**Innovación en las TIC’s y su incidencia en el direccionamiento estratégico para pymes de Santiago de Cali**

**Metodología**

Etapas:

1. Organización de datos.
2. Análisis de información sin tratamiento.
3. Análisis e imputación de datos faltantes.
4. Análisis de información imputada.

Se espera que los resultados obtenidos en las etapas 2 y 4 no varíen significativamente, es decir, conserven la estructura y las relaciones entre las variables estudiadas.

1. Ajuste y evaluación del modelo PLS-PM.

**Etapa 1. Organización de datos**

Codificación de variables de acuerdo a su escala de medición:

* Para variables categóricas utilizar las escalas, nominal y ordinal.
* Para variables con opción de respuesta múltiple, generar tantas variables nominales como opciones de respuesta se disponga.

Para todo el conjunto de variables en la base de datos se omitieron las categorías asociadas a mediciones negativas o ambiguas. Por ejemplo, categorías de respuesta: NO SABE, NO TIENE, NO APLICA; ya que se consideró que estas categorías de respuesta no aportan significativamente a la explicación de los constructos definidos.

Por otro lado, se añadió la restricción de conservar variables con un porcentaje de datos faltantes inferior al 15% (en términos absolutos, 16 datos faltantes como máximo por variable). Esto con el fin de realizar la posterior imputación de los mismos sin afectar la calidad de la información imputada y la estructura de la matriz de datos.

Finalmente, se eliminó del estudio la información correspondiente a una empresa dedicada a la *Reparación y fabricación de piezas para maquina industrial,* ya que no cuenta con red de datos, lo que implica que la sección de REDES en el cuestionario carece de todas las respuestas para las variables asociadas; por tanto, la muestra definitiva queda conformada por 106 PYMES de la ciudad.

**Etapa 2. Análisis de información sin tratamiento**

Especificación de constructos a analizar. Para estudiar el papel que desempeña la innovación tecnológica sobre el direccionamiento estratégico de la población objetivo, se definieron a priori los siguientes constructos o variables latentes (medidas a partir de ítems del cuestionario) y su papel en el modelo estructural:

Tabla 1. Variables a estudiar y su papel en el modelo estructural

|  |  |
| --- | --- |
| **Constructo** | **Papel a desempeñar** |
| Innovación en Hardware | Variables predictoras |
| Innovación en Software | Variables predictoras |
| Innovación en Redes y Comunicaciones | Variables predictoras |
| Direccionamiento estratégico | Variables de respuesta |
| Control | Variables de control |

Dado que el interés del estudio radica en la exploración de relaciones entre los constructos definidos, como primera medida se examinaron las relaciones entre todos los posibles pares de variables mediante ***Test de Independencia Chi-cuadrado*** tomando como nivel de significancia de la prueba, α=5%.

**Etapa 3. Análisis e imputación de datos faltantes**

En la presente etapa se verificó la proporción de datos faltantes por variable (imponiendo una restricción como máximo de 15% de datos faltantes) y por número de casos, con el propósito de determinar la influencia que se obtendría con la imputación de dicha información faltante.

Esta fase de análisis se desarrolló mediante procedimientos de imputación múltiple, la cual en términos generales permite reemplazar cada valor ausente por un conjunto de *M* valores plausibles, completando la matriz de datos original de *M* formas, es decir, a partir de la matriz original se generan *M* matrices completas.

Específicamente, se hizo uso de la técnica de imputación múltiple de datos: ***Multivariate imputation by chained equations*** (Van Buuren & Groothuis-Oudshoorn, 2011) la cual realiza la imputación de los valores faltantes a partir de regresiones secuenciales por variable, las cuales preservan las relaciones iniciales en los datos y su incertidumbre asociada, permitiendo manejar diferentes tipos de patrones de datos faltantes mediante un algoritmo iterativo que permite evaluar en cada paso la consistencia de las estimaciones obtenidas y manejar variables en diferentes escalas de medición (numéricas o categóricas). Aquí definimos un número de imputaciones múltiples *M*=20.

No obstante, para simplificar el número de análisis posteriores se propone una variación del método de Rubin (Rubin, 1987), reduciendo a priori el conjunto de 20 matrices de datos con los valores imputados a una sola, asignando para cada caso donde se presenta un dato faltante el valor máximo de las categorías estimadas en el proceso de imputación.

**Etapa 4. Análisis de información imputada**

Para comparar los resultados obtenidos con la información original e imputada de la matriz de asociación, se efectuó un ***Test de Mantel*** para analizar la correlación entre las matrices de asociación usando el coeficiente de correlación no paramétrico de Spearman.

