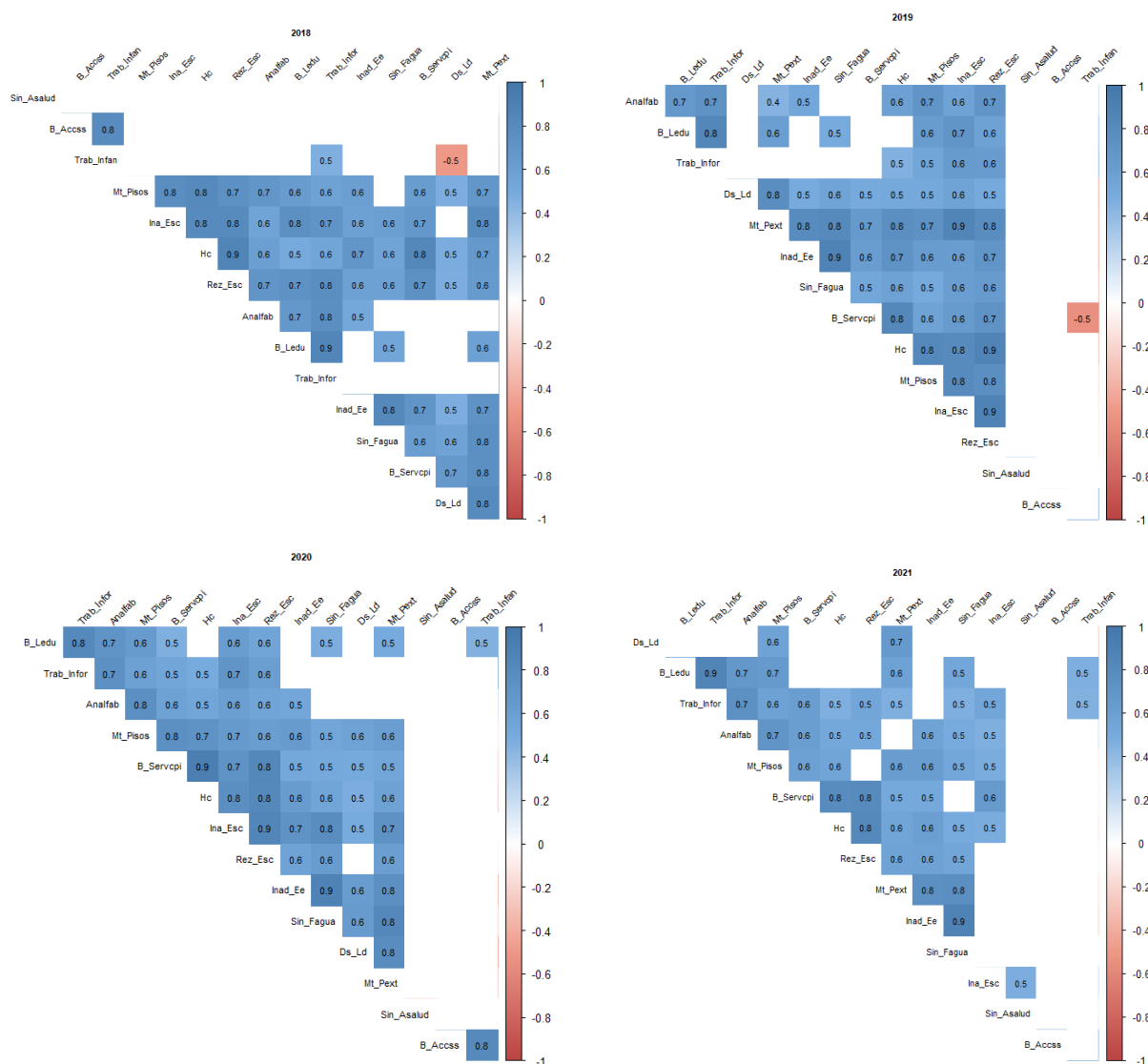
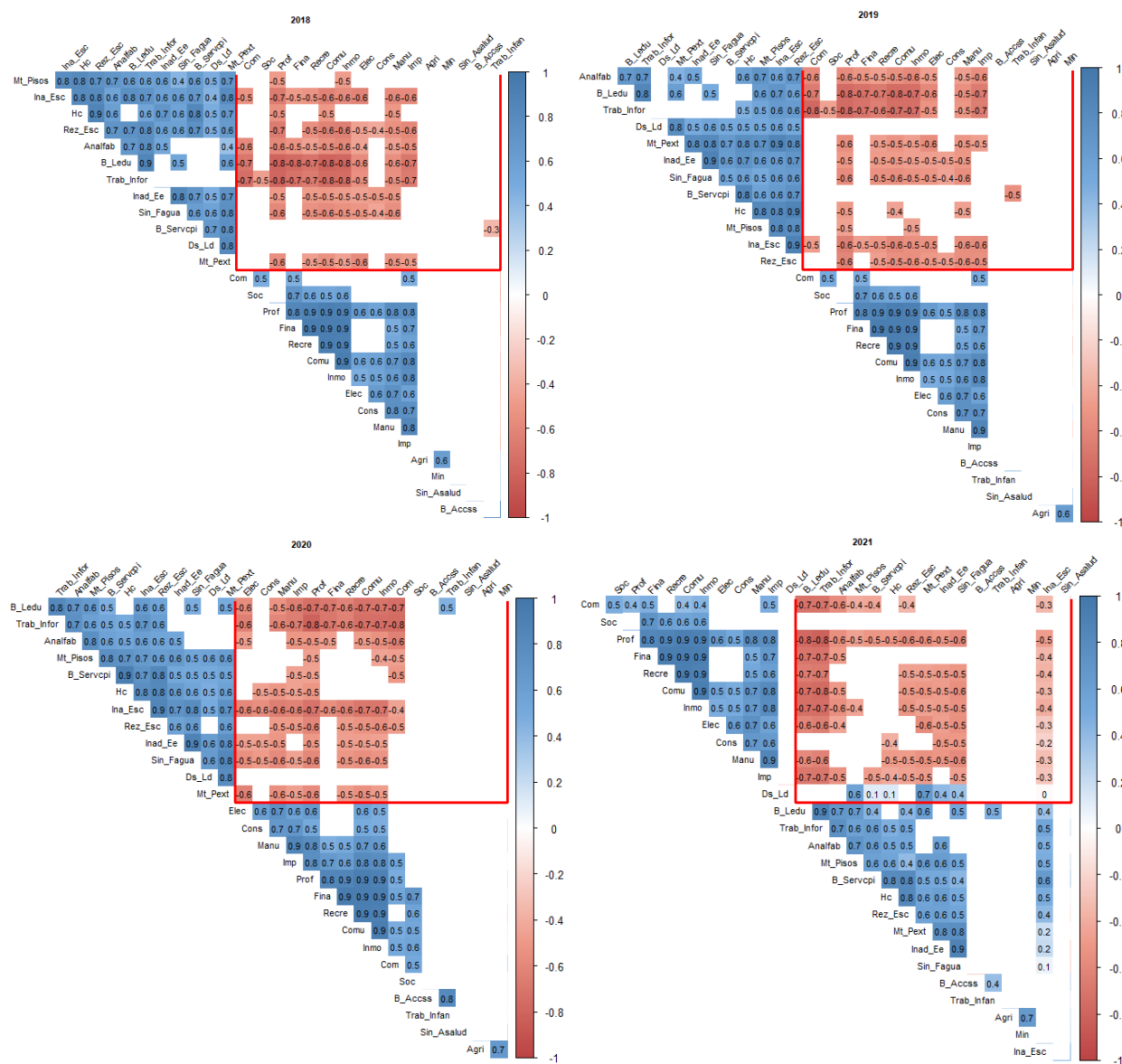


Figura 6-4.: Correlación entre las variables del IPM por año

La Figura 6-5 muestra la estructura de correlación entre las variables que conforman al índice de Pobreza Multidimensional (IPM) y el Producto Interno Bruto (PIB), con una comparación realizada a lo largo de los cuatro años evaluados (2018 al 2021). El área en el

cuadrante rojo representa la correlación que existe entre las dos estructuras de datos.



**Figura 6-5.:** Correlación entre las variables del IPM y PIB por año

A partir de la Figura 6-5 se evidencia una asociación negativa fuerte entre las variables de pobreza y de producción, lo que evidencia que cuando aumenta la producción de riqueza representa una disminución en los índices de pobreza.

Al observar la variable de Trabajo informal y Bajo logro educativo representa una disminución de sus índices con respecto a cada una de las variables del PIB, por lo que se podría decir que tienen una relación inversa. Cada año, se presenta un patrón similar de correlación, sin embargo, para el año 2020, se evidencia un cambio significativo de correlación entre la

variable de inasistencia escolar con las variables del PIB, ya que presentan correlaciones altas entre el 70 % y 60 %. Lo anterior da indicios para pensar que la productividad económica de un país contribuye a disminuir los índices de pobreza y sugiere que una forma de contribuir a la reducción de la pobreza consiste en fortalecer las capacidades productivas de las regiones.

En resumen, este análisis proporciona una visión completa del comportamiento de las dos estructuras de datos de tres vías, tanto de forma individual como en conjunto, revelando su interrelación. Además, permite identificar las fortalezas y debilidades de ciertos departamentos en términos de sus actividades económicas y producción, así como las disparidades y condiciones de pobreza que enfrentan.

Se observó una relación negativa moderada entre los dos conjuntos de datos. Por lo tanto, el objetivo de este proyecto es realizar evaluación metodológica para analizar simultáneamente las dos estructuras de datos en múltiples años o modos. Para lograr esto, se emplearán técnicas como el Análisis Factorial Múltiple (AFM), Mínimos Cuadrados Parciales (PLS) y Mínimos Cuadrados Parciales Multiway (N-PLS). El propósito es examinar la asociación entre estas dos estructuras de datos de tres vías y obtener una comprensión más completa de su relación.

Durante el análisis de la base del PIB, se ha observado que Bogotá es uno de los departamentos que más contribuye a la producción de Colombia en diversas actividades. La Figura 6-1 muestra que Bogotá sobresale en 9 de las 13 actividades económicas en términos del PIB. Debido a esta destacada participación, se considera a Bogotá como un caso suplementario en el análisis multivariante, ya que su inclusión tendría un impacto significativo en la inercia y dificultaría una visualización óptima de las dos estructuras de datos.

## **6.2. Análisis bajo una estructura de datos en dos vías**

### **6.2.1. Análisis Factorial Múltiple (AFM)**

Se realiza el análisis de la tabla completa de datos mediante el Análisis Factorial Múltiple, teniendo en cuenta que se tienen dos tablas globales, las cuales se despliegan según las Figuras 5-5, 5-6 y 5-4.

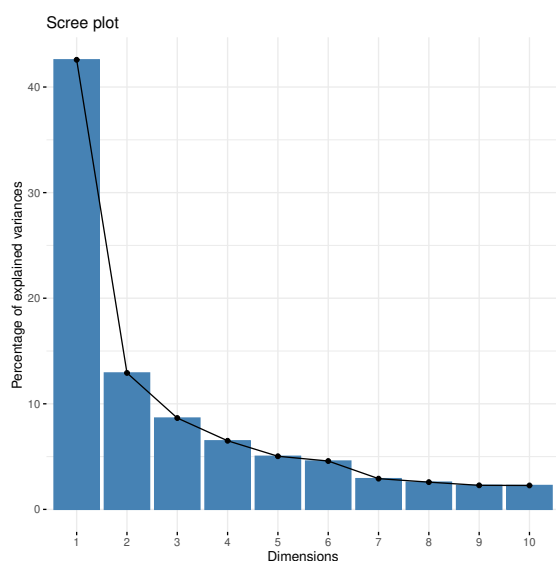
#### **AFM para la alternativa de despliegue vertical**

En primer lugar, se lleva a cabo un Análisis Factorial Múltiple (AFM) con un despliegue del modo tiempo en un sentido vertical, lo que implica que los individuos se incrementarán por filas. En este caso, se consideran un total de 128 individuos que corresponden a 32 departamentos (excluyendo a Bogotá) durante un período de 4 años (2018 a 2021). Los grupos

de variables se dividen en dos: el primer grupo comprende las 12 variables del conjunto de datos del PIB, mientras que el segundo grupo incluye las variables del Índice de Pobreza Multidimensional (IPM).

## Análisis de factores

En primer lugar, se procede a determinar el número de componentes a analizar mediante el análisis de los valores propios, los cuales representan la cantidad de variación explicada por cada componente.



| Componente | Valor propio | % de Varianza | % de Varianza Acumulada |
|------------|--------------|---------------|-------------------------|
| comp 1     | 1.75         | 42.58         | 42.58                   |
| comp 2     | 0.53         | 12.92         | 55.51                   |
| comp 3     | 0.36         | 8.65          | 64.16                   |
| comp 4     | 0.27         | 6.50          | 70.66                   |
| comp 5     | 0.21         | 5.03          | 75.69                   |
| comp 6     | 0.19         | 4.58          | 80.27                   |
| comp 7     | 0.12         | 2.91          | 83.19                   |
| comp 8     | 0.11         | 2.58          | 85.77                   |
| comp 9     | 0.09         | 2.28          | 88.05                   |
| comp 10    | 0.09         | 2.27          | 90.32                   |

**Tabla 6-3.:** Valores propios del AFM

**Figura 6-6.:** Gráfico de sedimentación de valores propios del AFM

En la Figura 6-6 y la Tabla 6-3, se puede observar que los dos primeros ejes factoriales explican el 55.51 % de la variabilidad total de los datos. El primer eje factorial es especialmente relevante, ya que representa el 42.52 % de la variabilidad total, mientras que el segundo eje explica el 12.92 %. Como se puede apreciar, el primer plano factorial captura una variabilidad significativa, por lo tanto, el análisis de los datos se centrará en el estudio de los departamentos, variables y grupos de variables en relación con este primer plano.

## Ponderación de los grupos de variables

El Análisis Factorial Múltiple (AFM) tiene como objetivo equilibrar la influencia de los grupos de variables asignando un peso a cada variable. Es importante que todas las variables dentro de un mismo grupo tengan el mismo peso para preservar la estructura interna de dicho grupo [Escofier y Pagès \(1992b\)](#). El peso asignado a las variables de un grupo se calcula como

el inverso del primer valor propio obtenido a través del Análisis de Componentes Principales (ACP) aplicado a ese grupo en particular. En la Tabla **6-4**, se muestran los primeros valores propios obtenidos de los ACP realizados individualmente en cada grupo, sus inversos y el número de variables activas en cada grupo.

| Grupo | Cantidad de variables | $\lambda_1^j$ | $1/\lambda_1^j$ |
|-------|-----------------------|---------------|-----------------|
| PIB   | 13                    | 6.86          | 0.14            |
| IPM   | 15                    | 6.76          | 0.15            |

**Tabla 6-4.:** Valores propios y ponderaciones de los ACP separados por grupo

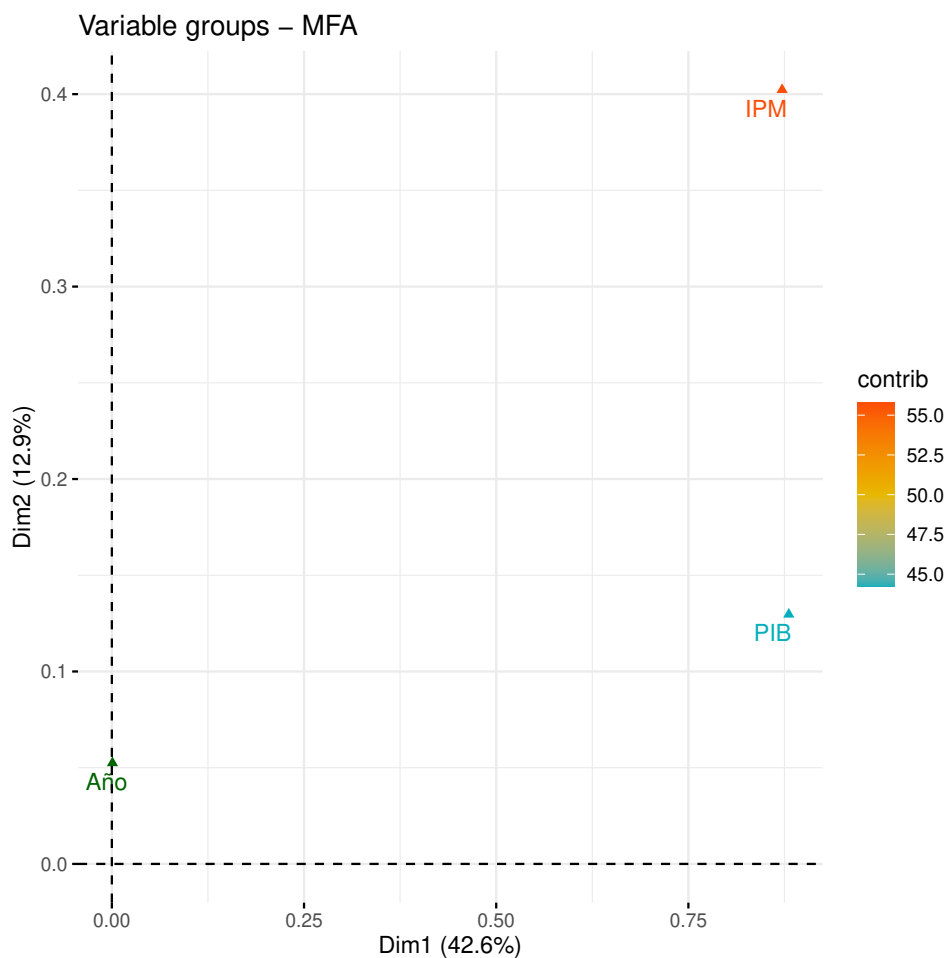
En la Tabla **6-4** se muestra una distribución de peso similar para ambas estructuras de datos. En contraste, la Tabla **6-5** revela la descomposición de la inercia en las dos primeras dimensiones. Por lo tanto, en la primera dimensión, las inercias de los dos grupos de datos (PIB e IPM) están cercanas a uno (1), lo que indica una fuerte relación entre esta dimensión y estos dos grupos. En cuanto a la segunda dimensión, aunque su inercia es de 0.40, presenta una relación más significativa con el grupo de variables del PIB. Además, es importante mencionar que las contribuciones en la primera dimensión son similares, lo que sugiere una fuerte relación y una contribución similar del PIB y el IPM en la caracterización de los departamentos.

| Grupo | Dim 1 | Contrib | Cos2 | Dim 2 | Contrib | Cos2 |
|-------|-------|---------|------|-------|---------|------|
| PIB   | 0.87  | 49.75   | 0.60 | 0.40  | 75.62   | 0.12 |
| IPM   | 0.88  | 50.24   | 0.67 | 0.13  | 24.37   | 0.01 |

**Tabla 6-5.:** Contribuciones del conjunto de datos del PIB y el IPM a la construcción de los ejes

### Tipología de los grupos

La Figura **6-7** ilustra la correlación entre grupos y dimensiones. Ambos grupos de variables, se encuentran similarmente representados en la dimensión 1, ya que su coordenada es muy similar. Por otro lado, para la dimensión 2, el grupo de variables del IPM está mejor representado, pues es el que tiene la coordenada y contribución más alta para eje factorial, por tanto, las diferencias en condiciones de pobreza se evidencian en la ubicación de los departamentos respecto al segundo eje factorial.



