

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE MATEMÁTICA



UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Construcción y análisis de indicadores para la caracterización de patrones de uso a partir de registros de actividad digital

Autor: Gabriel Moya Valenzuela

Profesor Guía: Gonzalo Contador

Supervisor Empresa: Cristobal Uslar

Memoria para optar al título de
Ingeniería Civil Matemática

Santiago, Chile

Octubre 2026



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Construcción y análisis de indicadores para la caracterización de patrones de uso a partir de registros de actividad digital

Nombre del candidato(a): Gabriel Francisco Manuel Moya Valenzuela

Carrera / Grado: Ingeniería Civil Matemática

Campus: San Joaquín

Departamento: Departamento de Matemáticas

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Gonzalo Contador, en mi calidad de profesor guía del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 06/10/2026

Firma: 

Estudiante o Candidato(a):

Fecha: 05 / 10 / 2026

Firma: 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Construcción y análisis de indicadores para la caracterización de patrones de uso a partir de registros de actividad digital

AUTOR: Gabriel Francisco Manuel Moya Valenzuela

TRABAJO DE MEMORIA, presentado como requisito parcial para optar al título profesional de Ingeniero Civil Matemático de la Universidad Técnica Federico Santa María.

COMISIÓN EVALUADORA

Gonzalo Contador
Universidad Técnica Federico Santa María

Nicolas Carreño
Universidad Técnica Federico Santa María

Cristobal Uslar
Supervisor Pares & Alvarez

Santiago, Chile
Octubre 2026

Resumen

Esta memoria analiza registros de actividad generados por AutoCAD, Revit y Time-Doctor con el objetivo de construir variables que permitan describir y comparar distintos patrones de uso entre usuarios. A partir de estos registros se definieron cinco variables: un perfil general de actividad, concentración horaria, diversidad documental, concentración documental y fragmentación temporal.

Para identificar perfiles se compararon K-Means y Gaussian Mixture Models (GMM), seleccionándose GMM por entregar grupos más adecuados para su interpretación. Luego se estudió el comportamiento de los indicadores y su relación con los perfiles mediante análisis gráficos y pruebas estadísticas.

Los resultados muestran que la diversidad y la concentración documental son los indicadores que presentan la relación más consistente con los perfiles generales. La concentración horaria y la fragmentación temporal también permiten describir diferencias entre usuarios, aunque su relación con los perfiles es menos uniforme entre plataformas.

En conjunto, el trabajo muestra que es posible transformar registros operacionales en variables interpretables a nivel de usuario, permitiendo identificar diferencias, relacionar distintas dimensiones de la actividad y caracterizar patrones de uso.

Abstract

This thesis analyzes activity records generated by AutoCAD, Revit and TimeDoctor in order to build variables that allow different user behavior patterns to be described and compared. From these records, five variables were defined: a general activity profile, hourly concentration, document diversity, document concentration and temporal fragmentation.

To identify activity profiles, K-Means and Gaussian Mixture Models (GMM) were compared, with GMM being selected because it produced groups that were more suitable for interpretation. The behavior of the indicators and their relationship with the identified profiles were then analyzed using graphical analysis and statistical tests.

The results show that document diversity and document concentration are the indicators with the most consistent relationship with the general activity profiles. Hourly concentration and temporal fragmentation also describe differences among users, although their relationship with the profiles is less consistent across platforms.

Overall, the study shows that operational records can be transformed into interpretable user-level variables, making it possible to identify differences, relate different dimensions of activity and characterize usage patterns.

Keywords: digital activity records, indicators, clustering, GMM, usage patterns, user analysis.

Agradecimientos

A mi familia y a la gente que me apoyó.

Índice general

1. Introducción	1
1.1. Contexto	1
1.2. Problema de estudio	2
1.3. Objetivos	3
1.3.1. Objetivo general	3
1.3.2. Objetivos específicos	3
1.4. Alcances y limitaciones	4
2. Datos utilizados	5
2.1. Descripción general de los datos	5
2.2. Preparación de los registros	5
2.3. Fuentes de información	6
2.3.1. AutoCAD	7
2.3.2. Revit	8
2.3.3. TimeDoctor	11
2.4. Consolidación de datos	12
2.4.1. Períodos de monitoreo	13
3. Fundamentos matemáticos y estadísticos	14
3.1. Notación general	14
3.2. K-Means	14
3.2.1. Notación y formulación	15
3.2.2. Función objetivo	16
3.2.3. Criterio de detención	16
3.3. Modelos de Mezclas Gaussianas	16
3.3.1. Notación y formulación	17

3.3.2. Distribución gaussiana multivariada	18
3.3.3. Modelo de mezcla	18
3.3.4. Probabilidades de pertenencia	18
3.4. Algoritmo Expectation-Maximization	19
3.4.1. Máxima verosimilitud	19
3.4.2. Etapa de Expectativa	20
3.4.3. Etapa de Maximización	21
3.4.4. Propiedades y convergencia	22
3.5. Coeficiente de silueta	22
3.6. Entropía de Shannon	23
3.6.1. Definición	23
3.6.2. Dominio y propiedades	24
3.6.3. Entropía normalizada	24
3.7. Índice de Herfindahl–Hirschman	25
3.7.1. Definición	25
3.7.2. Dominio y propiedades	26
3.8. Medidas descriptivas de dispersión	26
3.9. Prueba Chi-cuadrado de independencia	28
3.9.1. Tabla de contingencia y frecuencias esperadas	28
3.9.2. Estadístico Chi-cuadrado	28
3.9.3. Hipótesis y distribución del estadístico	29
3.10. Coeficiente V de Cramér	30
3.10.1. Definición	30
3.10.2. Dominio e interpretación	30
3.10.3. Relación con la prueba Chi-cuadrado	30
3.11. Prueba de Kruskal–Wallis	31
3.11.1. Rangos y formulación	31
3.11.2. Hipótesis de la prueba	32
3.11.3. Interpretación	32
3.12. Combinación de valores p mediante el método de Fisher	33
4. Construcción de variables	35
4.1. Variable 1: Perfil general de actividad	36
4.1.1. Descripción de la variable	36
4.1.2. Pregunta de investigación	36

4.1.3. Variables utilizadas	37
4.1.4. Construcción del espacio de características	38
4.1.5. Modelos de agrupamiento	38
4.1.6. Selección del número de clusters	39
4.2. Indicador 2: Concentración horaria	40
4.2.1. Descripción del indicador	40
4.2.2. Pregunta de investigación	40
4.2.3. Construcción de la variable	40
4.3. Indicador 3: Diversidad documental	41
4.3.1. Descripción del indicador	41
4.3.2. Pregunta de investigación	41
4.3.3. Aplicación al indicador de diversidad documental	42
4.4. Indicador 4: Concentración documental	42
4.4.1. Descripción e interpretación del indicador	42
4.4.2. Pregunta de investigación e hipótesis	43
4.4.3. Construcción del indicador	43
4.5. Indicador 5: Fragmentación temporal	44
4.5.1. Descripción del indicador	44
4.5.2. Pregunta de investigación	44
4.5.3. Construcción de la variable	45
4.5.4. Concepto de fragmentación temporal	46
5. Validación de las variables	47
5.1. Validación del perfil general de actividad	47
5.1.1. Distribución de las métricas agregadas	48
5.1.2. Distribución de usuarios por cluster	53
5.1.3. Comparación de los clusters en AutoCAD	54
5.1.4. Comparación de los clusters en Revit	57
5.1.5. Comparación de los clusters en TimeDoctor	60
5.1.6. Selección del método de agrupamiento	62
5.1.7. Respuesta a las preguntas planteadas	63
5.2. Validación de la concentración horaria	64
5.2.1. Distribución del indicador	64
5.2.2. Distribución de usuarios por categoría	66
5.2.3. Variabilidad del indicador	67

5.2.4. Respuesta a la pregunta de investigación	68
5.3. Validación de la diversidad documental	68
5.3.1. Distribución del indicador	69
5.3.2. Distribución de usuarios por categoría	71
5.3.3. Variabilidad del indicador	71
5.3.4. Respuesta a la pregunta de investigación	72
5.4. Validación de la concentración documental	73
5.4.1. Distribución del indicador	73
5.4.2. Distribución de usuarios por categoría	75
5.4.3. Variabilidad del indicador	76
5.4.4. Respuesta a la pregunta de investigación	77
5.5. Validación de la fragmentación temporal	77
5.5.1. Distribución del indicador	78
5.5.2. Distribución de usuarios por categoría	80
5.5.3. Variabilidad del indicador	80
5.5.4. Respuesta a la pregunta de investigación	81
6. Relación entre indicadores	83
6.1. Asociación entre variables categóricas	83
6.1.1. Prueba Chi-cuadrado de independencia	83
6.1.2. Coeficiente V de Cramér	86
6.1.3. Interpretación de las asociaciones	87
6.2. Comparación de indicadores entre perfiles	88
6.2.1. Concentración horaria en AutoCAD	89
6.2.2. Concentración horaria en Revit	89
6.2.3. Concentración horaria en TimeDoctor	90
6.2.4. Comparación general de la concentración horaria	90
6.2.5. Diversidad documental en AutoCAD	91
6.2.6. Diversidad documental en Revit	91
6.2.7. Diversidad documental en TimeDoctor	92
6.2.8. Comparación general de la diversidad documental	92
6.2.9. Concentración documental en AutoCAD	93
6.2.10. Concentración documental en Revit	94
6.2.11. Concentración documental en TimeDoctor	94
6.2.12. Comparación general de la concentración documental	95

6.2.13. Fragmentación temporal en AutoCAD	96
6.2.14. Fragmentación temporal en Revit	96
6.2.15. Fragmentación temporal en TimeDoctor	97
6.2.16. Comparación general de la fragmentación temporal	97
6.3. Resultados de la prueba de Kruskal–Wallis	98
6.3.1. Comparaciones por pares entre clusters	100
6.3.2. Concentración horaria	100
6.3.3. Diversidad documental	100
6.3.4. Concentración documental	101
6.3.5. Fragmentación temporal	102
6.4. Propuesta de caracterización de los perfiles	102
6.4.1. Puntaje documental	103
6.4.2. Resultados del puntaje documental	104
6.4.3. Distancia respecto del comportamiento general	106
6.4.4. Resultados de la distancia respecto del comportamiento general	107
7. Discusión	110
7.1. Interpretación general de los resultados	110
7.2. Discusión de las preguntas e hipótesis planteadas	111
7.3. Limitaciones del estudio	112
7.4. Uso responsable de los indicadores	113
8. Conclusiones	114
8.1. Conclusiones generales	114
8.2. Principales hallazgos	114
8.3. Aportes del trabajo	115
8.4. Trabajo futuro	115

Índice de cuadros

2.1. Percentiles superiores del tiempo de actividad por plataforma	6
2.2. Variables disponibles en la base consolidada de AutoCAD	8
2.3. Variables disponibles en la base consolidada de Revit	10
2.4. Variables disponibles en la base consolidada de TimeDoctor	12
4.1. Métricas utilizadas para construir el perfil general de actividad	38
4.2. Coeficiente de silueta obtenido mediante K-Means para distintos valores de K	39
4.3. Coeficiente de silueta obtenido mediante GMM para distintos valores de K	39
5.1. Promedios de las métricas por cluster en AutoCAD mediante K-Means.	55
5.2. Promedios de las métricas por cluster en AutoCAD mediante GMM.	56
5.3. Promedios de las métricas por cluster en Revit mediante K-Means.	57
5.4. Promedios de las métricas por cluster en Revit mediante GMM.	58
5.5. Promedios de las métricas por cluster en TimeDoctor mediante K-Means.	60
5.6. Promedios de las métricas por cluster en TimeDoctor mediante GMM.	61
5.7. Medidas descriptivas de dispersión del indicador de concentración horaria	67
5.8. Medidas descriptivas de dispersión del indicador de diversidad documental	72
5.9. Medidas descriptivas de dispersión del indicador de concentración documental	76
5.10. Medidas descriptivas de dispersión del indicador de fragmentación temporal	81
6.1. Resultados de la prueba Chi-cuadrado entre las variables categóricas	84
6.2. Combinación de los p -values de las pruebas Chi-cuadrado mediante el método de Fisher	85
6.3. Valores de V de Cramér en AutoCAD, Revit y TimeDoctor, ordenados según su valor promedio	86

6.4. Resultados de la prueba de Kruskal–Wallis y combinación de p -values mediante el método de Fisher	98
6.5. Comparaciones por pares para el indicador de diversidad documental . . .	101
6.6. Comparaciones por pares para el indicador de concentración documental .	102
6.7. Resultados del puntaje documental por plataforma y cluster.	104
6.8. Distancia respecto del comportamiento general por plataforma y cluster. .	108

Índice de figuras

5.1. Diagramas de caja de las variables de actividad.	49
5.2. Histogramas de las variables de actividad.	50
5.3. Distribuciones acumuladas de las variables de actividad.	52
5.4. Distribución de usuarios por cluster mediante K-Means y GMM en Auto-	
CAD, Revit y TimeDoctor.	53
5.5. Distribución del indicador de concentración horaria en AutoCAD.	65
5.6. Distribución del indicador de concentración horaria en Revit.	65
5.7. Distribución del indicador de concentración horaria en TimeDoctor.	66
5.8. Distribución de usuarios según las categorías de concentración horaria.	66
5.9. Distribución del indicador de diversidad documental en AutoCAD.	69
5.10. Distribución del indicador de diversidad documental en Revit.	70
5.11. Distribución del indicador de diversidad documental en TimeDoctor.	70
5.12. Distribución de usuarios según las categorías de diversidad documental.	71
5.13. Distribución del indicador de concentración documental en AutoCAD.	74
5.14. Distribución del indicador de concentración documental en Revit.	74
5.15. Distribución del indicador de concentración documental en TimeDoctor.	75
5.16. Distribución de usuarios según las categorías de concentración documental.	75
5.17. Distribución del indicador de fragmentación temporal en AutoCAD.	78
5.18. Distribución del indicador de fragmentación temporal en Revit.	79
5.19. Distribución del indicador de fragmentación temporal en TimeDoctor.	79
5.20. Distribución de usuarios según las categorías de fragmentación temporal.	80
6.1. Comparación del indicador de concentración horaria entre clusters en Au-	
toCAD.	89
6.2. Comparación del indicador de concentración horaria entre clusters en Revit.	89

6.3. Comparación del indicador de concentración horaria entre clusters en Ti-	
meDoctor	90
6.4. Comparación del indicador de diversidad documental entre clusters en Au-	
toCAD	91
6.5. Comparación del indicador de diversidad documental entre clusters en Revit	91
6.6. Comparación del indicador de diversidad documental entre clusters en Ti-	
meDoctor	92
6.7. Comparación del indicador de concentración documental entre clusters en	
AutoCAD	93
6.8. Comparación del indicador de concentración documental entre clusters en	
Revit	94
6.9. Comparación del indicador de concentración documental entre clusters en	
TimeDoctor	94
6.10. Comparación del indicador de fragmentación temporal entre clusters en	
AutoCAD	96
6.11. Comparación del indicador de fragmentación temporal entre clusters en Revit	96
6.12. Comparación del indicador de fragmentación temporal entre clusters en	
TimeDoctor	97
6.13. Distribución del puntaje documental por cluster en AutoCAD, Revit y Ti-	
meDoctor	106
6.14. Distribución de la distancia por cluster en AutoCAD, Revit y TimeDoctor	109

Nomenclatura

Símbolo	Descripción	Unidad
n	Número total de observaciones o usuarios considerados	-
d	Número de características utilizadas para representar cada observación	-
$\mathbf{x}_i$	Vector de características asociado al usuario u observación i	-
K	Número de clusters o componentes consideradas	-
k	Índice que identifica un cluster o componente, $k \in \{1, \dots, K\}$	-
C_k	Conjunto de observaciones asignadas al cluster k	-
n_k	Número de observaciones pertenecientes al cluster C_k	-
$\boldsymbol{\mu}_k$	Vector medio asociado al cluster o componente k	-
$\boldsymbol{\Sigma}_k$	Matriz de covarianza de la componente gaussiana k	-
π_k	Peso de la componente k en el modelo GMM	-
γ_{ik}	Probabilidad posterior de pertenencia de la observación i a la componente k	-
$s(i)$	Coeeficiente de silueta de la observación i	-
$\mathbf{q}_i$	Vector de proporciones de tiempo del usuario i entre los documentos utilizados	-
$H_{\text{norm}}(\mathbf{q}_i)$	Entropía normalizada de Shannon utilizada como indicador de diversidad documental	-
$HHI(\mathbf{q}_i)$	Índice de Herfindahl–Hirschman utilizado como indicador de concentración documental	-
P_{MM}	Proporción del tiempo registrada durante la mañana–madrugada	-
$I_{\text{Fragmentacin}}$	Índice normalizado de fragmentación temporal	-

Símbolo	Descripción	Unidad
Q_1	Primer cuartil de una distribución	-
Q_3	Tercer cuartil de una distribución	-
IQR	Rango intercuartílico, $Q_3 - Q_1$	-
P_{10}	Percentil 10 de una distribución	-
P_{90}	Percentil 90 de una distribución	-
χ^2	Estadístico de la prueba Chi-cuadrado de independencia	-
V	Coeficiente V de Cramér	-
H_{KW}	Estadístico de la prueba de Kruskal–Wallis	-
ν	Grados de libertad de una prueba estadística	-
α	Nivel de significancia utilizado en las pruebas estadísticas	-
p	Valor p asociado a una prueba estadística	-
S_i	Puntaje documental asociado al usuario i	-
S_k	Puntaje documental promedio del cluster C_k	-
d_i	Distancia del usuario i respecto del comportamiento general de la plataforma	-
D_k	Distancia del cluster C_k respecto del comportamiento general de la plataforma	-

Capítulo 1

Introducción

1.1. Contexto

La transformación digital corresponde a la incorporación de tecnologías digitales en una organización, generando cambios en la forma en que se desarrollan los procesos y se gestiona la información [1]. En empresas donde gran parte del trabajo se realiza mediante herramientas informáticas, estas tecnologías generan de manera continua registros asociados a las actividades realizadas por los usuarios. Como resultado, los datos operacionales han adquirido relevancia como fuente de información para comprender procesos, apoyar la toma de decisiones e identificar oportunidades de mejora.

Estos registros forman parte de la información generada por los sistemas utilizados en la operación de una organización. Un sistema de información puede entenderse como un conjunto de personas, procesos, datos y tecnologías que interactúan para recopilar, almacenar, procesar y distribuir información que apoye la operación y la toma de decisiones [2].

A partir del análisis de este tipo de registros se han desarrollado disciplinas como la minería de procesos (*Process Mining*) y la minería de tareas (*Task Mining*), orientadas a extraer información a partir de los eventos registrados por los sistemas. Este tipo de análisis permite reconstruir procesos, identificar patrones y estudiar la forma en que los usuarios interactúan con las herramientas digitales [3].

El uso de registros digitales para analizar procesos también se ha extendido al ámbito de la ingeniería. Plataformas como AutoCAD y Revit no solo permiten desarrollar pro-

yectos de diseño y modelamiento, sino que además generan información asociada al uso de documentos, vistas, proyectos y tiempos de interacción. De manera complementaria, herramientas como TimeDoctor registran información relacionada con el uso del computador, las aplicaciones utilizadas y la distribución temporal de la actividad.

La presente memoria se desarrolla en Pares & Álvarez, empresa chilena de ingeniería donde estas tres plataformas forman parte del trabajo diario de sus profesionales. En conjunto, los registros generados constituyen una fuente de información que actualmente no está siendo utilizada y que podría permitir estudiar el comportamiento del uso de las herramientas digitales desde una perspectiva basada en datos.