Posterior a esto, se realizó un proceso de cuantificación óptima de la matriz de datos imputados mediante un ***Análisis de Componentes Principales no Lineal*** (Meulman & Heiser, 1995), esto con el propósito de cuantificar las categorías de respuesta de las variables categóricas, lo cual facilitó en los posteriores análisis la utilización de técnicas para variables cuantitativas.

Finalmente, se realizó la aplicación del ***Análisis de Componentes Principales (PCA)***, para las variables que definen cada uno de los constructos con el fin de identificar asociaciones de interés que permitan refinar la definición de cada constructo.

**Etapa 5. Ajuste y evaluación del modelo PLS-PM**

El modelo PLS-PM el cual traduce ***Partial Least Squares – Path Modeling*** es una técnica de modelación que permite explorar relaciones entre constructos definidos como variables latentes o no observables, las cuales son calculadas a partir de ítems medidos. Hablar sobre los modelos de ecuaciones estructurales y que esta es su contraparte más simplista. Hablar también sobre cómo funciona el modelo, inputs y salidas. Así como la evaluación del modelo mediante los diferentes métodos propuestos en la literatura.

**RESULTADOS**

Como primera medida se presentan los resultados del test de independencia Chi-cuadrado, calculado para todos los posibles pares de variables. Como se puede observar en el heatmap de la Figura 1 existen pequeños parches de color rojo que corresponden a valores-p relativamente bajos, lo que indica asociaciones significativas entre grupos de variables, en este sentido utilizando un nivel de significancia α=0.05 obtenemos un porcentaje del 7.8% de asociaciones significativas entre todas las comparaciones por pares de variables, dando a entender que probablemente existen patrones de relación en los datos que es necesario explorar. Estos cálculos se realizan con la información original, es decir, sin realizar la imputación de valor faltante alguno.