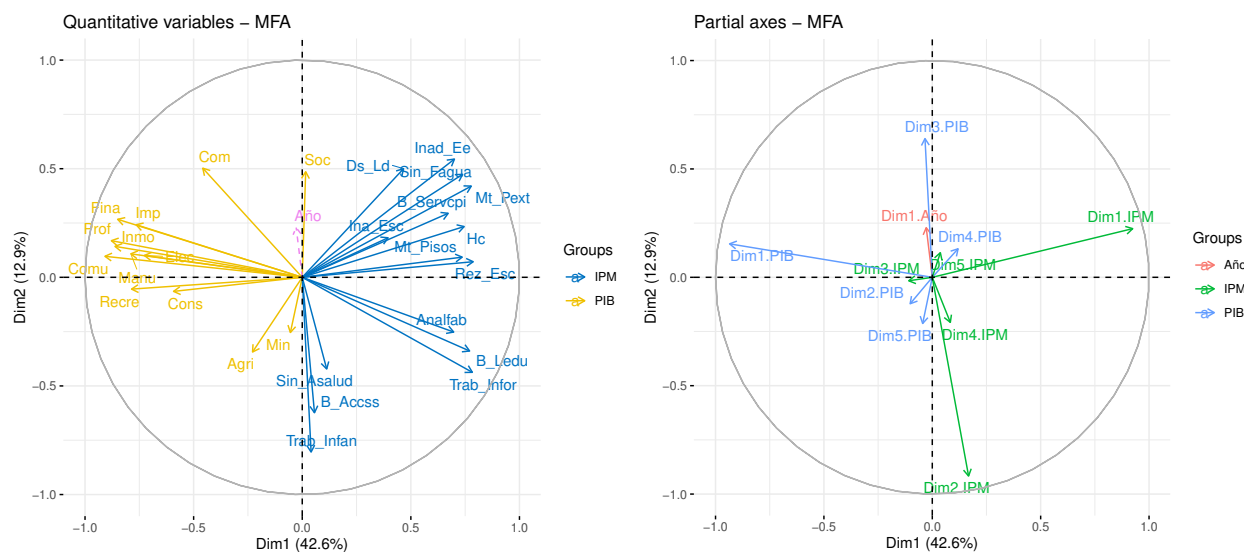
**Figura 6-7.:** Representación de los grupos en el análisis global

La dimensionalidad de cada grupo y su relación se mide a través de los coeficientes Lg y RV, respectivamente. En la Tabla **6-6**, se puede observar que los coeficientes Lg tienen valores superiores a 0.50, lo que indica una relación moderada entre los grupos. No es ni cercana a cero ni muy alta. Por otro lado, al examinar los coeficientes RV, se observa que el coeficiente entre los grupos de variables del PIB e IPM es 0.53, lo cual sugiere una relación moderada entre las variables que pertenecen a estos grupos.

| Grupo | Lg   |      |      | RV   |      |      |
|-------|------|------|------|------|------|------|
|       | PIB  | IPM  | AFM  | PIB  | IPM  | AFM  |
| PIB   | 1.15 | 0.64 | 1.02 | 1.00 | 0.53 | 0.87 |
| IPM   | 0.64 | 1.25 | 1.08 | 0.53 | 1.00 | 0.88 |
| AFM   | 1.02 | 1.08 | 1.20 | 0.87 | 0.88 | 1.00 |

**Tabla 6-6.:** Coeficientes Lg y RV

### Representación de las variables



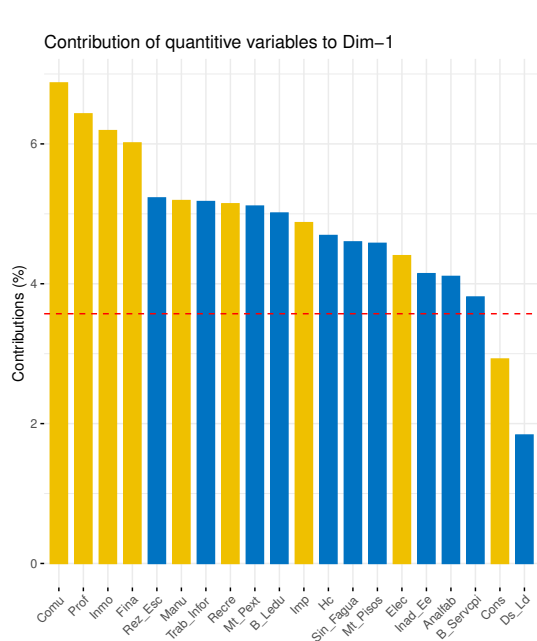
**Figura 6-8.:** Circulo de correlación de las variables en el análisis global

La Figura 6-8 presenta la correlación entre las variables de cada grupo y las dimensiones. Con respecto a la dimensión uno mide por su parte positiva las condiciones de pobreza, ya que en este se concentran las variables del IPM, mientras que por el lado negativo se aglomeran las variables del PIB. Esta relación inversa, sugiere que a mayores indicadores de productividad, las condiciones de pobreza tenderían a disminuir. Asimismo, se observa que departamentos muy pobres y poco productivos se ubican a la derecha de la dimension 1, mientras que los más productivos y menos pobres están a la izquierda

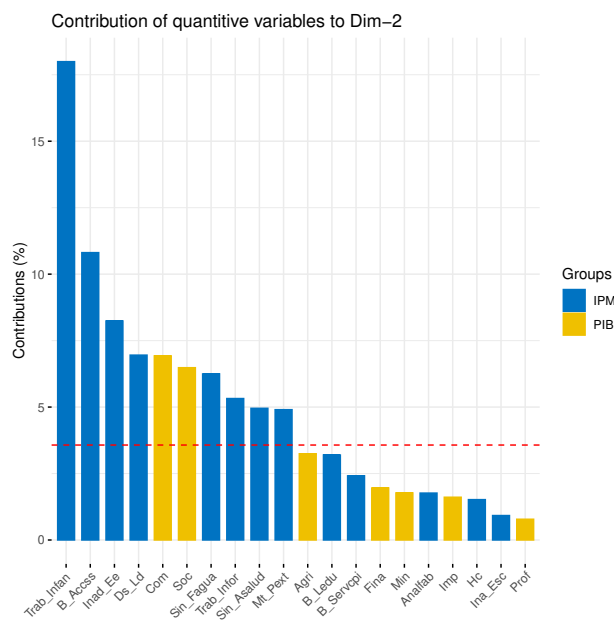
Por otro lado, se observa una agrupación de variables dentro de cada grupo, evidenciando por ejemplo para el PIB, la alta correlación entre variables del sector primario (Minería y Agricultura), las variables del sector secundario (Manufactura, Construcción, Manufactura, Electricidad) y variables del tercer sector (Financiero, Inmobiliario, entre otras), cabe destacar que la variable que representa la sociedad, se encuentra alejada de las otras variables y además, presenta una alta relación inversa entre variables de pobreza como el acceso a la salud y trabajo infantil, lo que sugiere que una menor inversión en administración pública y defensa; planes de seguridad social de afiliación obligatoria; educación; actividades de atención de la salud humana y de servicios sociales, aumenta condiciones de trabajo infantil y el acceso a la salud.

En la dimensión dos, en su dirección positiva se agrupan variables de pobreza que indican

el acceso a la salud, trabajo informal, bajo logro educativo y trabajo infantil, mientras que en la parte negativa se relacionan variables del PIB con respecto a la minería y agricultura. A partir de esto, se observa que la minería y la agricultura no presentan una correlación fuerte con las variables de condiciones de pobreza, lo que sugiere que los departamentos que generan un impacto económico para el país no tienen las condiciones e inversiones sociales suficientes para mejorar la calidad de vida de su población.



**Figura 6-9.:** Contribución de las variables a la inercia a la primera dimensión



**Figura 6-10.:** Contribución de las variables a la inercia a la segunda dimensión

En las Figuras 6-9 y 6-10, se detalla la contribución de las variables a las dimensiones, para la primera dimensión, las variables que más aportan a su construcción corresponde a las variables del conjunto de datos del PIB, estas son profesión, comunicaciones, Inmobiliaria y el sector financiero, perteneciendo estas al sector de servicios. Del grupo de variables del IPM las variables que más aportan a esta dimensión son Rezago Escolar y Trabajo informal. En cuanto la segunda dimensión, se observa que la contribución de las variables del grupo de IPM realizan un aporte más significativo, ya que las tres priemras son trabajo infantil, Barreras de acceso a servicios de salud e inadecuado eliminación de excretas.

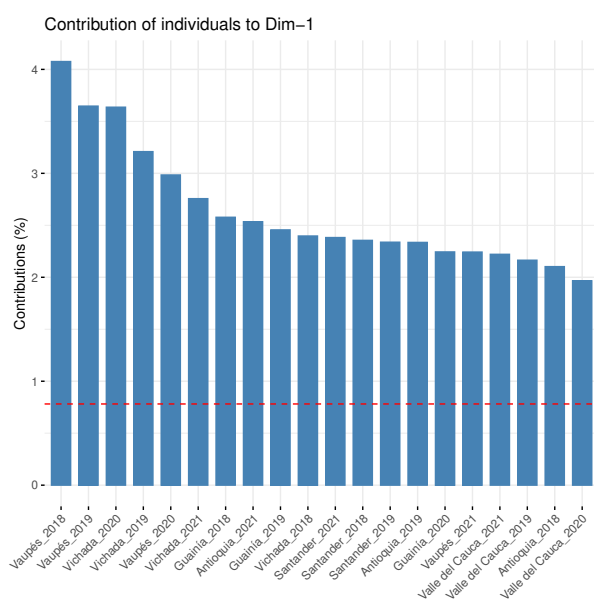
### Representación de los individuos

La interpretación de los departamentos en los diferentes años medidos (individuos) se presenta en el plano factorial, en la Figura 6-11 y sus respectivas contribuciones en las Figuras

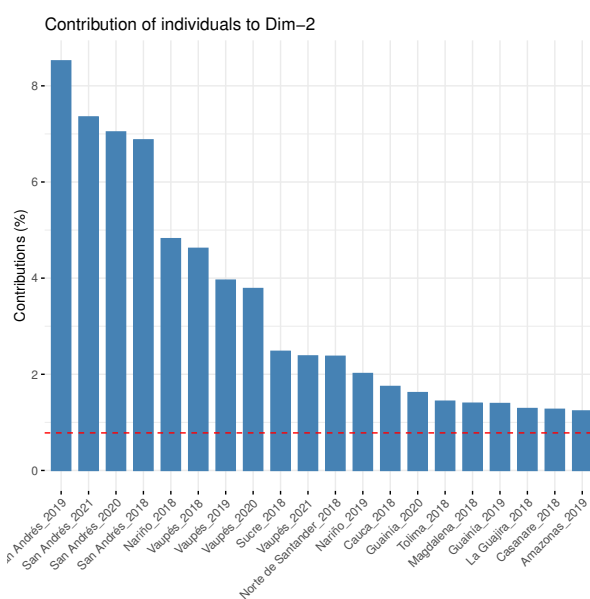
Con respecto al II y III cuadrante, los cuales se caracterizan por estar asociados a variables de productividad, los departamentos de Santander, Valle del Cauca, Atlántico, Bolívar, Risaralda, se caracterizan por tener mayores índices de asociación con sectores profesionales, financieros, comercio e inversión social. Asimismo, se puede observar la alta asociación del departamento de San Andrés con la productividad en comercio e inversión social. Los departamentos de Tolima, Huila, Quindío se caracterizan por su productividad en Agricultura mientras que Casanare y Meta con actividades de minería. Lo que sugiere que son departamentos más productivos con mayor aporte al PIB y a su vez, presentan menos índices de

pobreza.

Asimismo, se evidencia la trayectoria de los departamentos en los cuatro años medidos, en los cuales se observa los efectos de la pandemia de la Covid-19, ya que los años 2018 y 2019 presentaban un crecimiento, pero en todos los departamentos se identifica un patrón para el año 2020, ya que fue marcado por la pandemia, la productividad y las condiciones de pobreza para este año se vieron grandemente afectadas, donde para la productividad se presentó una reducción y en cuanto a niveles de pobreza se presentó un notable aumento.



**Figura 6-12.:** Contribución de los territorios a la primera dimensión



**Figura 6-13.:** Contribución de los territorios a la segunda dimensión

De acuerdo a las Figuras 6-12 y 6-13, los departamentos de Vaupés para los años 2018, 2019 y 2020 y Vichada para los años 2020, 2019 y 2021 presentan mayor contribución a la conformación de la primera dimensión, seguido por el departamento de Guainía 2018 y Antioquia 2021. Particularmente estos departamentos, Vaupés y Vichada, se caracterizan por ser departamentos con altos índices de pobreza.

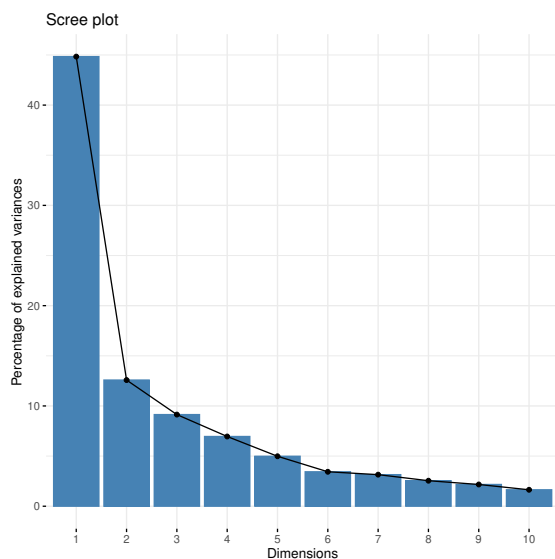
Por otro lado, la contribución para la segunda dimensión la realiza principalmente el departamento de San Andrés en sus cuatro años evaluados, seguido por Nariño en el 2018 y Vaupés en los años 2018 y 2019.

### AFM para la alternativa de despliegue horizontal

En esta segunda alternativa, se aplica la metodología AFM a partir de un despliegue horizontal de los años, por tanto, se tendrán 32 individuos que corresponde a los departamentos (excluyendo a Bogotá), 112 variables, que corresponden a las variables de cada conjunto de datos del PIB e IPM y 8 grupos que corresponden a los años (2018 a 2021) para cada conjunto de datos.

### Análisis de factores

Igualmente, se realiza el análisis de los factores a partir de los valores propios que representan la variación explicada.



| Componente | Valor propio | % de Varianza | % de Varianza Acumulada |
|------------|--------------|---------------|-------------------------|
| comp 1     | 7.01         | 44.84         | 44.84                   |
| comp 2     | 1.97         | 12.58         | 57.42                   |
| comp 3     | 1.43         | 9.14          | 66.56                   |
| comp 4     | 1.09         | 6.95          | 73.51                   |
| comp 5     | 0.78         | 4.98          | 78.49                   |
| comp 6     | 0.54         | 3.44          | 81.93                   |
| comp 7     | 0.49         | 3.15          | 85.08                   |
| comp 8     | 0.40         | 2.55          | 87.62                   |
| comp 9     | 0.34         | 2.17          | 89.79                   |
| comp 10    | 0.26         | 1.64          | 91.43                   |

**Tabla 6-7.:** Valores propios del AFM alternativa de despliegue horizontal

**Figura 6-14.:** Gráfico de sedimentación de valores propios del AFM alternativa de despliegue horizontal

En la Figura 6-14 y Tabla 6-7, se puede observar que los dos primeros ejes factoriales explican el 57.42 % de la variabilidad total de los datos. El primer eje factorial es especialmente relevante, ya que representa el 44.84 % de la variabilidad total, mientras que el segundo eje explica el 12.58 %. Como se puede apreciar, el primer plano factorial captura una variabilidad significativa, por lo tanto, el análisis de los datos se centrará en el estudio de los departamentos, variables y grupos de variables en relación con este primer plano.

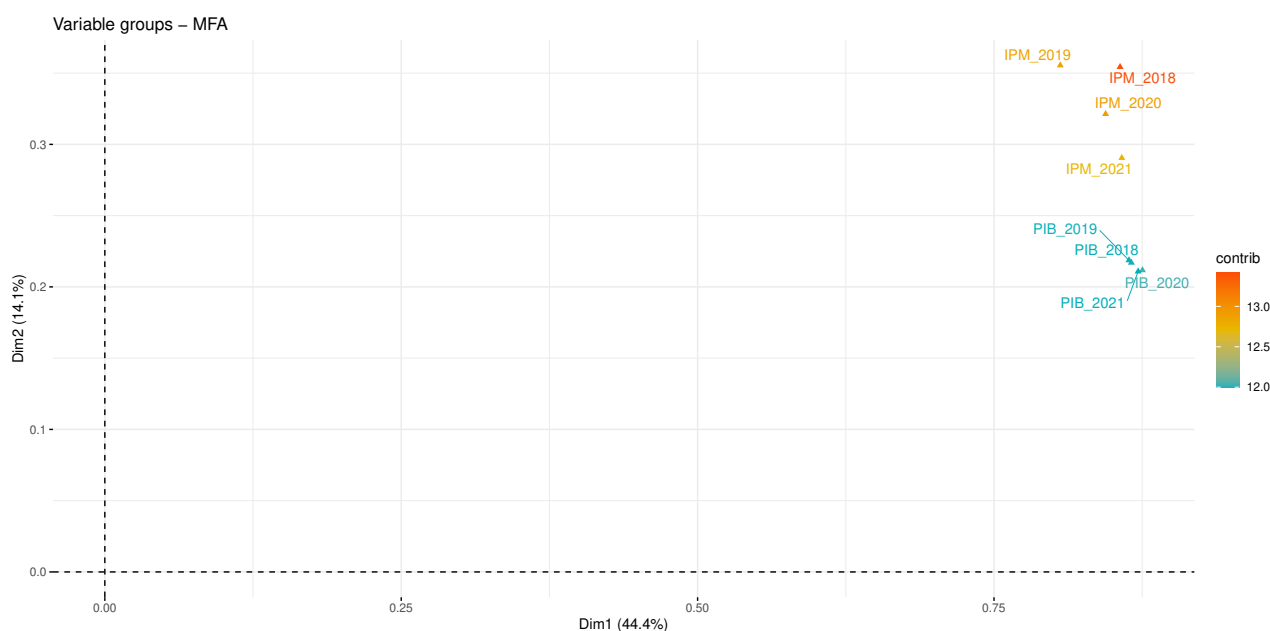
### Ponderación de los grupos de variables

| Grupo    | No. de variables | $\lambda_1^j$ | Coordenadas |       | Contribuciones |       | Cosenos |       |
|----------|------------------|---------------|-------------|-------|----------------|-------|---------|-------|
|          |                  |               | Dim 1       | Dim 2 | Dim 1          | Dim 2 | Dim 1   | Dim 2 |
| PIB_2018 | 13               | 6.93          | 0.88        | 0.15  | 12.61          | 7.83  | 0.68    | 0.02  |
| PIB_2019 | 13               | 6.93          | 0.88        | 0.15  | 12.62          | 7.67  | 0.68    | 0.02  |
| PIB_2020 | 13               | 6.96          | 0.89        | 0.14  | 12.75          | 6.88  | 0.70    | 0.02  |
| PIB_2021 | 13               | 6.93          | 0.89        | 0.14  | 12.66          | 6.96  | 0.69    | 0.02  |
| IPM_2018 | 15               | 7.74          | 0.87        | 0.36  | 12.45          | 18.15 | 0.63    | 0.11  |
| IPM_2019 | 15               | 7.71          | 0.83        | 0.37  | 11.87          | 18.82 | 0.58    | 0.12  |
| IPM_2020 | 15               | 7.55          | 0.88        | 0.35  | 12.53          | 17.78 | 0.64    | 0.10  |
| IPM_2021 | 15               | 6.59          | 0.88        | 0.31  | 12.51          | 15.91 | 0.60    | 0.08  |

**Tabla 6-8.:** Valores propios, Coordenadas, contribuciones y ponderaciones del conjunto de datos del PIB y el IPM a la construcción de los ejes

En la Tabla 6-8 se muestra una distribución de peso similar para ambas estructuras de datos. En contraste, revela la descomposición de la inercia en las dos primeras dimensiones. Similar a la alternativa uno, las contribuciones sugieren que existe una fuerte relación y una contribución similar del PIB y el IPM en la caracterización de los departamentos.

### Tipología de los grupos



**Figura 6-15.:** Representación de los grupos en el análisis global

La Figura 6-15 ambos grupos de variables, se encuentran similarmente representados en la

dimensión 1, ya que su coordenada es muy similar. Por otro lado, para la dimensión 2, el grupo de variables del IPM está mejor representado, pues es el que tiene la coordenada y contribución más alta para el eje factorial, además, es importante mencionar que en esta figura se evidencia el recorrido de los conjuntos de datos en los cuatro años evaluados. El conjunto de variables del PIB en el año 2020, presenta mayor contribución para la construcción del primer eje factorial y para el segundo eje factorial el conjunto de variables para el IPM en el 2019 presenta mayor contribución para la construcción de este eje.

En la Tabla **6-9** se puede apreciar que estos coeficientes son superiores a 0.50, lo que indica una relación moderada entre los grupos. No es ni cercana a cero ni muy alta. Además, la relación es constante entre los diferentes años y los conjuntos de datos.

| Grupos   | PIB_2018 | PIB_2019 | PIB_2020 | PIB_2021 | IPM_2018 | IPM_2019 | IPM_2020 | IPM_2021 | MFA  |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|------|
| PIB_2018 | 1.15     | 1.15     | 1.14     | 1.14     | 0.66     | 0.61     | 0.67     | 0.67     | 1.03 |
| PIB_2019 | 1.15     | 1.15     | 1.15     | 1.14     | 0.66     | 0.61     | 0.67     | 0.67     | 1.03 |
| PIB_2020 | 1.14     | 1.15     | 1.15     | 1.14     | 0.67     | 0.62     | 0.68     | 0.68     | 1.03 |
| PIB_2021 | 1.14     | 1.14     | 1.14     | 1.15     | 0.66     | 0.61     | 0.68     | 0.68     | 1.03 |
| IPM_2018 | 0.66     | 0.66     | 0.67     | 0.66     | 1.21     | 1.14     | 1.14     | 1.14     | 1.04 |
| IPM_2019 | 0.61     | 0.61     | 0.62     | 0.61     | 1.14     | 1.19     | 1.14     | 1.13     | 1.00 |
| IPM_2020 | 0.67     | 0.67     | 0.68     | 0.68     | 1.14     | 1.14     | 1.20     | 1.17     | 1.05 |
| IPM_2021 | 0.67     | 0.67     | 0.68     | 0.68     | 1.14     | 1.13     | 1.17     | 1.28     | 1.06 |
| MFA      | 1.03     | 1.03     | 1.03     | 1.03     | 1.04     | 1.00     | 1.05     | 1.06     | 1.18 |

**Tabla 6-9.:** Coeficientes Lg

Por otro lado, los coeficientes RV consignados en la Tabla **6-10**, se observa que los coeficientes entre los grupos de variables del PIB e IPM son superiores a 0.48, lo cual sugiere una relación moderada entre las variables que pertenecen a estos grupos.