No obstante, disponer de grandes volúmenes de información no garantiza la obtención de conocimiento útil. Los registros disponibles corresponden principalmente a eventos individuales y datos operacionales que, en su forma original, no necesariamente representan de manera adecuada los fenómenos que se desean analizar. Por esta razón, es necesario procesar y transformar los datos originales en variables que sinteticen características relevantes y faciliten su posterior análisis e interpretación [4].

1.2. Problema de estudio

Las plataformas AutoCAD, Revit y TimeDoctor generan una gran cantidad de registros asociados a la actividad de los usuarios durante su trabajo. Sin embargo, estos registros describen principalmente eventos individuales y no permiten caracterizar de manera directa cómo se desarrolla el uso de estas herramientas ni identificar si existen diferencias relevantes entre los usuarios.

Por ejemplo, los registros permiten conocer cuánto tiempo estuvo activo un documento, qué aplicación fue utilizada o en qué momento ocurrió una determinada interacción. No obstante, esta información se encuentra distribuida entre una gran cantidad de eventos que deben ser analizados de manera conjunta para representar características más generales del comportamiento del usuario al usar el equipo de trabajo.

Una alternativa para explorar la estructura presente en estos datos corresponde a las técnicas de agrupamiento no supervisado, las cuales permiten identificar grupos de obser-

vaciones con características similares sin requerir categorías definidas previamente [5]. En esta investigación, estas técnicas se utilizan para explorar la existencia de distintos perfiles de actividad entre los usuarios. Sin embargo, identificar grupos no resulta suficiente para comprender las características que los diferencian ni otras dimensiones del comportamiento que pueden estar presentes en los registros.

Por esta razón, la investigación considera también la construcción de indicadores que representen distintas dimensiones del uso de las plataformas, tales como la distribución temporal de la actividad, la diversidad y concentración en el uso de documentos y la continuidad del trabajo. Estos indicadores permiten resumir los registros originales en variables interpretables y posteriormente analizar su distribución, sus relaciones y su comportamiento respecto de los perfiles identificados.

En este contexto, la pregunta que guía esta investigación es la siguiente:

¿Es posible construir y validar indicadores, a partir de los registros de actividad de AutoCAD, Revit y TimeDoctor, que permitan identificar y caracterizar patrones de uso entre los usuarios?

1.3. Objetivos

1.3.1. Objetivo general

Construir y validar un conjunto de indicadores, a partir de los registros de actividad generados por AutoCAD, Revit y TimeDoctor, que permitan identificar y caracterizar patrones de uso entre los usuarios.

1.3.2. Objetivos específicos

- Consolidar y preprocesar los registros provenientes de AutoCAD, Revit y TimeDoctor, con el fin de obtener bases de datos consistentes para su análisis.
- Construir un perfil general de actividad mediante técnicas de agrupamiento no supervisado, a partir de métricas agregadas de comportamiento de los usuarios.

- Construir indicadores que representen distintas dimensiones del comportamiento de uso, considerando la distribución horaria de la actividad, la diversidad y concentración documental y la fragmentación temporal.
- Analizar la distribución de los indicadores contruidos para evaluar su comportamiento e interpretabilidad entre los usuarios.
- Analizar la relación entre los indicadores y su comportamiento respecto de los perfiles identificados mediante técnicas estadísticas.
- Comparar los resultados obtenidos en AutoCAD, Revit y TimeDoctor para identificar similitudes y diferencias en los patrones de uso observados.

1.4. Alcances y limitaciones

Esta investigación se desarrolla a partir de registros de actividad generados por AutoCAD, Revit y TimeDoctor en el contexto de una empresa de ingeniería. El análisis se limita, por lo tanto, a la actividad registrada por estas plataformas durante el período de estudio y no busca representar la totalidad de las tareas realizadas por los usuarios.

Para resguardar la identidad de las personas, se utiliza información anonimizada y los resultados se analizan desde una perspectiva agregada. El propósito de la investigación es identificar y caracterizar patrones de uso de las herramientas digitales, sin evaluar el desempeño individual de los usuarios.

En consecuencia, los indicadores contruidos no deben interpretarse como métricas de productividad laboral ni como medidas de rendimiento. Su alcance corresponde a una propuesta metodológica para transformar registros de actividad en variables que permitan describir y analizar distintas dimensiones del comportamiento de uso de las plataformas.

Capítulo 2

Datos utilizados

2.1. Descripción general de los datos

Esta investigación utiliza registros de actividad provenientes de AutoCAD, Revit y TimeDoctor. En conjunto, estas fuentes permiten observar distintos aspectos del uso de herramientas digitales por parte de los usuarios [6, 7, 8].

Debido a que cada plataforma registra información de manera diferente, los datos fueron trabajados por separado antes de construir las variables utilizadas en la investigación. Este proceso incluyó la consolidación de las distintas fuentes, la revisión de los tiempos registrados y la preparación de las tablas finales utilizadas en el análisis.

En este capítulo se presentan las características de cada fuente, el tratamiento realizado sobre los registros y la forma en que se relacionaron los usuarios entre las distintas plataformas.

2.2. Preparación de los registros

Antes de construir las variables, se revisaron los tiempos de actividad para identificar registros con duraciones atípicamente altas. En AutoCAD y Revit se utilizó la variable `timelapse`, mientras que en TimeDoctor se consideró `timeSec`; ambas se representan como t más adelante.

Inicialmente se evaluó el criterio de valores atípicos extremos basado en el rango inter-

cuartílico, considerando como tales los registros superiores a $Q_3 + 3 \times IQR$, siguiendo el enfoque de Tukey [9]. Sin embargo, este criterio eliminaba entre un 11 % y un 15 % de los registros, por lo que se consideró demasiado restrictivo para los datos disponibles.

Por esta razón, se analizaron los percentiles superiores de la duración registrada en cada plataforma. En los tres casos se observó un aumento importante entre los percentiles 99,5 y 99,9, por lo que se utilizó el percentil 99,5 de cada plataforma como límite para identificar los valores atípicos.

Cuadro 2.1: Percentiles superiores del tiempo de actividad por plataforma

Plataforma	P99	P99,5	P99,9
AutoCAD	1 h 46 min	2 h 41 min	6 h 34 min
Revit	1 h 29 min	3 h 49 min	18 h 28 min
TimeDoctor	50 min 33 s	1 h 29 min	4 h 47 min

Luego de eliminar los registros superiores a este límite, las tablas utilizadas para construir las variables quedaron compuestas por 1.578.214 registros de AutoCAD, 1.316.513 de Revit y 310.563 de TimeDoctor.

2.3. Fuentes de información

AutoCAD, Revit y TimeDoctor entregan información distinta sobre la actividad de los usuarios. Mientras AutoCAD y Revit registran principalmente interacciones asociadas al diseño y modelamiento de proyectos, TimeDoctor registra información más general sobre el uso del computador y las aplicaciones utilizadas.

Los datos están formados por registros individuales asociados a un usuario. Según la plataforma, estos registros incluyen información sobre duración, documentos, aplicaciones, proyectos y fechas de inicio y término.

A continuación se describe cada fuente considerando la cantidad de registros y usuarios, el período disponible y las características de las variables que contiene.

2.3.1. AutoCAD

AutoCAD corresponde a una herramienta de diseño asistido por computador (CAD), utilizada para la creación y edición de dibujos y documentación técnica en dos y tres dimensiones [6].

La información disponible se encontraba originalmente distribuida en tres tablas: registros de actividad, usuarios y equipos. Luego de su consolidación, la base de AutoCAD quedó compuesta por 1.578.214 registros, 11 variables y 212 usuarios. El período registrado abarca desde el 8 de marzo de 2024 hasta el 6 de enero de 2026.

Las variables disponibles describen la duración y temporalidad de cada interacción, el documento y programa asociados al registro y la información necesaria para relacionar a los usuarios con las demás plataformas. La Tabla 2.2 presenta las 11 variables que componen la base y las principales características de sus valores.

Cuadro 2.2: Variables disponibles en la base consolidada de AutoCAD

Información	Descripción	Dominio o cardinalidad
Duración	Tiempo asociado a cada interacción registrada.	$0 \leq t \leq 863\,783$, $t \in \mathbb{N}$
Documento	Archivo o ruta asociada a la interacción.	$ \mathcal{D}_{\text{doc}} = 123\,619$
Programa	Programa desde el cual se generó el registro.	$ \mathcal{D}_{\text{prog}} = 4$
Versión del sistema	Versión del sistema utilizado para registrar la actividad.	$ \mathcal{D}_{\text{sis}} = 11$
Versión de la aplicación	Versión de AutoCAD o de la aplicación asociada al registro.	$ \mathcal{D}_{\text{app}} = 36$
Origen	Fuente desde la cual se generó la información.	$ \mathcal{D}_{\text{orig}} = 2$
ID TimeDoctor	Identificador utilizado para relacionar al usuario con la base de TimeDoctor.	Identificador en $\mathbb{N}$
Inicio	Fecha y hora en que comienza el registro de actividad.	$t_{\text{inicio}} \in \mathcal{T}_A$
Término	Fecha y hora en que finaliza el registro de actividad.	$t_{\text{trmino}} \in \mathcal{T}_A$
ID Revit	Identificador utilizado para relacionar al usuario con la base de Revit.	Identificador en $\mathbb{Z}$
Usuario	Identificador anonimizado de la persona asociada al registro.	$ U_A = 212$

Los identificadores utilizados para vincular las distintas plataformas no representan características del comportamiento de los usuarios, sino que permiten mantener la correspondencia entre las bases. Para la construcción posterior de las variables resultan especialmente relevantes la duración de los registros, los documentos utilizados y las fechas y horas asociadas a cada interacción.

2.3.2. Revit

Revit corresponde a una plataforma de modelado de información para la construcción (BIM), orientada al diseño, modelado y documentación de proyectos [7].

La información disponible se encontraba originalmente distribuida en tres tablas: registros de actividad, usuarios y equipos. Luego de su consolidación, la base de Revit quedó compuesta por 1.316.513 registros, 16 variables y 110 usuarios. El período registrado abarca desde el 11 de marzo de 2024 hasta el 6 de enero de 2026.

Además de información sobre la duración y temporalidad de las interacciones, Revit incorpora datos relacionados con los archivos, proyectos y vistas utilizadas durante el trabajo. También incluye información del programa e identificadores que permiten relacionar a los usuarios con las demás plataformas. La Tabla [2.3](#) presenta las variables disponibles y las principales características de sus valores.

Cuadro 2.3: Variables disponibles en la base consolidada de Revit

Información	Descripción	Dominio o cardinalidad
Duración	Tiempo asociado a cada interacción registrada.	$0 \leq t \leq 1\,779\,094$, $t \in \mathbb{N}$
Tipo de archivo	Tipo o extensión del archivo utilizado.	$ \mathcal{D}_{\text{tipo_doc}} = 5$
Documento	Archivo asociado a la interacción.	$ \mathcal{D}_{\text{doc}} = 7\,576$
ID de vista	Identificador de la vista asociada al registro.	Identificador numérico.
Vista	Nombre de la vista utilizada.	$ \mathcal{D}_{\text{vista}} = 24\,401$
Tipo de vista	Tipo de vista utilizada dentro del modelo.	$ \mathcal{D}_{\text{tipo_vista}} = 19$
Proyecto	Proyecto asociado al registro.	$ \mathcal{D}_{\text{proyecto}} = 53$
Programa	Programa desde el cual se generó el registro.	$ \mathcal{D}_{\text{prog}} = 1$
Versión del sistema	Versión del sistema utilizado para registrar la actividad.	$ \mathcal{D}_{\text{sis}} = 5$
Versión de la aplicación	Versión de Revit asociada al registro.	$ \mathcal{D}_{\text{app}} = 5$
Origen	Fuente desde la cual se generó la información.	$ \mathcal{D}_{\text{orig}} = 1$
ID TimeDoctor	Identificador utilizado para relacionar al usuario con la base de TimeDoctor.	Identificador en $\mathbb{Z}$
Inicio	Fecha y hora en que comienza el registro de actividad.	$t_{\text{inicio}} \in \mathcal{T}_R$
Término	Fecha y hora en que finaliza el registro de actividad.	$t_{\text{trmino}} \in \mathcal{T}_R$
ID AutoCAD	Identificador utilizado para relacionar al usuario con la base de AutoCAD.	Identificador en $\mathbb{Z}$
Usuario	Identificador anonimizado de la persona asociada al registro.	$ U_R = 110$

El identificador de vista se encuentra almacenado como una variable numérica, pero corresponde a un identificador y no a una medida cuantitativa. Además, las variables asociadas al programa y al origen presentan un único valor dentro de la base, por lo que no aportan variabilidad entre los registros. Para la construcción posterior de las variables

resultan especialmente relevantes la duración, los documentos utilizados y las fechas y horas asociadas a cada interacción.

2.3.3. TimeDoctor

TimeDoctor es una plataforma orientada al seguimiento de la actividad digital realizada durante la jornada laboral, incluyendo información sobre las aplicaciones utilizadas y el uso del computador [8].

La información disponible se encontraba distribuida en distintos archivos asociados a usuarios, equipos y registros de actividad. Luego de su consolidación, la base de TimeDoctor quedó compuesta por 310.563 registros, 10 variables y 170 usuarios. El período registrado abarca desde el 3 de diciembre de 2025 hasta el 1 de enero de 2026.

A diferencia de AutoCAD y Revit, TimeDoctor registra actividad general del computador. La base contiene información sobre las aplicaciones y ventanas utilizadas, la duración de cada registro y sus fechas y horas de inicio y término. También incluye información asociada al usuario e identificadores que permiten relacionarlo con las demás plataformas. La Tabla 2.4 presenta las variables disponibles y las principales características de sus valores.

Cuadro 2.4: Variables disponibles en la base consolidada de TimeDoctor

Información	Descripción	Dominio o cardinalidad
Tipo de actividad	Tipo de registro generado por la plataforma.	$ \mathcal{D}_{\text{tipo}} = 10$
Aplicación	Aplicación o elemento asociado al registro de actividad.	$ \mathcal{D}_{\text{app}} = 8\,658$
Duración	Tiempo asociado a cada registro de actividad.	$1 \leq t \leq 453\,582$, $t \in \mathbb{N}$
Ventana activa	Título anonimizado de la ventana o aplicación activa.	$ \mathcal{D}_{\text{ventana}} = 149\,364$
Fecha de contratación	Información asociada a la fecha de contratación del usuario.	$ \mathcal{D}_{\text{contratacin}} = 158$
ID AutoCAD	Identificador utilizado para relacionar al usuario con la base de AutoCAD.	Identificador en $\mathbb{Z}$
ID Revit	Identificador utilizado para relacionar al usuario con la base de Revit.	Identificador en $\mathbb{Z}$
Inicio	Fecha y hora en que comienza el registro de actividad.	$t_{\text{inicio}} \in \mathcal{T}_T$
Término	Fecha y hora en que finaliza el registro de actividad.	$t_{\text{trmino}} \in \mathcal{T}_T$
Usuario	Identificador anonimizado de la persona asociada al registro.	$ U_T = 170$

2.4. Consolidación de datos

La información de cada plataforma se encontraba distribuida en distintas tablas y archivos, por lo que fue necesario reunir los datos relevantes antes de comenzar el análisis. Este proceso se realizó por separado para AutoCAD, Revit y TimeDoctor, manteniendo la estructura propia de cada fuente.

Para relacionar usuarios entre plataformas se utilizaron los identificadores creados durante la consolidación: `id_acad`, `id_revit` e `id_td`. Estos permiten establecer correspondencias entre una persona registrada en una plataforma y su identificador en otra, mientras que `user_id` se mantiene como identificador del usuario dentro de cada base.

Si U_A , U_R y U_T representan los conjuntos de usuarios de AutoCAD, Revit y TimeDoctor, respectivamente, se obtiene

$$|U_A \cap U_R| = 108, \quad |U_A \cap U_T| = 22, \quad |U_R \cap U_T| = 14,$$

y

$$|U_A \cap U_R \cap U_T| = 11.$$

Considerando las correspondencias entre las tres fuentes, el conjunto completo contiene 385 usuarios distintos.

Como resultado de este proceso se obtuvieron tres tablas consolidadas, una por plataforma, que mantienen sus variables originales de interés y permiten relacionar a los usuarios cuando existe una correspondencia entre las fuentes. Estas tablas constituyen el punto de partida para la construcción posterior de las variables e indicadores a nivel de usuario.

2.4.1. Períodos de monitoreo

Los períodos de observación disponibles no son iguales entre las tres plataformas. En AutoCAD, los registros abarcan desde el 8 de marzo de 2024 hasta el 6 de enero de 2026, mientras que en Revit el período se extiende desde el 11 de marzo de 2024 hasta el 6 de enero de 2026. En TimeDoctor, en cambio, los registros disponibles corresponden al período comprendido entre el 3 de diciembre de 2025 y el 1 de enero de 2026.

Esta diferencia debe considerarse al comparar los resultados entre plataformas, ya que las variables construidas resumen la actividad observada durante períodos de distinta duración. Por esta razón, las comparaciones entre AutoCAD, Revit y TimeDoctor se interpretan en términos de los patrones y distribuciones observadas en cada plataforma, y no como una comparación directa de la magnitud total de la actividad registrada.

Capítulo 3

Fundamentos matemáticos y estadísticos

En este capítulo se presentan los principales fundamentos matemáticos y estadísticos utilizados en el desarrollo de la investigación.

3.1. Notación general

Se considera un conjunto de n observaciones,

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\},$$

donde cada observación está representada por un vector de características

$$\mathbf{x}_i \in \mathbb{R}^d, \quad i = 1, \dots, n,$$

y d corresponde al número de características consideradas.

3.2. K-Means

K-Means es un método de agrupamiento que particiona un conjunto de observaciones en K grupos, buscando que las observaciones pertenecientes a un mismo grupo sean similares entre sí [10].

3.2.1. Notación y formulación

Para los métodos de agrupamiento se utilizará K para representar el número de clusters.

El conjunto de observaciones asignadas al cluster k se denotará por C_k , con

$$k \in \{1, \dots, K\}.$$

Los clusters forman una partición de $\mathcal{X}$, por lo que

$$C_k \cap C_\ell = \emptyset, \quad k \neq \ell,$$

y

$$\bigcup_{k=1}^K C_k = \mathcal{X}.$$

La cantidad de observaciones pertenecientes al cluster k se denotará por

$$n_k = |C_k|,$$

mientras que su centro o vector medio se representará mediante

$$\boldsymbol{\mu}_k \in \mathbb{R}^d.$$

El centroide del cluster C_k se define como

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i.$$

La distancia entre una observación $\mathbf{x}_i$ y un centroide $\boldsymbol{\mu}_k$ se mide mediante la distancia euclidiana,

$$\|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2 = \sqrt{\sum_{j=1}^d (x_{ij} - \mu_{kj})^2}.$$

3.2.2. Función objetivo

K-Means busca encontrar una partición de las observaciones que minimice la suma de las distancias cuadráticas entre cada observación y el centroide de su cluster. La función objetivo se define como

$$J = \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2.$$

Por lo tanto, el problema de optimización puede expresarse como

$$\min_{C_1, \dots, C_K} J.$$

Dado un conjunto de centroides, cada observación se asigna al cluster cuyo centroide se encuentra a menor distancia:

$$\mathbf{x}_i \in C_k \quad \text{si} \quad k = \arg \min_{\ell \in \{1, \dots, K\}} \|\mathbf{x}_i - \boldsymbol{\mu}_\ell\|_2^2.$$

Luego de realizar las asignaciones, los centroides se recalculan utilizando la media de las observaciones de cada cluster. Este procedimiento de asignación y actualización se repite de manera iterativa hasta alcanzar un criterio de convergencia.