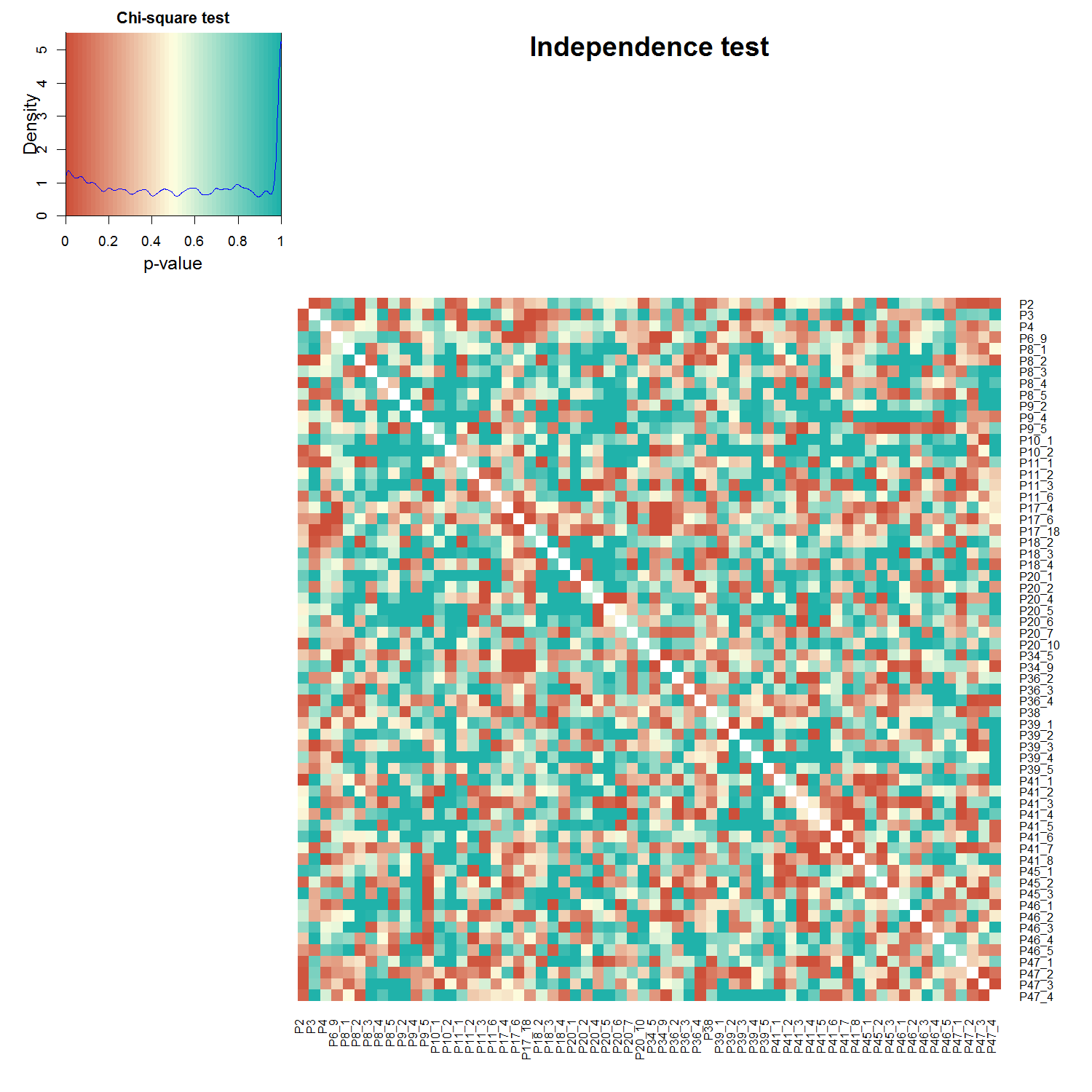


Figura 1. Valores-p correspondientes al test de independencia Chi-cuadrado entre todos los posibles pares de variables, la diagonal superior e inferior contiene los mismos valores

Por otro lado, el número de datos faltantes se evaluó desde diferentes puntos de vista, por un lado se tiene que el porcentaje de datos faltantes por variables respecto al número total de columnas[[1]](#footnote-1) equivale a un 18% de las variables; mientras la relación por casos alcanza un cuarto del total de filas de la matriz, en este sentido si se ajusta el modelo únicamente con las filas que contienen la totalidad de la información se está perdiendo un 25% de empresas que pueden aportar en la explicación de las relaciones teóricas planteadas (ver Figura 2).

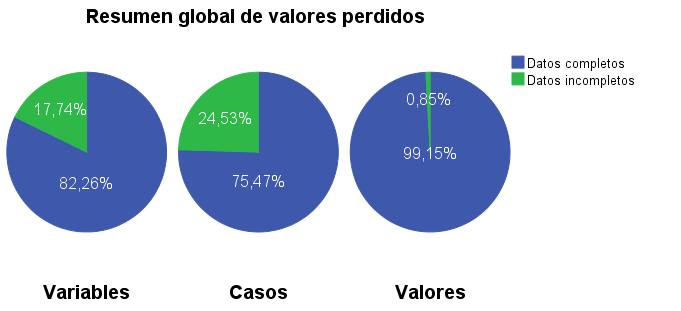


Figura 2. Porcentaje de datos faltantes desde diferentes perspectivas

Paso siguiente se realizó la exploración del patrón de datos faltantes como se exhibe en la Figura 3. Los espacios en blanco representan los vacíos de información en la matriz de datos. Como puede observarse la distribución de los valores faltantes en la matriz de información no parece presentar un patrón definido, por tanto, el proceso de imputación de datos faltantes se maneja desde un marco de referencia aleatorio.