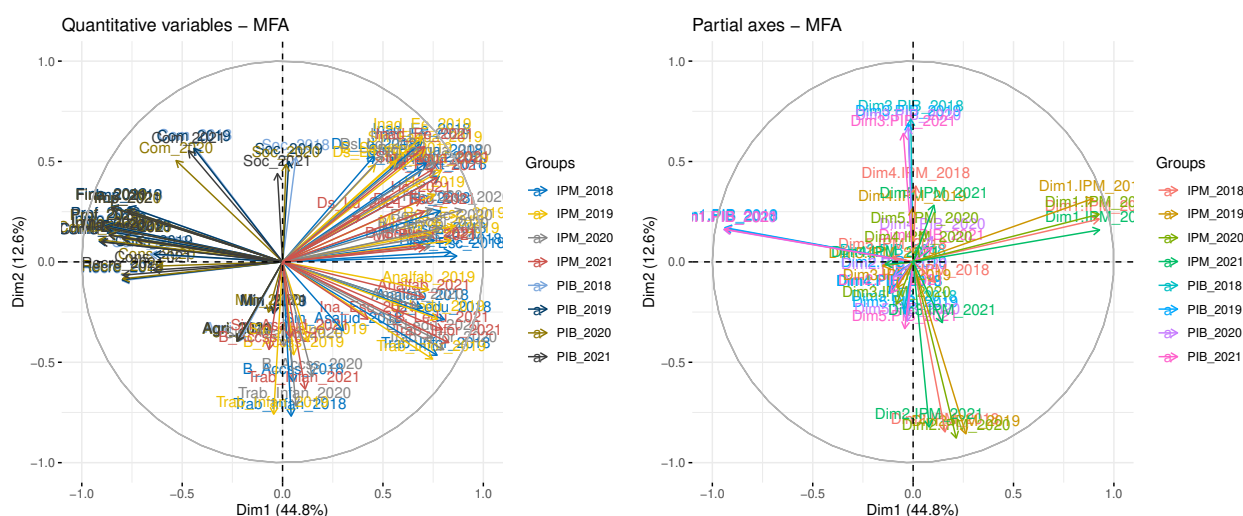
| Grupos   | PIB_2018 | PIB_2019 | PIB_2020 | PIB_2021 | IPM_2018 | IPM_2019 | IPM_2020 | IPM_2021 | MFA  |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|------|
| PIB_2018 | 1.00     | 1.00     | 1.00     | 0.99     | 0.56     | 0.52     | 0.57     | 0.56     | 0.88 |
| PIB_2019 | 1.00     | 1.00     | 1.00     | 1.00     | 0.56     | 0.52     | 0.57     | 0.56     | 0.88 |
| PIB_2020 | 1.00     | 1.00     | 1.00     | 1.00     | 0.57     | 0.53     | 0.58     | 0.56     | 0.89 |
| PIB_2021 | 0.99     | 1.00     | 1.00     | 1.00     | 0.56     | 0.52     | 0.57     | 0.56     | 0.88 |
| IPM_2018 | 0.56     | 0.56     | 0.57     | 0.56     | 1.00     | 0.95     | 0.94     | 0.91     | 0.87 |
| IPM_2019 | 0.52     | 0.52     | 0.53     | 0.52     | 0.95     | 1.00     | 0.95     | 0.92     | 0.85 |
| IPM_2020 | 0.57     | 0.57     | 0.58     | 0.57     | 0.94     | 0.95     | 1.00     | 0.95     | 0.88 |
| IPM_2021 | 0.56     | 0.56     | 0.56     | 0.56     | 0.91     | 0.92     | 0.95     | 1.00     | 0.86 |
| MFA      | 0.88     | 0.88     | 0.89     | 0.88     | 0.87     | 0.85     | 0.88     | 0.86     | 1.00 |

**Tabla 6-10.:** Coeficientes RV

### Representación de las variables

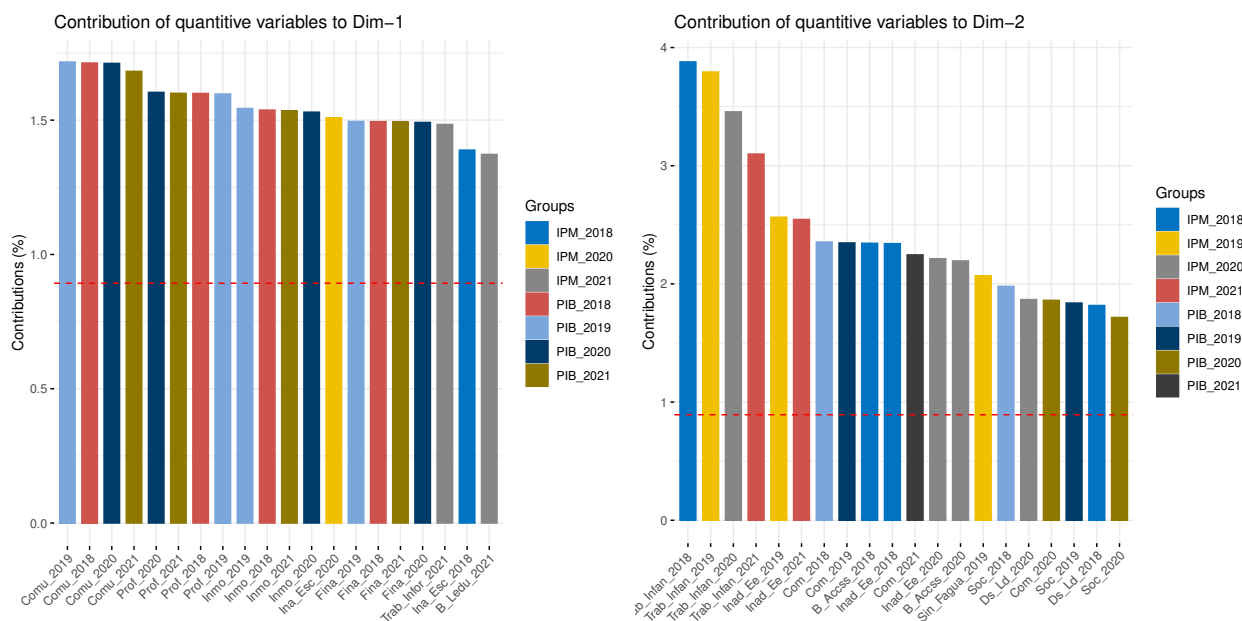
La Figura **6-16** se presenta al igual que en la alternativa uno, la dimensión uno mide las condiciones de pobreza, ya que al lado derecho se concentran las variables del IPM en sus respectivos años, mientras que por el lado negativo se aglomeran las variables del PIB. Esta

relación inversa, sugiere que a mayores indicadores de productividad, las condiciones de pobreza tenderían a disminuir o al revés. Para la construcción del primer eje factorial, se observa mayor contribución del conjunto de variables del IPM en los diferentes años evaluados y el PIB 2019 y 2020.



**Figura 6-16.:** Circulo de correlación de las variables en el análisis global

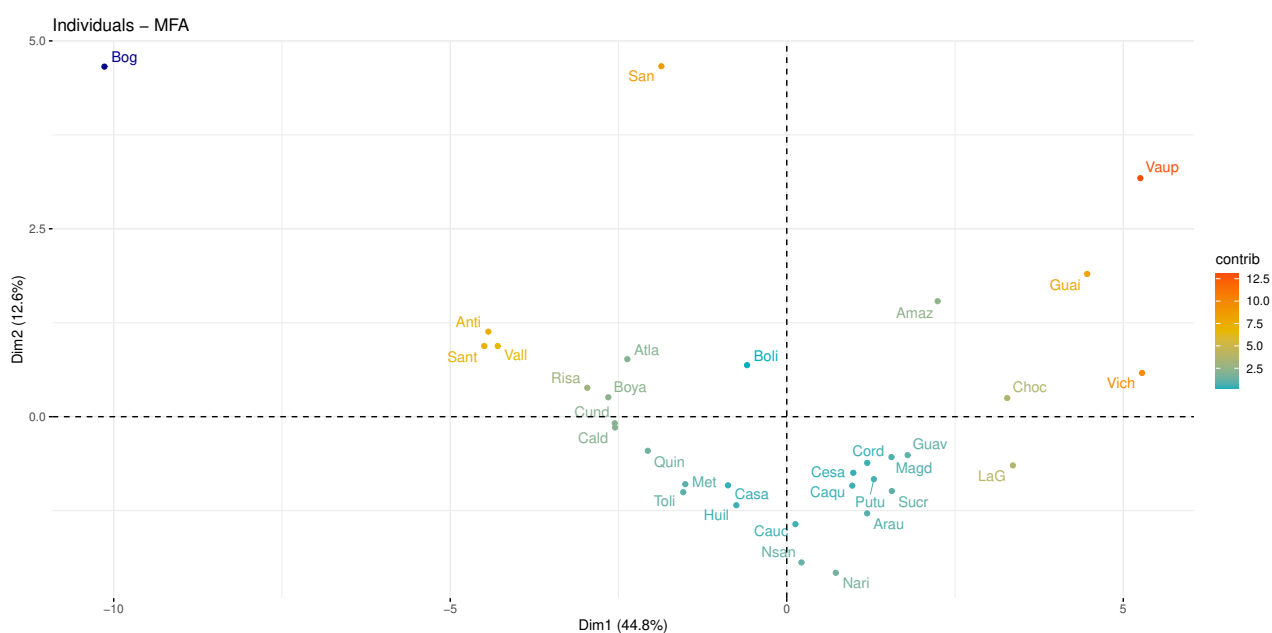
Lo anterior se evidencia igualmente en las Figuras 6-17 y 6-18, donde la contribución de las variables a las dimensiones, para la primera dimensión, las variables que más aportan a su construcción corresponde a las variables del conjunto de datos del PIB, estas son profesión, comunicaciones e Inmobiliaria, perteneciendo estas al sector de servicios. Del grupo de variables del IPM las variables que más aportan a esta dimensión son Trabajo informal y bajo logro educativo. En cuanto la segunda dimensión, se observa que la contribución de las variables del grupo de IPM son más notables, ya que las tres primeras son trabajo infantil, sin embargo, a diferencia de la primera alternativa, la variable de Agricultura perteneciente al conjunto de datos del PIB en sus diferentes años resulta significativa para la construcción del segundo eje factorial.



**Figura 6-17.:** Contribución de las variables a la inercia del primer eje factorial

**Figura 6-18.:** Contribución de las variables a la inercia del segundo eje factorial

### Representación de los individuos



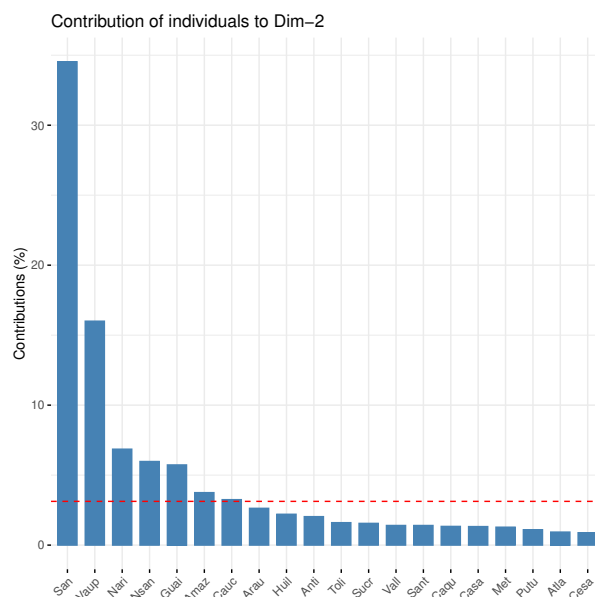
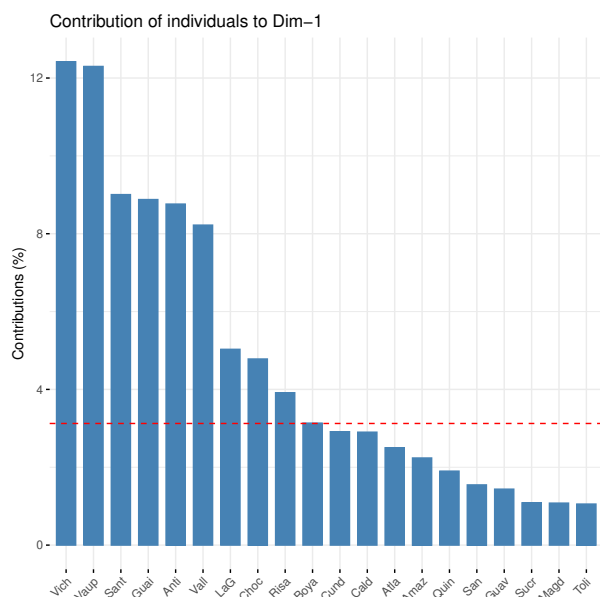
**Figura 6-19.:** Representación de los territorios en el análisis global

En la Figura 6-19, se puede observar la agrupación de los departamentos en el cuadrante I y

el cuadrante IV. En el cuadrante I, se encuentran aquellos departamentos asociados a variables de pobreza. Departamentos como Vaupés, Vichada, Guainía, Amazonas y La Guajira presentan condiciones de pobreza que están relacionadas con aspectos como las condiciones de los hogares, el hacinamiento, el rezago escolar, el desempleo y el bajo acceso a servicios públicos. Por otro lado, en el cuadrante IV se encuentran departamentos como Nariño, Norte de Santander, Cesar, Cauca, entre otros. Estos departamentos se caracterizan por condiciones de pobreza relacionadas con el trabajo infantil, la falta de acceso a la salud, el trabajo informal y el bajo logro educativo.

En relación al II y III cuadrante, que se caracterizan por estar asociados a variables de productividad, se observa que los departamentos de Santander, Valle del Cauca, Atlántico, Bolívar y Risaralda presentan mayores índices de asociación con sectores profesionales, financieros, comerciales e inversión social. Además, se destaca la alta asociación del departamento de San Andrés con la productividad en el sector comercial y la inversión social. Por otro lado, los departamentos de Tolima, Huila y Quindío se caracterizan por su productividad en el sector agrícola, mientras que Casanare y Meta muestran actividades de minería como principales impulsores de su productividad.

Es importante resaltar que, a diferencia de la primer alternativa de análisis, en esta segunda alternativa no se permite observar la trayectoria de los departamentos.



**Figura 6-20.:** Contribución de los territorios a la primer componente

**Figura 6-21.:** Contribución de los territorios a la segunda componente

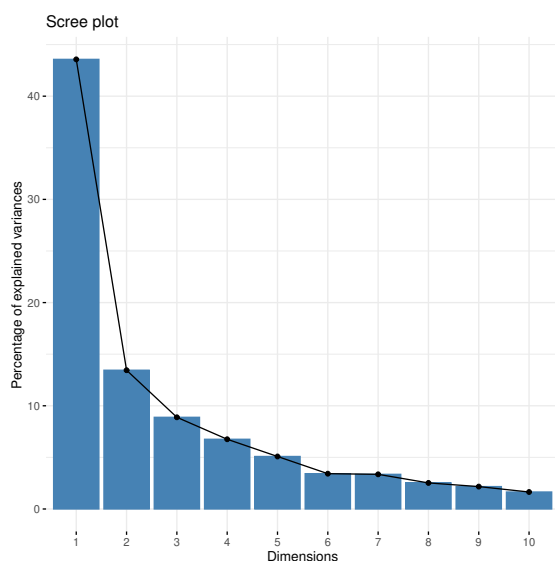
Finalmente, en las Figuras 6-20 y 6-21, el departamento de Vichada, presenta mayor contri-

bución a la conformación de la primera dimensión, seguido por los departamentos de Vaupés y Santander. Particularmente estos departamentos, se caracterizan por ser departamentos con altos índices de pobreza. Por otro lado, las contribuciones más altas para la segunda dimensión las aporta los departamentos de San Andrés, Vaupés y Nariño.

### AFM para la alternativa de despliegue combinada

En la tercera alternativa, se utiliza un Análisis Factorial Múltiple (AFM) con un despliegue horizontal de las variables. En este caso, se considerarán 32 individuos, que corresponden a los departamentos (excluyendo a Bogotá), y 28 grupos. Cada grupo incluirá los datos de los cuatro años para cada variable del PIB y del IPM, fusionados en un solo conjunto de datos. Se realiza el mismo análisis que en las anteriores alternativas.

### Análisis de factores



| Componente | Valor propio | % de Varianza | % de Varianza Acumulada |
|------------|--------------|---------------|-------------------------|
| comp 1     | 13.10        | 43.57         | 43.57                   |
| comp 2     | 4.04         | 13.45         | 57.01                   |
| comp 3     | 2.67         | 8.88          | 65.90                   |
| comp 4     | 2.03         | 6.76          | 72.66                   |
| comp 5     | 1.53         | 5.09          | 77.74                   |
| comp 6     | 1.03         | 3.42          | 81.17                   |
| comp 7     | 1.01         | 3.36          | 84.53                   |
| comp 8     | 0.76         | 2.53          | 87.05                   |
| comp 9     | 0.65         | 2.17          | 89.22                   |
| comp 10    | 0.50         | 1.65          | 90.87                   |

**Tabla 6-11.:** Valores propios del AFM alternativa de despliegue combinada

**Figura 6-22.:** Gráfico de sedimentación de valores propios del AFM alternativa de despliegue combinada

En la Figura 6-22 y la Tabla 6-11, se puede observar que los dos primeros ejes factoriales explican el 57.01 % de la variabilidad total de los datos. El primer eje factorial es especialmente relevante, ya que representa el 43.57 % de la variabilidad total, mientras que el segundo eje explica el 13.45 %. Como se puede apreciar, el primer plano factorial captura una variabilidad significativa, por lo tanto, el análisis de los datos se centrará en el estudio

de los departamentos, variables y grupos de variables en relación con este primer plano.

### Ponderación de los grupos de variables

| Grupo      | Coordenadas |       | Contribuciones |       | Cosenos |       |
|------------|-------------|-------|----------------|-------|---------|-------|
|            | Dim 1       | Dim 2 | Dim 1          | Dim 2 | Dim 1   | Dim 2 |
| Agri       | 0.05        | 0.11  | 0.41           | 2.64  | 0.00    | 0.01  |
| Min        | 0.00        | 0.04  | 0.02           | 1.06  | 0.00    | 0.00  |
| Manu       | 0.61        | 0.03  | 4.64           | 0.72  | 0.37    | 0.00  |
| Elec       | 0.52        | 0.03  | 3.96           | 0.66  | 0.27    | 0.00  |
| Cons       | 0.40        | 0.00  | 3.09           | 0.08  | 0.16    | 0.00  |
| Com        | 0.22        | 0.27  | 1.71           | 6.63  | 0.05    | 0.07  |
| Comu       | 0.81        | 0.02  | 6.21           | 0.46  | 0.66    | 0.00  |
| Fina       | 0.71        | 0.08  | 5.42           | 1.97  | 0.50    | 0.01  |
| Inmo       | 0.73        | 0.03  | 5.58           | 0.68  | 0.53    | 0.00  |
| Prof       | 0.76        | 0.04  | 5.81           | 1.03  | 0.58    | 0.00  |
| Soc        | 0.00        | 0.18  | 0.01           | 4.52  | 0.00    | 0.03  |
| Recre      | 0.63        | 0.01  | 4.79           | 0.15  | 0.39    | 0.00  |
| Imp        | 0.57        | 0.09  | 4.36           | 2.14  | 0.33    | 0.01  |
| Analfab    | 0.51        | 0.04  | 3.92           | 1.04  | 0.26    | 0.00  |
| B_ledu     | 0.62        | 0.11  | 4.74           | 2.62  | 0.39    | 0.01  |
| B_servcpi  | 0.56        | 0.15  | 4.27           | 3.72  | 0.31    | 0.02  |
| B_accss    | 0.01        | 0.56  | 0.09           | 13.74 | 0.00    | 0.27  |
| Ds_ld      | 0.25        | 0.27  | 1.94           | 6.75  | 0.06    | 0.07  |
| Hc         | 0.61        | 0.09  | 4.68           | 2.15  | 0.38    | 0.01  |
| Inad_ee    | 0.50        | 0.32  | 3.83           | 7.85  | 0.25    | 0.10  |
| Ina_esc    | 0.84        | 0.05  | 6.42           | 1.22  | 0.66    | 0.00  |
| Mt_pext    | 0.64        | 0.20  | 4.89           | 4.93  | 0.41    | 0.04  |
| Mt_pisos   | 0.57        | 0.01  | 4.36           | 0.36  | 0.33    | 0.00  |
| Rez_esc    | 0.71        | 0.03  | 5.44           | 0.81  | 0.51    | 0.00  |
| Sin_fagua  | 0.56        | 0.22  | 4.27           | 5.54  | 0.31    | 0.05  |
| Sin_asalud | 0.03        | 0.14  | 0.25           | 3.35  | 0.00    | 0.02  |
| Trab_infan | 0.01        | 0.73  | 0.04           | 17.98 | 0.00    | 0.52  |
| Trab_infor | 0.63        | 0.21  | 4.84           | 5.19  | 0.40    | 0.04  |

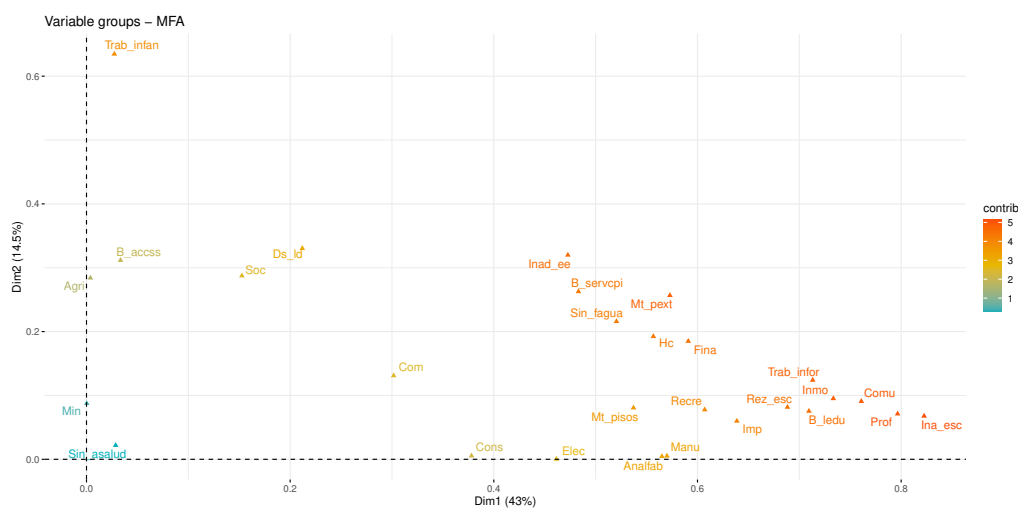
**Tabla 6-12.:** Valores propios, ponderaciones, y contribuciones del conjunto de datos del PIB y el IPM a la construcción de los ejes

La Tabla **6-12** revela la descomposición de la inercia en las dos primeras dimensiones. Por lo tanto, en la primera dimensión, las inercias de las variables de los dos grupos de datos

(PIB e IPM) están cercanas a uno (1), lo que indica una fuerte relación entre esta dimensión y algunas variables, sin embargo, se observan variables que representan una relación baja e incluso cero con esta dimensión, como es el caso de la variable de Agricultura y Minas, las cuales presentan coordenadas de 0.0, además, es importante resaltar que la primera dimensión tiene una relación moderada con las variables de ambos conjuntos de datos.