3.2.3. Criterio de detención

El procedimiento iterativo de K-Means se mantiene hasta que se cumple un criterio de convergencia. Entre los criterios habituales se encuentran que las asignaciones de las observaciones a los clusters no cambien o que la variación de los centroides entre dos iteraciones consecutivas sea menor que una tolerancia establecida.

Adicionalmente, puede establecerse un número máximo de iteraciones para asegurar la finalización del procedimiento.

3.3. Modelos de Mezclas Gaussianas

Un Modelo de Mezclas Gaussianas (Gaussian Mixture Model, GMM) supone que las observaciones provienen de una combinación de K distribuciones gaussianas. A diferencia

de K-Means, donde cada observación es asignada directamente a un único cluster, en un GMM la pertenencia se representa de manera probabilística [10].

3.3.1. Notación y formulación

Para cada observación se introduce una variable latente Z_i , que representa la componente gaussiana que genera dicha observación,

$$Z_i \in \{1, \dots, K\}.$$

Para cada componente k se consideran los parámetros

$$\pi_k, \quad \boldsymbol{\mu}_k, \quad \boldsymbol{\Sigma}_k,$$

donde

$$\pi_k \in [0, 1]$$

corresponde al peso de la componente,

$$\boldsymbol{\mu}_k \in \mathbb{R}^d$$

corresponde a su vector de las medias, y

$$\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$$

corresponde a su matriz de covarianza.

Los pesos de las componentes satisfacen

$$\sum_{k=1}^K \pi_k = 1,$$

y representan las probabilidades a priori

$$P(Z_i = k) = \pi_k.$$

La matriz de covarianza satisface

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k^T,$$

y se considera definida positiva.

3.3.2. Distribución gaussiana multivariada

Condicionado a que la observación pertenece a la componente k , se supone que

$$\mathbf{x}_i \mid Z_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Por lo tanto, su función de densidad está dada por

$$p(\mathbf{x}_i \mid Z_i = k) = \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

donde

$$\mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right).$$

3.3.3. Modelo de mezcla

La densidad de una observación $\mathbf{x}_i$ se obtiene considerando todas las componentes posibles,

$$p(\mathbf{x}_i) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

De esta forma, cada componente contribuye a la distribución total de acuerdo con su peso π_k .

3.3.4. Probabilidades de pertenencia

La probabilidad posterior de que la observación $\mathbf{x}_i$ pertenezca a la componente k se denotará por

$$\gamma_{ik} = P(Z_i = k \mid \mathbf{x}_i).$$

Aplicando el teorema de Bayes,

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{\ell=1}^K \pi_\ell \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)}.$$

Para cada observación se cumple

$$0 \leq \gamma_{ik} \leq 1,$$

y

$$\sum_{k=1}^K \gamma_{ik} = 1.$$

Por lo tanto, una observación puede presentar distintos grados de pertenencia a las componentes del modelo, en lugar de ser asignada inicialmente de manera determinista a un único grupo.

3.4. Algoritmo Expectation-Maximization

El algoritmo Expectation-Maximization (EM) es un procedimiento iterativo utilizado para estimar los parámetros de modelos que incorporan variables latentes [12]. En esta investigación, se utiliza para estimar los parámetros del Modelo de Mezclas Gaussianas definido en la Sección 3.3.

3.4.1. Máxima verosimilitud

Sea

$$\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$$

el conjunto de parámetros del modelo de mezcla.

Para una observación $\mathbf{x}_i$, la densidad definida por el GMM es

$$p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Suponiendo que las n observaciones son independientes, la función de verosimilitud del conjunto de datos $\mathcal{X}$ está dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{x}_i \mid \boldsymbol{\theta}),$$

es decir,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n \left[\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right].$$

El objetivo es encontrar los parámetros que maximicen esta función,

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}).$$

En la práctica se utiliza la log-verosimilitud, ya que transforma el producto de probabilidades en una suma y conserva el mismo punto de máximo:

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^n \log \left[\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right].$$

Por lo tanto,

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}).$$

La presencia de una suma de componentes gaussianas dentro del logaritmo dificulta la maximización directa de la log-verosimilitud. El algoritmo EM aborda este problema utilizando la variable latente Z_i definida anteriormente y alternando entre una etapa de estimación de las probabilidades de pertenencia y una etapa de actualización de los parámetros del modelo [10].

3.4.2. Etapa de Expectativa

En la etapa de Expectativa (E) se mantienen fijos los parámetros obtenidos en la iteración t , denotados por

$$\boldsymbol{\theta}^{(t)},$$

y se calculan las probabilidades posteriores de pertenencia de cada observación $\mathbf{x}_i$ a

cada componente k .

Utilizando la notación definida para el GMM, estas probabilidades se expresan como

$$\gamma_{ik}^{(t)} = P\left(Z_i = k \mid \mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right),$$

y se calculan mediante

$$\gamma_{ik}^{(t)} = \frac{\pi_k^{(t)} \mathcal{N}\left(\mathbf{x}_i \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{\ell=1}^K \pi_\ell^{(t)} \mathcal{N}\left(\mathbf{x}_i \mid \boldsymbol{\mu}_\ell^{(t)}, \boldsymbol{\Sigma}_\ell^{(t)}\right)}.$$

Para cada observación se cumple

$$\sum_{k=1}^K \gamma_{ik}^{(t)} = 1.$$

A partir de estas probabilidades se define la cantidad efectiva de observaciones asociada a cada componente,

$$N_k^{(t)} = \sum_{i=1}^n \gamma_{ik}^{(t)}.$$

3.4.3. Etapa de Maximización

En la etapa de Maximización (M) se mantienen fijas las probabilidades de pertenencia calculadas en la etapa anterior y se actualizan los parámetros del modelo.

El vector de medias de la componente k se actualiza mediante

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^n \gamma_{ik}^{(t)} \mathbf{x}_i.$$

La matriz de covarianza se actualiza como

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^n \gamma_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T.$$

Finalmente, el peso de mezcla de cada componente se actualiza mediante

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{n}.$$

Los nuevos pesos continúan satisfaciendo

$$\pi_k^{(t+1)} \geq 0, \quad \sum_{k=1}^K \pi_k^{(t+1)} = 1.$$

3.4.4. Propiedades y convergencia

Las etapas E y M se aplican sucesivamente, generando una secuencia de estimaciones

$$\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$$

Una propiedad fundamental del algoritmo EM es que cada iteración no disminuye la log-verosimilitud del modelo, es decir,

$$\ell(\boldsymbol{\theta}^{(t+1)}) \geq \ell(\boldsymbol{\theta}^{(t)}).$$

Sin embargo, EM no garantiza alcanzar el máximo global de la log-verosimilitud. La solución obtenida puede corresponder a un máximo local o a otro punto estacionario y puede depender de los valores utilizados para inicializar los parámetros.

En la implementación práctica, las iteraciones se mantienen hasta satisfacer un criterio de convergencia, por ejemplo, cuando la variación de la log-verosimilitud entre iteraciones consecutivas es suficientemente pequeña, o hasta alcanzar un número máximo de iteraciones.

3.5. Coeficiente de silueta

El coeficiente de silueta permite evaluar qué tan similar es una observación a las observaciones de su propio cluster en comparación con las observaciones de otros clusters [13].

Sea $\mathbf{x}_i \in C_k$ una observación perteneciente al cluster C_k .

Cuando $n_k > 1$, la distancia promedio entre $\mathbf{x}_i$ y las demás observaciones de su propio cluster se denotará por $a(i)$ y se define como

$$a(i) = \frac{1}{n_k - 1} \sum_{\substack{\mathbf{x}_j \in C_k \\ j \neq i}} \|\mathbf{x}_i - \mathbf{x}_j\|_2.$$

Para cada cluster C_ℓ distinto de C_k , se considera la distancia promedio

$$d(i, C_\ell) = \frac{1}{n_\ell} \sum_{\mathbf{x}_j \in C_\ell} \|\mathbf{x}_i - \mathbf{x}_j\|_2.$$

La menor de estas distancias se denotará por

$$b(i) = \min_{\ell \neq k} d(i, C_\ell).$$

Para observaciones pertenecientes a clusters con más de un elemento, el coeficiente de silueta se define como

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

En el caso particular de un cluster formado por una única observación, $n_k = 1$, no existen otras observaciones dentro del mismo grupo con las cuales calcular $a(i)$. Para este caso se adopta la convención

$$s(i) = 0.$$

En consecuencia, el coeficiente de silueta satisface

$$s(i) \in [-1, 1].$$

3.6. Entropía de Shannon

La entropía de Shannon es una medida que cuantifica el grado de incertidumbre o diversidad asociado a una distribución de probabilidades [14].

3.6.1. Definición

Sea

$$\mathbf{q} = (q_1, q_2, \dots, q_m)$$

una distribución de probabilidad discreta sobre m elementos, donde

$$q_j \geq 0, \quad \sum_{j=1}^m q_j = 1.$$

La entropía de Shannon se define como

$$H(\mathbf{q}) = - \sum_{j=1}^m q_j \ln(q_j),$$

adoptando la convención

$$0 \ln(0) = 0.$$

3.6.2. Dominio y propiedades

La entropía satisface

$$0 \leq H(\mathbf{q}) \leq \ln(m).$$

El valor mínimo

$$H(\mathbf{q}) = 0$$

se alcanza cuando toda la probabilidad se concentra en un único elemento.

Por otra parte, el valor máximo

$$H(\mathbf{q}) = \ln(m)$$

se alcanza cuando la distribución es uniforme, es decir,

$$q_j = \frac{1}{m}, \quad j = 1, \dots, m.$$

Por lo tanto, valores bajos de entropía representan distribuciones más concentradas, mientras que valores altos representan distribuciones más uniformes entre los elementos.

3.6.3. Entropía normalizada

Debido a que el valor máximo de $H(\mathbf{q})$ depende del número de elementos m , para $m \geq 2$ puede definirse una versión normalizada de la entropía [16] como

$$H_{\text{norm}}(\mathbf{q}) = \frac{H(\mathbf{q})}{\ln(m)}.$$

Esta transformación satisface

$$0 \leq H_{\text{norm}}(\mathbf{q}) \leq 1.$$

Un valor cercano a 0 representa una distribución altamente concentrada, mientras que un valor cercano a 1 representa una distribución aproximadamente uniforme.

Cuando $m = 1$, se tiene $H(\mathbf{q}) = 0$ y la expresión normalizada anterior no está definida debido a que $\ln(1) = 0$. En este caso, se define

$$H_{\text{norm}}(\mathbf{q}) = 0,$$

ya que una distribución formada por un único elemento corresponde al nivel mínimo de diversidad.

3.7. Índice de Herfindahl–Hirschman

El índice de Herfindahl–Hirschman (HHI) es una medida de concentración definida sobre una distribución de proporciones [15].

3.7.1. Definición

Sea

$$\mathbf{q} = (q_1, q_2, \dots, q_m)$$

una distribución de proporciones, donde

$$q_j \geq 0, \quad \sum_{j=1}^m q_j = 1.$$

El índice de Herfindahl–Hirschman se define como

$$\text{HHI}(\mathbf{q}) = \sum_{j=1}^m q_j^2.$$

3.7.2. Dominio y propiedades

El índice satisface

$$\frac{1}{m} \leq \text{HHI}(\mathbf{q}) \leq 1.$$

El valor mínimo

$$\text{HHI}(\mathbf{q}) = \frac{1}{m}$$

se alcanza cuando la distribución es uniforme, es decir,

$$q_j = \frac{1}{m}, \quad j = 1, \dots, m.$$

Por otra parte, el valor máximo

$$\text{HHI}(\mathbf{q}) = 1$$

se alcanza cuando toda la proporción se concentra en un único elemento.

Por lo tanto, valores altos del índice representan una mayor concentración, mientras que valores bajos representan una distribución más uniforme entre los elementos.

Además, dado que se elevan al cuadrado las proporciones, los elementos con mayor peso tienen una influencia mayor sobre el valor final del índice.

3.8. Medidas descriptivas de dispersión

Para complementar el análisis gráfico de las distribuciones se utilizan medidas descriptivas que permiten cuantificar qué tan dispersos se encuentran los valores de una variable entre los usuarios. Entre estas medidas se consideran el recorrido y el rango intercuartílico, utilizados habitualmente en el análisis exploratorio de datos [9]. De manera complementaria, en este trabajo se considera también la diferencia entre los percentiles 90 y 10.

Sea $x_{\min}$ el menor valor observado y $x_{\max}$ el mayor. El recorrido se define como

$$R = x_{\max} - x_{\min}.$$

Esta medida representa la amplitud total de los valores observados. Un recorrido mayor indica que existe una mayor distancia entre los valores extremos de la distribución. Sin

embargo, al depender directamente del mínimo y del máximo, puede verse afectado por la presencia de valores extremos.

Para describir la dispersión de la parte central de la distribución se utiliza el rango intercuartílico. Sean Q_1 y Q_3 el primer y tercer cuartil, correspondientes a los percentiles 25 y 75, respectivamente. El rango intercuartílico se define como

$$IQR = Q_3 - Q_1.$$

El intervalo comprendido entre Q_1 y Q_3 contiene al 50 % central de las observaciones. Por lo tanto, un valor pequeño de IQR indica que esta parte de los usuarios presenta valores relativamente cercanos entre sí, mientras que un valor mayor indica una mayor variabilidad en la zona central de la distribución.

De manera complementaria, se considera la diferencia entre los percentiles 90 y 10,

$$R_{80} = P_{90} - P_{10}.$$

El intervalo comprendido entre P_{10} y P_{90} contiene al 80 % central de las observaciones. Esta medida permite considerar una parte más amplia de la distribución, excluyendo solamente el 10 % de los valores más bajos y el 10 % de los más altos. Al igual que con el IQR , valores mayores de R_{80} representan una mayor dispersión entre los usuarios.

La interpretación se realiza considerando ambas medidas de manera conjunta. El IQR describe la variabilidad del 50 % central de los usuarios, mientras que R_{80} amplía el análisis al 80 % central. Si ambas medidas presentan valores elevados, la variabilidad se encuentra presente en una parte amplia de la distribución. En cambio, un IQR relativamente pequeño acompañado de un R_{80} mayor indica que los usuarios centrales presentan valores similares, pero existe una mayor separación al incorporar observaciones más alejadas del centro.

En este trabajo estas medidas no se utilizan con puntos de corte fijos para determinar si la dispersión es baja o alta. Su interpretación se realiza comparando sus magnitudes entre AutoCAD, Revit y TimeDoctor para un mismo indicador. De esta forma, permiten complementar la evidencia gráfica y cuantificar las diferencias observadas en los valores de los indicadores entre los usuarios.

3.9. Prueba Chi-cuadrado de independencia

La prueba Chi-cuadrado de independencia permite evaluar si existe asociación entre dos variables categóricas a partir de las frecuencias observadas en una tabla de contingencia [17].

3.9.1. Tabla de contingencia y frecuencias esperadas

Sean X e Y dos variables categóricas con dominios

$$X \in \{1, \dots, r\}, \quad Y \in \{1, \dots, c\},$$

donde r y c corresponden al número de categorías de cada variable.

Se denotará por O_{ij} la frecuencia observada correspondiente a la categoría i de X y la categoría j de Y , con

$$i \in \{1, \dots, r\}, \quad j \in \{1, \dots, c\}.$$

El número total de observaciones está dado por

$$n = \sum_{i=1}^r \sum_{j=1}^c O_{ij}.$$

Las frecuencias marginales se definen como

$$O_{i\cdot} = \sum_{j=1}^c O_{ij}, \quad O_{\cdot j} = \sum_{i=1}^r O_{ij}.$$

Bajo el supuesto de independencia entre X e Y , la frecuencia esperada en la celda (i, j) está dada por

$$E_{ij} = \frac{O_{i\cdot} O_{\cdot j}}{n}.$$

3.9.2. Estadístico Chi-cuadrado

El estadístico Chi-cuadrado compara las frecuencias observadas con las frecuencias esperadas bajo la hipótesis de independencia y se define como

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

El estadístico satisface

$$\chi^2 \geq 0.$$

Si las frecuencias observadas coinciden con las esperadas, entonces

$$\chi^2 = 0.$$

Valores mayores de χ^2 indican una mayor discrepancia entre las frecuencias observadas y aquellas que se esperarían bajo independencia.

3.9.3. Hipótesis y distribución del estadístico

La prueba considera las hipótesis

$$H_0 : X \text{ e } Y \text{ son independientes,}$$

$$H_1 : X \text{ e } Y \text{ no son independientes.}$$

Bajo H_0 y cuando se cumplen las condiciones para utilizar la aproximación Chi-cuadrado, el estadístico sigue aproximadamente una distribución Chi-cuadrado con

$$\nu = (r - 1)(c - 1)$$

grados de libertad, es decir,

$$\chi^2 \sim \chi_\nu^2.$$

A partir del valor observado del estadístico se obtiene el valor p . Para un nivel de significancia α , se rechaza la hipótesis nula cuando

$$p < \alpha.$$

En caso contrario, no existe evidencia estadística suficiente para rechazar la hipótesis de independencia.

3.10. Coeficiente V de Cramér

El coeficiente V de Cramér es una medida de asociación entre dos variables categóricas. Se construye a partir del estadístico Chi-cuadrado definido en la Sección 3.9 y permite cuantificar la intensidad de la relación entre las variables [17].

3.10.1. Definición

Sean X e Y dos variables categóricas con r y c categorías, respectivamente. El coeficiente V de Cramér se define como

$$V = \sqrt{\frac{\chi^2}{n \min(r-1, c-1)}}.$$

En esta expresión, χ^2 corresponde al estadístico Chi-cuadrado de independencia, n al número total de observaciones, y r y c al número de categorías de las variables comparadas.

3.10.2. Dominio e interpretación

El coeficiente satisface

$$0 \leq V \leq 1.$$

Un valor

$$V = 0$$

indica ausencia de asociación entre las variables, mientras que valores mayores representan una asociación de mayor intensidad.

Valores cercanos a 1 indican una asociación fuerte entre las variables categóricas.

El coeficiente V de Cramér mide la intensidad de la asociación, pero no su dirección. Por lo tanto, a diferencia de otras medidas de asociación utilizadas para variables numéricas, no distingue entre relaciones positivas y negativas.

3.10.3. Relación con la prueba Chi-cuadrado

La prueba Chi-cuadrado y el coeficiente V de Cramér entregan información complementaria.

La prueba Chi-cuadrado permite evaluar si existe evidencia estadística para rechazar la hipótesis de independencia entre las variables, mientras que V de Cramér permite cuantificar la magnitud de dicha asociación.

Por esta razón, un resultado estadísticamente significativo en la prueba Chi-cuadrado no implica necesariamente una asociación de gran intensidad. En el análisis, ambos resultados deben interpretarse de manera conjunta.

3.11. Prueba de Kruskal–Wallis

La prueba de Kruskal–Wallis es una prueba no paramétrica utilizada para comparar G grupos independientes. A diferencia de métodos paramétricos basados directamente en los valores observados, esta prueba utiliza los rangos de las observaciones y no requiere suponer normalidad en los datos. En esta investigación se utiliza para evaluar si la distribución de un indicador presenta diferencias entre los clusters obtenidos [18].

3.11.1. Rangos y formulación

Sean G grupos independientes, con n_g observaciones en el grupo g , donde

$$g \in \{1, \dots, G\}.$$

El número total de observaciones está dado por

$$N = \sum_{g=1}^G n_g.$$

Las N observaciones se consideran conjuntamente y se ordenan de menor a mayor, asignándoles rangos entre 1 y N . Se denotará por R_g la suma de los rangos correspondientes a las observaciones del grupo g .