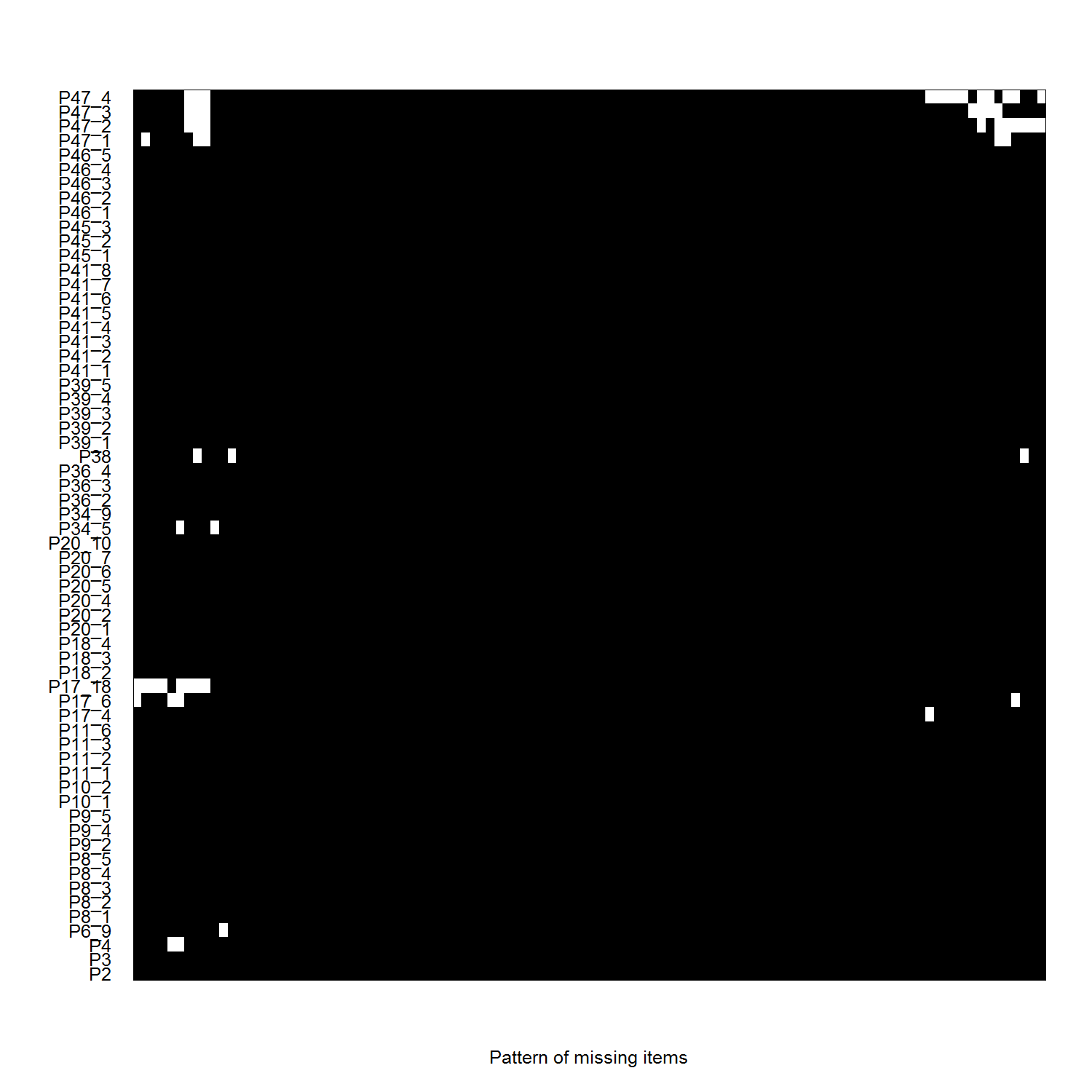


Figura 3. Patrón de datos faltantes en la matriz de datos

Para efectuar la imputación de los valores faltantes mediante el procedimiento *Multivariate imputation by chained equations* se construyeron un total de 50 matrices de datos a través de 5 iteraciones múltiples. Con este método se logra determinar la estabilidad de las imputaciones y se selecciona como matriz de análisis final el valor con mayor frecuencia por dato faltante[[2]](#footnote-2).

Con la nueva matriz de datos incluyendo los valores imputados se obtuvo un porcentaje de asociaciones significativas del 8.0% entre pares de variables (ver Figura 4).