En cuanto a la segunda dimensión, se observa una relación más fuerte con variables del conjunto de datos del IPM, aunque la mayor inercia fue de 0.63, presentando una relación más significativa con este conjunto de datos. Es importante mencionar que las contribuciones en la primera dimensión son similares, lo que sugiere una fuerte relación y una contribución similar del PIB y el IPM en la caracterización de los departamentos, Sin embargo, la variable Trabajo Infantil perteneciente al grupo de IPM con una contribución a la segunda dimensión del 14.55.

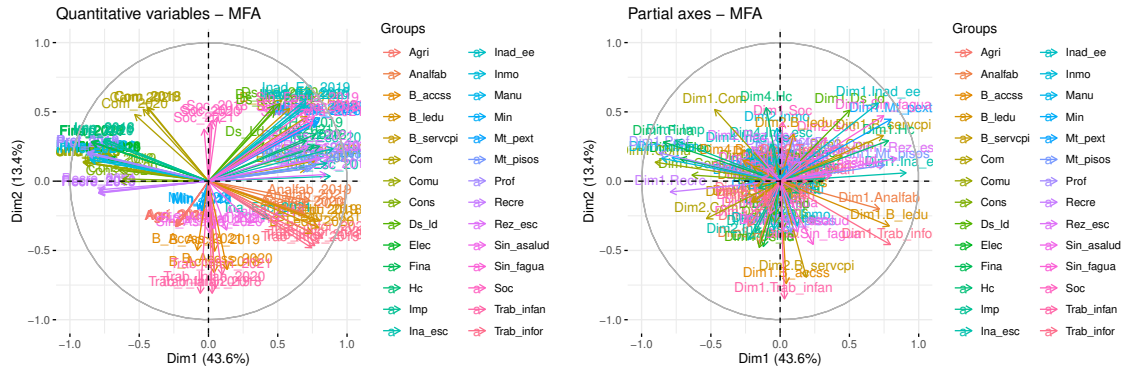
### Tipología de los grupos



**Figura 6-23.:** Representación de los grupos en el análisis global

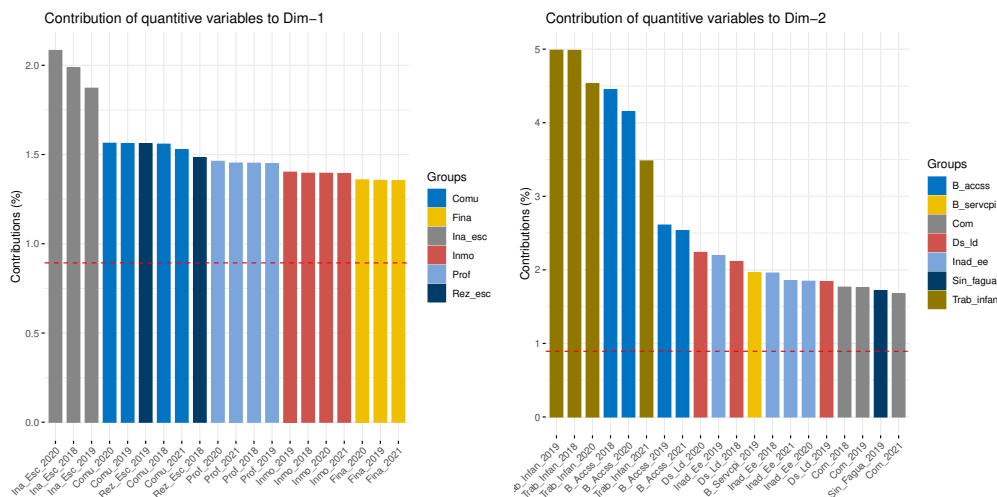
La Figura 6-23 ilustra la correlación entre las variables y las dimensiones, complementa el análisis de la Tabla 6-8 y de las alternativas de desglose vertical y despliegue horizontal. Se observa que las variables que pertenecen a ambos conjuntos de datos presentan mayor contribución con la primera dimensión. Por otro lado, la variable Trabajo Infantil tiene la mayor contribución para la segunda dimensión.

### Representación de las variables



**Figura 6-24.:** Circulo de correlación de las variables en el análisis global alternativa de despliegue combinada

En la Figura 6-24 donde se presenta la correlación entre las variables y las dimensiones. Al igual que la alternativa 1 y 2, se observa que en la primera dimensión se mide las condiciones de pobreza, ya que en este se concentran las variables del IPM, mientras que por el lado negativo se aglomeran las variables del PIB. Esta relación inversa, sugiere que cuando se disminuye la inversión o producción en alguna variable disminuye o aumenta las condiciones de pobreza.



**Figura 6-25.:** Contribución de las variables a la inercia del primer eje factorial

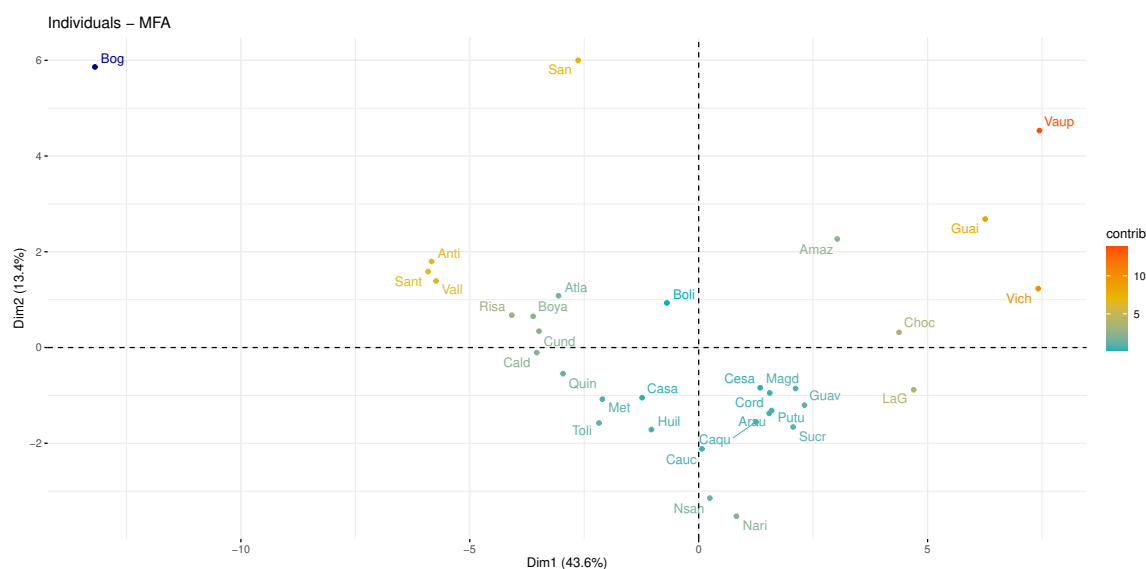
**Figura 6-26.:** Contribución de las variables a la inercia del segundo eje factorial

En las Figuras **6-25** y **6-26**, se detalla la contribución de las variables a la construcción de los ejes factoriales, para el primer eje factorial o primera dimensión, las variables que más aportan a su construcción corresponde a variables de ambos conjuntos de datos, primeramente, Inasistencia Escolar en los años 2020, 2018 y 2019, seguida por la actividad de Comunicación en los años 2020 y 2019.

En cuanto a la segunda dimensión, se observa que predomina las variables del conjunto de datos del IPM, aportando mayor contribución la variable de Trabajo Infantil de los años 2019, 2018 y 2020.

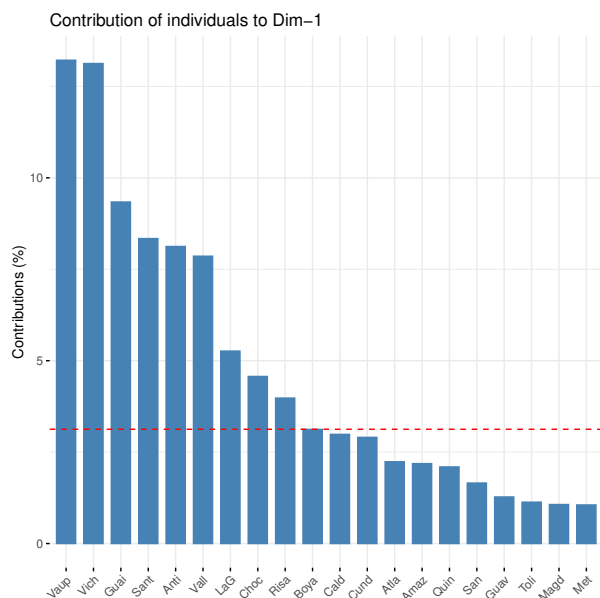
### Representación de los individuos

En la Figura **6-27**, la caracterización de los departamentos sigue la tendencia de las alternativas anteriores, ya que se agrupan en el primer cuadrante los departamentos que presentan altos índices de pobreza, en el segundo y tercer cuadrante, se caracterizan los departamentos por estar asociados a variables de productividad y en el cuarto cuadrante se encuentran departamentos como Nariño, Norte de Santander, Cesar, Cauca, entre otros.

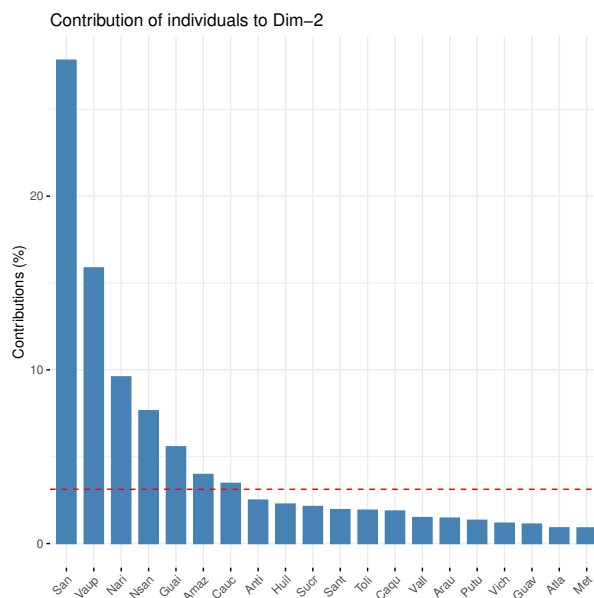


**Figura 6-27.:** Representación de los territorios en el análisis global

Es importante resaltar que, a diferencia de la primer alternativa de análisis, en esta tercer alternativa no es posible observar la trayectoria de los departamentos en los diferentes años medidos.



**Figura 6-28.:** Contribución de los territorios a la primer componente



**Figura 6-29.:** Contribución de los territorios a la segunda componente

Finalmente, en las Figuras 6-28 y 6-29, la contribución a la construcción de los ejes sigue la tendencia de la segunda alternativa. Donde Vaupés presenta mayor contribución a la conformación de la primera y la segunda dimensión, seguido por el departamento de Vichada.

### 6.2.2. Análisis comparativo de las alternativas a través de AFM

El Análisis Factorial Múltiple (AFM), al igual que las técnicas multivariantes, es una técnica cuyo objetivo es maximizar la varianza contenida en la totalidad de la tabla de datos, permitiendo entender y resumir la estructura de un conjunto de datos complejo. En el AFM, las variables se agrupan en bloques que representan diferentes aspectos de los datos, los cuales se normalizan para asegurar que cada bloque tenga igual peso en el análisis. A partir de lo anterior, en esta técnica las variables se agrupan en dos grandes tablas, las cuales corresponden a las ramas de actividad del PIB y los indicadores del IPM. La tabla de datos en cada una de las alternativas es un gran tabla compuesta por los dos conjuntos de datos, por tanto, el objetivo de esta técnica fue analizar la varianza máxima de los dos conjuntos de datos.

En general, se observó que la técnica de AFM proporcionó valores de inercia explicada en el rango del 55 % al 58 % en las tres alternativas propuestas. En lo que respecta a los resultados de LG y RV, se observó que para las tres alternativas, existe una relación moderada entre las dos estructuras de datos.

La alternativa de despliegue vertical en los dos primeros ejes factoriales obtuvo el 55.51 %

de la variabilidad total de los datos. La alternativa de despliegue horizontal proporcionó una variabilidad total del 57.42 % y la alternativa por despliegue combinado proporcionó el 57.01 %. Lo que se sugiere a partir de esta comparación es que el método de despliegue horizontal y el despliegue combinado tienden a capturar una mayor variabilidad total de los datos en comparación con el despliegue vertical, aunque las diferencias en los valores no sean extremadamente significativas.

Dado que los valores de variabilidad total de los datos no presenten diferencias significativas en las alternativas de despliegue, indica que todas las tres alternativas son razonablemente efectivas para capturar la variabilidad en los dos conjuntos de datos. Sin embargo, el ligero aumento en la variabilidad explicada por las alternativas de despliegue horizontal y combinado sugiere que estas son ligeramente mejores para capturar la estructura de los datos.

La alternativa de despliegue horizontal puede ser especialmente útil cuando hay muchas variables que aportan información adicional relevante, mientras que el despliegue combinado puede ofrecer un equilibrio entre los dos enfoques al capturar tanto más observaciones como más variables y su tercera condición. Esto puede ser ventajoso en situaciones donde se necesita una comprensión más completa y detallada de la relación entre las variables y el aumento de la tercera condición de los bloques de datos.

### **Alternativa de despliegue vertical**

- **Ventajas:** Esta alternativa permite analizar la trayectoria de los departamentos y su comportamiento a lo largo de la tercera condición, los años. Igualmente, permite observar la varianza acumulada en la totalidad de la tabla y en cada grupo que se conforma, lo que permite identificar patrones comunes en los departamentos según sus condiciones de productividad y pobreza, a la vez, proporciona una visión más completa de las condiciones de cada departamento a medida que se aumenta un año o en este contexto, se aumenta el tercer modo de los conjuntos de datos.
- **Desventajas:** No permite ver la trayectoria de las variables. El aumento de observaciones, indica un tamaño de información más amplia, lo cual puede introducir desafíos en la gestión de grandes volúmenes de datos y el riesgo de ruido en los datos.

### **Alternativa de despliegue horizontal y combinado**

- **Ventajas:** Estas alternativas se toman de manera similar, en ambas, permite analizar la trayectoria de las variables y su comportamiento a lo largo de los años medidos, lo que implica que puede aumentar la riqueza y variedad de los datos, mejorando

la capacidad de análisis y la robustez de los resultados. Permite identificar patrones comunes y relaciones entre las variables de los diferentes bloques.

- Desventajas: No permite observar la trayectoria de los individuos. El aumento en el número de variables y observaciones puede resultar en una alta dimensionalidad, lo cual puede requerir más recursos computacionales generando una sobrecarga computacional. Con más variables, hay un mayor riesgo de sobreajuste, donde se puede capturar el ruido en lugar de los patrones subyacentes. La inclusión de múltiples bloques de datos puede resultar en factores que son más difíciles de interpretar, especialmente si las variables de los diferentes bloques no están claramente relacionadas.

La elección entre las diferentes alternativas de despliegue debe basarse en la naturaleza de los datos disponibles, su complejidad y los objetivos específicos de la investigación. Si el objetivo es analizar la trayectoria de los individuos en la tercera condición de los bloques, se recomienda emplear la alternativa de despliegue vertical. En contraste, si el objetivo es examinar la trayectoria de las variables en la tercera condición de los bloques, las alternativas de despliegue horizontal y combinado resultan ser más adecuadas. Estas estrategias permiten capturar diferentes dimensiones de la variabilidad en los datos y proporcionan una visión más detallada y precisa en función de los objetivos del estudio.

## 6.3. Asociación entre dos estructuras de datos en dos vías (Despliegue)

### 6.3.1. Mínimos Cuadrados Parciales (PLS)

Se aplicará la metodología del algoritmo de PLS2, el cual combina las metodologías de Análisis de Componentes Principales (ACP) y el Análisis de Regresión Múltiple.

El método PLS2 se emplea cuando se dispone de una amplia cantidad de predictores, como ocurre en este estudio en particular, donde se cuenta con un total de 28 variables. En este enfoque, es necesario seleccionar un conjunto de datos que represente la matriz de variables predictoras, se ha optado por tomar el conjunto de variables del Producto Interno Bruto (PIB) como una de las matrices, mientras que la otra estará compuesta por las variables de respuesta, específicamente las relacionadas con el Índice de Pobreza Multidimensional (IPM).

Se realiza la combinación lineal  $u_1$  de las columnas (variables) de la matriz estandarizada del Índice de Pobreza Multidimensional (IPM), denominada  $Y_0$ , y la combinación lineal  $t_1$  de las columnas de la matriz estandarizada del Producto Interno Bruto (PIB), denominada  $X_0$ . El objetivo es maximizar la covarianza entre  $u_1$  y  $t_1$ , estas dos nuevas variables presentan la mayor correlación posible y resumen la información contenida en  $X_0$  y  $Y_0$ . Este proceso

permite determinar la primera componente.

Luego, para determinar la segunda componente, se sustituyen las matrices  $X_0$  y  $Y_0$  por las nuevas matrices residuales  $X_1$  y  $Y_1$ . Se procede a obtener dos nuevas componentes,  $t_2$  y  $u_2$ , que maximizan la covarianza entre ellas.

El método PLS proporciona una estadística crucial para la interpretación de las variables predictoras, conocida como VIP (Variable Importance for Projection). Este valor resume la contribución de una variable al modelo, es decir, su importancia en la proyección. El VIP es aplicable tanto al conjunto de variables respuesta o dependientes como a las variables predictoras, y representa el valor de cada variable explicativa en el ajuste del modelo. A continuación, se presentan los resultados del análisis de PLS2 aplicado a las alternativas propuestas.

### PLS para alternativa despliegue vertical

Al igual que con el método AFM, se lleva a cabo el análisis PLS. Inicialmente, se realiza un despliegue vertical de cada conjunto de datos, lo que implica que los individuos se incrementan en filas. Para el conjunto de datos del IPM, se consideran un total de 128 individuos (excluyendo a Bogotá), que corresponden a 33 departamentos durante un período de 4 años (2018 a 2021), y 15 variables. Por otro lado, para el conjunto de datos del PIB, se tienen un total de 128 individuos (excluyendo a Bogotá) y 13 variables.

### Análisis de componentes

|          | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $u_6$ | $u_7$ | $u_8$ | $u_9$ | $u_{10}$ |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| $t_1$    | 0.80  | 0.00  | -0.00 | -0.00 | -0.00 | -0.00 | -0.00 | 0.00  | -0.00 | 0.00     |
| $t_2$    | 0.04  | 0.74  | 0.00  | 0.00  | -0.00 | -0.00 | -0.00 | 0.00  | 0.00  | 0.00     |
| $t_3$    | 0.27  | -0.05 | 0.47  | 0.00  | 0.00  | 0.00  | 0.00  | -0.00 | -0.00 | -0.00    |
| $t_4$    | 0.21  | -0.09 | 0.33  | 0.40  | -0.00 | 0.00  | 0.00  | -0.00 | -0.00 | -0.00    |
| $t_5$    | 0.01  | -0.09 | 0.11  | 0.13  | 0.54  | -0.00 | -0.00 | 0.00  | 0.00  | 0.00     |
| $t_6$    | 0.04  | 0.21  | 0.05  | -0.02 | -0.06 | 0.44  | -0.00 | 0.00  | -0.00 | -0.00    |
| $t_7$    | 0.15  | 0.02  | 0.19  | 0.24  | 0.14  | -0.11 | 0.42  | -0.00 | -0.00 | 0.00     |
| $t_8$    | 0.03  | -0.02 | 0.03  | 0.02  | -0.08 | 0.01  | -0.05 | 0.47  | 0.00  | 0.00     |
| $t_9$    | -0.09 | 0.11  | -0.12 | -0.15 | 0.06  | 0.12  | -0.15 | -0.23 | 0.29  | 0.00     |
| $t_{10}$ | -0.04 | -0.07 | -0.02 | -0.07 | -0.15 | -0.03 | -0.15 | 0.02  | -0.02 | 0.41     |

**Tabla 6-13.:** Correlación entre T y U

En la Tabla **6-13** se evidencian las correlaciones entre las primeras diez componentes latentes T (variables predictoras X) y U (variables de respuesta Y). Se observa que las primeras dos

componentes presentan correlaciones considerablemente altas, siendo del 80 % para  $t_1$  con  $u_1$  y 74 % para  $t_2$  con  $u_2$ . Estas altas correlaciones indican que las primeras dos componentes capturan la mayor parte de la variabilidad compartida entre las variables predictoras y las variables de respuesta.

Las correlaciones en las componentes posteriores disminuyen significativamente, esta disminución rápida sugiere que las componentes adicionales capturan cada vez menos variabilidad compartida, lo que indica una reducción en su importancia relativa.

### Varianza explicada

|    | $R_X^2$ | $R_{Xcum}^2$ | $R_Y^2$ | $R_{Ycum}^2$ |
|----|---------|--------------|---------|--------------|
| t1 | 0.52    | 0.52         | 0.27    | 0.27         |
| t2 | 0.12    | 0.64         | 0.08    | 0.35         |

**Tabla 6-14.:** Coeficientes  $R^2$

En la Tabla **6-14** se observa que la primer componente  $t_1$  explica una proporción significativa de la variabilidad en las variables predictoras  $X$ , capturando el 53 % de la varianza total. Además, este mismo componente explica el 27 % de la varianza en las variables de respuesta  $Y$ . Cuando se incorpora el segundo componente  $t_2$ , la varianza explicada en las variables  $X$  incrementa en un 12 %, alcanzando una varianza acumulada del 64 %. De manera similar, para las variables  $Y$ , el segundo componente agrega un 8 % adicional de varianza explicada, llevando la varianza acumulada a un 35 %.