El estadístico de Kruskal–Wallis se define como

$$H_{\text{KW}} = \frac{12}{N(N+1)} \sum_{g=1}^G \frac{R_g^2}{n_g} - 3(N+1).$$

El estadístico satisface

$$H_{\text{KW}} \geq 0.$$

Valores pequeños de H_{KW} indican que los rangos se distribuyen de manera similar entre los grupos, mientras que valores mayores indican una mayor diferencia entre ellos.

3.11.2. Hipótesis de la prueba

La prueba considera las hipótesis

H_0 : los grupos presentan la misma distribución,

H_1 : al menos uno de los grupos presenta una distribución diferente.

Bajo la hipótesis nula y para tamaños muestrales suficientemente grandes, el estadístico H_{KW} se aproxima a una distribución Chi-cuadrado con

$$\nu = G - 1$$

grados de libertad, es decir,

$$H_{KW} \approx \chi_{G-1}^2.$$

A partir del estadístico se obtiene un valor p . Para un nivel de significancia α , se rechaza la hipótesis nula cuando

$$p < \alpha.$$

En caso contrario, no existe evidencia estadística suficiente para rechazar H_0 .

3.11.3. Interpretación

El rechazo de H_0 permite concluir que existe evidencia de diferencias entre los grupos considerados, pero no permite determinar directamente cuáles grupos difieren entre sí.

Por esta razón, cuando la prueba de Kruskal–Wallis resulta significativa, el análisis puede complementarse mediante comparaciones por pares. En esta investigación, estas comparaciones se realizan aplicando nuevamente la prueba de Kruskal–Wallis a cada par de clusters, considerando en cada caso $G = 2$ grupos.

3.12. Combinación de valores p mediante el método de Fisher

Cuando una misma hipótesis se evalúa de manera independiente sobre distintos conjuntos de datos, es posible combinar los p -values obtenidos con el objetivo de resumir la evidencia estadística en un único resultado global. Existen distintos procedimientos para realizar esta combinación, cuya elección depende de las características de las pruebas que originan los p -values [19].

Sean

$$p_1, p_2, \dots, p_m$$

los p -values obtenidos a partir de m pruebas independientes. Bajo la hipótesis nula conjunta, se considera que

$$p_i \sim U(0, 1), \quad i = 1, \dots, m.$$

El método de Fisher combina estos valores mediante el estadístico

$$X_F = -2 \sum_{i=1}^m \ln(p_i).$$

Bajo la hipótesis nula y suponiendo independencia entre las pruebas, este estadístico sigue una distribución Chi-cuadrado con $2m$ grados de libertad,

$$X_F \sim \chi_{2m}^2.$$

Por lo tanto, una vez calculado el valor observado de X_F , el p -value combinado se obtiene a partir de la probabilidad de observar un valor igual o mayor bajo esta distribución,

$$p_{\text{comb}} = P(\chi_{2m}^2 \geq X_F).$$

En esta investigación se consideran los resultados obtenidos en AutoCAD, Revit y TimeDoctor, por lo que, al combinar los tres p -values correspondientes a una misma comparación,

$$m = 3,$$

y el estadístico de Fisher se evalúa respecto de una distribución Chi-cuadrado con

$$2m = 6$$

grados de libertad,

$$X_F \sim \chi_6^2.$$

La elección del método de Fisher se basa en el criterio presentado por Heard y Rubin-Delanchy [19]. Los autores señalan que este método resulta adecuado cuando los p -value provienen de pruebas cuyos estadísticos son positivos y tienden a tomar valores mayores bajo la hipótesis alternativa. Esta situación corresponde a las pruebas utilizadas en esta investigación, ya que tanto el estadístico Chi-cuadrado como el estadístico de Kruskal–Wallis son no negativos y valores mayores representan una mayor discrepancia respecto de la hipótesis nula.

Por esta razón, el método de Fisher se utiliza para combinar, por separado, los p -values obtenidos en AutoCAD, Revit y TimeDoctor para una misma comparación en la prueba Chi-cuadrado y para un mismo indicador en la prueba de Kruskal–Wallis.

Se utiliza un nivel de significancia de $\alpha = 0,05$. Cuando

$$p_{\text{comb}} < 0,05,$$

se considera que existe evidencia conjunta contra la hipótesis nula al considerar los resultados de las tres plataformas. Este resultado global complementa los p -values individuales y no permite concluir, por sí solo, que la hipótesis nula sea rechazada en cada plataforma de manera independiente.

Para la aplicación del método de Fisher en esta investigación, se asume independencia entre los p -values obtenidos en AutoCAD, Revit y TimeDoctor. Este supuesto permite combinar los resultados de las tres plataformas y evaluar la evidencia conjunta asociada a una misma hipótesis.

Capítulo 4

Construcción de variables

A partir de las bases preparadas en el Capítulo 2, los registros fueron agrupados por usuario con el objetivo de transformar la información contenida en eventos individuales en variables que permitieran describir características más generales de su actividad.

La selección de estas variables surgió a partir de una etapa exploratoria de los datos, en la que se revisó qué información podía obtenerse de manera consistente a partir de AutoCAD, Revit y TimeDoctor. En particular, los registros disponibles permitían observar aspectos relacionados con la cantidad y duración de la actividad, el uso de documentos, la distribución temporal de los registros y la continuidad entre actividades.

A partir de esta exploración se definieron variables que resumieran estos aspectos a nivel de usuario y que, al mismo tiempo, fueran interpretables y comparables entre las tres plataformas. La intención no fue asumir previamente que cada indicador permitiría distinguir comportamientos relevantes, sino construir medidas que representaran características observables en los registros y evaluar posteriormente su comportamiento en los datos.

Se consideraron cinco variables. La primera corresponde a un perfil general de actividad, construido a partir de un conjunto de métricas agregadas y técnicas de agrupamiento no supervisado. Su objetivo es representar de manera conjunta distintas características de la actividad de cada usuario.

Las cuatro variables restantes corresponden a indicadores que describen aspectos espe-

cíficos de los registros. Estos permiten estudiar cómo se distribuye la actividad a lo largo del día, cómo se reparte el tiempo entre los documentos utilizados y qué tan continua o interrumpida es la actividad registrada.

Las variables consideradas son:

- Perfil general de actividad.
- Concentración horaria.
- Diversidad documental.
- Concentración documental.
- Fragmentación temporal.

La construcción de estas variables constituye una primera etapa del análisis. En el Capítulo 5 se estudia su distribución y variabilidad entre los usuarios, mientras que en el Capítulo 6 se analiza su relación con los perfiles generales de actividad y entre los propios indicadores. De esta forma, su utilidad no se establece únicamente a partir de su definición, sino también a partir del comportamiento que presentan posteriormente en los datos.

4.1. Variable 1: Perfil general de actividad

4.1.1. Descripción de la variable

El perfil general de actividad busca representar, de manera conjunta, distintas características de la actividad registrada para cada usuario durante el período de estudio.

Para construirlo, los registros se agruparon por usuario y se calcularon métricas relacionadas con la cantidad de registros, los documentos utilizados y los tiempos de actividad. Estas métricas se utilizaron posteriormente en los métodos de agrupamiento para identificar usuarios con características similares [10].

4.1.2. Pregunta de investigación

A partir de esta variable se plantea la siguiente pregunta e hipótesis:

Pregunta: ¿Es posible identificar grupos de usuarios con características diferenciadas a partir de las métricas agregadas de su actividad?

Hipótesis: Las métricas agregadas de actividad permiten identificar distintos perfiles de usuarios, caracterizados por diferentes combinaciones en la cantidad de registros, los documentos utilizados y los tiempos de actividad.

4.1.3. Variables utilizadas

Para construir el perfil general de actividad, los registros fueron agrupados por usuario y se calcularon métricas a partir de los tiempos registrados, los documentos utilizados y las fechas asociadas a cada interacción.

Sea R_u el conjunto de registros asociados al usuario u y t_r la duración del registro r , según la definición presentada en el Capítulo 2. La Tabla [4.1](#) presenta las métricas obtenidas, su definición, su dominio y el rango de valores común observado en AutoCAD, Revit y TimeDoctor.

Cuadro 4.1: Métricas utilizadas para construir el perfil general de actividad

Medida	Definición y origen	Dominio	Rango común observado
Cantidad de registros	Número total de registros asociados al usuario.	$\mathbb{N}$	[49, 5 779]
Documentos distintos	Cantidad de documentos diferentes utilizados por el usuario.	$\mathbb{N}$	[11, 473]
Tiempo total	Suma de la duración de todos los registros del usuario: $\sum_{r \in R_u} t_r$.	$\mathbb{R}_{\geq 0}$	[1 884, 647 010] s
Tiempo promedio por registro	Promedio de la duración de los registros asociados al usuario.	$\mathbb{R}_{\geq 0}$	[44,96, 7 624,02] s
Promedio de documentos diarios	Promedio de la cantidad de documentos utilizados diariamente.	$\mathbb{R}_{\geq 0}$	[0,483, 2,375]
Variabilidad de documentos diarios	Desviación estándar de la cantidad de documentos utilizados diariamente.	$\mathbb{R}_{\geq 0}$	[1,576, 12,644]
Proporción en el documento principal	Proporción del tiempo total correspondiente al documento de mayor uso.	[0, 1]	[0,127, 0,937]
Promedio del máximo diario	Promedio de los mayores tiempos de actividad registrados en cada día.	$\mathbb{R}_{\geq 0}$	[142, 27 158,88] s
Variabilidad del máximo diario	Desviación estándar de los mayores tiempos de actividad registrados en cada día.	$\mathbb{R}_{\geq 0}$	[171,77, 103 922,36] s

4.1.4. Construcción del espacio de características

Cada usuario fue representado mediante un vector $\mathbf{x}_i$ formado por las métricas definidas anteriormente. Estos vectores constituyen los datos de entrada para los métodos de agrupamiento.

Debido a que las métricas presentan escalas diferentes, fue necesario estandarizarlas antes de aplicar los algoritmos. De esta forma, se evita que las variables con valores numéricamente mayores tengan más peso en el cálculo de las distancias [11].

4.1.5. Modelos de agrupamiento

Para identificar perfiles de comportamiento se aplicaron los métodos K-Means y Modelos de Mezclas Gaussianas (GMM), cuyos fundamentos se presentan en las Secciones 3.2 y 3.3, respectivamente.

Los parámetros del GMM fueron estimados mediante el algoritmo Expectation-Maximization, descrito en la Sección 3.4.

4.1.6. Selección del número de clusters

Para determinar el número de clusters se evaluaron valores de K entre 2 y 6, tanto para K-Means como para GMM. Para cada configuración se calculó el coeficiente de silueta definido en la Sección 3.5.

Los resultados obtenidos con K-Means se presentan en la Tabla 4.2.

Plataforma	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
AutoCAD	0,422	0,503	0,538	0,388	0,370
Revit	0,354	0,428	0,413	0,423	0,418
TimeDoctor	0,369	0,497	0,294	0,398	0,403

Cuadro 4.2: Coeficiente de silueta obtenido mediante K-Means para distintos valores de K

En K-Means, el mayor coeficiente se obtuvo con $K = 4$ para AutoCAD y con $K = 3$ para Revit y TimeDoctor. Para GMM, los resultados fueron los siguientes:

Plataforma	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
AutoCAD	0,316	0,322	0,217	0,162	0,181
Revit	0,360	0,292	0,184	0,150	0,045
TimeDoctor	0,420	0,287	0,125	0,003	-0,065

Cuadro 4.3: Coeficiente de silueta obtenido mediante GMM para distintos valores de K

En este caso, el mayor coeficiente se obtuvo con $K = 3$ para AutoCAD y con $K = 2$ para Revit y TimeDoctor.

Dado que el valor que maximiza el coeficiente no fue el mismo en todos los casos, se seleccionó $K = 3$ como una configuración común para las tres plataformas y ambos métodos. Esta elección permite mantener la misma cantidad de grupos en los análisis posteriores y facilita la comparación de los resultados entre AutoCAD, Revit y TimeDoctor.

4.2. Indicador 2: Concentración horaria

4.2.1. Descripción del indicador

El indicador de concentración horaria describe cómo se distribuye el tiempo de actividad de cada usuario a lo largo del día. Para ello, se compara la proporción de tiempo registrada durante la madrugada y mañana con la registrada durante la tarde y noche.

A partir de esta distribución se busca identificar si la actividad se concentra principalmente en uno de estos períodos o si se reparte de manera más equilibrada entre ambos.

4.2.2. Pregunta de investigación

A partir de este indicador se plantea la siguiente pregunta e hipótesis:

Pregunta: ¿Los usuarios presentan diferencias en la forma en que distribuyen su actividad a lo largo del día?

Hipótesis: Los usuarios presentan distintos niveles de concentración horaria, reflejados en diferentes proporciones de actividad durante la mañana-madrugada y la tarde-noche.

4.2.3. Construcción de la variable

Para construir el indicador, cada registro se asignó a uno de dos períodos según su hora de inicio: mañana-madrugada o tarde-noche.

Luego, para cada usuario se sumó el tiempo registrado en ambos períodos. Si T_{MM} corresponde al tiempo acumulado durante la mañana-madrugada y T_{TN} al tiempo acumulado durante la tarde-noche, entonces el tiempo total del usuario es

$$T_{Total} = T_{MM} + T_{TN}.$$

A partir de estos valores se calcularon las proporciones

$$P_{MM} = \frac{T_{MM}}{T_{Total}}, \quad P_{TN} = \frac{T_{TN}}{T_{Total}},$$

con

$$T_{\text{MM}}, T_{\text{TN}} \geq 0, \quad T_{\text{Total}} > 0.$$

Por lo tanto,

$$P_{\text{MM}}, P_{\text{TN}} \in [0, 1], \quad P_{\text{MM}} + P_{\text{TN}} = 1.$$

Como ambas proporciones son complementarias, se utilizó P_{MM} como indicador de concentración horaria. Su dominio es

$$P_{\text{MM}} \in [0, 1].$$

Valores cercanos a 1 indican una mayor concentración de la actividad durante la mañana–mañana, mientras que valores cercanos a 0 indican una mayor concentración durante la tarde–noche.

4.3. Indicador 3: Diversidad documental

4.3.1. Descripción del indicador

El indicador de diversidad documental describe qué tan uniformemente se distribuye el tiempo de actividad de un usuario entre los documentos que utiliza.

Una mayor diversidad documental indica que el tiempo se encuentra distribuido de forma más uniforme entre los documentos utilizados, mientras que una menor diversidad indica que una mayor proporción del tiempo se concentra en algunos de ellos.

4.3.2. Pregunta de investigación

A partir de este indicador se plantea la siguiente pregunta e hipótesis:

Pregunta: ¿Los usuarios presentan diferencias en la diversidad con que distribuyen su tiempo entre los documentos utilizados?

Hipótesis: Los usuarios presentan distintos niveles de diversidad documental, existiendo casos en que el tiempo se reparte de manera más uniforme entre los documentos utilizados y otros en que una mayor proporción del tiempo se concentra en algunos de ellos.

4.3.3. Aplicación al indicador de diversidad documental

Para cada usuario se identificaron los m documentos utilizados y el tiempo acumulado en cada uno de ellos. Sea T_j el tiempo registrado en el documento j y T_{Total} el tiempo total del usuario. La proporción de tiempo asociada a cada documento se define como

$$q_j = \frac{T_j}{T_{\text{Total}}}, \quad j = 1, \dots, m.$$

De esta forma, cada usuario queda representado por la distribución

$$\mathbf{q} = (q_1, q_2, \dots, q_m),$$

donde

$$q_j \in [0, 1], \quad \sum_{j=1}^m q_j = 1.$$

Sobre esta distribución se calculó la entropía normalizada de Shannon definida en la Sección 3.6. El indicador de diversidad documental satisface

$$H_{\text{norm}}(\mathbf{q}) \in [0, 1].$$

Valores cercanos a 1 indican que el tiempo se reparte de manera más uniforme entre los documentos utilizados, mientras que valores cercanos a 0 indican que una mayor proporción del tiempo se concentra en algunos de ellos.

4.4. Indicador 4: Concentración documental

4.4.1. Descripción e interpretación del indicador

El indicador de concentración documental busca representar cuánto del tiempo de actividad de un usuario se concentra en algunos de los documentos que utiliza.

Valores altos indican que una mayor proporción del tiempo se concentra en pocos documentos, mientras que valores bajos indican que el tiempo se distribuye de manera más equilibrada entre ellos.

Aunque este indicador se construye a partir de la misma distribución de proporciones que la diversidad documental, ambos representan aspectos distintos. La diversidad documental, medida mediante la entropía normalizada de Shannon, considera qué tan uniformemente se distribuye el tiempo entre el conjunto de documentos utilizados. La concentración documental, en cambio, se construye mediante el índice de Herfindahl–Hirschman y da mayor peso a los documentos que concentran una mayor proporción del tiempo. Por esta razón, ambos indicadores se analizan de manera complementaria y no corresponden simplemente a medidas inversas entre sí.

4.4.2. Pregunta de investigación e hipótesis

A partir de este indicador se plantea la siguiente pregunta e hipótesis:

Pregunta: ¿Los usuarios presentan diferencias en el grado de concentración de su tiempo entre los documentos utilizados?

Hipótesis: Los usuarios presentan distintos niveles de concentración documental, existiendo casos en que una mayor proporción del tiempo se concentra en algunos documentos y otros en que el tiempo se distribuye de manera más equilibrada entre los documentos utilizados.

4.4.3. Construcción del indicador

Para este indicador se utiliza la misma distribución de proporciones de tiempo definida para la diversidad documental,

$$\mathbf{q} = (q_1, q_2, \dots, q_m),$$

donde q_j representa la proporción del tiempo total del usuario asociada al documento j , con

$$q_j \in [0, 1], \quad \sum_{j=1}^m q_j = 1.$$

Sobre esta distribución se calcula el índice de Herfindahl–Hirschman definido en la Sección [3.7](#).

$$\text{HHI}(\mathbf{q}) = \sum_{j=1}^m q_j^2.$$

Para un usuario que utiliza m documentos, el indicador satisface

$$\frac{1}{m} \leq \text{HHI}(\mathbf{q}) \leq 1.$$

Valores cercanos a 1 indican que una proporción importante del tiempo se concentra en una parte reducida de los documentos utilizados, mientras que valores cercanos a $\frac{1}{m}$ indican que el tiempo se distribuye de manera más uniforme entre ellos.

4.5. Indicador 5: Fragmentación temporal

4.5.1. Descripción del indicador

El indicador de fragmentación temporal describe qué tan continua o interrumpida es la actividad registrada de cada usuario durante el período de estudio.

Su objetivo es diferenciar a los usuarios cuya actividad se desarrolla de forma más continua de aquellos que presentan interrupciones más frecuentes o prolongadas entre un registro y el siguiente.

4.5.2. Pregunta de investigación

A partir de este indicador se plantea la siguiente pregunta e hipótesis:

Pregunta: ¿Los usuarios presentan diferencias en la continuidad de su actividad durante el período de estudio?

Hipótesis: Los usuarios presentan distintos niveles de fragmentación temporal, con casos en que la actividad se desarrolla de forma más continua y otros en que presenta interrupciones más frecuentes o prolongadas.