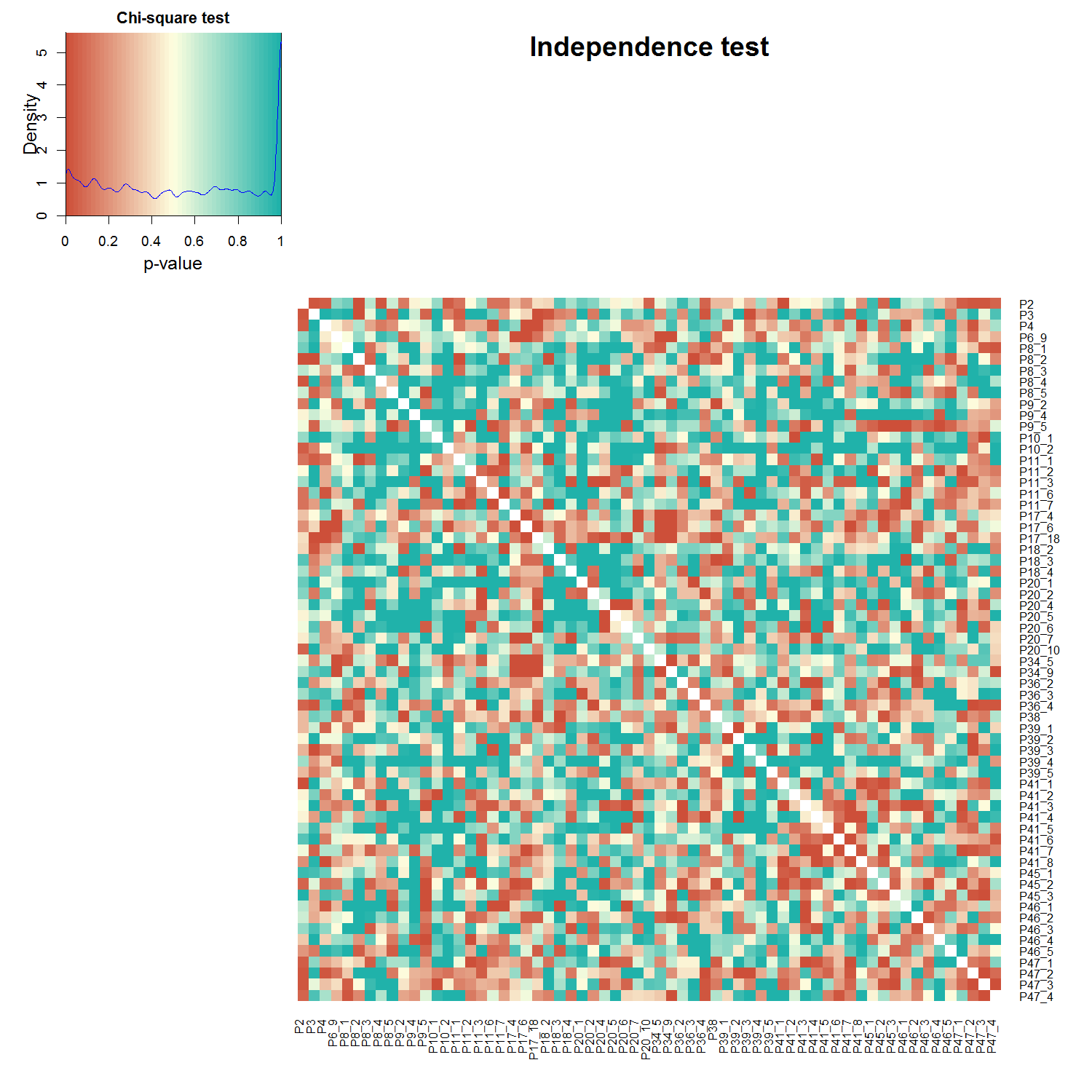


Figura 4. Valores-p correspondientes al test de independencia Chi-cuadrado entre todos los posibles pares de variables sobre la matriz de datos imputados

Adicional a esto se estudió la similaridad de las matrices de asociación a través del Test de Mantel realizando 10.000 permutaciones para obtener una aproximación realista de la significancia del test.

Mantel statistic based on Spearman's rank correlation rho

Mantel statistic r: 0.969

Significance: 9.999e-05

Obteniendo una correlación entre las matrices de asociación de datos aproximada del 97%, lo que indica una mínima perdida en la estructura de asociaciones entre las matrices de asociación de datos originales e imputados.

Para estudiar la asociación entre variables desde el punto de vista de cada uno de los constructos, se aplicó un PCA para los constructos que constituyeron las variables explicativas del modelo.