Este patrón sugiere que, aunque el primer componente es particularmente efectivo en capturar la mayor parte de la variabilidad en los datos, los componentes adicionales contribuyen de manera decreciente. En resumen, los primeros dos componentes del modelo PLS son capaces de capturar una proporción sustancial de la variabilidad tanto en  $X$  como en  $Y$ , destacando la eficiencia del modelo en la representación de la estructura subyacente de los datos.

### Poder explicativo sobre $Y$

Los valores VIP permiten clasificar las variables  $x_j$  en función de su poder explicativo sobre  $y_j$  en el espacio de proyección de la nube de puntos. Aquellas variables que presenten valores VIP altos ( $> 1$ ) se consideran las que contribuyen en mayor medida a la construcción de  $Y$ .

| Variable | $t_1$ | $t_2$ |
|----------|-------|-------|
| Agri     | 0.50  | 0.67  |
| Min      | 0.18  | 0.37  |
| Manu     | 1.07  | 0.96  |
| Elec     | 1.06  | 0.94  |
| Cons     | 0.74  | 0.77  |
| Com      | 0.81  | 1.24  |
| Comu     | 1.34  | 1.19  |
| Fina     | 1.26  | 1.15  |
| Inmo     | 1.27  | 1.13  |
| Prof     | 1.30  | 1.15  |
| Soc      | 0.07  | 1.02  |
| Recre    | 1.21  | 1.13  |
| Imp      | 1.04  | 0.93  |

**Tabla 6-15.:** Estadísticos VIP

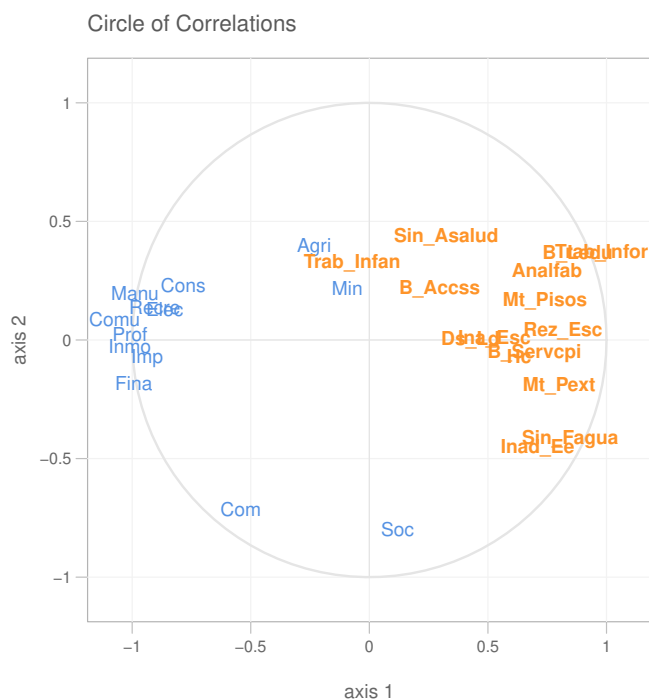
En la Tabla **6-15** se muestran los valores del VIP correspondientes a cada variable del PIB, las cuales se consideran variables predictoras en este análisis. Se observa que las actividades económicas del sector profesional y científico, Comunicaciones, Inmobiliario y el sector de recreación son aquellos que presentan la mayor contribución en la construcción de Y (IPM). Sin embargo, en general, las demás variables tienen valores cercanos a uno, lo que indica que todas las variables predictoras están aportando al análisis de manera significativa. Aunque las actividades económicas relacionadas con Minas e inversión social tienen una contribución muy pequeña.

### Representación de las variables

En la Figura **6-30** representa las correlaciones entre las variables  $u_1$  y  $t_1$  construidas, las cuales, resumen la información de la matriz del PIB con la matriz de respuesta IPM. Se aprecia que el conjunto de variables relacionadas con la pobreza presenta una correlación inversa con el conjunto de variables relacionadas con la productividad. Esto sugiere que un mayor crecimiento en las actividades económicas contribuirá a la disminución de la pobreza.

Se observa en el conjunto de datos del PIB una agrupación de variables en tres bloques principales, indicando una correlación significativa entre las variables de cada bloque. El primer bloque comprende actividades económicas como la minería (Min) y la agricultura (Agri). El segundo bloque incluye actividades como las manufacturas (Manu), electricidad (Elec), construcción (Cons), comunicaciones (Comu), actividades financieras (Fina) y profesionales (Prof), entre otras. En el tercer bloque se agrupan las actividades relacionadas con el comercio (Com) y los servicios sociales (Soc). En contraste, las variables del conjunto de datos del Índice de Pobreza Multidimensional (IPM) muestran una mayor cohesión entre sí, lo que

sugiere una menor variabilidad interna.



**Figura 6-30.:** Circulo de correlaciones del método PLS

Este patrón sugiere que la minería y la agricultura no presentan una correlación significativa con las variables asociadas a las condiciones de pobreza, indicando que los departamentos con un alto impacto económico en estas áreas no cuentan con las condiciones e inversiones sociales necesarias para mejorar la calidad de vida de su población. Por consiguiente, estas actividades no contribuyen de manera significativa a la reducción de la pobreza. En contraste, las actividades relacionadas con los servicios profesionales y recreativos, la construcción y el suministro eléctrico muestran una mayor efectividad en la disminución de las condiciones de pobreza. Las actividades comerciales, tanto al por mayor como al por menor, juegan un papel crucial en la reducción de indicadores específicos de pobreza, tales como el trabajo infantil, la falta de aseguramiento en salud y las barreras de acceso a servicios de salud.

Finalmente, lo anterior a estructura y correlación de las variables económicas y de pobreza sugieren que para reducir de manera efectiva la pobreza, es necesario focalizar las inversiones y políticas en sectores que demuestren una mayor capacidad para mejorar las condiciones de vida, como los servicios profesionales, recreativos, construcción y comercio, en lugar de solo impulsar actividades tradicionales como la minería y la agricultura.

## Representación de los individuos

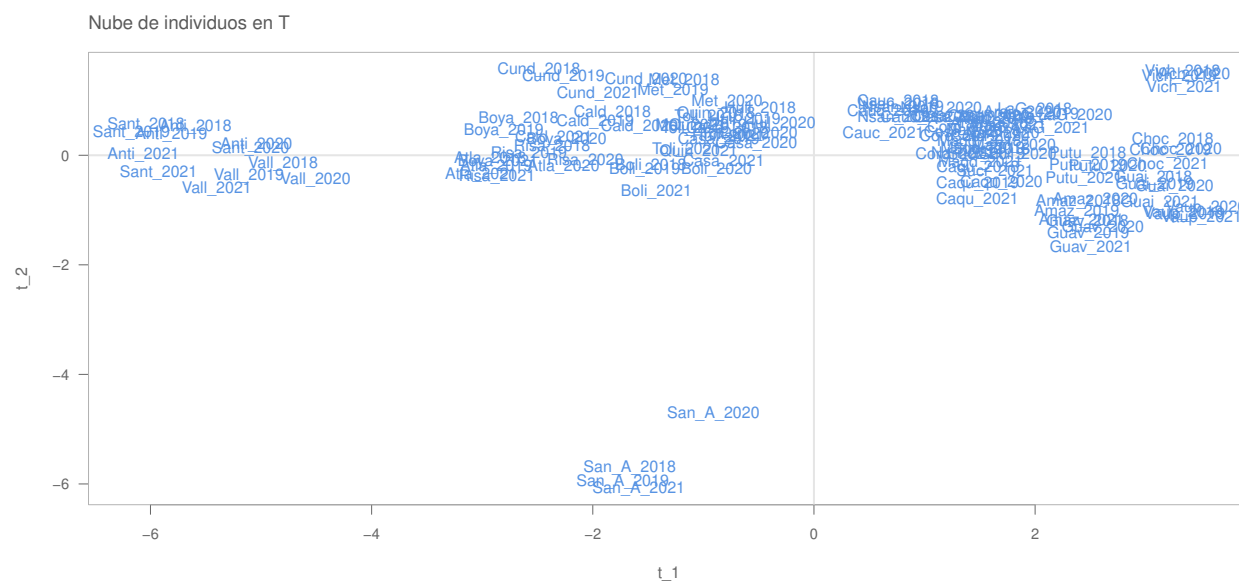


Figura 6-31.: Nube de individuos en T

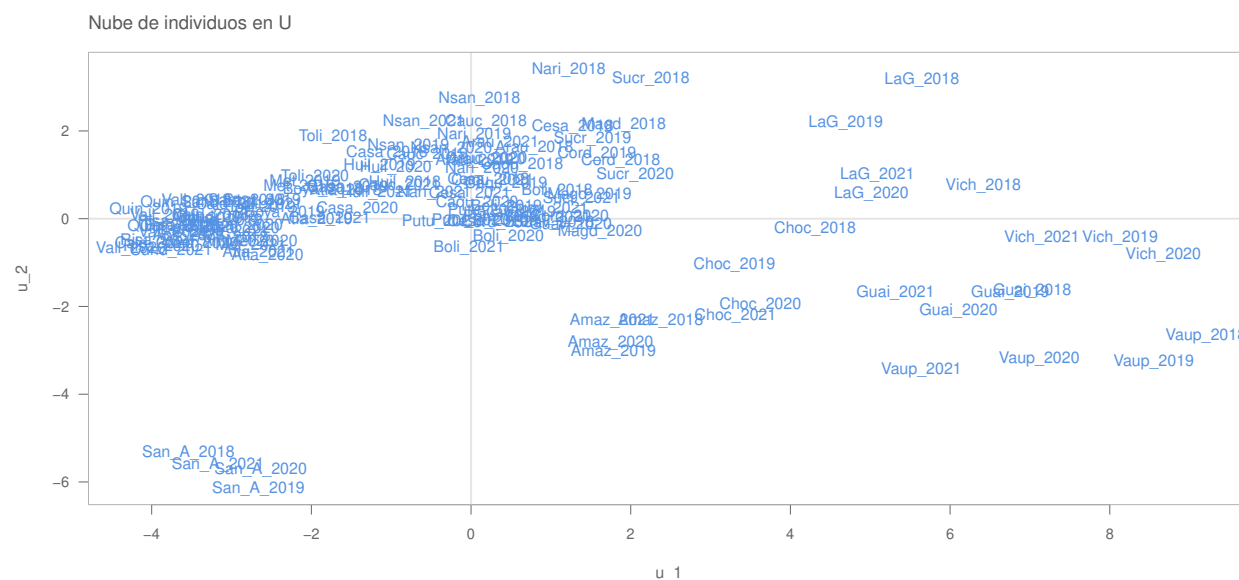


Figura 6-32.: Nube de individuos en U

Las Figuras 6-31 y 6-32 presentan la proyección de los departamentos sobre los planos de las componentes U de Y y T de X, la Figura 6-31 se encuentra en el espacio de las variables

predictoras (PIB), es decir, el plano está construido a partir de las variables  $t_1$  y  $t_2$  y en la Figura 6-32 en el espacio de las variables de respuesta (IPM).

Con respecto a la Figura 6-31 se destaca que el departamento de San Andrés se encuentra notablemente distanciado de los demás departamentos, lo que indica sus altos niveles de productividad en actividades de Comercio al por mayor y al por menor en comparación con los demás departamentos. Esto implica que es probable que presenten índices de pobreza más bajos. Además, es importante mencionar que se pueden observar la formación de grupos de departamentos en el plano. En el cuadrante I se encuentran las regiones que presentan condiciones extremas de pobreza, como Vichada, Vaupés, La Guajira, Guainía, entre otros. Por otro lado, se forma un grupo de departamentos con una productividad medianamente alta, como Cundinamarca, Caldas, Boyacá y Atlántico.

Con respecto a la Figura 6-32 se destaca que departamentos como Vaupés, Vichada y Guainía se encuentran distanciados de los demás departamentos, que a su vez son departamentos que se encuentran caracterizados por condiciones de pobreza extrema. Asimismo, es importante mencionar que al igual que en el AFM la primer alternativa permite observar la trayectoria de los departamentos.

### PLS para alternativa de despliegue horizontal

La segunda alternativa consiste en un despliegue horizontal de los años, por tanto se incrementan las variables. Para el caso del conjunto de datos del IPM se consideran un total de 32 individuos (excluyendo a Bogotá), que corresponden a los departamentos y 60 variables durante un período de 4 años (2018 a 2021). Por otro lado, para el conjunto de datos del PIB se consideran un total de 32 individuos (excluyendo a Bogotá) y 52 variables.

### Análisis de componentes

|          | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $u_6$ | $u_7$ | $u_8$ | $u_9$ | $u_{10}$ |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| $t_1$    | 0.81  | -0.00 | -0.00 | -0.00 | -0.00 | 0.00  | -0.00 | 0.00  | -0.00 | -0.00    |
| $t_2$    | 0.05  | 0.79  | 0.00  | -0.00 | -0.00 | 0.00  | -0.00 | -0.00 | 0.00  | 0.00     |
| $t_3$    | 0.26  | -0.07 | 0.48  | -0.00 | -0.00 | 0.00  | -0.00 | -0.00 | -0.00 | -0.00    |
| $t_4$    | 0.21  | -0.13 | 0.37  | 0.50  | -0.00 | -0.00 | 0.00  | 0.00  | -0.00 | 0.00     |
| $t_5$    | -0.03 | -0.07 | 0.08  | 0.14  | 0.66  | 0.00  | -0.00 | -0.00 | 0.00  | 0.00     |
| $t_6$    | 0.19  | 0.13  | 0.24  | 0.21  | 0.07  | 0.52  | 0.00  | 0.00  | -0.00 | -0.00    |
| $t_7$    | -0.08 | 0.07  | -0.12 | -0.14 | 0.03  | -0.03 | 0.45  | 0.00  | -0.00 | -0.00    |
| $t_8$    | -0.13 | 0.13  | -0.21 | -0.28 | 0.01  | -0.24 | 0.23  | 0.42  | -0.00 | -0.00    |
| $t_9$    | -0.03 | -0.01 | -0.04 | 0.03  | 0.17  | 0.06  | 0.20  | -0.04 | 0.58  | 0.00     |
| $t_{10}$ | 0.04  | 0.10  | -0.00 | 0.03  | 0.14  | 0.20  | 0.01  | -0.04 | 0.20  | 0.55     |

**Tabla 6-16.:** Correlación entre T y U

En la Tabla **6-16** se evidencian las correlaciones entre las primeras diez componentes T y U. Se observa que las primeras dos componentes presentan correlaciones significativamente altas, siendo del 81 % para  $t_1$  con  $u_1$  y 79 % para  $t_2$  con  $u_2$ . Estas altas correlaciones indican que las primeras dos componentes capturan la mayor parte de la variabilidad compartida entre las variables predictoras y las variables de respuesta para esta alternativa de despliegue.

Las correlaciones en las componentes posteriores disminuyen significativamente, esta disminución rápida sugiere que las componentes adicionales capturan cada vez menos variabilidad compartida, lo que indica una reducción en su importancia relativa.

En relación con las correlaciones observadas en la alternativa de despliegue vertical (Tabla **6-13**), se ha identificado un incremento en las correlaciones para la alternativa despliegue horizontal, lo que sugiere que este despliegue presenta un mejor ajuste y asociación entre los dos conjuntos de datos.

### Varianza explicada

|    | $R_X^2$ | $R_{Xcum}^2$ | $R_Y^2$ | $R_{Ycum}^2$ |
|----|---------|--------------|---------|--------------|
| t1 | 0.53    | 0.53         | 0.30    | 0.30         |
| t2 | 0.12    | 0.65         | 0.09    | 0.39         |

**Tabla 6-17.:** Coeficientes  $R^2$

En la Tabla **6-17** se observa que la primer componente  $t_1$  explica una proporción significativa de la variabilidad en las variables predictoras  $X$ , capturando el 53 % de la varianza total. Además, este mismo componente explica el 30 % de la varianza en las variables de respuesta  $Y$ . Cuando se incorpora el segundo componente  $t_2$ , la varianza explicada en las variables  $X$  incrementa en un 12 %, alcanzando una varianza acumulada del 65 %. De manera similar, para las variables  $Y$ , el segundo componente agrega un 9 % adicional de varianza explicada, llevando la varianza acumulada a un 39 %.

Este patrón sugiere que, al igual que en la alternativa de despliegue vertical, el primer componente es efectivo al capturar gran parte de la variabilidad de los datos, la adición de componentes genera una adición no muy significativa pero que igualmente, son capaces de capturar una proporción sustancial de la variabilidad tanto en  $X$  como en  $Y$ , destacando la eficiencia del modelo.

### Poder explicativo sobre $Y$

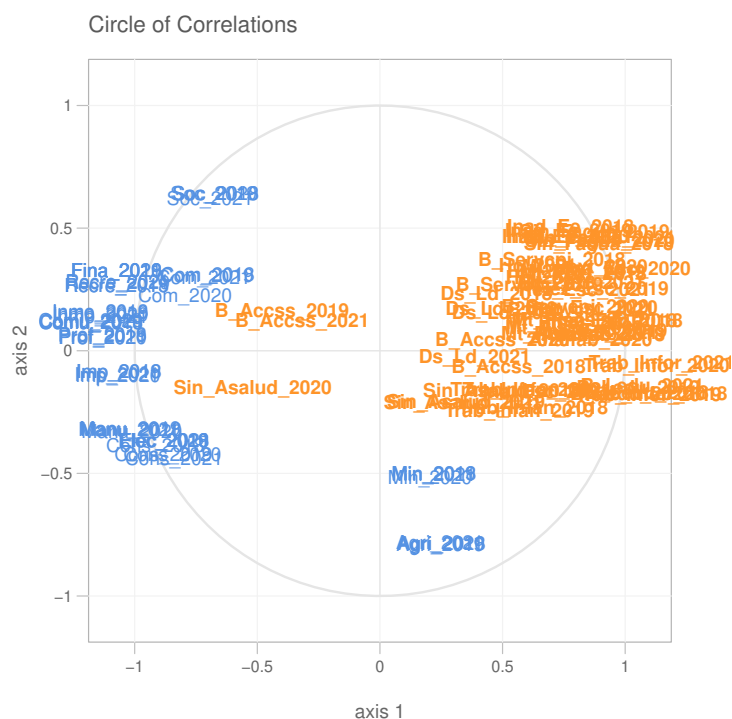
| Variable   | t1   | t2   | Variable   | t1   | t2   |
|------------|------|------|------------|------|------|
| Agri_2018  | 0.47 | 0.66 | Agri_2020  | 0.49 | 0.66 |
| Min_2018   | 0.16 | 0.34 | Min_2020   | 0.21 | 0.33 |
| Manu_2018  | 1.05 | 0.94 | Manu_2020  | 1.07 | 0.96 |
| Elec_2018  | 1.05 | 0.93 | Elec_2020  | 1.06 | 0.94 |
| Cons_2018  | 0.86 | 0.80 | Cons_2020  | 0.82 | 0.79 |
| Com_2018   | 0.78 | 1.27 | Com_2020   | 0.93 | 1.25 |
| Comu_2018  | 1.33 | 1.19 | Comu_2020  | 1.33 | 1.18 |
| Fina_2018  | 1.25 | 1.13 | Fina_2020  | 1.25 | 1.13 |
| Inmo_2018  | 1.26 | 1.12 | Inmo_2020  | 1.26 | 1.12 |
| Prof_2018  | 1.29 | 1.14 | Prof_2020  | 1.29 | 1.15 |
| Soc_2018   | 0.03 | 1.05 | Soc_2020   | 0.07 | 1.03 |
| Recre_2018 | 1.20 | 1.14 | Recre_2020 | 1.20 | 1.13 |
| Imp_2018   | 1.03 | 0.93 | Imp_2020   | 1.04 | 0.93 |
| Agri_2019  | 0.48 | 0.67 | Agri_2021  | 0.49 | 0.66 |
| Min_2019   | 0.16 | 0.34 | Min_2021   | 0.15 | 0.33 |
| Manu_2019  | 1.06 | 0.95 | Manu_2021  | 1.08 | 0.96 |
| Elec_2019  | 1.06 | 0.94 | Elec_2021  | 1.05 | 0.93 |
| Cons_2019  | 0.81 | 0.76 | Cons_2021  | 0.84 | 0.83 |
| Com_2019   | 0.79 | 1.28 | Com_2021   | 0.84 | 1.27 |
| Comu_2019  | 1.33 | 1.19 | Comu_2021  | 1.31 | 1.16 |
| Fina_2019  | 1.25 | 1.13 | Fina_2021  | 1.25 | 1.13 |
| Inmo_2019  | 1.26 | 1.12 | Inmo_2021  | 1.26 | 1.11 |
| Prof_2019  | 1.29 | 1.14 | Prof_2021  | 1.29 | 1.14 |
| Soc_2019   | 0.04 | 1.04 | Soc_2021   | 0.14 | 0.96 |
| Recre_2019 | 1.20 | 1.14 | Recre_2021 | 1.21 | 1.13 |
| Imp_2019   | 1.03 | 0.93 | Imp_2021   | 1.04 | 0.93 |

**Tabla 6-18.:** Estadísticos VIP

En la Tabla **6-18** se muestran los valores del VIP para esta segunda alternativa, corresponde a cada variable del PIB, las cuales son las variables predictoras en este análisis. Se observa que las actividades económicas del sector profesional y científico, Comunicaciones, Inmobiliario y el sector de recreación son aquellos que presentan la mayor contribución en la construcción del IPM, al igual que en la alternativa uno. Sin embargo, en general, las demás variables tienen valores cercanos a uno, lo que indica que todas las variables predictoras están aportando al análisis de manera significativa. Aunque la actividad económica relacionada con Minas y agricultura tienen contribuciones pequeñas en la construcción de Y para todos los años evaluados.