4.5.3. Construcción de la variable

Para cada usuario, los registros se ordenaron cronológicamente y se calculó el tiempo entre el término de un registro y el inicio del siguiente. Para dos registros consecutivos, este intervalo se define como

$$\Delta t_i = t_{\text{inicio}}^{(i)} - t_{\text{trmino}}^{(i-1)},$$

donde $t_{\text{inicio}}^{(i)}$ corresponde al inicio del registro actual y $t_{\text{trmino}}^{(i-1)}$ al término del registro anterior. Para los intervalos considerados como interrupciones se tiene

$$\Delta t_i \geq 0.$$

A partir de estos intervalos se calcularon dos medidas para cada usuario: el número promedio de interrupciones por día, I_{da} , y el tiempo promedio entre actividades, $\bar{t}_{\text{muerto}}$. Ambas satisfacen

$$I_{\text{da}} \in \mathbb{R}_{\geq 0}, \quad \bar{t}_{\text{muerto}} \in \mathbb{R}_{\geq 0}.$$

El índice de fragmentación se definió como

$$F = I_{\text{da}} \bar{t}_{\text{muerto}},$$

por lo que

$$F \in \mathbb{R}_{\geq 0}.$$

De esta forma, F considera tanto la frecuencia de las interrupciones como su duración. Un valor mayor representa una actividad más fragmentada.

Para facilitar la comparación entre usuarios, el índice se normalizó mediante

$$I_{\text{Fragmentacin}} = \frac{F - F_{\text{mín}}}{F_{\text{máx}} - F_{\text{mín}}},$$

donde $F_{\text{mín}}$ y $F_{\text{máx}}$ corresponden al menor y mayor valor observado, respectivamente. De esta forma, siempre que $F_{\text{máx}} > F_{\text{mín}}$, el dominio del indicador final es

$$I_{\text{Fragmentacin}} \in [0, 1].$$

Valores cercanos a 0 representan una actividad más continua dentro del conjunto analizado, mientras que valores cercanos a 1 representan una mayor fragmentación temporal.

En el cálculo también se consideraron los fines de semana y días feriados, con el fin de evitar que períodos en los que normalmente no existe actividad laboral fueran contabilizados como interrupciones entre registros. De esta forma, la fragmentación busca representar las interrupciones ocurridas durante los períodos en que efectivamente podía existir actividad registrada.

4.5.4. Concepto de fragmentación temporal

En esta investigación, la fragmentación temporal representa el grado de interrupción de la actividad de un usuario. Para medirla se consideran de forma conjunta la frecuencia de las interrupciones y su duración.

Esta distinción es importante, ya que un mismo tiempo total sin actividad puede producirse de formas diferentes. Por ejemplo, un usuario puede presentar muchas interrupciones breves, mientras que otro puede tener pocas interrupciones de mayor duración.

Al combinar ambas componentes en un solo indicador, distintas combinaciones pueden dar lugar a valores similares de fragmentación. Por esta razón, el indicador permite comparar el nivel general de fragmentación entre usuarios, pero no permite identificar por sí solo si este valor se debe principalmente a la cantidad de interrupciones o a su duración.

Capítulo 5

Validación de las variables

En este capítulo se analizan las variables construidas en el capítulo anterior para evaluar su comportamiento en AutoCAD, Revit y TimeDoctor.

Para el perfil general de actividad se comparan los resultados obtenidos mediante K-Means y GMM. Se revisa la cantidad de usuarios asignados a cada cluster y los valores promedio de las métricas que los caracterizan. A partir de esta comparación se interpretan los grupos obtenidos y se selecciona el método que se utilizará en los análisis posteriores.

Para los cuatro indicadores se estudian sus distribuciones con el objetivo de observar las diferencias entre usuarios y plataformas. A partir de estas distribuciones se definen categorías con fines descriptivos, que permiten interpretar los distintos niveles observados en cada indicador.

Finalmente, el análisis se complementa con el rango intercuartílico (IQR) y la diferencia entre los percentiles 90 y 10 ($P_{90} - P_{10}$). Estas medidas permiten cuantificar la variabilidad de los indicadores y complementar la evidencia gráfica al responder las preguntas de investigación planteadas para cada uno.

5.1. Validación del perfil general de actividad

La validación del perfil general de actividad se realizó en dos etapas. Primero, se analizaron las distribuciones de las métricas utilizadas en su construcción para observar si presentan comportamientos similares entre AutoCAD, Revit y TimeDoctor.

Luego, se compararon los agrupamientos obtenidos mediante K-Means y GMM. Para cada método se analizó la cantidad de usuarios por cluster y los valores promedio de las métricas que caracterizan cada grupo. A partir de esta comparación se interpretaron los perfiles obtenidos y se evaluó cuál de los dos métodos entrega una representación más adecuada del comportamiento observado.

5.1.1. Distribución de las métricas agregadas

Antes de comparar los agrupamientos obtenidos mediante K-Means y GMM, se analizó la distribución de las métricas utilizadas para construir el perfil general de actividad. El objetivo de esta revisión es observar cómo se distribuyen sus valores entre los usuarios y determinar qué similitudes y diferencias existen entre AutoCAD, Revit y TimeDoctor.

Para ello se utilizaron tres representaciones complementarias. Los diagramas de caja permiten comparar la posición central, la dispersión y la presencia de valores alejados del resto; los histogramas permiten observar la forma de las distribuciones y las zonas donde se concentra una mayor cantidad de usuarios; y las distribuciones acumuladas permiten comparar directamente la proporción de usuarios que se encuentra bajo distintos valores de cada métrica [9].

En las tres figuras, los resultados se presentan por separado para AutoCAD, Revit y TimeDoctor, manteniendo la misma estructura para las nueve métricas analizadas.

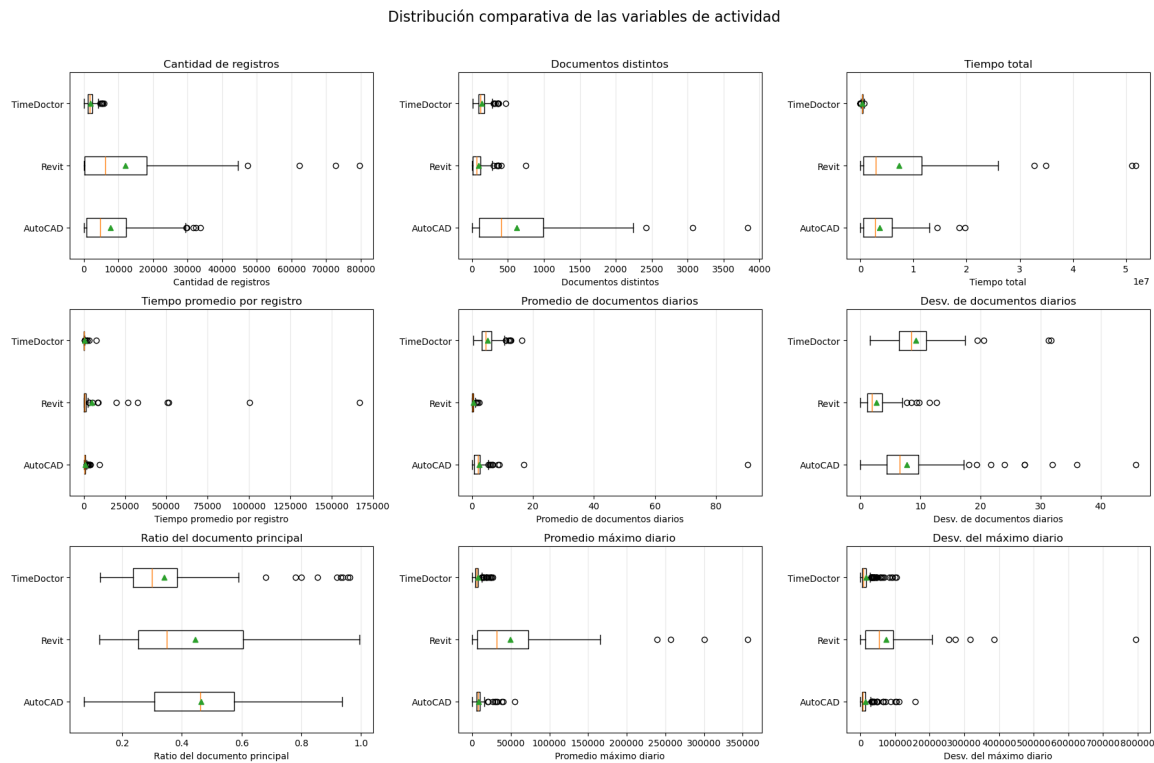


Figura 5.1: Diagramas de caja de las variables de actividad.

La Figura 5.1 presenta los diagramas de caja de las nueve métricas. En cada subfigura se comparan las tres plataformas. La línea naranja corresponde a la mediana, el triángulo verde al promedio y los círculos negros a valores atípicos.

En cantidad de registros, documentos distintos y tiempo total se observa una alta dispersión entre los usuarios y la presencia de valores considerablemente alejados del resto. AutoCAD alcanza valores mayores en documentos distintos, mientras que Revit presenta los valores más altos en cantidad de registros y tiempo total. TimeDoctor, en cambio, muestra rangos más reducidos en estas métricas.

Las métricas asociadas a la duración de la actividad también presentan una dispersión desigual. En el tiempo promedio por registro, el promedio del máximo diario y su desviación estándar, la mayor parte de los usuarios se concentra en valores relativamente bajos, mientras que un grupo reducido alcanza valores mucho mayores. Este comportamiento es especialmente visible en Revit.

En las métricas asociadas al uso diario de documentos, AutoCAD y TimeDoctor presentan valores más próximos entre sí en el promedio y la desviación de documentos diarios, mientras que Revit se concentra en valores menores. El ratio del documento principal muestra una mayor cercanía entre las tres plataformas, aunque se mantienen diferencias en su dispersión.

Por lo tanto, los diagramas de caja muestran que las plataformas comparten algunas características, como la presencia de distribuciones asimétricas y usuarios alejados de los valores centrales, pero también presentan diferencias importantes en la magnitud y dispersión de determinadas métricas.

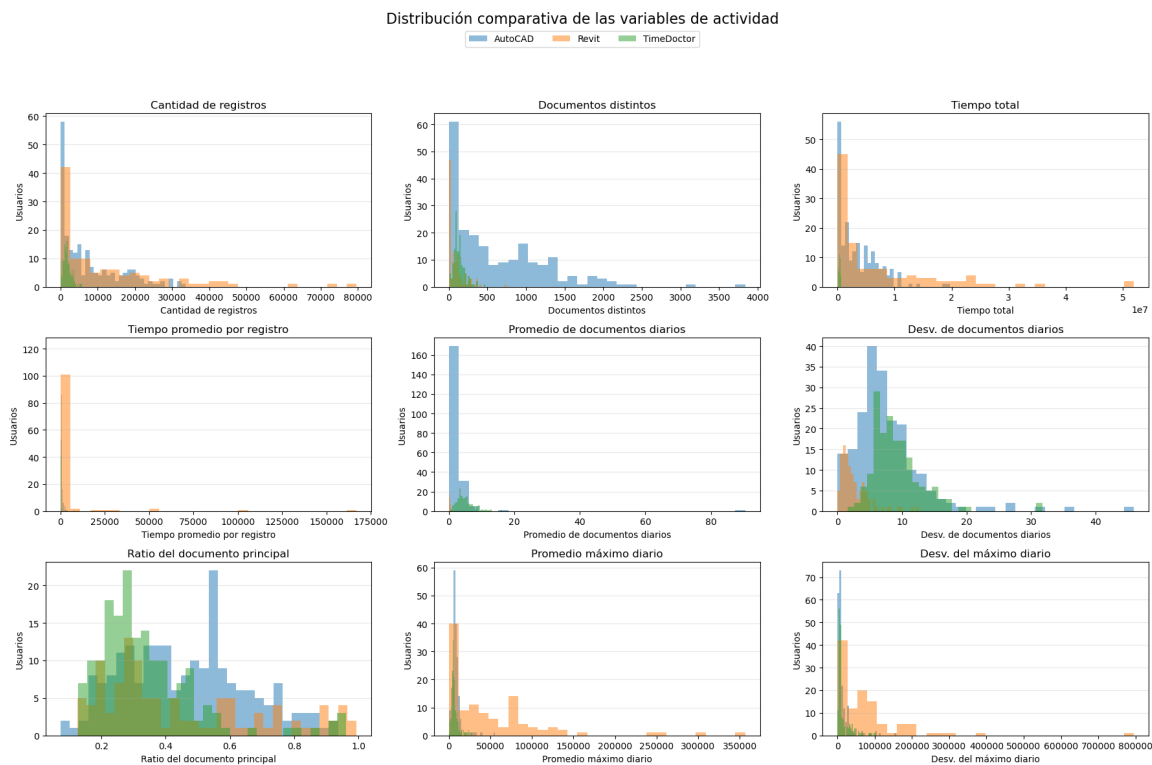


Figura 5.2: Histogramas de las variables de actividad.

La Figura 5.2 permite observar con mayor claridad la forma de las distribuciones. Debido a que la cantidad de usuarios es distinta en cada plataforma, la comparación se centra principalmente en la ubicación y forma de las distribuciones, y no en la altura de las barras.

En cantidad de registros, documentos distintos y tiempo total, la mayor parte de los usuarios se concentra en los valores bajos y la frecuencia disminuye a medida que estos

aumentan. Esta forma se observa en las tres plataformas, aunque los rangos son distintos: AutoCAD se extiende hacia valores mayores en documentos distintos y Revit en cantidad de registros y tiempo total.

Una forma similar aparece en el tiempo promedio por registro, el promedio del máximo diario y la desviación del máximo diario. Las distribuciones presentan una concentración importante en valores bajos y una cola hacia valores mayores. Nuevamente, Revit muestra la mayor extensión hacia valores altos.

En el promedio y la desviación de documentos diarios se observa una mayor superposición entre AutoCAD y TimeDoctor, mientras que Revit mantiene una distribución desplazada hacia valores menores. El ratio del documento principal presenta una mayor coincidencia en los rangos ocupados por las tres plataformas.

Los histogramas permiten entonces precisar que varias de las diferencias observadas en los diagramas de caja no se deben únicamente a casos aislados, sino también a la forma asimétrica de las distribuciones. En varias métricas, la mayoría de los usuarios se concentra en una zona reducida y un grupo menor se extiende hacia valores considerablemente mayores.

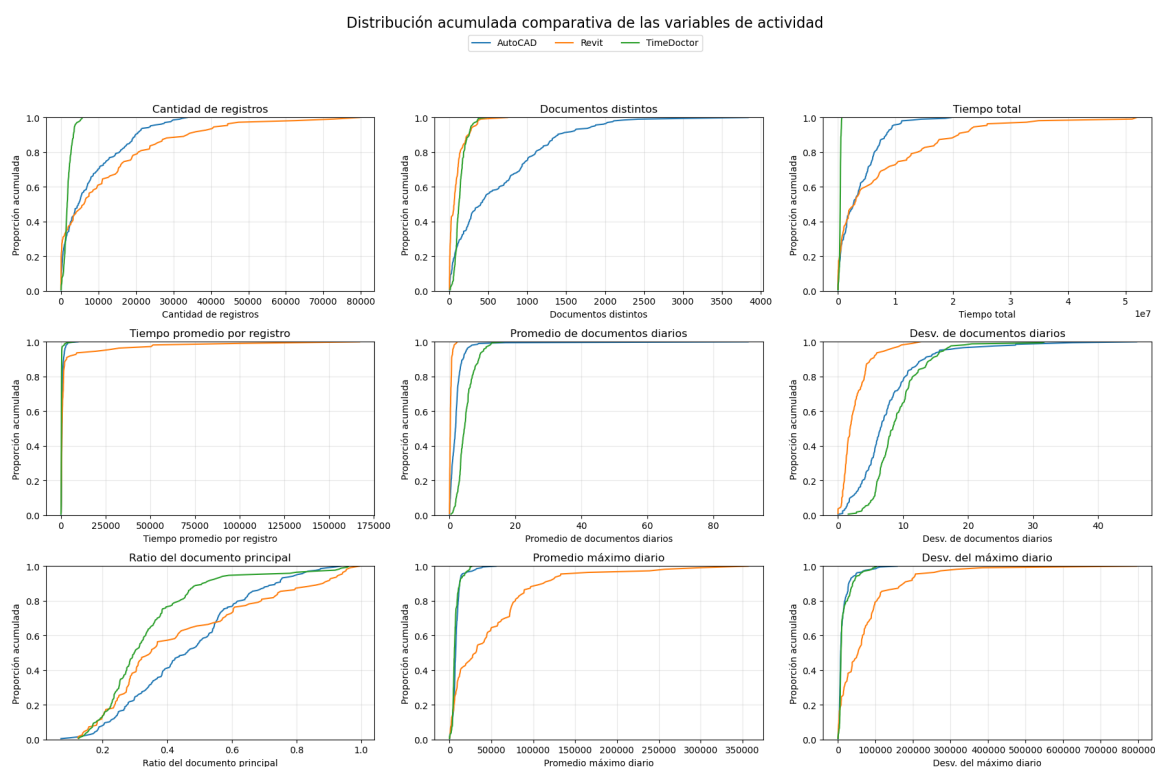


Figura 5.3: Distribuciones acumuladas de las variables de actividad.

La Figura 5.3 presenta las distribuciones acumuladas de las nueve métricas. En cada subfigura, el eje vertical representa la proporción de usuarios cuyo valor es menor o igual al indicado en el eje horizontal. De esta forma, curvas cercanas reflejan distribuciones similares, mientras que desplazamientos entre ellas muestran diferencias en los valores alcanzados por los usuarios.

En cantidad de registros, documentos distintos y tiempo total se observan diferencias claras entre plataformas. TimeDoctor acumula una alta proporción de usuarios en valores más bajos de cantidad de registros y tiempo total, mientras que Revit se extiende hacia valores mayores en ambas métricas. En documentos distintos, AutoCAD presenta la mayor extensión.

En las métricas relacionadas con la duración de la actividad, AutoCAD y TimeDoctor presentan curvas relativamente cercanas en parte importante de sus rangos. Revit sigue una tendencia similar en los valores bajos, pero se extiende hacia valores mayores debido a un grupo reducido de usuarios.

En las métricas relacionadas con el uso diario de documentos se observan mayores desplazamientos entre las plataformas. El promedio y la desviación de documentos diarios presentan diferencias más claras, mientras que el ratio del documento principal mantiene rangos más próximos, aunque las proporciones acumuladas no coinciden completamente.

En conjunto, las Figuras 5.1, 5.2 y 5.3 muestran que las métricas no presentan el mismo comportamiento en las tres plataformas. Existen similitudes en la forma de varias distribuciones, especialmente la concentración de usuarios en valores bajos y la presencia de colas hacia valores mayores, pero también diferencias en los rangos, la dispersión y la ubicación de los valores centrales.

Esta revisión muestra que las métricas contienen variabilidad entre los usuarios y que su comportamiento depende, en parte, de la plataforma analizada. Estas diferencias constituyen el punto de partida para evaluar si las combinaciones de métricas permiten identificar perfiles generales de actividad mediante los métodos de agrupamiento.

5.1.2. Distribución de usuarios por cluster

Definido $K = 3$, se comparó la cantidad de usuarios asignados a cada cluster por K-Means y GMM en AutoCAD, Revit y TimeDoctor. Esta comparación permite observar cómo cada método distribuye a los usuarios antes de analizar las características de los grupos obtenidos.