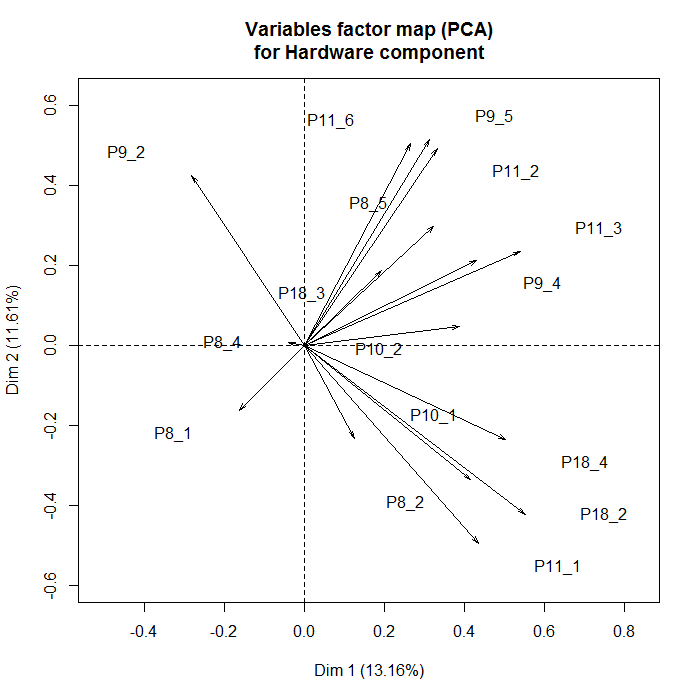


Figura 5. Mapa factorial de variables, componente HARDWARE

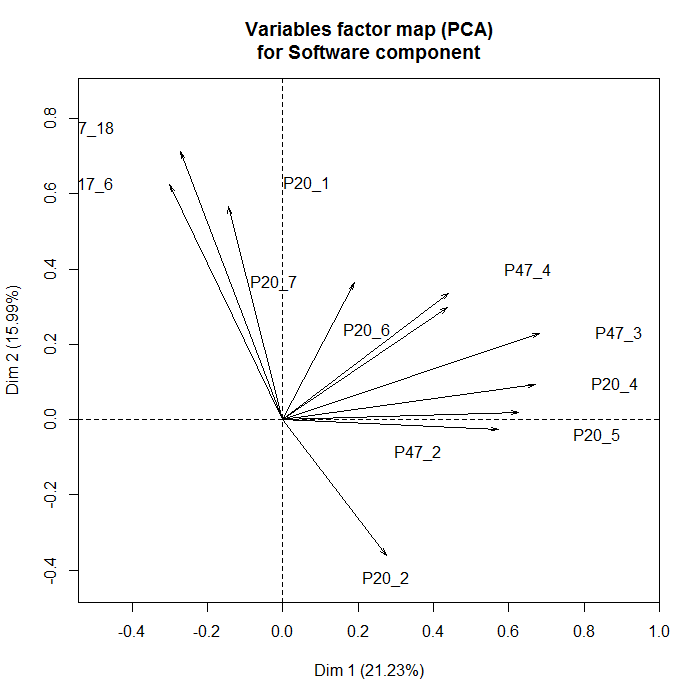


Figura 6. Mapa factorial de variables, componente SOFTWARE

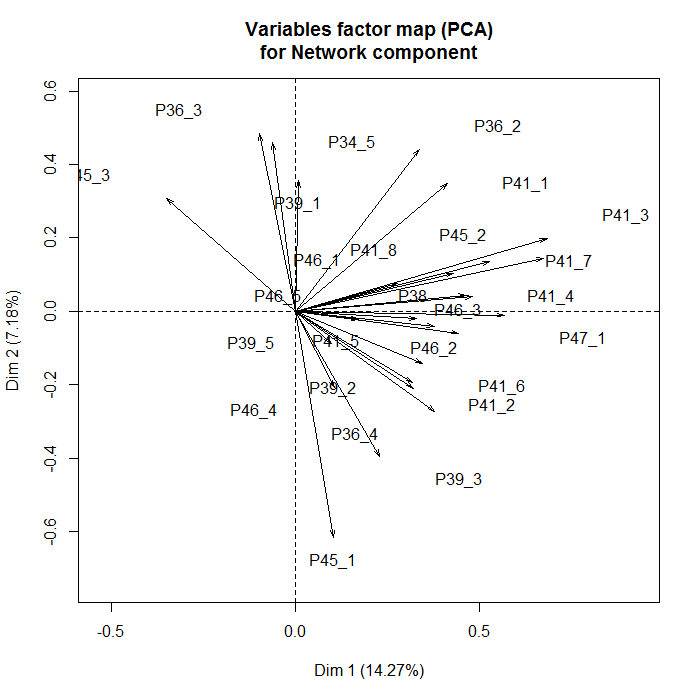


Figura 7. Mapa factorial de variables, componente REDES DE COMUNICACIÓN

Finalmente, en términos del ajuste del modelo obtenemos cargas positivas para todas las variables respecto a sus constructos, indicando así una correlación positiva entre cada una de las variables que definen la respectiva variable latente; excepto para dos variables de la componente de innovación en hardware (ver Figura 8).