## Representación de las variables

En el círculo de correlaciones de la Figura 6-33, se aprecia igualmente una relación inversa entre las variables de pobreza y productividad, sin embargo, a diferencia de la primer alternativa, en esta se puede observar la trayectoria de las variables en los diferentes años.



**Figura 6-33.:** Círculo de correlaciones del método PLS

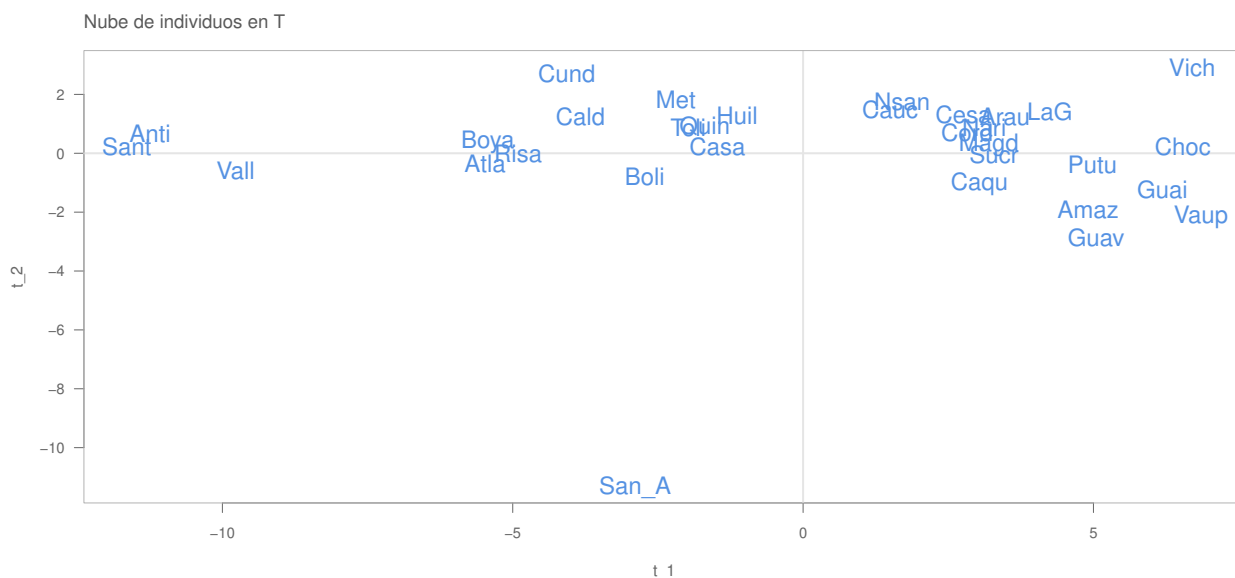
Se observa igualmente, la agrupación entre las variables del IPM, y las subdivisiones entre las variables del conjunto de datos del PIB. Sin embargo, se observa un cambio importante en las variables de Barreras de acceso a servicios de salud en los años 2019 y 2021 y la variable Sin aseguramiento en salud para el año 2020. Este comportamiento es interesante, ya que se relacionan con la salud, teniendo en cuenta que en el 2020 se enmarcó la pandemia de la Covid-19 y estas presentan correlaciones significativas con las actividades económicas del PIB a excepción de las actividades de Minería y Agricultura. Este hallazgo sugiere que la pandemia de Covid-19 impactó notablemente las condiciones de acceso a servicios de salud y el aseguramiento en salud, reflejándose en las correlaciones observadas con diversas actividades económicas. La falta de correlación con las actividades de Minería y Agricultura indica que estos sectores no están tan directamente afectados por las variables de salud, a diferencia de otros sectores económicos. Esto subraya la importancia de considerar factores externos y eventos globales, como la pandemia, al analizar las interacciones entre la economía y las

condiciones sociales y de salud.

Al igual que en la alternativa de despliegue vertical, la alternativa de despliegue horizontal revela un patrón consistente en el comportamiento de las actividades de minería y agricultura, las cuales no presentan una correlación significativa con las variables asociadas a las condiciones de pobreza en los años considerados. Asimismo, las actividades relacionadas con los servicios profesionales y recreativos, la construcción y el suministro eléctrico demuestran una mayor efectividad en la reducción de las condiciones de pobreza en los mismos períodos.

Esto indica que, a lo largo de los años medidos, las actividades de minería y agricultura no contribuyen de manera significativa a la mejora de las condiciones de vida en términos de pobreza. En contraste, los sectores de servicios profesionales y recreativos, construcción y suministro eléctrico tienen un impacto más positivo y efectivo en la reducción de la pobreza, subrayando la importancia de estos sectores en el desarrollo socioeconómico y la necesidad de enfocar políticas e inversiones hacia ellos para mejorar las condiciones de vida de la población.

### Representación de los individuos



**Figura 6-34.:** Nube de individuos en T

Con respecto a las Figuras 6-34 y 6-35 presentan el comportamiento de los departamentos, al igual que en la primera alternativa, en la Figura 6-34 se destaca igualmente que en esta alternativa, el comportamiento de los individuos es similar que en la alternativa uno. San Andrés se encuentra notablemente distanciado de los demás departamentos, lo que indica

sus altos niveles de productividad en comparación con los demás. Esto implica que es probable que presenten índices de pobreza más bajos. Además, es importante mencionar que se pueden observar la formación de grupos de departamentos en el plano. En el cuadrante I se encuentran las regiones que presentan condiciones extremas de pobreza, como Vichada, Vaupés, La Guajira, Guainía, entre otros. Por otro lado, se forma un grupo de departamentos con una productividad medianamente alta, como Cundinamarca, Caldas, Boyacá y Atlántico.

Con respecto a la Figura 6-35 se destaca que departamentos como Vaupés, Vichada y Guainía se encuentran distanciados de los demás departamentos, que a su vez son departamentos que se encuentran caracterizados por condiciones de pobreza extrema. A diferencia de la primer alternativa, no es posible observar la trayectoria de los departamentos.

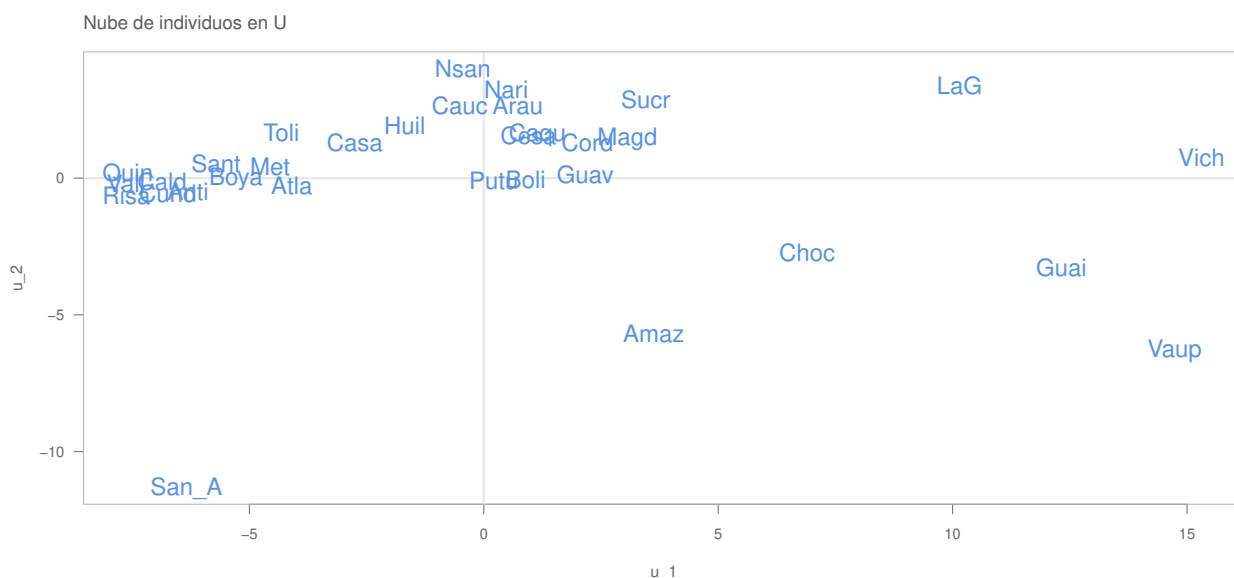


Figura 6-35.: Nube de individuos en U

### 6.3.2. Análisis comparativo de las alternativas a través de PLS

La técnica de Mínimos Cuadrados Parciales (PLS en sus siglas en inglés) tiene como objetivo principal maximizar la correlación entre dos conjuntos de datos, permitiendo así comprender cómo se comporta un conjunto de datos en función de otro. A diferencia de métodos como la regresión lineal múltiple que busca minimizar los errores de predicción, el PLS se enfoca en maximizar la correlación entre las variables predictoras ( $Y$ ) y las variables de respuesta ( $X$ ). Esto permite capturar las relaciones más fuertes entre los dos conjuntos de datos.

A partir de lo anterior, en esta técnica, las variables se agrupan en dos grandes tablas que corresponden a las ramas de actividad del PIB y los indicadores del IPM. La tabla de datos correspondiente a las variables predictoras ( $X$ ) está compuesta por el conjunto de datos del

PIB, mientras que la tabla de datos correspondiente a las variables de respuesta ( $Y$ ) está compuesta por el conjunto de datos del IPM. En cada una de las alternativas, se realiza el despliegue de cada cubo de datos en matrices bidimensionales. Esto indica que el mismo despliegue realizado para el cubo de datos del IPM ( $Y$ ) se aplica al cubo de datos del PIB ( $X$ ).

En general, se observó una alta capacidad para explicar la variabilidad en los datos en cada alternativa propuesta. Específicamente, los valores de varianza explicada por  $R^2$  fueron del 64 % y 65 % para las variables del conjunto de datos  $X$ , indicando que una proporción significativa de la variabilidad en  $X$  es capturada por los componentes del modelo. Para las variables del conjunto de datos  $Y$ , los valores de  $R^2$  fueron del 35 % y 39 %, reflejando que, aunque en menor medida que en  $X$ , la técnica PLS sigue siendo efectivo en la representación de la estructura de las variables de respuesta.

El análisis comparativo entre las alternativas de despliegue vertical y horizontal presenta diferencias en la capacidad de explicación de la variabilidad en los datos del Producto Interno Bruto (PIB) y el Índice de Pobreza Multidimensional (IPM). La alternativa de despliegue vertical en los dos primeros componentes obtuvo una proporción de varianza explicada del 64 % para las variables del PIB y el 35 % para las variables del IPM. La alternativa de despliegue horizontal proporcionó unos coeficientes de  $R^2$  del 65 % para las variables del PIB y el 39 % para las variables del IPM. Estos resultados sugieren que, aunque ambas alternativas son efectivas, el despliegue horizontal puede ofrecer una mejor capacidad explicativa, especialmente para las variables del IPM, lo cual podría ser determinante en la elección de la técnica de despliegue según los objetivos específicos del estudio.

La proporción de varianza explicada por las dos alternativas de despliegue no presentan diferencias significativas, por tanto, las dos alternativas son efectivas para capturar la asociación entre las dos estructuras de datos, sin embargo, el ligero aumento en la variabilidad explicada por la alternativa de despliegue horizontal sugiere que esta captura mayor información especialmente para el conjunto de datos de  $Y$ .

### Alternativa de despliegue vertical

- **Ventajas:** Esta alternativa permite analizar la trayectoria de los departamentos y su comportamiento a lo largo de la tercera condición, los años. Igualmente, permite analizar patrones comunes en los departamentos según sus condiciones de productividad y pobreza en los diferentes años medidos.
- **Desventajas:** Al igual que en la técnica de AFM, esta alternativa no permite ver la trayectoria de las variables. Además, el aumento de individuos puede conducir al aumento del costo computacional, por tanto, con grandes cantidades de datos el análisis requerirá una mayor capacidad computacional.

### Alternativa de despliegue horizontal

- **Ventajas:** Permite analizar la trayectoria de las variables y su comportamiento a lo largo de los años medidos, lo que implica que se captura más información potencialmente relevante sobre las relaciones entre las variables predictoras y las de respuesta, mejorando la capacidad de análisis y la robustez de los resultados.
- **Desventajas:** Aumentar las variables puede dificultar la identificación de las relaciones clave entre las variables. Igualmente, esta alternativa conduce al aumento del costo computacional.

La elección entre las diferentes alternativas de despliegue dependerá de los objetivos específicos de la investigación. Ya que, si el objetivo es analizar la trayectoria de los individuos en la tercera condición de los bloques, se recomienda emplear la alternativa de despliegue vertical, sin embargo tener presente que, bajo el caso de estudio esta alternativa castiga la proporción de variabilidad explicada para las variables de  $Y$ . En contraste, si el objetivo es examinar la trayectoria de las variables en la tercera condición de los bloques, la alternativa de despliegue horizontal resulta ser más adecuada y esta a su vez, presenta mejores resultados de la variabilidad explicada. Estas estrategias permiten capturar diferentes dimensiones de la variabilidad en los datos y proporcionan una visión más detallada y precisa en función de los objetivos del estudio.

## 6.4. Asociación entre dos estructuras de datos en tres vías

### 6.4.1. Mínimos Cuadrados Parciales Multiway (N-PLS)

N-way PLS es un método que estudia la relación entre estructuras de datos de tres vías o  $N$  vías ( $\mathbf{X}$ ) y cualquier otra estructura de datos  $\mathbf{Y}$ . Basicamente, es una extensión del PLS a estructuras de más de dos vías (N-way) cumpliendo el mismo objetivo, intentar maximizar la correlación entre las matrices N-way  $\mathbf{X}$  y  $\mathbf{Y}$  ([Hervás et al., 2018](#)).

En particular para este caso, la estructura de datos de tres vías  $\mathbf{X}$  estará representada por las variables del conjunto de datos del PIB, mientras que la estructura de datos  $\mathbf{Y}$  es el arreglo de tres vías que representa el IPM. Este método busca encontrar espacios latentes a partir de nuevas variables como  $\mathbf{W}_j$  y  $\mathbf{W}_k$  que permitan maximizar la correlación entre las estructuras de datos del PIB con el IPM.

Se realiza una validación cruzada para definir el número de componentes para el análisis NPLS. Igualmente, se realiza la descomposición de la estructura del PIB  $\mathbf{X}$  y del IPM  $\mathbf{Y}$ ,

y prosigue con el análisis de N-way PLS. A continuación, se presentan los resultados del análisis de este método.

### Validación cruzada

La validación cruzada (CV por su nombre en inglés) para el modelo sNPLS se desarrolla a partir de la función de **cv\_snpls**, el cual es un método utilizado para evaluar la capacidad predictiva y la robustez del modelo.

El proceso consiste en dividir el conjunto de datos en varios conjuntos más pequeños llamados folds. Luego, se ajusta el modelo sNPLS en cada combinación posible de estos folds, dejando uno de ellos como conjunto de prueba y los demás como conjunto de entrenamiento. Se calcula el error de predicción para cada ajuste del modelo y se promedian para obtener una estimación global del rendimiento del modelo. Esto ayuda a evaluar cómo se generaliza el modelo a nuevos datos y a identificar posibles problemas de sobreajuste.

Este procedimiento es especialmente útil porque ayuda a controlar la complejidad del modelo y a evitar el ajuste excesivo al conjunto de datos de entrenamiento.

Se realiza validación cruzada con los siguientes parametros:

- **Número de componentes a probar: 3**
- **Número de muestras: 20**

La mejor solución según el proceso de validación cruzada se encontró en la muestra número **463** de la cual se obtiene que la mejor la cantidad de componentes a utilizar en el modelo para NPLS es 1 componente, lo anterior se presenta en la Tabla **6-19**

| Muestra | Ncomp | Threshold_j | Threshold_k |
|---------|-------|-------------|-------------|
| 463     | 1     | 0.62        | 0.02        |

**Tabla 6-19.:** Validación Cruzada para la selección del número de componentes

Para el desarrollo de esta técnica se toman dos componentes.

### Contribución de las variables y el modo tiempo a los componentes

La contribución de las variables y el modo tiempo se mide a través del ratio de selectividad, esta métrica proporciona una medida de cuanta contribución aporta cada componente del modelo en la explicación de la variabilidad de los datos de entrada.

Esto permite evaluar la importancia relativa de cada componente en la explicación de la variabilidad de los datos y ayuda a identificar qué componentes contribuyen de manera más significativa a la relación entre las variables predictoras y las variables de respuesta.

| Variable | Componente 1 | Componente 2 |
|----------|--------------|--------------|
| Agri     | 0.00         | 0.00         |
| Min      | 0.00         | 0.00         |
| Manu     | 1.85         | 1.88         |
| Elec     | 1.64         | 1.66         |
| Cons     | 1.85         | 1.83         |
| Com      | 1.05         | 1.18         |
| Comu     | 1.51         | 1.53         |
| Fina     | 1.55         | 1.58         |
| Inmo     | 1.54         | 1.57         |
| Prof     | 1.53         | 1.55         |
| Soc      | 0.00         | 1.18         |
| Recre    | 1.52         | 1.53         |
| Imp      | 1.86         | 1.91         |

**Tabla 6-20.:** Contribuciones a las componentes de las actividades económicas

A partir de lo anterior, en la Tabla **6-20** se observa que para las actividades económicas Agricultura (Agri) y Minería (Min) la primera componente la contribución es 0. Esto sugiere que el Componente 1 no está contribuyendo a la explicación de la variabilidad en esas variables o viceversa. Sin embargo, para las variables Manufactura (Manu), Suministro de electricidad (Elec), Construcción (Cons), Información y comunicaciones (Comu), actividades financieros (Fina), actividades Inmobiliarias (Inmo), Actividades profesionales (Prof), y Actividades artísticas y de recreación (Recre), las contribuciones son superiores a 1, lo que indica que el Componente 1 tiene una contribución significativa en la explicación de la variabilidad en esas actividades económicas. Por otro lado, las actividades de inversión social (Soc) y comercio (Com) contribuyen en la construcción del segundo componente, siendo la variable de inversión social la única que contribuye específicamente al segundo componente.

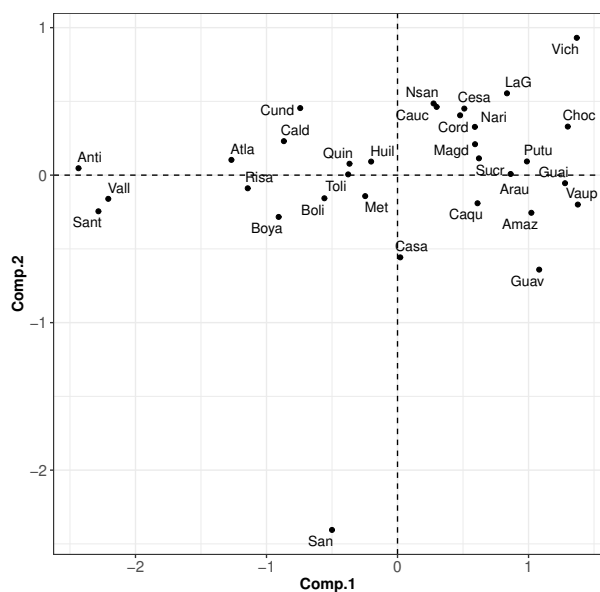
| Variable | Componente 1 | Componente 2 |
|----------|--------------|--------------|
| 2018     | 1.52         | 1.28         |
| 2019     | 1.52         | 1.27         |
| 2020     | 1.51         | 1.35         |
| 2021     | 1.51         | 1.32         |

**Tabla 6-21.:** Contribuciones a las componentes del modo tiempo

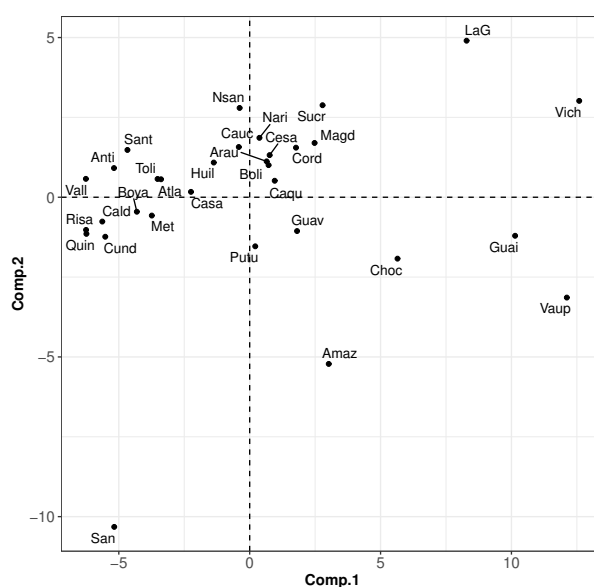
Asimismo, en la Tabla **6-21** se observa que la contribución del modo tiempo, es superior a 1 en las dos componentes y significativamente similares lo que sugiere que las dos componentes

tienen una contribución significativa en la explicación de la variabilidad de cada año. Igualmente, se observa que en los años medidos, Colombia no presenta una variación significativa en los índices de pobreza durante el 2018 al 2021.