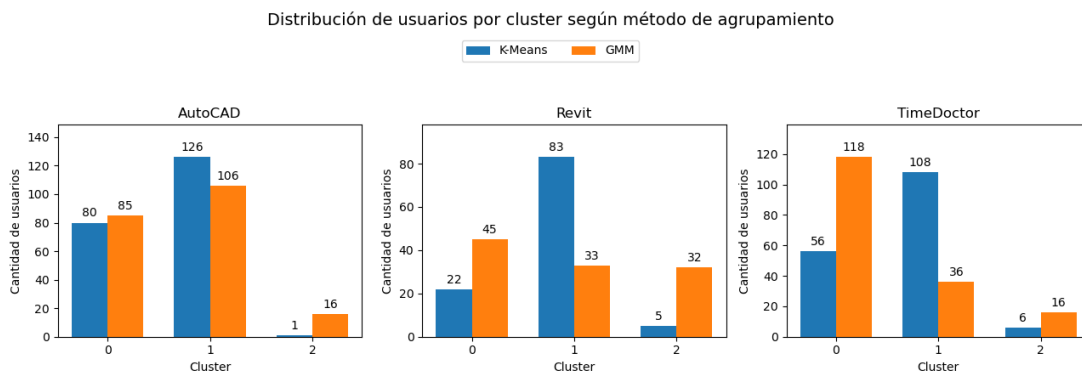


Figura 5.4: Distribución de usuarios por cluster mediante K-Means y GMM en AutoCAD, Revit y TimeDoctor.

Como se observa en la Figura 5.4, K-Means presenta un comportamiento similar en las tres plataformas: concentra a la mayoría de los usuarios en dos clusters y deja un tercer grupo con muy pocos casos. Este tercer cluster contiene 1 usuario en AutoCAD, 5 en Revit y 6 en TimeDoctor.

Con GMM la distribución cambia. En AutoCAD el tercer cluster aumenta a 16 usuarios; en Revit los usuarios se reparten de forma más equilibrada entre los tres grupos; y en TimeDoctor, aunque existe un cluster dominante, el tercer grupo también aumenta a 16 usuarios.

Por lo tanto, una primera diferencia entre ambos métodos es que K-Means tiende a separar un grupo muy pequeño en las tres plataformas, mientras que GMM distribuye una mayor cantidad de usuarios entre los tres clusters. Sin embargo, esta diferencia por sí sola no permite elegir un método, por lo que a continuación se revisan las métricas que caracterizan a cada grupo.

5.1.3. Comparación de los clusters en AutoCAD

Para comparar los agrupamientos obtenidos mediante K-Means y GMM, se analizaron los valores promedio de las métricas dentro de cada cluster. El objetivo es observar qué características utiliza cada método para separar a los usuarios y si los grupos obtenidos pueden interpretarse como perfiles diferenciados.

Las Tablas 5.1 y 5.2 presentan los resultados para AutoCAD. Los números asignados a los clusters corresponden únicamente a etiquetas dentro de cada método, por lo que un mismo número de cluster no representa necesariamente el mismo grupo en K-Means y GMM.

Indicador	Cluster 0	Cluster 1	Cluster 2
Cantidad de usuarios	80	126	1
Cantidad de registros	15.888,46	2.499,52	3,00
Documentos distintos	1.249,90	227,65	2,00
Tiempo total	6.723.671,36	1.680.590,34	28.648,00
Tiempo promedio por registro	443,13	890,76	9.549,33
Promedio de documentos diarios	4,19	1,33	0,20
Desviación estándar de documentos diarios	11,74	5,23	0,71
Ratio del documento principal	0,36	0,52	0,89
Promedio del máximo timelapse diario	8.233,48	8.529,29	13.869,00
Desviación estándar del máximo timelapse diario	12.867,26	14.760,94	16.619,84

Cuadro 5.1: Promedios de las métricas por cluster en AutoCAD mediante K-Means.

En K-Means, la principal diferencia entre los clusters 0 y 1 se observa en la cantidad de registros, los documentos distintos y el tiempo total. El cluster 0 presenta valores considerablemente mayores en estas tres métricas, mientras que el cluster 1 reúne a usuarios con valores más bajos.

El cluster 2 presenta una situación diferente, ya que está compuesto por un solo usuario. Este caso muestra valores muy bajos en cantidad de registros y documentos distintos, pero valores elevados en tiempo promedio por registro, ratio del documento principal y máximo diario. Al estar formado por una única observación, sus promedios corresponden directamente a las características de ese usuario, por lo que resulta difícil interpretarlo como un perfil general de actividad.

Indicador	Cluster 0	Cluster 1	Cluster 2
Cantidad de usuarios	85	106	16
Cantidad de registros	8.454,29	8.014,41	1.117,38
Documentos distintos	672,62	655,14	128,75
Tiempo total	4.836.899,46	3.025.582,79	1.114.281,88
Tiempo promedio por registro	912,28	447,96	2.012,98
Promedio de documentos diarios	1,46	2,18	9,20
Desviación estándar de documentos diarios	6,40	8,44	10,07
Ratio del documento principal	0,55	0,38	0,59
Promedio del máximo timelapse diario	9.341,76	6.015,25	19.723,20
Desviación estándar del máximo timelapse diario	18.415,23	5.466,09	47.573,68

Cuadro 5.2: Promedios de las métricas por cluster en AutoCAD mediante GMM.

En GMM, los clusters 0 y 1 presentan valores relativamente cercanos en cantidad de registros y documentos distintos, mientras que el cluster 2 presenta valores diferentes en ambas métricas.

El cluster 0 se caracteriza por presentar el mayor tiempo total, mientras que el cluster 1 muestra valores menores en las métricas asociadas a la duración de la actividad y el menor ratio del documento principal. Además, el cluster 1 presenta un promedio y una desviación de documentos diarios mayores que el cluster 0.

El cluster 2 presenta un comportamiento claramente distinto. Aunque registra una menor cantidad de registros, documentos distintos y tiempo total, alcanza los valores más altos en tiempo promedio por registro, promedio y desviación de documentos diarios, ratio del documento principal, promedio del máximo timelapse diario y desviación del máximo timelapse diario. Además, este grupo está compuesto por 16 usuarios, por lo que sus características corresponden a un conjunto de observaciones y no a un único caso aislado.

GMM forma grupos que se diferencian por distintas combinaciones de las métricas analizadas, y no únicamente por la cantidad de registros, documentos utilizados o tiempo total.

En conjunto, los resultados muestran que ambos métodos generan grupos diferentes, pero lo hacen de distinta forma. En K-Means, la separación está marcada principalmente

por las diferencias en cantidad de registros, documentos distintos y tiempo total, además de generar un tercer cluster formado por un solo usuario. En GMM, en cambio, los usuarios están en grupos con tamaños más uniformes y las diferencias entre ellos se observan en distintas combinaciones de las métricas analizadas.

Estos resultados muestran que las métricas agregadas permiten identificar usuarios con características diferentes, en concordancia con la hipótesis planteada para el perfil general de actividad. Además, la distribución y la interpretación de los grupos obtenidos mediante GMM entregan un primer antecedente para su posterior selección como método de agrupamiento. Esta comparación se repite a continuación en Revit y TimeDoctor antes de realizar la elección final.

5.1.4. Comparación de los clusters en Revit

Las Tablas 5.3 y 5.4 presentan los resultados para Revit. Los números asignados a los clusters corresponden únicamente a etiquetas dentro de cada método, por lo que un mismo número de cluster no representa necesariamente el mismo grupo en K-Means y GMM.

Indicador	Cluster 0	Cluster 1	Cluster 2
Cantidad de usuarios	22	83	5
Cantidad de registros	34.359,00	6.831,81	19,60
Documentos distintos	248,95	56,34	2,60
Tiempo total	22.764.466,36	3.591.849,46	1.045.342,40
Tiempo promedio por registro	808,84	1.658,16	77.769,85
Promedio de documentos diarios	0,52	0,32	0,18
Desviación estándar de documentos diarios	3,96	2,40	0,46
Ratio del documento principal	0,24	0,47	0,96
Promedio del máximo timelapse diario	59.177,86	33.731,11	263.652,60
Desviación estándar del máximo timelapse diario	78.139,06	54.780,85	381.792,41

Cuadro 5.3: Promedios de las métricas por cluster en Revit mediante K-Means.

En K-Means, el cluster 0 reúne a los usuarios con los mayores valores promedio en cantidad de registros, documentos distintos y tiempo total. También presenta el mayor promedio y desviación estándar de documentos diarios. El cluster 1, que concentra la mayor cantidad de usuarios, presenta valores menores en estas métricas y un mayor tiempo

promedio por registro y ratio del documento principal que el cluster 0.

El cluster 2 está compuesto por solo 5 usuarios y presenta un comportamiento diferente. Este grupo registra los menores valores en cantidad de registros, documentos distintos, tiempo total y promedio de documentos diarios. Sin embargo, alcanza valores muy superiores a los otros clusters en tiempo promedio por registro, ratio del documento principal, promedio del máximo timelapse diario y desviación estándar del máximo timelapse diario.

Por lo tanto, en Revit K-Means separa un grupo con niveles altos de registros, documentos y tiempo total, un grupo mayoritario con valores intermedios y un tercer grupo pequeño caracterizado por pocos registros y documentos, pero por valores muy elevados en varias métricas relacionadas con la duración de la actividad.

Indicador	Cluster 0	Cluster 1	Cluster 2
Cantidad de usuarios	45	33	32
Cantidad de registros	21.516,91	10.071,21	700,78
Documentos distintos	162,11	78,88	8,38
Tiempo total	11.253.670,24	8.485.265,91	554.360,66
Tiempo promedio por registro	544,77	1.108,86	15.098,87
Promedio de documentos diarios	0,60	0,25	0,13
Desviación estándar de documentos diarios	3,84	2,57	0,97
Ratio del documento principal	0,36	0,28	0,74
Promedio del máximo timelapse diario	31.687,01	55.246,97	67.837,27
Desviación estándar del máximo timelapse diario	44.349,63	80.582,34	109.996,27

Cuadro 5.4: Promedios de las métricas por cluster en Revit mediante GMM.

En GMM, los tres clusters presentan diferencias claras en varias de las métricas analizadas. El cluster 0 reúne a los usuarios con los mayores valores promedio en cantidad de registros, documentos distintos y tiempo total. También presenta el mayor promedio y desviación estándar de documentos diarios. En cambio, registra el menor tiempo promedio por registro y los valores más bajos en el promedio y la desviación estándar del máximo timelapse diario.

El cluster 1 presenta valores intermedios en cantidad de registros, documentos distintos y tiempo total. Se diferencia además por presentar el menor ratio del documento princi-

pal, mientras que sus valores asociados al máximo timelapse diario son mayores que los del cluster 0, pero menores que los del cluster 2.

El cluster 2 presenta los menores valores en cantidad de registros, documentos distintos, tiempo total, promedio de documentos diarios y desviación estándar de documentos diarios. Sin embargo, alcanza los valores más altos en tiempo promedio por registro, ratio del documento principal, promedio del máximo timelapse diario y desviación estándar del máximo timelapse diario.

La solución obtenida mediante GMM identifica tres grupos con características diferenciadas. El cluster 0 presenta una mayor cantidad de registros y documentos junto con un mayor tiempo total; el cluster 1 ocupa una posición intermedia en estas métricas y presenta la menor proporción asociada al documento principal; mientras que el cluster 2 combina una menor cantidad de registros y documentos con mayores valores en las métricas relacionadas con la duración de los registros y los máximos diarios. Además, los usuarios se distribuyen de forma más equilibrada entre los tres clusters que en la solución obtenida mediante K-Means.

Al comparar ambas soluciones aparecen diferencias importantes entre K-Means y GMM. K-Means forma un grupo mayoritario, un grupo de menor tamaño con valores altos en cantidad de registros, documentos y tiempo total, y un tercer cluster pequeño compuesto por usuarios con valores extremos en varias métricas.

GMM, en cambio, distribuye a los usuarios de forma más equilibrada entre los tres clusters y permite identificar grupos con combinaciones distintas de las métricas analizadas. Un grupo presenta mayores niveles de registros, documentos y tiempo total; otro mantiene valores intermedios y una menor concentración en el documento principal; y un tercero se caracteriza por una menor cantidad de actividad registrada, pero por mayores valores en las métricas asociadas a la duración de los registros y a los máximos diarios.

El caso de Revit reproduce la tendencia observada en AutoCAD: las métricas agregadas permiten formar grupos con características diferenciadas. Además, la distribución e interpretación de los grupos obtenidos mediante GMM aportan nueva evidencia a favor de este método.

5.1.5. Comparación de los clusters en TimeDoctor

Las Tablas 5.5 y 5.6 presentan los resultados para TimeDoctor. Los números asignados a los clusters corresponden únicamente a etiquetas dentro de cada método, por lo que un mismo número de cluster no representa necesariamente el mismo grupo en K-Means y GMM.

Métrica	Cluster 0	Cluster 1	Cluster 2
Cantidad de usuarios	56	108	6
Cantidad de registros	2.804,04	1.418,78	306,67
Documentos distintos	222,45	100,17	57,17
Tiempo total	407.607,70	340.980,48	419.230,83
Tiempo promedio por registro	156,31	255,91	2.759,25
Promedio de documentos diarios	8,03	3,68	1,97
Desviación estándar de documentos diarios	13,28	7,40	5,78
Ratio del documento principal	0,25	0,36	0,92
Promedio del máximo timelapse diario	5.564,60	7.034,21	24.120,82
Desviación estándar del máximo timelapse diario	9.977,61	14.923,99	89.618,87

Cuadro 5.5: Promedios de las métricas por cluster en TimeDoctor mediante K-Means.

En K-Means, el cluster 0 reúne a los usuarios con los mayores valores promedio en cantidad de registros, documentos distintos, promedio de documentos diarios y desviación estándar de documentos diarios. Además, presenta el menor tiempo promedio por registro, el menor ratio del documento principal y los valores más bajos en las métricas asociadas al máximo timelapse diario.

El cluster 1 concentra la mayor cantidad de usuarios y presenta valores intermedios en la mayoría de las métricas. En comparación con el cluster 0, registra una menor cantidad de registros y documentos distintos, pero un mayor tiempo promedio por registro y un mayor ratio del documento principal.

El cluster 2 está compuesto por solo 6 usuarios y presenta los menores valores en cantidad de registros, documentos distintos, promedio de documentos diarios y desviación estándar de documentos diarios. Sin embargo, alcanza el mayor tiempo total y valores considerablemente superiores en tiempo promedio por registro, ratio del documento prin-

cial, promedio del máximo timelapse diario y desviación estándar del máximo timelapse diario.

Por lo tanto, en TimeDoctor K-Means separa un grupo con mayor cantidad de registros y documentos utilizados, un grupo mayoritario con valores intermedios y un tercer grupo pequeño que, pese a registrar menos eventos y documentos, presenta tiempos de actividad considerablemente mayores.

Métrica	Cluster 0	Cluster 1	Cluster 2
Cantidad de usuarios	118	36	16
Cantidad de registros	1.940,53	1.775,81	1.198,94
Documentos distintos	146,32	142,36	76,69
Tiempo total	392.307,65	308.798,39	297.391,44
Tiempo promedio por registro	240,73	173,61	1.143,17
Promedio de documentos diarios	5,05	5,99	2,98
Desviación estándar de documentos diarios	9,00	11,55	6,22
Ratio del documento principal	0,32	0,28	0,62
Promedio del máximo timelapse diario	7.175,41	4.781,83	12.324,59
Desviación estándar del máximo timelapse diario	15.121,18	6.229,30	43.731,00

Cuadro 5.6: Promedios de las métricas por cluster en TimeDoctor mediante GMM.

En GMM, los clusters 0 y 1 presentan valores relativamente cercanos en cantidad de registros y documentos distintos, mientras que el cluster 2 presenta valores menores en ambas métricas.

El cluster 0 concentra la mayor cantidad de usuarios y presenta el mayor tiempo total. Además, muestra valores intermedios en tiempo promedio por registro, promedio de documentos diarios, desviación estándar de documentos diarios y ratio del documento principal. En las métricas asociadas al máximo timelapse diario también presenta valores intermedios entre los otros dos grupos.

El cluster 1 presenta el mayor promedio y desviación estándar de documentos diarios, junto con el menor tiempo promedio por registro, el menor ratio del documento principal y los valores más bajos en el promedio y la desviación estándar del máximo timelapse diario.

El cluster 2 presenta los menores valores en cantidad de registros, documentos distintos, tiempo total, promedio de documentos diarios y desviación estándar de documentos diarios. Sin embargo, alcanza los valores más altos en tiempo promedio por registro, ratio del documento principal, promedio del máximo *timelapse* diario y desviación estándar del máximo *timelapse* diario.

GMM vuelve a formar grupos diferenciados por distintas combinaciones de las métricas analizadas. Los clusters 0 y 1 presentan niveles relativamente cercanos de registros y documentos, pero se distinguen en la forma en que se distribuye su actividad, mientras que el cluster 2 combina una menor cantidad de registros y documentos con valores más altos en las métricas relacionadas con la duración de la actividad.

La comparación en TimeDoctor también evidencia diferencias claras entre ambas soluciones. K-Means concentra a la mayor parte de los usuarios en un cluster, mientras que genera un tercer grupo pequeño compuesto por 6 usuarios con valores muy altos en varias métricas relacionadas con la duración de la actividad.

GMM, en cambio, forma grupos que se diferencian por distintas combinaciones de las métricas analizadas. Los clusters 0 y 1 presentan cantidades relativamente cercanas de registros y documentos, pero se distinguen en los tiempos de actividad, el uso diario de documentos y la concentración en el documento principal. El cluster 2 presenta una menor cantidad de registros y documentos, junto con valores más altos en las métricas asociadas a la duración de la actividad.

TimeDoctor confirma la misma conclusión general. Las métricas agregadas permiten distinguir usuarios con características diferentes y GMM vuelve a producir grupos más interpretables y menos condicionados por casos extremos. Esto entrega un último antecedente para seleccionar este método como representación del perfil general de actividad.

5.1.6. Selección del método de agrupamiento

A partir de la comparación realizada en AutoCAD, Revit y TimeDoctor, se seleccionó GMM para representar el perfil general de actividad.

Si bien K-Means obtuvo valores más altos del coeficiente de silueta, este criterio no se

consideró de forma aislada para seleccionar el método. También se revisó la cantidad de usuarios asignados a cada cluster y las características que permitían diferenciar e interpretar los grupos obtenidos.

En las tres plataformas, K-Means generó un tercer cluster de tamaño reducido, compuesto por 1 usuario en AutoCAD, 5 en Revit y 6 en TimeDoctor. Estos grupos presentaron valores extremos en varias de las métricas analizadas, por lo que su interpretación como perfiles generales de actividad resultó más limitada.

Con GMM, en cambio, los tres clusters estuvieron formados por un mayor número de usuarios y pudieron distinguirse a partir de diferentes combinaciones de las métricas. En AutoCAD, Revit y TimeDoctor se observaron diferencias asociadas tanto a la cantidad de registros y documentos como a los tiempos de actividad, el uso diario de documentos y el ratio del documento principal. De esta forma, los grupos obtenidos no quedaron definidos únicamente por casos con valores extremos en determinadas métricas.

Por estas razones, se consideró que GMM entregaba una representación más adecuada para construir e interpretar los perfiles generales de actividad, aun cuando K-Means presentara un mejor coeficiente de silueta. En los análisis posteriores se utilizarán, por lo tanto, los clusters obtenidos mediante GMM, cuyos parámetros fueron estimados utilizando el algoritmo EM.

Una vez seleccionado GMM, las etiquetas de los clusters fueron homologadas entre AutoCAD, Revit y TimeDoctor a partir de las características observadas en las métricas promedio de cada grupo. De esta forma, los clusters con comportamientos comparables mantienen la misma etiqueta en las tres plataformas. Este reordenamiento no modifica la asignación de los usuarios ni los grupos obtenidos por el modelo, sino únicamente la numeración utilizada para facilitar su comparación posterior.