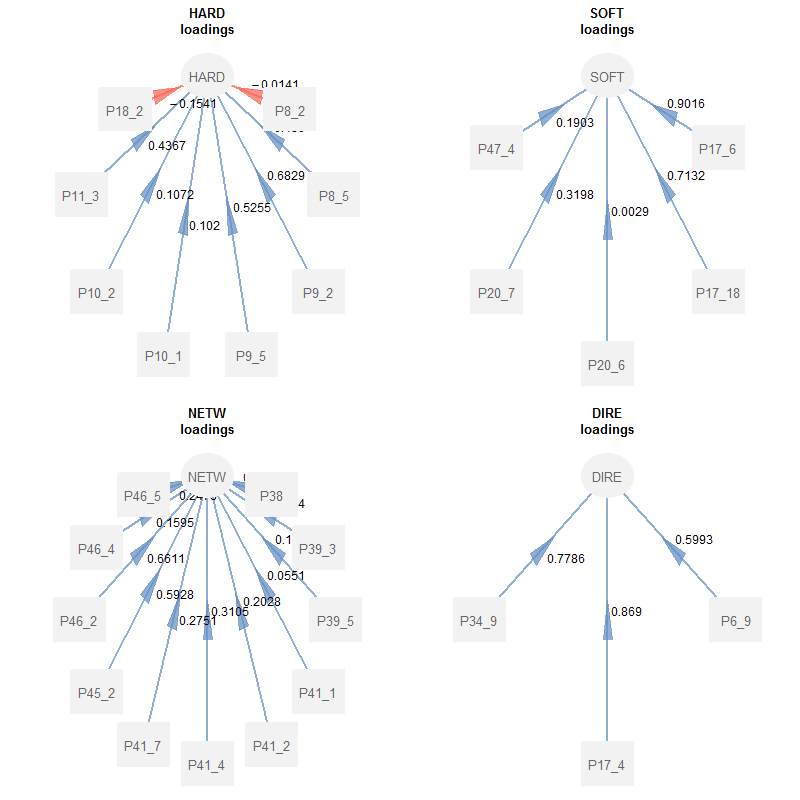


Figura 8. Cargas factoriales de las variables observadas al constructo.

Además de analizar las cargas de las variables con su respectivo constructo también se presenta a continuación las cargas cruzadas es decir la correlación que presenta cada una de las variables medidas con los demás constructos definidos, esto nos permite concluir de igual manera la adecuación de dichas variables para la explicación de las variables latentes (ver Figura 9).