### Representación de los individuos



**Figura 6-36.:** Gráfico de puntuación de los dos primeros componentes en la matriz T

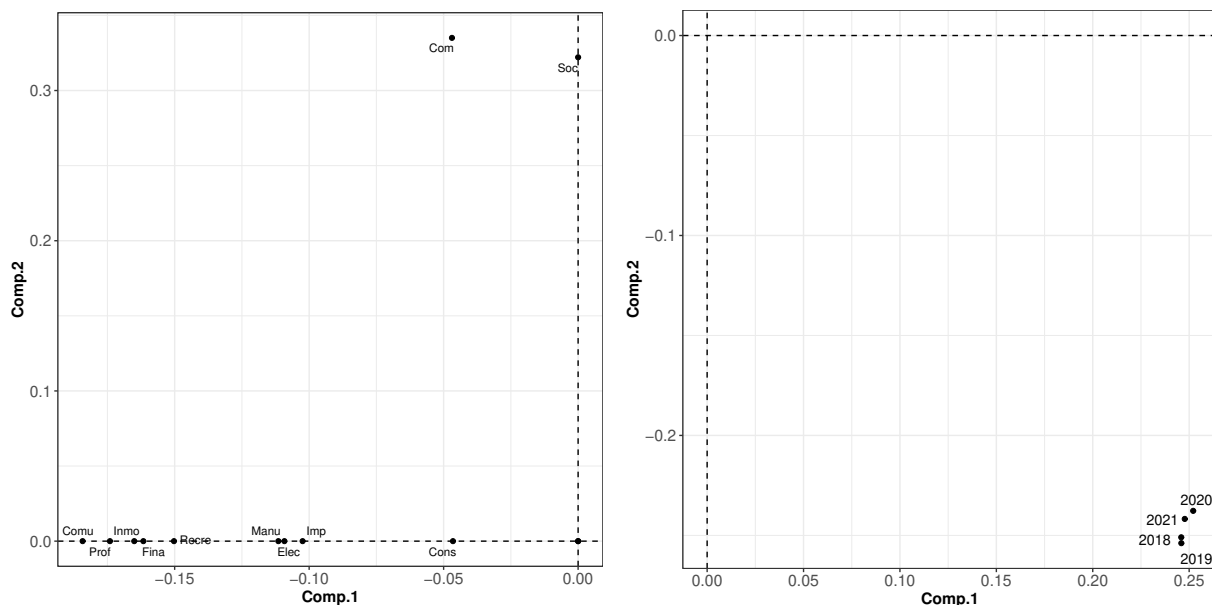


**Figura 6-37.:** Gráfico de puntuación de los dos primeros componentes en la matriz U

La Figura 6-36 ilustra la distribución de los departamentos en los dos primeros componentes. Se observa la agrupación por regiones, así como la marcada tendencia del departamento de San Andrés a alejarse de la agrupación general. En contraste, la Figura 6-37, relacionada con las puntuaciones de Y, muestra los niveles de pobreza en los años desde el 2018 al 2021, señalando que Vaupés, La Guajira, Vichada y Chocó presentaron índices de pobreza considerablemente altos en comparación con el resto de los departamentos.

Por otro lado, en la Figura 6-38 se muestran las variables del PIB, mientras que la Figura 6-39 presenta los años que contribuyen a predecir la pobreza para cada uno de los dos componentes. La Figura 6-38 revela que la primera componente de las variables del PIB está relacionada con sectores como Profesión, Comunicaciones, Manufactura, Inmobiliario e Impuestos, entre otros.

### Representación de las variables y el modo tiempo



**Figura 6-38.:** Gráfico de ponderaciones de la matriz  $W_j$  **Figura 6-39.:** Gráfico de ponderaciones de la matriz  $W_k$

Por otro lado, la Figura 6-39 revela que, para el primer componente, se observa un efecto aproximadamente igual en todos los años. En relación a las variables del PIB, se observa una mayor influencia del año 2020, mientras que para el segundo componente, se destaca que las variables de agricultura están más influenciadas por el año 2019. Todos los años muestran una influencia similar en todas las variables para predecir los índices de pobreza. No obstante, el año 2020 exhibe una coordenada destacada en ambos componentes.

### Valores predichos

La Tabla 6-22 presenta valores predichos con el modelo construido a partir de la técnica de NPLS, estos valores se presentan con el objetivo de visualizar como la pobreza se ve impactada por la productividad económica, es decir, que por medio de estos valores se da un panorama de comparación con respecto a los valores que se predijeron y los valores originales. Se eligieron departamentos que presentan un aporte significativo en la productividad del país y a su vez, presentaron altos niveles de pobreza. Igualmente, se seleccionan indicadores del IPM que tienen una relación más directa con la pobreza como lo son el Bajo logro educativo (B\_Ledu), Desempleo de larga duración (Ds\_Ld), Hacinamiento crítico (Hc), Inasistencia escolar(Ina\_Esc) y Trabajo informal (Trab\_Infor).

Por lo anterior, los departamentos como Arauca, Casanare, Cauca, Chocó, Guaviare, La Guajira, Magdalena, Meta, Norte de Santander, Putumayo, Sucre, Tolima y Vichada presentaron altos valores en los indicadores de pobreza y a su vez son departamentos que se

caracterizan por su riqueza Minera y Agropecuaria. Estas actividades económicas han demostrado no aportar a reducir la pobreza.

| Departamentos | B_Ledu | Ds_Ld | Hc    | Ina_Esc | Trab_Infor |
|---------------|--------|-------|-------|---------|------------|
| Anti          | -0.67  | 7.50  | -5.50 | 6.96    | 65.45      |
| Arau          | 21.48  | 32.17 | 11.03 | 53.27   | 14.63      |
| Casa          | 26.39  | 50.40 | 2.10  | 5.93    | 14.77      |
| Cauc          | 36.09  | 9.87  | 14.85 | 10.03   | 29.92      |
| Choc          | 2.34   | 17.50 | 37.11 | 31.58   | 29.48      |
| Guav          | 9.69   | 5.74  | 32.48 | 81.93   | 1.73       |
| LaG           | 19.69  | 31.69 | 11.37 | 60.43   | 16.90      |
| Magd          | 13.39  | 31.68 | 1.81  | 10.51   | 56.24      |
| Met           | 16.95  | 10.08 | 80.80 | 6.30    | 8.30       |
| Nsan          | 11.75  | 84.23 | 55.66 | 14.14   | 17.32      |
| Putu          | 22.97  | 12.24 | 11.37 | 30.41   | 14.69      |
| Sucr          | 86.30  | 11.62 | 5.39  | 30.72   | 31.32      |
| Toli          | 11.96  | 21.65 | 5.64  | 10.59   | 19.69      |
| Vich          | 13.83  | 23.92 | 37.48 | 18.11   | 96.42      |

**Tabla 6-22.:** Valores predichos para algunos indicadores del IPM

Es decir, un departamento en el cual los indicadores de pobreza son altos implica que no hay condiciones para suplir las condiciones de sus habitantes y por consiguiente, tendría un efecto en la productividad, ya que no tendría las condiciones para aportar a la productividad nacional. En contraste con departamentos como Antioquia, Valle del Cauca, Cundinamarca, entre otros, presentan bajos niveles de pobreza y altos niveles en productividad en algunas actividades económicas por lo que podría traducirse en que un crecimiento económico en las actividades productivas y un gran aporte al PIB provoca un efecto de mejoría en las condiciones y la calidad de vida de los habitantes, en otras palabras, bajos indicadores de pobreza en el IPM.

Los valores predichos para departamentos como Antioquia, en su mayoría presenta valores negativos, lo que indica que la alta productividad de estas regiones contribuye a la disminución de los niveles de pobreza, a diferencia de los otros departamentos, los cuales presentan altos valores de los indicadores predichos.

## 6.5. Análisis Comparativo de los resultados

Inicialmente, realiza un análisis comparativo de los resultados obtenidos para el caso de estudio, por tanto, para analizar la asociación de las estructuras de datos de tres vías correspondientes a la generación de riqueza (PIB) y condiciones de pobreza (IPM) en Colombia

para el periodo del 2018 al 2021, a partir de las tres alternativas de despliegue que se documentaron.

| Alternativas                    | AFM    | PLS  |
|---------------------------------|--------|------|
| Despliegue vertical             | 55.5 % | 80 % |
| Despliegue horizontal           | 57.4 % | 81 % |
| Despliegue horizontal combinado | 57.0 % | NA   |

**Tabla 6-23.:** Resultados comparativos de % de varianza obtenida por AFM y Correlación obtenida por PLS

- **Alternativa de despliegue vertical:** Se aplican las técnicas de AFM y PLS, observándose una inercia del 55.5 % para AFM y una correlación del 80 % para PLS. Esta alternativa permitió analizar la trayectoria de los departamentos y su comportamiento a lo largo de los años evaluados. Esta alternativa permitió identificar patrones comunes en los departamentos según sus condiciones de productividad y pobreza en las dos técnicas aplicadas, a la vez, proporciona una visión más completa de las condiciones de cada departamento a medida que se aumenta un año en los cubos de datos.
- **Alternativa de despliegue horizontal:** En esta alternativa se aplican las técnicas de AFM y PLS, en los cuales se observa que para el caso de AFM la inercia obtenida fue del 57.4 %, mientras que con el método de PLS se observa una correlación del 81 %. En esta alternativa se observa la trayectoria de los grupos de variables en cada conjunto de datos, su comportamiento en los diferentes años evaluados, sin embargo, no es posible realizar la visualización de la trayectoria de los individuos. En general, el uso de la alternativa de despliegue horizontal facilita la implementación de las técnicas como AFM y PLS, además, la visualización de los gráficos y relaciones con estas dos técnicas son de fácil interpretabilidad. Por otro lado, se presentaría un gran desafío para la escalabilidad de los datos, ya que con grandes volúmenes de datos, la eficiencia computacional puede verse afectada por la gran cantidad de variables que se puedan presentar e igualmente, podría presentar una redundancia en las variables, lo que podría complicar el procesamiento para AFM y PLS.
- **Alternativa de despliegue combinado:** En esta opción se emplea exclusivamente el método de AFM, ya que esta técnica permite realizar una división de los grupos de variables por los años medidos. Se observa una inercia del 57.0 %. En esta alternativa se observa la trayectoria de las variables en los diferentes años evaluados, sin embargo, se castiga la visualización de la trayectoria de los individuos.

En términos de inercia y correlación, las alternativas de despliegue horizontal y despliegue combinado logran valores más altos en comparación con la alternativa de despliegue vertical,

indicando así una mejor explicación de la variabilidad de los datos y una mayor correlación entre las dos estructuras de datos utilizadas en el caso de estudio.

En general, la aplicación de las diferentes alternativas propuestas para analizar la asociación de las dos estructuras de datos de tres vías, a partir de las técnicas estadísticas de AFM y PLS permiten examinar diferentes aspectos de los datos. La primera alternativa, que corresponde a un despliegue vertical, se centra en los individuos (departamentos para el caso de estudio), lo que permite observar la trayectoria específica de estos en la tercera condición de los cubos de datos. La segunda alternativa, que corresponde a un despliegue horizontal, se enfoca en la representación de la trayectoria de las variables en la tercera condición de los cubos de datos (Actividades económicas del PIB e indicadores del IPM para el caso de estudio). Asimismo, la tercer alternativa, que corresponde a un despliegue horizontal combinado, igualmente se centra en la representación y trayectoria de las variables que componen los cubos de datos.

La elección entre las diferentes alternativas de despliegue dependerá de los objetivos específicos de la investigación y en la naturaleza de los datos disponibles. Ya que, si el objetivo es analizar la trayectoria de los individuos en la tercera condición de los bloques, se recomienda emplear la alternativa de despliegue vertical. En contraste, si el objetivo es examinar la trayectoria de las variables en la tercera condición de los bloques, las alternativas de despliegue horizontal y combinado resultan ser más adecuadas. Estas estrategias permiten capturar diferentes dimensiones de la variabilidad en los datos y proporcionan una visión más detallada y precisa en función de los objetivos del estudio. En el caso de estudio de esta investigación, las alternativas que presentaban un despliegue horizontal presentaron mejores resultados de la variabilidad explicada, se recomienda desarrollar un análisis similar con otros datos para observar si la elección sería la misma.

Por otro lado, las tres técnicas utilizadas para desarrollar la metodología propuesta para el análisis de las dos estructuras de datos de tres vías arrojan resultados similares, aunque con objetivos distintos.

El **Análisis de Factores Múltiples (AFM)** tiene como objetivo maximizar la varianza contenida en la totalidad de la tabla de datos. Por otro lado, a partir de los coeficientes LG (Lagrangeano) y RV (V de Rogers), se analiza la relación entre grupos y componentes. En este estudio, el método proporcionó valores de inercia explicada en el rango del 55 % al 58 % en las tres alternativas propuestas. En lo que respecta a los resultados de LG y RV, se observó que para las tres alternativas, existe una relación moderada entre las dos estructuras de datos. El **PLS (Partial Least Squares)** tiene como objetivo maximizar la covarianza entre una tabla y otra, es decir, permite comprender cómo se comporta un conjunto de variables en función de otro conjunto de variables. Este método arrojó una correlación del 80 % y 81 % para la primera y segunda alternativa, respectivamente. Estos resultados indi-

can que estas estructuras de datos están interrelacionadas. El ***N-PLS***, al igual que el PLS, busca maximizar la covarianza entre dos estructuras de datos de tres vías. Este método, a diferencia de los anteriores, permitió visualizar el comportamiento de la matriz de datos del Índice de Pobreza Multidimensional (IPM) con la estructura de datos del Producto Interno Bruto (PIB) para los cuatro años evaluados.

A pesar que el AFM, PLS y *N-PLS* tienen objetivos metodológicos diferentes, los resultados obtenidos no presentan diferencias significativas, los círculos de correlación y las nubes de individuos en los tres métodos brindan información similar, indicando una relación negativa entre las dos estructuras de datos. Los planos factoriales por su parte, no presentan una diferencia significativa en las tres técnicas, ya que la agrupación y patrones siguen siendo similares.

Asimismo, la metodología aplicada al caso de estudio, confirma la relación existente entre ambos cubos de datos, demostrando la influencia de la productividad en las condiciones de vida de la población. Esto sugiere que cuando se presentan niveles altos de productividad en los departamentos, se espera que contribuyan a la disminución de los índices de pobreza, y viceversa.

En terminos generales, la aplicación de las diferentes técnicas propuestas (AFM, PLS, *N-PLS*) dependerá de los objetivos del estudio y de los datos disponibles, ya que las tres alternativas presentan elementos diferentes y a su vez, puede ser utilizada de manera diferente según sus propiedades y utilidades para la interpretación y aplicación. Por lo anterior, la técnica de AFM permite la aplicación de cualquiera de las alternativas propuestas, ya que es una técnica muy flexible, además, dado que el AFM es una técnica que combina la metodología de Análisis de Componentes Principales (ACP) y Análisis de Correspondencia Múltiple (ACM), es posible utilizar estructuras de datos que combinen tanto datos cuantitativos y cualitativos. Proporciona gráficos y representaciones visuales intuitivos que son fáciles de interpretar, lo que ayuda a entender las relaciones entre los tres modos.

La técnica PLS por su parte, es más estricta en el uso de datos cuantitativos tanto para la matriz de  $X$  como para la matriz  $Y$ , sin embargo, este análisis permite confirmar la asociación y correlación existente entre una matriz y otra, además, su flexibilidad en la aplicación de las alternativas de despliegue propuestas permite una interpretación fácil de la relación entre individuos y variables. Además, esta técnica es eficaz para manejar datos con colinealidad alta, lo cual es común en los cubos de datos, e igualmente, a partir de esta, es posible proporcionar predicciones de los datos.

Finalmente, la técnica de *N-PLS* es la técnica más apropiada para analizar la correlación de dos cubos de datos, ya que es adecuado para analizar datos tensoriales o estructurados

en múltiples dimensiones, como las estructuras de tres vías. Captura la complejidad y las interacciones en estructuras de  $N$  dimensiones, lo que proporciona una visión más completa entre las dimensiones de los conjuntos de datos. Al igual que PLS,  $N$ -PLS al ser una extensión del PLS, también permite realizar predicciones. Sin embargo, su aplicación y análisis aún presenta desafíos que se deben mejorar, ya que los elementos interpretativos pueden ser complejos a la hora de identificar la influencia de la tercera condición en los individuos y variables de los cubos de datos. Igualmente, NPLS no permite una visualización simultánea de los tres modos en un mismo plano factorial, sino antes bien, desarrolla un plano factorial por cada modo que hace parte del cubo, lo que presenta una ventaja para el análisis de cada modo de manera individual pero una desventaja para la visualización de los tres en un mismo plano factorial.

La elección entre AFM, PLS y  $N$ -PLS como técnicas para analizar la asociación entre dos estructuras de datos de tres vías dependerá de varios factores, incluyendo la complejidad de los datos, la disponibilidad de los datos y los objetivos del análisis. Cada técnica ofrece ventajas y desventajas específicas.

## 7. Conclusiones y recomendaciones

### 7.1. Conclusiones

La recolección y análisis de datos se han convertido en aspectos fundamentales en diversos campos y con el aumento de la complejidad de los conjuntos de datos, el análisis de datos de tres vías se ha vuelto crucial. Este tipo de datos, representados en una forma N-dimensional difieren de las estructuras de datos convencionales, que suelen ser matrices bidimensionales. Para abordar estos desafíos, se han desarrollado diversas metodologías, como el modelo de Tucker3, PARAFAC y N-PLS.

A través del análisis bibliométrico llevado a cabo, se observa que las técnicas de análisis de datos de tres vías aún no han alcanzado un uso generalizado debido a su complejidad. Por lo tanto, resulta fundamental promover y fomentar la aplicación de estas técnicas para su utilización más amplia en diversas áreas. Asimismo, en el contexto colombiano se observa una escasa implementación y contribución en este campo, lo que resalta la necesidad de impulsar investigaciones y aplicaciones que impulsen el avance científico en el análisis de estas estructuras de datos complejas.

Por otra parte, a partir del análisis de los artículos, se desarrollaron enfoques y técnicas de aplicación para este proyecto de investigación. Se observó que el análisis de datos de tres vías generalmente se realiza mediante diversas alternativas de despliegue que permiten un tratamiento eficaz de las complejas estructuras de datos. Tras una revisión exhaustiva, se propusieron en este estudio alternativas de despliegue específicas, cada una con objetivos y enfoques distintos. La primera alternativa se centra en los departamentos (individuos), lo que permite examinar la trayectoria específica de estos. La segunda alternativa se enfoca en los grupos de datos al separarlos según los años evaluados, mientras que la tercera alternativa analiza las variables en general y su trayectoria a lo largo de los años de estudio.

A partir del caso de aplicación, que corresponde al tercer objetivo de esta investigación, se puede afirmar que en los tres métodos se observa una agrupación entre departamentos basada en sus respectivas variables. Departamentos como Vaupés, Vichada, Guainía, Amazonas y La Guajira presentan condiciones de pobreza relacionadas con aspectos como las condiciones de los hogares, el hacinamiento, el rezago escolar, el desempleo y el limitado acceso a servicios públicos, lo cual refleja la compleja realidad del país.

De manera similar, departamentos como Nariño, Norte de Santander, Cesar, Cauca, entre otros, se caracterizan por condiciones de pobreza vinculadas al trabajo infantil, la falta de acceso a la salud, el empleo informal y el bajo rendimiento educativo. Por otro lado, se identificaron departamentos que se distinguen por variables de productividad, como Santander, Valle del Cauca, Atlántico, Bolívar y Risaralda, los cuales presentan una fuerte asociación con sectores profesionales, financieros, comerciales y de inversión social. San Andrés, por su parte, se destaca en los tres métodos y alternativas debido a su tendencia a apartarse de la agrupación general de departamentos, y muestra una alta asociación con los sectores comercial e inversión social.

En relación a la productividad en sectores específicos, los departamentos de Tolima, Huila y Quindío se destacan por su productividad en el sector agrícola, mientras que Casanare y Meta muestran una fuerte presencia de actividades mineras como impulsores principales de su productividad. Estos hallazgos representan un panorama detallado de la situación económica y las condiciones de pobreza en los diferentes departamentos, identificando sus principales impulsores y constantes desafíos.

Este análisis confirma la relación existente entre ambos conjuntos de datos, demostrando la influencia de la productividad en las condiciones de vida de la población. Esto sugiere que cuando se presentan niveles altos de productividad en los departamentos, se espera que contribuyan a la disminución de los índices de pobreza, y viceversa.

Por último, la aplicación de técnicas para analizar estructuras de datos de tres vías resulta sumamente importante, ya que permite no solo analizar una estructura individual, sino también estudiar la asociación entre dos de estas estructuras. Este enfoque evita la necesidad de llevar a cabo procedimientos innecesarios al contar con una metodología para analizar directamente estas complejas estructuras de datos. Además, la implementación en el software R desarrollada por (Hervás et al., 2018) demostró ser sumamente útil, dado que la metodología original de este código se basó en una estructura de tres vías ( $\mathbf{X}$ ) y una matriz  $\mathbf{Y}$ . A partir de este código, se realizaron ajustes para permitir su aplicación a las dos estructuras de datos de tres vías tanto para  $\mathbf{X}$  como para  $\mathbf{Y}$ .

## 7.2. Recomendaciones

En este trabajo, el análisis se limitó a explorar la relación entre la pobreza y la productividad en Colombia durante los años 2018 a 2021, debido a la disponibilidad de los datos. No obstante, se propone ampliar y enriquecer este análisis con estructuras de datos que abarquen un período temporal más amplio en su tercer modo. La intención es captar una perspectiva más completa de los cambios y las tendencias a lo largo del tiempo, específicamente en la

relación entre la pobreza y la productividad. Este enfoque es crucial considerando que estos cambios suelen presentarse en ciclos gubernamentales, que generalmente se extienden a lo largo de varios años.

Asimismo, para futuras investigaciones, los métodos pueden complementarse eficazmente con el Análisis de Clúster como una técnica de agrupación posterior, además, evaluar la posibilidad de realizar un análisis comparativo a partir de la implementación de medidas como RV de Escoufier entre coordenadas del AFM vs coordenadas del PLS.

Se recomienda aplicar otras técnicas como t-SNE, una herramienta de Machine Learning, así como STATIS y Análisis de Correlación Canónica, para evaluar el grado de correlación entre las componentes derivadas de dos estructuras de tres vías. Estas técnicas podrían enriquecer las metodologías aplicables a este tipo de estructuras de datos, que son cada vez más comunes.