5.1.7. Respuesta a las preguntas planteadas

Los resultados obtenidos muestran que las métricas agregadas permiten identificar grupos de usuarios con características diferentes. En AutoCAD, Revit y TimeDoctor se observaron diferencias entre los clusters en varias de las métricas analizadas, incluyendo la

cantidad de registros, los documentos utilizados, los tiempos de actividad y el uso diario de documentos. Por lo tanto, los resultados respaldan la hipótesis planteada para el perfil general de actividad.

Aunque las magnitudes específicas de las métricas cambian entre plataformas, se observaron características comunes que permitieron establecer una correspondencia entre los perfiles obtenidos mediante GMM. A partir de esta correspondencia, las etiquetas fueron homologadas para facilitar la comparación posterior entre AutoCAD, Revit y TimeDoctor.

Finalmente, la comparación entre K-Means y GMM mostró que ambos métodos permiten formar grupos diferenciados, pero GMM produjo clusters con una distribución de usuarios más adecuada para su interpretación y menos condicionada por grupos muy pequeños asociados a valores extremos. Por esta razón, los clusters obtenidos mediante GMM fueron seleccionados como perfiles generales de actividad para los análisis posteriores.

5.2. Validación de la concentración horaria

La validación de la concentración horaria se realiza a partir de la distribución del indicador P_{MM} en AutoCAD, Revit y TimeDoctor.

El análisis tiene dos objetivos. Primero, observar si existen diferencias entre los usuarios y entre las plataformas que permitan responder la pregunta de investigación y evaluar la hipótesis planteada. Segundo, analizar la distribución del indicador para definir categorías que faciliten la interpretación y comparación de los distintos niveles observados.

Para ello, se analizan distintas representaciones de la distribución del indicador y, a partir de lo observado, se definen y posteriormente se interpretan las categorías obtenidas. Finalmente, el análisis se complementa con medidas descriptivas de dispersión para cuantificar la variabilidad observada entre los usuarios.

5.2.1. Distribución del indicador

Para analizar la concentración horaria se revisó la distribución de P_{MM} en AutoCAD, Revit y TimeDoctor. En cada figura, la subfigura izquierda muestra la cantidad de usuarios según el valor del indicador, mientras que la subfigura derecha muestra la proporción

acumulada de usuarios. La línea azul representa la distribución de cada plataforma y las líneas verticales discontinuas marcan los valores 0,4 y 0,6.

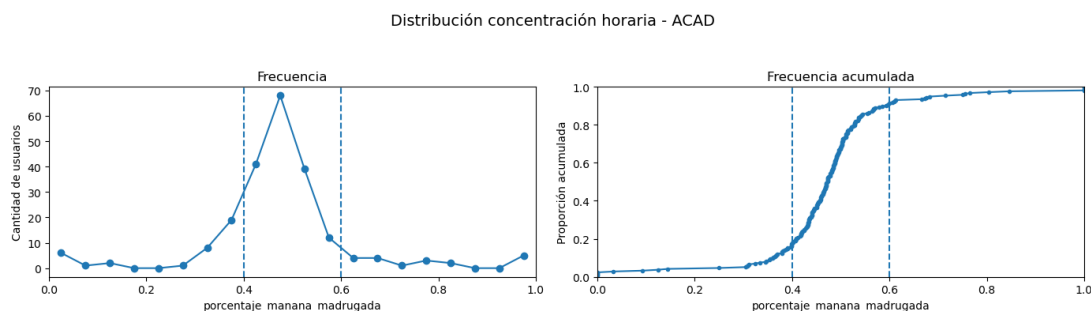


Figura 5.5: Distribución del indicador de concentración horaria en AutoCAD.

En AutoCAD, la mayor parte de los usuarios se concentra alrededor de $P_{MM} = 0,5$. Hacia ambos extremos la cantidad de usuarios disminuye, por lo que los casos con una concentración marcada hacia alguno de los dos períodos son menos frecuentes.

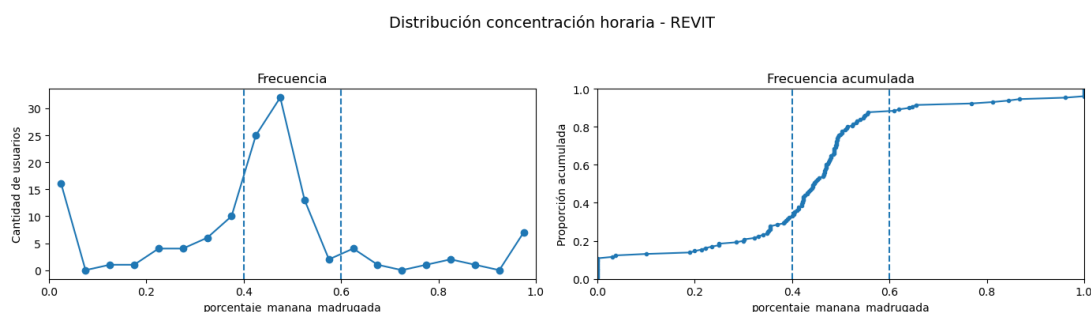


Figura 5.6: Distribución del indicador de concentración horaria en Revit.

En Revit también se observa una concentración alrededor de 0,5, aunque la distribución es más dispersa. Además, existe un grupo de usuarios con valores cercanos a 0, asociado a una mayor actividad durante la tarde-noche.

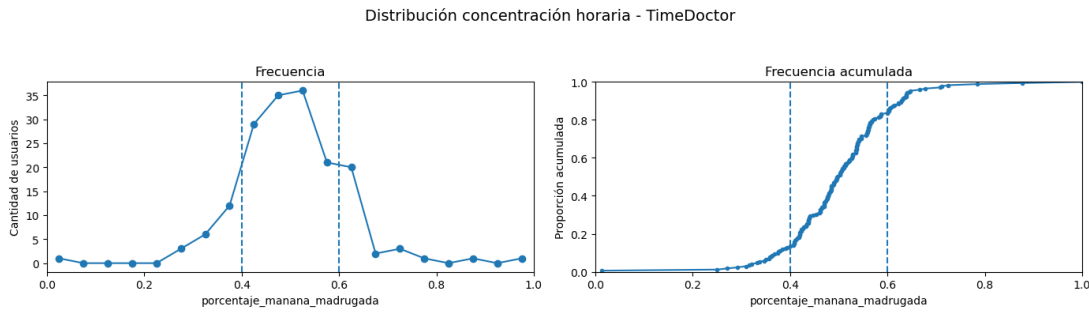


Figura 5.7: Distribución del indicador de concentración horaria en TimeDoctor.

En TimeDoctor la mayor parte de los usuarios vuelve a concentrarse en valores cercanos a 0,5, mientras que los valores extremos son menos frecuentes.

En conjunto, las tres plataformas presentan un comportamiento similar: existe una zona central donde se concentra la mayor parte de los usuarios y grupos menores hacia ambos extremos. A partir de esta distribución se definieron tres categorías. Los valores inferiores a 0,4 corresponden a una mayor actividad durante la tarde-noche; los valores entre 0,4 y 0,6 representan una actividad más equilibrada; y los valores superiores a 0,6 corresponden a una mayor actividad durante la mañana-madrugada.

5.2.2. Distribución de usuarios por categoría

Una vez definidos los rangos de P_{MM} , los usuarios fueron clasificados en *Tarde-noche*, *Equilibrada* y *Mañana-madrugada*. La Figura 5.8 muestra la cantidad de usuarios en cada categoría para AutoCAD, Revit y TimeDoctor.

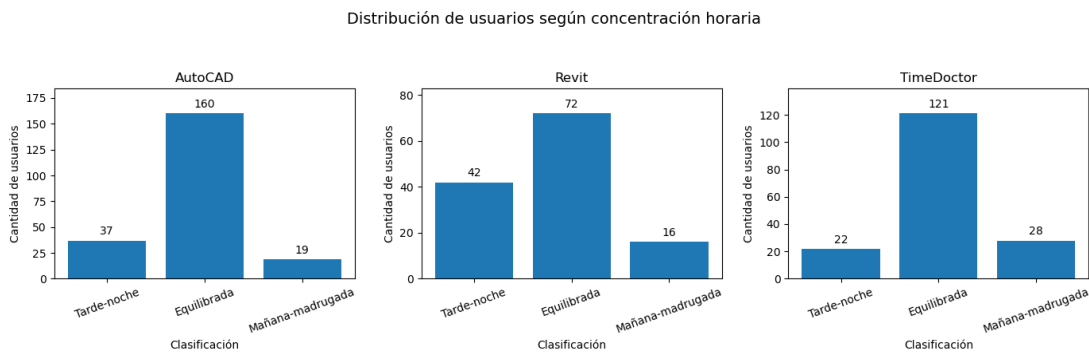


Figura 5.8: Distribución de usuarios según las categorías de concentración horaria.

En las tres plataformas predomina la categoría *Equilibrada*, con 160 usuarios en AutoCAD, 72 en Revit y 121 en TimeDoctor. Esto es consistente con las distribuciones observadas anteriormente, donde la mayor concentración se encontraba entre $P_{MM} = 0,4$ y $P_{MM} = 0,6$.

Las categorías de mayor concentración hacia alguno de los períodos reúnen menos usuarios. Revit presenta una mayor presencia de usuarios en *Tarde-noche*, mientras que en TimeDoctor la categoría *Mañana-madrugada* tiene mayor presencia que en las otras plataformas. Aun así, en los tres casos se mantienen las tres categorías, mostrando que existen distintos niveles de concentración horaria entre los usuarios.

5.2.3. Variabilidad del indicador

Para complementar el análisis gráfico y la distribución de usuarios por categoría, se analizaron los cuartiles y percentiles del indicador P_{MM} en cada plataforma. Los resultados se presentan en la Tabla [5.7](#).

Indicador	Plataforma	Q_1	Q_3	IQR	P_{10}	P_{90}	$P_{90} - P_{10}$
Concentración horaria	AutoCAD	0,4320	0,5112	0,0793	0,3795	0,5731	0,1936
Concentración horaria	Revit	0,3876	0,4930	0,1055	0,2500	0,5502	0,3002
Concentración horaria	TimeDoctor	0,4353	0,5597	0,1244	0,3761	0,6275	0,2514

Cuadro 5.7: Medidas descriptivas de dispersión del indicador de concentración horaria

En AutoCAD, el 50 % central de los usuarios presenta valores de P_{MM} entre $Q_1 = 0,4320$ y $Q_3 = 0,5112$, con un $IQR = 0,0793$. Al considerar el 80 % central, los valores se encuentran entre $P_{10} = 0,3795$ y $P_{90} = 0,5731$, con una diferencia de 0,1936. Esto muestra que la mayor parte de los usuarios se concentra en valores cercanos al equilibrio entre ambos períodos del día.

En Revit, el 50 % central se encuentra entre $Q_1 = 0,3876$ y $Q_3 = 0,4930$, con un $IQR = 0,1055$. El 80 % central se extiende desde $P_{10} = 0,2500$ hasta $P_{90} = 0,5502$, alcanzando la mayor diferencia entre percentiles de las tres plataformas, con $P_{90} - P_{10} = 0,3002$. La mayor extensión hacia valores bajos indica que existe una mayor presencia de usuarios cuya actividad se concentra hacia la tarde-noche.

En TimeDoctor, el 50 % central de los usuarios se encuentra entre $Q_1 = 0,4353$ y $Q_3 = 0,5597$, con el mayor IQR , igual a 0,1244. Al considerar el 80 % central, los valores se extienden desde $P_{10} = 0,3761$ hasta $P_{90} = 0,6275$, con una diferencia de 0,2514. Esto muestra una mayor variabilidad entre los usuarios centrales, aunque sin una extensión hacia los extremos tan marcada como en Revit.

La comparación entre plataformas muestra que la concentración horaria varía de manera distinta en cada caso. AutoCAD se concentra más alrededor del equilibrio, Revit presenta una mayor extensión hacia la tarde-noche y TimeDoctor muestra mayor variabilidad en la zona central.

5.2.4. Respuesta a la pregunta de investigación

Los resultados muestran que los usuarios presentan diferencias en la forma en que distribuyen su actividad a lo largo del día. En las tres plataformas predomina una distribución cercana al equilibrio, aunque también existen usuarios con mayor actividad durante la tarde-noche o durante la mañana-madrugada.

Las medidas de dispersión respaldan estas diferencias. AutoCAD presenta los valores más concentrados alrededor del equilibrio, mientras que Revit muestra la mayor amplitud entre los percentiles 10 y 90, principalmente hacia valores bajos de P_{MM} . TimeDoctor, por su parte, presenta la mayor dispersión dentro del 50 % central de los usuarios.

La evidencia gráfica y las medidas de dispersión respaldan la hipótesis planteada: los usuarios presentan distintos niveles de concentración horaria. Las categorías propuestas resumen estos niveles y permiten compararlos entre plataformas.

5.3. Validación de la diversidad documental

La validación de la diversidad documental se realiza a partir de la distribución del indicador $H_{\text{norm}}(\mathbf{q})$ en AutoCAD, Revit y TimeDoctor. En este caso, $\mathbf{q}$ corresponde al vector formado por las proporciones de tiempo que cada usuario dedica a cada uno de sus documentos únicos.

El análisis tiene dos objetivos. Primero, observar si existen diferencias entre los usuarios y entre las plataformas que permitan responder la pregunta de investigación y evaluar la hipótesis planteada. Segundo, analizar la distribución del indicador para definir categorías que faciliten la interpretación y comparación de los distintos niveles de diversidad documental observados.

Para ello, se analizan distintas representaciones de la distribución del indicador y, a partir de lo observado, se definen y posteriormente se interpretan las categorías obtenidas. Finalmente, el análisis se complementa mediante el rango intercuartílico (IQR) y la diferencia entre los percentiles 90 y 10 ($P_{90} - P_{10}$), con el fin de cuantificar la variabilidad del indicador entre los usuarios.

5.3.1. Distribución del indicador

Para analizar la diversidad documental se revisó la distribución de $H_{\text{norm}}(\mathbf{q})$ en AutoCAD, Revit y TimeDoctor. En cada figura, la subfigura izquierda muestra la cantidad de usuarios según el valor del indicador, mientras que la subfigura derecha muestra la proporción acumulada de usuarios. La línea azul representa la distribución de cada plataforma y las líneas verticales discontinuas marcan los valores 0,4 y 0,6.

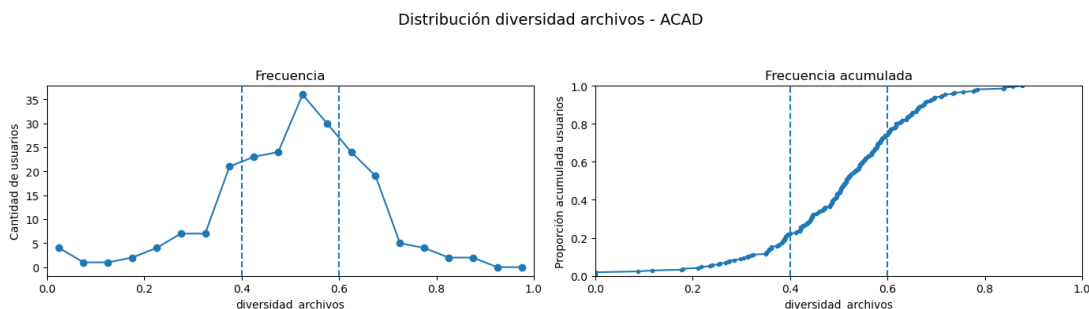


Figura 5.9: Distribución del indicador de diversidad documental en AutoCAD.

En AutoCAD, la mayor parte de los usuarios se concentra en valores intermedios del indicador, principalmente entre 0,4 y 0,6. La frecuencia disminuye hacia ambos extremos, aunque también se observan usuarios con valores superiores a 0,6, asociados a una mayor diversidad documental.

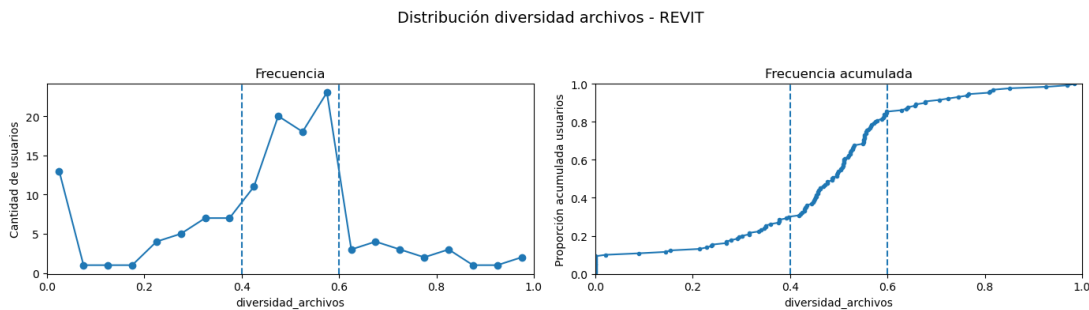


Figura 5.10: Distribución del indicador de diversidad documental en Revit.

En Revit también existe una concentración importante entre 0,4 y 0,6, pero la distribución es más dispersa. Se observa además un grupo de usuarios con valores cercanos a 0 y otros con valores superiores a 0,6, por lo que la diversidad documental presenta una mayor variación entre los usuarios.

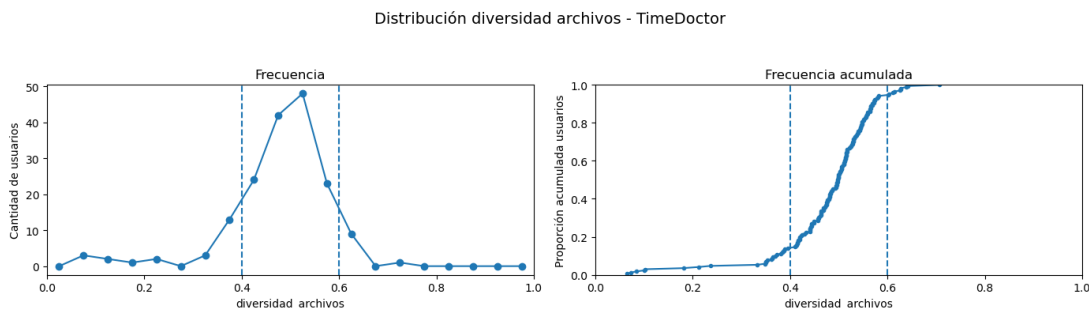


Figura 5.11: Distribución del indicador de diversidad documental en TimeDoctor.

En TimeDoctor la distribución se concentra con mayor claridad alrededor de valores intermedios, principalmente entre 0,4 y 0,6. Los valores muy bajos o muy altos son menos frecuentes que en las otras plataformas.

En conjunto, las tres plataformas muestran una concentración principal en valores intermedios de $H_{\text{norm}}(\mathbf{q})$, junto con grupos menores hacia ambos extremos. A partir de este comportamiento se definieron tres categorías: valores inferiores a 0,4 corresponden a *Baja diversidad*; valores entre 0,4 y 0,6 a *Diversidad media*; y valores superiores a 0,6 a *Alta diversidad*.

5.3.2. Distribución de usuarios por categoría

Una vez definidos los rangos de $H_{\text{norm}}(\mathbf{q})$, los usuarios fueron clasificados en *Baja diversidad*, *Diversidad media* y *Alta diversidad*. La Figura 5.12 muestra la cantidad de usuarios en cada categoría para AutoCAD, Revit y TimeDoctor.