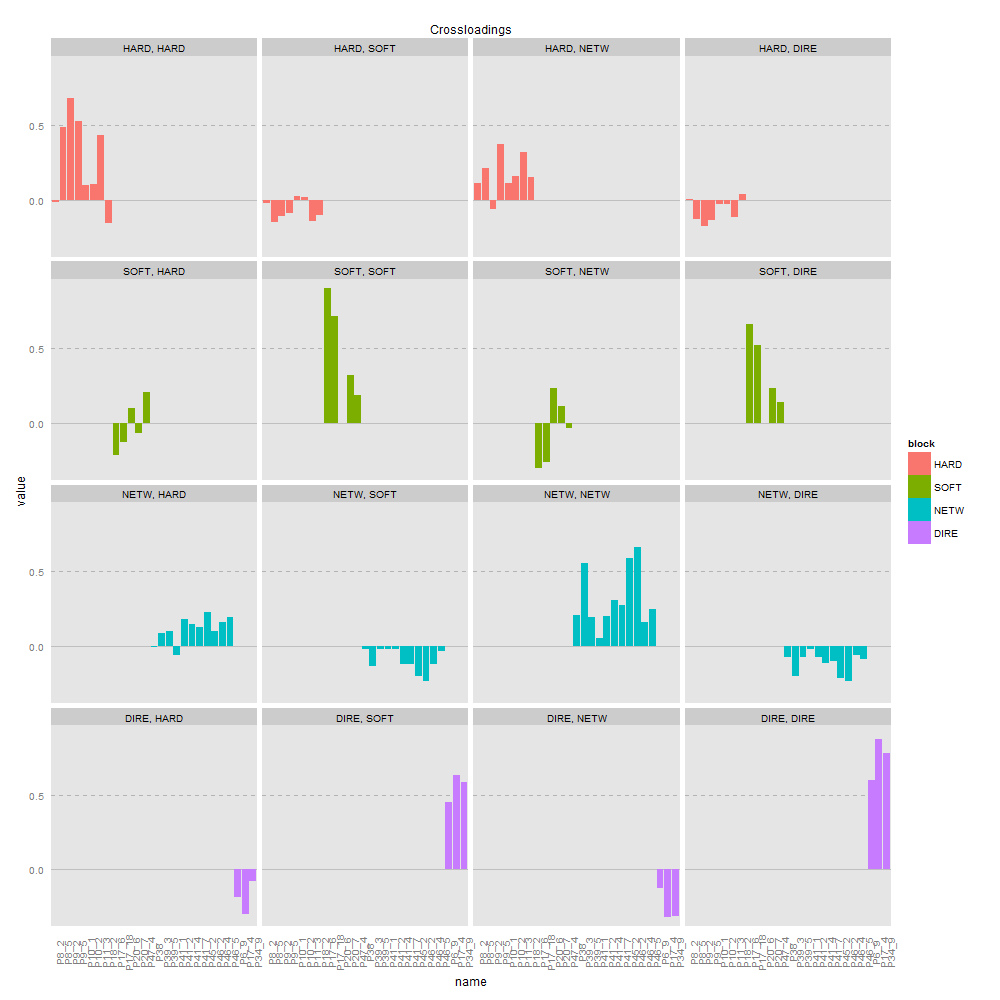


Figura 9. Cargas cruzadas entre los diferentes constructos

Es de notar que las variables que definen cada uno de los constructos muestran los niveles más altos de asociación a diferencia de la asociación que se presenta entre estos y los demás constructos, lo que permite concluir la adecuación de los cargas y variables utilizadas.

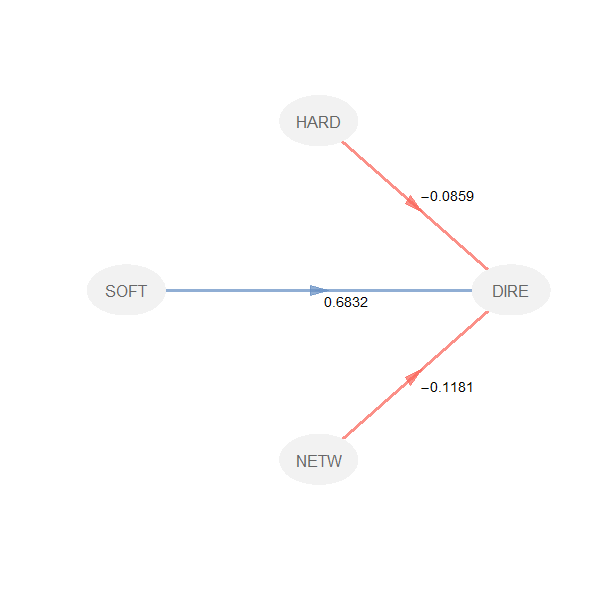


Figura 10. Coeficientes del modelo estructural

Evaluando el modelo estructural en la Figura 10, observamos el aporte positivo de la innovación en software sobre el direccionamiento estratégico de las empresas estudiadas este presenta una correlación del 68% a diferencia de la innovación en hardware y redes que aunque presentan estimaciones negativas del aporte al direccionamiento estratégico en base a los valores-p obtenidos de la significancia de los parámetros del modelo estos no pueden descartarse ser estadísticamente distintos de cero, por tanto tener un aporte bajo o nulo (Tabla 1).

Ajustar un modelo de regresión Ridge para tener en cuenta el efecto de la multicolinealidad entre las variables explicativas del modelo final, cuyas estimaciones no difieren significativamente. Probar con técnicas de regularización para ver si cambia sustancialmente los resultados del modelo. Todo esto con el fin de minimizar el efecto de la multicolinealidad entre los regresores.

Listo, al utilizar regresión Ridge obtenemos una reducción en el error de predicción del 11% en comparación con el modelo de regresión lineal convencional. Por su parte, utilizando una corrección mediante regresión LASSO obtenemos una reducción del 40%.

Tabla 2. Resumen de los parámetros del modelo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Estimate* | *Std. Error* | *t-value* | *p-value* |
| *Innovación en hardware* | -0.086 | 0.069 | -1.252 | 0.213 |
| *Innovación en software* | 0.683 | 0.069 | 9.905 | 0 |
| *Innovación en redes* | -0.118 | 0.071 | -1.669 | 0.098 |

**BIBLIOGRAFÍA**

Meulman, J. J., & Heiser, W. J. (1995). *Chapter 3 Principal Components Analysis With Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data*.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. *Harvard University*, (JOHN WILEY & SONS), 15–19. http://doi.org/10.1002/9780470316696

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Multivariate Imputation by Chained Equations. *Journal Of Statistical Software*, *45*(3), 1–67. http://doi.org/10.1177/0962280206074463

1. Esto quiere decir: el número de variables que contienen al menos un dato faltante respecto al número total de columnas. [↑](#footnote-ref-1)
2. Por cada dato faltante se obtienen múltiples valores imputados y se selecciona como valor definitivo la categoría con mayor frecuencia. [↑](#footnote-ref-2)