Dada la disponibilidad de datos, resulta pertinente explorar el desarrollo de análisis en estructuras de datos de tres vías que incluyan la consideración de datos faltantes. Adicionalmente, sería valioso investigar la asociación dentro de estas estructuras utilizando técnicas como PLS en combinación con modelos generalizados.

Además, dado que la técnica NPLS aún es relativamente nueva, es imperativo continuar su investigación y desarrollo tanto en el ámbito teórico como en su implementación práctica. Esto incluye la mejora de la librería sNPLS en R y la incorporación de medidas complementarias como contribuciones, valores propios, cosenos cuadrados, entre otras.

# Referencias

- Abascal, E., y Landaluce, M. (2002). Análisis factorial múltiple como técnica de estudio de la estabilidad de los resultados de un análisis de componentes principales.
- Abdi, H., Williams, L. J., Valentin, D., y Bennani-Dosse, M. (2012). Statis and distatis: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2), 124–167.
- ACM. (2021). *Minería en cifras* (Reporte técnico). Asociación Colombiana de Minería.
- Ajno, L. G. C. (2020). Descomposición tensorial tucker3 aplicado a tablas de contingencias de tres vías. *Revista Varianza*, 20–28.
- Alciaturi, C. E., Escobar, M. E., De La Cruz, C., y Rincón, C. (2003). Partial least squares (pls) regression and its application to coal analysis. *Rev. Téc. Ing. Univ. Zulia*, 26(3), 197–204.
- Arce Antón, M., y cols. (2021). Estrategias quimiométricas aplicadas a técnicas cromatográficas para el análisis de monómeros y aditivos.
- Aria, M., y Cuccurullo, C. (2021). Bibliometrix: Comprehensive science mapping analysis. *R Core Team: Vienna, Austria*.
- Arias, D. (2022). *Análisis de tablas múltiples para la caracterización de la relación entre la pobreza y la actividad económica de los departamentos de colombia en el año 2018*. (Tesis de grado, Universidad del Valle.)
- Arias Ramírez, R., Sánchez, L., y Rodríguez, M. (2020). Pobreza y desigualdad en costa rica: una mirada más allá de la distribución de los ingresos. *Revista Estudios del Desarrollo Social: Cuba y América Latina*, 8(1).
- Barahona, I. (2018). Poverty in mexico: Its relationship to social and cultural indicators. *Social Indicators Research*, 135(2), 599–627.
- Bautista, G. (2009). *Comparación de los métodos tucker 3 y análisis factorial múltiple para el análisis de datos tres vías* (Tesis Doctoral no publicada). PhD thesis, Tese de Mestrado, Facultad de Ciencias da Universidad Nacional . . . .

- Braeken, J., y van Assen, M. (2017). An empirical kaiser criterion. *Psychological Methods*, 22, 450–466.
- Bro, R. (1996). Multiway calibration. multilinear pls. *Journal of chemometrics*, 10(1), 47–61.
- Bro, R. (1997). Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2), 149–171. Descargado de <https://www.sciencedirect.com/science/article/pii/S0169743997000324> doi: [https://doi.org/10.1016/S0169-7439\(97\)00032-4](https://doi.org/10.1016/S0169-7439(97)00032-4)
- Bro, R., y Kiers, H. A. (2003). A new efficient method for determining the number of components in parafac models. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(5), 274–286.
- Bro, R., y Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, 6(9), 2812–2831.
- Bro-Jorgensen, R. (1998). Multiway analysis in the food industry. models, algorithms and applications.
- Camacho, J., Picó, J., y Ferrer, A. (2007, 07). Self-tuning run to run optimization of fed-batch processes using unfold-pls. *AIChE Journal*, 53, 1789 - 1804. doi: 10.1002/aic.11205
- Camiz, S. (2011). Elections in the province of brescia: concurrent use of three-way data analysis techniques. *REVCIUNI*, 14(2), 26–42.
- Campo, J., y Sarmiento, V. (2011). Un modelo de corrección de errores para la relación entre el consumo de energía y el pib en colombia (1970-2009). *Revista Finanzas y Política Económica*, 3(1), 59–70.
- Cardona, D., González, J., Rivera, M., y Cárdenas, E. (2013). Aplicación de la regresión lineal en un problema de pobreza. *Interacción*, 12, 73–84.
- Carroll, J., y Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, (35), 283–319.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1, 245–276.
- Chessel, D., Dufour, A.-B., Dray, S., y cols. (2009). Analysis of ecological data: exploratory and euclidean methods in environmental sciences. *Version 1.4-14. 2 October 2010; 1: 4, 11.*

- Cuarita Ajno, L. G. (2020, nov.). Descomposición tensorial tucker3 aplicado a tablas de contingencias de tres vías. *REVISTA VARIANZA*, 17(17), 20–28. Descargado de <https://ojs.umsa.bo/ojs/index.php/revistavarianza/article/view/414>
- DANE. (2022a). *Boletín técnico: Producto interno bruto (pib) 2020*. (Boletín Técnico DANE, Febrero 2022. <http://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-trimestrales/pib-informacion-tecnica>)
- DANE. (2022b). *Pobreza multidimensional - resultados 2021*. ([https://www.dane.gov.co/files/investigaciones/condiciones\\_vida/pobreza/2021/presentacion-rueda-de-prensa-pobreza-multidimensional-21.pdf](https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2021/presentacion-rueda-de-prensa-pobreza-multidimensional-21.pdf))
- DANE. (2022c). Proyecciones de población.
- Decaëns, T., Jiménez, J. J., y Rossi, J.-P. (2009). A null-model analysis of the spatio-temporal distribution of earthworm species assemblages in colombian grasslands. *Journal of Tropical Ecology*, 25(4), 415–427. doi: 10.1017/S0266467409006075
- Deming, S., Michotte, Y., Massart, D., Kaufman, L., y Vandeginste, B. (1988). *Chemometrics: A textbook*. Elsevier Science. Descargado de <https://books.google.com.co/books?id=G8JMac7OCtAC>
- DeSarbo, W. S. (1984). An application of parafac to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. *Research Methods for Multimode Data Analysis* p, 602–642.
- Dolédéc, S., y Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshwater biology*, 31(3), 277–294.
- Escofier, B., y cols. (1984). L’analyse factorielle multiple. *Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche*, 42, 3–68.
- Escofier, B., y Pagès, J. (1992a). *Análisis factoriales simples y múltiples: objetivos, métodos e interpretación*. Universidad del País Vasco Bilbao.
- Escofier, B., y Pagès, J. (1992b). *Análisis factoriales simples y múltiples: objetivos, métodos e interpretación*. España: Universidad del País Vasco.
- Escofier, B., y Pagès, J. (1998). Analyses factorielles simples et multiples. *Dunod, Paris*, 284, 104–107.
- Feki-Sahnoun, W., Hamza, A., Béjaoui, B., Mahfoudi, M., Rebai, A., y Bel Hassen, M. (2018). Multi-table approach to assess the biogeography of phytoplankton: ecological and management implications. *Hydrobiologia*, 815, 229–251.

- García A, C. (2021). *Análisis factorial múltiple para describir las condiciones de salud sentidas de la población priorizada de la ciudad de cali en el año 2018*. (Tesis de Master). Descargado de <https://hdl.handle.net/10893/20874>
- Garthwaite, P. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425), 122–127. Descargado de <https://www.tandfonline.com/doi/abs/10.1080/01621459.1994.10476452> doi: 10.1080/01621459.1994.10476452
- Gaviria, C. A. (2016). Regresión por mínimos cuadrados parciales pls aplicada a datos variedad valuados. *Facultad de Ciencias, Escuela de Estadística Medellín*.
- González, V. M. (2014). *Análisis conjunto de múltiples tablas de datos mixtos mediante pls* (Tesis Doctoral no publicada). University of Science and Technology of Languedoc.
- Gourvénec, S., Tomasi, G., Durville, C., Di Crescenzo, E., Saby, C., Massart, D., ... Openheim, G. (2005). Cubatch, a matlab® interface for n-mode data analysis. *Chemometrics and Intelligent Laboratory Systems*, 77(1-2), 122–130.
- Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1), 33–51.
- Gutiérrez, Y. Á. T., Ramos, M. C. M., y Morinson, M. L. V. (2022). Análisis multivariado de los factores socio-económicos asociados al rendimiento en las pruebas saber pro 2016: El caso de los estudiantes de licenciatura en matemáticas en colombia: Multivariate analysis of the socioeconomic factors associated to the performance in the pro saber 2016 tests: The case of the students of the bachelor of mathematics in colombia. *South Florida Journal of Development*, 3(1), 928–939.
- Hanafi, M., Qannari, E. M., y Jaillais, B. (2020). Multi-block and three-way data analysis.
- Harshman, R. A. (1970). Foundations of the parafac procedure: models and conditions for an “explanatory” multi-mode factor analysis. *UCLA Working Papers in Phonetics*, (16), 1-84. (Citado por Kiers (1991).)
- Henrion, R. (1994). N-way principal component analysis : Theory, algorithms and applications. *Chemometrics and Intelligent Laboratory Systems*, 25, 1-23. Descargado de <https://api.semanticscholar.org/CorpusID:123430826>
- Hernández, E. J. M., Ibañez, L. F. S., y Monterrosa, A. R. D. (2021). Perspectiva multidimensional de la pobreza en los hogares colombianos. *Sociedad & Economía*(44), 3.
- Herrera, J., y Díaz, G. (2002). *Metodología de un modelo arima condicionado para el pronóstico del pib* (Inf. Téc.). Departamento Nacional de Planeación.

- Hervás, D., Prats-Montalbán, J., Lahoz, A., y Ferrer, A. (2018). Sparse n-way partial least squares with r package snpls. *Chemometrics and Intelligent Laboratory Systems*, 179, 54-63. Descargado de <https://www.sciencedirect.com/science/article/pii/S0169743918302387> doi: <https://doi.org/10.1016/j.chemolab.2018.06.005>
- Hoo, K., Piovoso, M., Dahl, K., MacGregor, J., y Nomikos, P. (1994, 01). Multiway pca applied to an industrial batch process. En (Vol. 94, p. 1294 - 1298 vol.2). doi: 10.1109/ACC.1994.752268
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.
- Husson, F., Josse, J., Le, S., y Mazet, J. (2013). Factominer: multivariate exploratory data analysis and data mining with r. *R package version*, 1(1.29).
- Jaffrenou, P.-A. (1978). *Sur l'analyse des familles finies de variables vectorielles: Bases algebriques et application a la description statistique* (Tesis Doctoral no publicada).
- Jiang, J.-H., Wu, H.-L., Li, Y., y Yu, R.-Q. (1999). Alternating coupled vectors resolution (acover) method for trilinear analysis of three-way data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 13(6), 557–578.
- Kamstrup-Nielsen, M. H., Johnsen, L. G., y Bro, R. (2013). Core consistency diagnostic in parafac2. *Journal of Chemometrics*, 27(5), 99–105.
- Kassambara, A. (2016). Factoextra: extract and visualize the results of multivariate data analyses. *R package version*, 1.
- Kroonenberg, P. (1983). *Three-mode principal component analysis: Theory and applications* (Vol. 42). doi: 10.2307/2531268
- Kroonenberg, P. (1989). The analysis of multiple tables in factorial ecology. iii. three-mode principal component analysis: "analyse triadique complete". *ACTA OECOL.(OECOL. GEN.)*, 10(3), 245–256.
- Kroonenberg, P. M. (1983). *Three-mode principal component analysis: Theory and applications* (Vol. 2). DSWO press.
- kumar Maji, J., Shukla, V., y Prajapati, P. (2014). Quantitative determination in three-way calibration strategies with hyphenated-data (chromatography-spectroscopy) of polyherbal-herbomineral formulation. , 5.
- Lavit, C., Bernard, M.-C., Pérez-Hugalde, C., y Escoufier, Y. (1988). *Analyse conjointe de tableaux quatitatifs*. Editorial Masson.

- Lavit, C., Escoufier, Y., Sabatier, R., y Traissac, P. (1994). The act (statis method). *Computational Statistics and Data Analysis*, (18), 97–119.
- Leardi, R. (2005). *Multi-way analysis with applications in the chemical sciences, age smilde, rasmus bro and paul geladi, wiley, chichester, 2004, isbn 0-471-98691-7, 381 pp.* Wiley Online Library.
- Lebart, L., Morineau, A., y Piron, M. (1995). *Statistique exploratoire multidimensionnelle* (Vol. 3). Dunod Paris.
- Luo, L., Bao, S., Mao, J., y Tang, D. (2016). Quality prediction and quality-relevant monitoring with multilinear pls for batch processes. *Chemometrics and Intelligent Laboratory Systems*, 150, 9-22. Descargado de <https://www.sciencedirect.com/science/article/pii/S0169743915002749> doi: <https://doi.org/10.1016/j.chemolab.2015.11.004>
- Márquez Ruiz, C. (2017). Modelo de regresión pls.
- Marrugo-Arnedo, C., Del Risco-Serje, K., Marrugo-Arnedo, V., Herrera-Llamas, J., y Pérez-Valbuena, G. (2015). Determinants of poverty in the colombian caribbean region. *Revista de Economía del Caribe*(15).
- Marrugo-Arnedo, V. (2013). Crecimiento económico y desarrollo humano en colombia (2000-2010). *Revista de economía del Caribe*(11), 127–143.
- Martens, M. (2001, 10). Multivariate analysis of quality. an introduction. *Measurement Science Technology - MEAS SCI TECHNOL*, 12, 1746-1746. doi: 10.1088/0957-0233/12/10/708
- Martí, R. B., y Baldrich, J. F. (2015). Fundamentals of parafac. En *Data handling in science and technology* (Vol. 29, pp. 7–35). Elsevier.
- Medina-Hernández, E. J., Fernández-Gómez, M. J., y Barrera-Mellado, I. (2021). Analysis of time use surveys using co-statis: A multiway data analysis of gender inequalities in time use in colombia. *Sustainability*, 13(23), 13073.
- Mejia, R. (2023). *Comportamiento de la mortalidad según las agrupaciones de causas, lista 6/67 ops-oms durante los años 2018-2021 a nivel departamental en colombia.* (Thesis de grado, Universidad del Valle.)
- Nomikos, P., y MacGregor, J. F. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40(8), 1361–1375.
- Ochoa, A. F. (2018). Análisis de correspondencias múltiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo nipals (acmpdd).

- Olaya, J., Ippolito, K., Moreno, G., González, V., Klinger, R., Parra, C. M., y Marmolejo, L. (2013). Asociaciones entre la composición socioeconómica familiar y la generación urbana de residuos sólidos domiciliarios. *Revista EIA*, 10(20), 127–137.
- Pearson, K. (1901). Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2), 559.
- Plaza, A. (2022). *Análisis multivariado de datos para la evaluación de la dinámica de la pobreza monetaria en colombia por departamentos durante el período 2012-2020*. (Thesis de grado, Universidad del Valle.)
- Preda, C., Saporta, G., y Lévéder, C. (2007). Pls classification of functional data. *Computational Statistics*, 22(2), 223–235.
- Quinde, V., Bucaram, R., Bucaram, M., y Silvera, C. (2019). Producto interno bruto en américa latina y el caribe: Relaciones entre crecimiento económico y sustentabilidad ambiental. *Revista Venezolana de Gerencia*, 24(87), 769–784.
- R Core Team. (2020). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de <https://www.R-project.org/>
- Robert, P., y Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 25(3), 257–265.
- Ruiz, J., y Bonilla, J. (2018, 11). *Análisis multivariante de datos en tres vías para caracterizar la dinámica económica de los departamentos en colombia en el periodo 2000-2015*. (Trabajo de grado, Universidad del Valle)
- Sánchez, G. (2016). plsdepot: Partial least squares (pls) data analysis methods version 0.1.17. *R package*.
- Siberchicot, A., Julien-Laferrrière, A., Dufour, A.-B., Thioulouse, J., y Dray, S. (2017). adegraphics: an s4 lattice-based package for the representation of multivariate data. *R Journal*, 9(2).
- Singh, K. P., Malik, A., Singh, V. K., y Sinha, S. (2006). Multi-way data analysis of soils irrigated with wastewater—a case study. *Chemometrics and Intelligent Laboratory Systems*, 83(1), 1–12.
- Slimani, N., Guilbert, E., Ayni, F. E., Jrad, A., Boumaiza, M., y Thioulouse, J. (2017). The use of statico and costatis, two exploratory three-ways analysis methods: an application to the ecology of aquatic heteroptera in the medjerda watershed (tunisia). *Environmental and Ecological Statistics*, 24, 269–295.

- Smilde, A. K. (2001). Comments on three-way analyses used for batch process data. *A Journal of the Chemometrics Society*, 15, 19–27.
- Stanimirova, I., Walczak, B., Massart, D., Simeonov, V., Saby, C., y Di Crescenzo, E. (2004). Statis, a three-way method for data analysis. application to environmental data. *Chemometrics and intelligent laboratory systems*, 73(2), 219–233.
- Swiercz, M., y Mroczkowska, H. (2020). Multiway pca for early leak detection in a pipeline system of a steam boiler—selected case studies. *Sensors*, 20(6), 1561.
- Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods.
- Thioulouse, J., y Chessel, D. (1987). Les analyses multitableaux en ecologie factorielle. *Acta Ecologica, (Ecologia Generalis)*, 8(4), 463–480.
- Thioulouse, J., Simier, M., y Chessel, D. (2004). Simultaneous analysis of a sequence of paired ecological tables. *Ecology*, 85(1), 272–283.
- Vaquer Rivera, L., y cols. (2022). Aplicaciones del método pls en quimiometría.
- Vargas, D. (2018). Relación entre el pib per cápita y la pobreza monetaria a nivel departamental en colombia 2002-2014. *Universidad de La Salle. Facultad de Ciencias Económicas y Sociales. Economía*.
- Vega-Vilca, J. C., y Guzmán, J. (2011). Regresión pls y pca como solución al problema de multicolinealidad en regresión múltiple. *Revista de Matemática: Teoría y Aplicaciones*, 18(1), 9–20.
- Virguez, J. (2019). *Modelo de pronóstico pib financiero*. (Thesis en especialización en Estadística Aplicada. Fundación Universitaria Los Libertadores. Sede Bogotá)
- Westerhuis, J., Kourti, T., y MacGregor, J. (1999). Comparing alternative approaches for multivariate statistical analysis of batch process data. *A Journal of the Chemometrics Society*, 13, 397–413.
- Wold, S., Sjöström, M., y Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2), 109–130.



# A. Anexo A: Diccionario de datos de las matrices IPM y PIB

## A.1. Diccionario de datos - Producto Interno Bruto

| Variable | Descripción                                                                                                                                                                                                                                                                  | Tipo de dato | Unidad de medida           | Años      | Fuente |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|----------------------------|-----------|--------|
| Agri     | Agricultura, ganadería, caza, silvicultura y pesca                                                                                                                                                                                                                           | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Min      | Explotación de minas y canteras                                                                                                                                                                                                                                              | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Manu     | Industrias manufactureras                                                                                                                                                                                                                                                    | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Elec     | Suministro de electricidad, gas, vapor y aire acondicionado; distribución de agua; evacuación y tratamiento de aguas residuales, gestión de desechos y actividades de saneamiento ambiental                                                                                  | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Cons     | Construcción                                                                                                                                                                                                                                                                 | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Com      | Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas; transporte y almacenamiento; alojamiento y servicios de comida                                                                                                                     | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Comu     | Información y comunicaciones                                                                                                                                                                                                                                                 | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Fina     | Actividades financieras y de seguros                                                                                                                                                                                                                                         | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Inmo     | Actividades inmobiliarias                                                                                                                                                                                                                                                    | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Prof     | Actividades profesionales, científicas y técnicas; actividades de servicios administrativos y de apoyo                                                                                                                                                                       | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Soc      | Administración pública y defensa; planes de seguridad social de afiliación obligatoria; educación; actividades de atención de la salud humana y de servicios sociales                                                                                                        | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Recre    | Actividades artísticas, de entretenimiento y recreación y otras actividades de servicios; actividades de los hogares individuales en calidad de empleadores; actividades no diferenciadas de los hogares individuales como productores de bienes y servicios para uso propio | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |
| Imp      | Impuestos                                                                                                                                                                                                                                                                    | Númerico     | Miles de millones de pesos | 2018-2021 | DANE   |

**Tabla A-1.:** Diccionario de datos de las variables de la matriz del PIB

## A.2. Diccionario de datos - Índice de Pobreza Multidimensional

| Variable   | Descripción                                              | Tipo de dato | Unidad de medida | Años      | Fuente |
|------------|----------------------------------------------------------|--------------|------------------|-----------|--------|
| ANALFAB    | Analfabetismo                                            | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| B_LEDU     | Bajo logro educativo                                     | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| B_SERVCPI  | Barreras a servicios para cuidado de la primera infancia | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| B_ACCSS    | Barreras de acceso a servicios de salud                  | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| DS_LD      | Desempleo de larga duración                              | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| HC         | Hacinamiento crítico                                     | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| INAD_EE    | Inadecuada eliminación de excretas                       | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| INA_ESC    | Inasistencia escolar                                     | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| MT_PEXT    | Material inadecuado de paredes exteriores                | Número       | Porcentaje       | 2018-2021 | DANE   |
| MT_PISOS   | Material inadecuado de pisos                             | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| REZ_ESC    | Rezago escolar                                           | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| SIN_FAGUA  | Sin acceso a fuente de agua mejorada                     | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| SIN_ASALUD | Sin aseguramiento en salud                               | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| TRAB_INFAN | Trabajo infantil                                         | Numérico     | Porcentaje       | 2018-2021 | DANE   |
| TRAB_INFOR | Trabajo informal                                         | Numérico     | Porcentaje       | 2018-2021 | DANE   |

**Tabla A-2.:** Diccionario de datos de las variables de la matriz del IPM