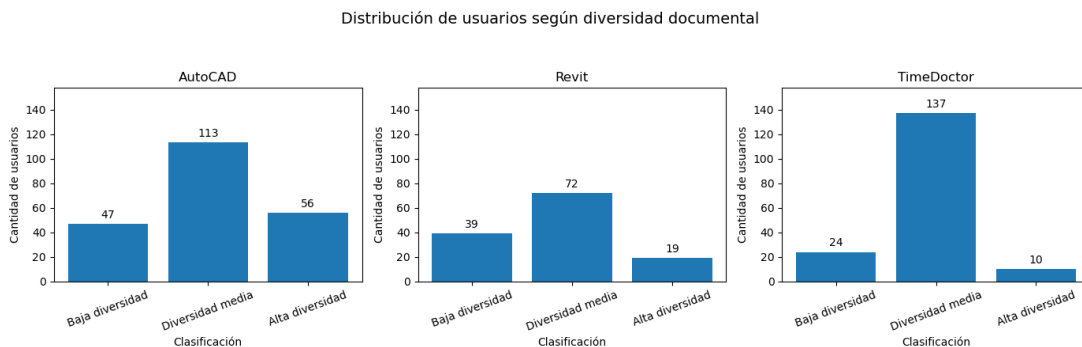


Figura 5.12: Distribución de usuarios según las categorías de diversidad documental.

En las tres plataformas predomina la categoría *Diversidad media*, lo que es consistente con las distribuciones observadas anteriormente, donde la mayor parte de los usuarios presenta valores de $H_{\text{norm}}(\mathbf{q})$ entre 0,4 y 0,6. Esta concentración es especialmente marcada en TimeDoctor.

También se observan usuarios en las categorías de *Baja diversidad* y *Alta diversidad*, aunque en menor cantidad. Esto muestra que, además del comportamiento predominante, existen usuarios cuyo tiempo se distribuye de manera menos uniforme entre los documentos utilizados y otros que lo reparten de forma más equilibrada entre ellos.

5.3.3. Variabilidad del indicador

Para complementar el análisis gráfico y la distribución de usuarios por categoría, se analizaron los cuartiles y percentiles del indicador $H_{\text{norm}}(\mathbf{q})$ en cada plataforma. Los resultados se presentan en la Tabla 5.8.

Indicador	Plataforma	Q_1	Q_3	IQR	P_{10}	P_{90}	$P_{90} - P_{10}$
Diversidad documental	AutoCAD	0,4260	0,5978	0,1718	0,3498	0,6655	0,3156
Diversidad documental	Revit	0,4254	0,5603	0,1348	0,2930	0,6562	0,3632
Diversidad documental	TimeDoctor	0,4431	0,5391	0,0960	0,3729	0,5696	0,1967

Cuadro 5.8: Medidas descriptivas de dispersión del indicador de diversidad documental

En AutoCAD, el 50 % central de los usuarios presenta valores de $H_{\text{norm}}(\mathbf{q})$ entre $Q_1 = 0,4260$ y $Q_3 = 0,5978$, con un $IQR = 0,1718$. Al considerar el 80 % central, los valores se extienden desde $P_{10} = 0,3498$ hasta $P_{90} = 0,6655$, con una diferencia de 0,3156. Esto muestra una variabilidad importante en la forma en que los usuarios distribuyen su tiempo entre los documentos.

En Revit, el 50 % central se encuentra entre $Q_1 = 0,4254$ y $Q_3 = 0,5603$, con un $IQR = 0,1348$. El 80 % central se extiende desde $P_{10} = 0,2930$ hasta $P_{90} = 0,6562$, alcanzando la mayor diferencia entre percentiles de las tres plataformas, con $P_{90} - P_{10} = 0,3632$. Esto indica que existen usuarios con valores más alejados de la zona central, especialmente hacia niveles bajos de diversidad documental.

En TimeDoctor, el 50 % central de los usuarios se encuentra entre $Q_1 = 0,4431$ y $Q_3 = 0,5391$, con el menor IQR , igual a 0,0960. Al considerar el 80 % central, los valores van desde $P_{10} = 0,3729$ hasta $P_{90} = 0,5696$, con una diferencia de 0,1967. Esto muestra que los usuarios presentan valores de diversidad documental más cercanos entre sí que en AutoCAD y Revit.

La variabilidad de la diversidad documental tampoco es igual entre plataformas. AutoCAD presenta la mayor dispersión en el 50 % central, Revit la mayor extensión al considerar el 80 % y TimeDoctor concentra sus valores en un rango más reducido.

5.3.4. Respuesta a la pregunta de investigación

Los resultados muestran que los usuarios presentan diferencias en la forma en que distribuyen su tiempo entre los documentos utilizados. En las tres plataformas predomina la diversidad media, aunque también existen usuarios con valores bajos y altos de $H_{\text{norm}}(\mathbf{q})$.

Las medidas de dispersión respaldan estas diferencias. AutoCAD presenta la mayor

variabilidad dentro del 50 % central de los usuarios, mientras que Revit muestra la mayor extensión al considerar el 80 % central. TimeDoctor, en cambio, presenta los valores más concentrados, indicando una mayor similitud entre sus usuarios en este indicador.

En la diversidad documental, ambos análisis apuntan a la misma conclusión: los usuarios distribuyen su tiempo entre documentos de formas diferentes. Las categorías definidas permiten resumir esa variación y compararla entre plataformas.

5.4. Validación de la concentración documental

La validación de la concentración documental se realiza a partir de la distribución del indicador $HHI(\mathbf{q})$ en AutoCAD, Revit y TimeDoctor. En este caso, $\mathbf{q}$ corresponde al vector formado por las proporciones de tiempo que cada usuario dedica a cada uno de sus documentos únicos.

El análisis tiene dos objetivos. Primero, observar si existen diferencias entre los usuarios y entre las plataformas que permitan responder la pregunta de investigación y evaluar la hipótesis planteada. Segundo, analizar la distribución del indicador para definir categorías que faciliten la interpretación y comparación de los distintos niveles de concentración documental observados.

Para ello, se analizan distintas representaciones de la distribución del indicador y, a partir de lo observado, se definen y posteriormente se interpretan las categorías obtenidas. Finalmente, el análisis se complementa mediante el rango intercuartílico (IQR) y la diferencia entre los percentiles 90 y 10 ($P_{90} - P_{10}$), con el fin de cuantificar la variabilidad del indicador entre los usuarios.

5.4.1. Distribución del indicador

Para analizar la concentración documental se revisó la distribución de $HHI(\mathbf{q})$ en AutoCAD, Revit y TimeDoctor. En cada figura, la subfigura izquierda muestra la cantidad de usuarios según el valor del indicador, mientras que la subfigura derecha muestra la proporción acumulada de usuarios. La línea azul representa la distribución de cada plataforma y las líneas verticales discontinuas marcan los valores 0,25 y 0,50.

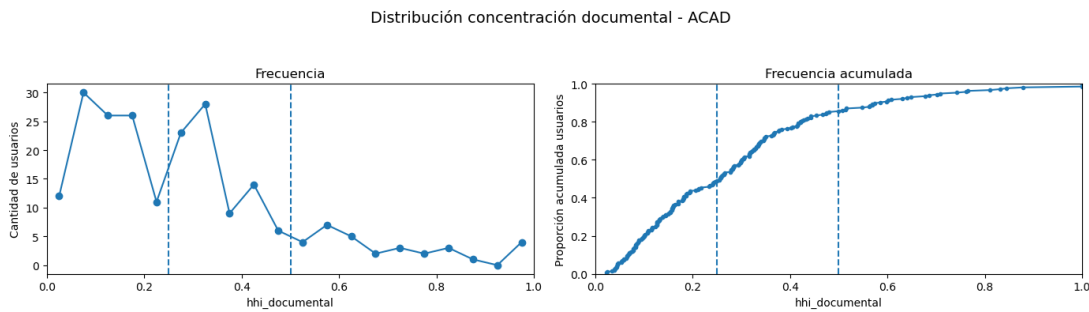


Figura 5.13: Distribución del indicador de concentración documental en AutoCAD.

En AutoCAD, una parte importante de los usuarios presenta valores bajos de $HHI(\mathbf{q})$, aunque también se observa una cantidad considerable de usuarios entre 0,25 y 0,50. A partir de 0,50 la frecuencia disminuye, por lo que los casos con alta concentración documental son menos frecuentes.

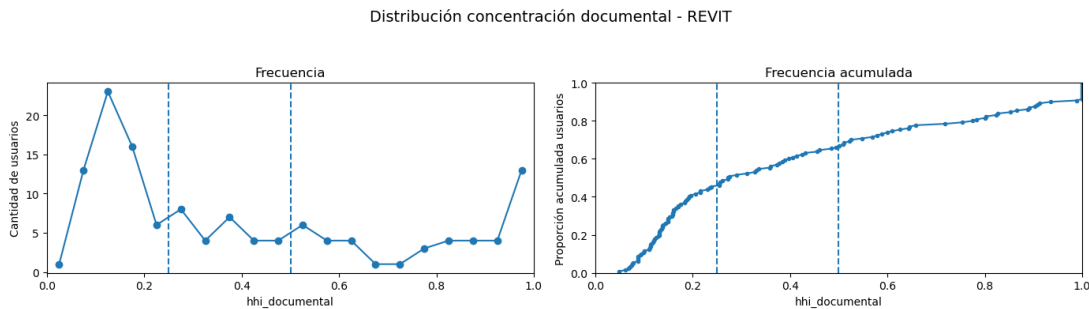


Figura 5.14: Distribución del indicador de concentración documental en Revit.

En Revit la distribución se encuentra más extendida a lo largo del rango del indicador. Se observan usuarios con valores bajos, medios y altos de $HHI(\mathbf{q})$, incluyendo un grupo con valores cercanos a 1. Esto muestra una mayor presencia de usuarios cuyo tiempo se concentra en uno o pocos documentos.

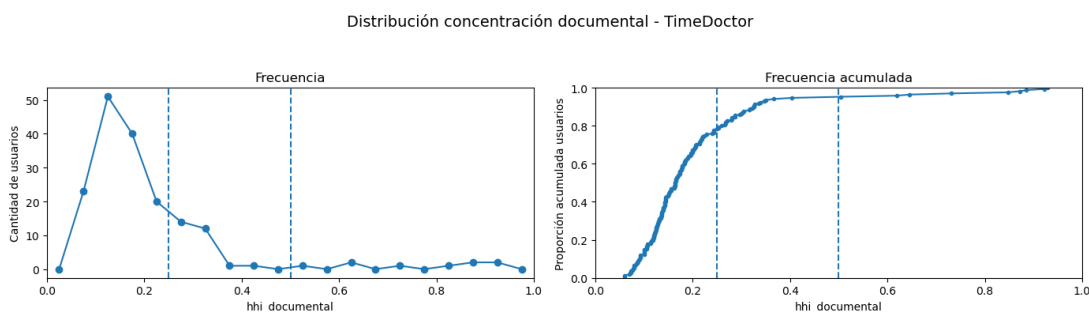


Figura 5.15: Distribución del indicador de concentración documental en TimeDoctor.

En TimeDoctor la mayor parte de los usuarios se concentra en valores inferiores a 0,25. A medida que aumenta $HHI(q)$, la cantidad de usuarios disminuye de forma clara, siendo poco frecuentes los valores superiores a 0,50.

En conjunto, las tres plataformas presentan distintos niveles de concentración documental, aunque con diferencias en su distribución. AutoCAD y TimeDoctor concentran una mayor cantidad de usuarios en valores bajos, mientras que Revit presenta una distribución más amplia. A partir de este comportamiento se definieron tres categorías: valores inferiores a 0,25 corresponden a *Baja concentración*; valores entre 0,25 y 0,50 a *Concentración media*; y valores superiores a 0,50 a *Alta concentración*.

5.4.2. Distribución de usuarios por categoría

Una vez definidos los rangos de $HHI(q)$, los usuarios fueron clasificados en *Baja concentración*, *Concentración media* y *Alta concentración*. La Figura 5.16 muestra la cantidad de usuarios en cada categoría para AutoCAD, Revit y TimeDoctor.

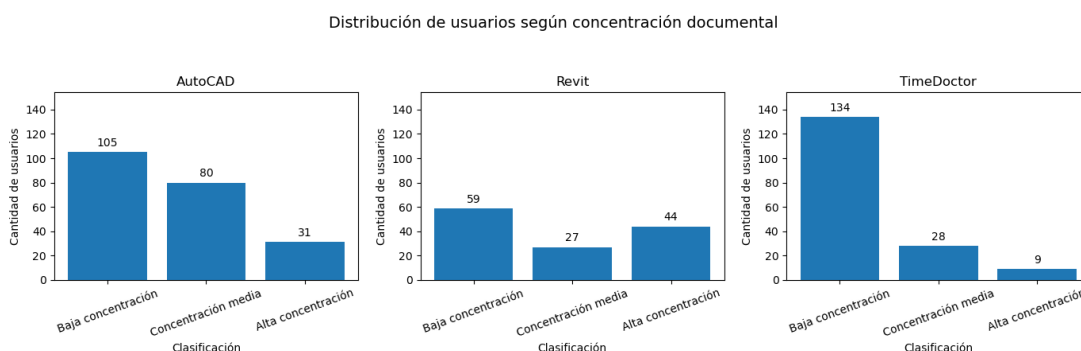


Figura 5.16: Distribución de usuarios según las categorías de concentración documental.

En las tres plataformas predomina la categoría *Baja concentración*, con 105 usuarios en AutoCAD, 59 en Revit y 134 en TimeDoctor. Esta tendencia es especialmente marcada en TimeDoctor, donde la mayor parte de los usuarios presenta valores bajos de $HHI(\mathbf{q})$.

AutoCAD también presenta una cantidad importante de usuarios en *Concentración media*, mientras que en Revit se observa una mayor presencia de usuarios en *Alta concentración*. Esto es consistente con la distribución observada anteriormente, donde Revit presenta valores más extendidos a lo largo del rango del indicador.

5.4.3. Variabilidad del indicador

Para complementar el análisis gráfico y la distribución de usuarios por categoría, se analizaron los cuartiles y percentiles del indicador $HHI(\mathbf{q})$ en cada plataforma. Los resultados se presentan en la Tabla 5.9.

Indicador	Plataforma	Q_1	Q_3	IQR	P_{10}	P_{90}	$P_{90} - P_{10}$
Concentración documental	AutoCAD	0,1165	0,3654	0,2488	0,0672	0,5633	0,4961
Concentración documental	Revit	0,1322	0,4592	0,3271	0,0924	0,7849	0,6925
Concentración documental	TimeDoctor	0,1241	0,2231	0,0990	0,0917	0,3252	0,2336

Cuadro 5.9: Medidas descriptivas de dispersión del indicador de concentración documental

En AutoCAD, el 50 % central de los usuarios presenta valores de $HHI(\mathbf{q})$ entre $Q_1 = 0,1165$ y $Q_3 = 0,3654$, con un $IQR = 0,2488$. Al considerar el 80 % central, los valores se extienden desde $P_{10} = 0,0672$ hasta $P_{90} = 0,5633$, con una diferencia de 0,4961. Esto muestra que existen diferencias importantes entre los usuarios, desde niveles bajos hasta niveles más altos de concentración documental.

En Revit, el 50 % central se encuentra entre $Q_1 = 0,1322$ y $Q_3 = 0,4592$, con el mayor IQR de las tres plataformas, igual a 0,3271. El 80 % central se extiende desde $P_{10} = 0,0924$ hasta $P_{90} = 0,7849$, con una diferencia de 0,6925. Estos valores muestran la mayor variabilidad entre usuarios y una extensión marcada hacia niveles altos de concentración documental.

En TimeDoctor, el 50 % central de los usuarios se encuentra entre $Q_1 = 0,1241$ y $Q_3 = 0,2231$, con un $IQR = 0,0990$. Al considerar el 80 % central, los valores van desde

$P_{10} = 0,0917$ hasta $P_{90} = 0,3252$, con una diferencia de 0,2336. Esto muestra una distribución más concentrada, con usuarios que presentan valores más similares y principalmente bajos del indicador.

La concentración documental presenta una dispersión distinta según la plataforma: Revit muestra la mayor variabilidad, AutoCAD ocupa una posición intermedia y TimeDoctor concentra sus valores en un rango más reducido.

5.4.4. Respuesta a la pregunta de investigación

Los resultados muestran que los usuarios presentan diferencias en la forma en que concentran su tiempo entre los documentos utilizados. En las tres plataformas se observan usuarios con valores bajos, medios y altos de $HHI(\mathbf{q})$, aunque con distintas proporciones en cada caso.

Las medidas de dispersión respaldan estas diferencias. Revit presenta la mayor variabilidad y una mayor extensión hacia valores altos de concentración documental, AutoCAD muestra un comportamiento intermedio y TimeDoctor presenta valores más concentrados y principalmente bajos.

Para la concentración documental, los gráficos y las medidas de dispersión muestran distintos grados de concentración entre los usuarios. La clasificación propuesta sintetiza estas diferencias para su posterior comparación.

5.5. Validación de la fragmentación temporal

La validación de la fragmentación temporal se realiza a partir de la distribución del indicador $I_{\text{Fragmentacin}}$ en AutoCAD, Revit y TimeDoctor.

El análisis tiene dos objetivos. Primero, observar si existen diferencias entre los usuarios y entre las plataformas que permitan responder la pregunta de investigación y evaluar la hipótesis planteada. Segundo, analizar la distribución del indicador para definir categorías que faciliten la interpretación y comparación de los distintos niveles de fragmentación temporal observados.

Para ello, se analizan distintas representaciones de la distribución del indicador y, a partir de lo observado, se definen y posteriormente se interpretan las categorías obtenidas. Finalmente, el análisis se complementa mediante el rango intercuartílico (IQR) y la diferencia entre los percentiles 90 y 10 ($P_{90} - P_{10}$), con el fin de cuantificar la variabilidad del indicador entre los usuarios.

5.5.1. Distribución del indicador

Para analizar la fragmentación temporal se revisó la distribución de $I_{\text{Fragmentacin}}$ en AutoCAD, Revit y TimeDoctor. En cada figura, la subfigura izquierda muestra la cantidad de usuarios según el valor del indicador, mientras que la subfigura derecha muestra la proporción acumulada de usuarios. La línea azul representa la distribución de cada plataforma y las líneas verticales discontinuas marcan los puntos de corte definidos en cada caso.

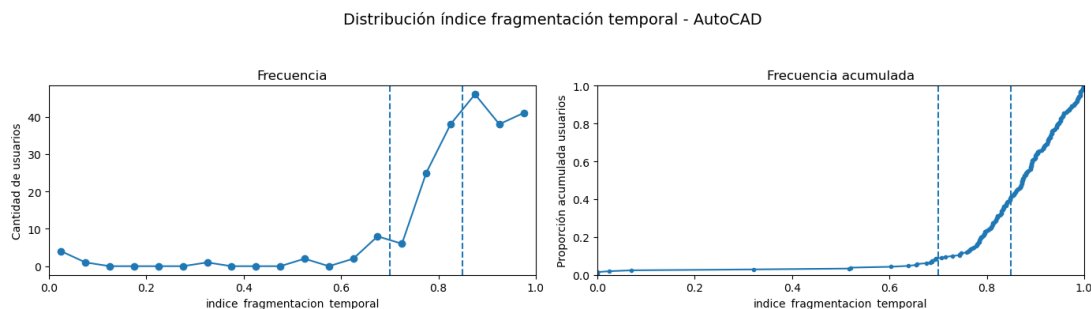


Figura 5.17: Distribución del indicador de fragmentación temporal en AutoCAD.

En AutoCAD, la mayor parte de los usuarios presenta valores altos de $I_{\text{Fragmentacin}}$, principalmente por sobre 0,7. Los valores bajos son poco frecuentes y la mayor concentración se encuentra entre 0,8 y 1.

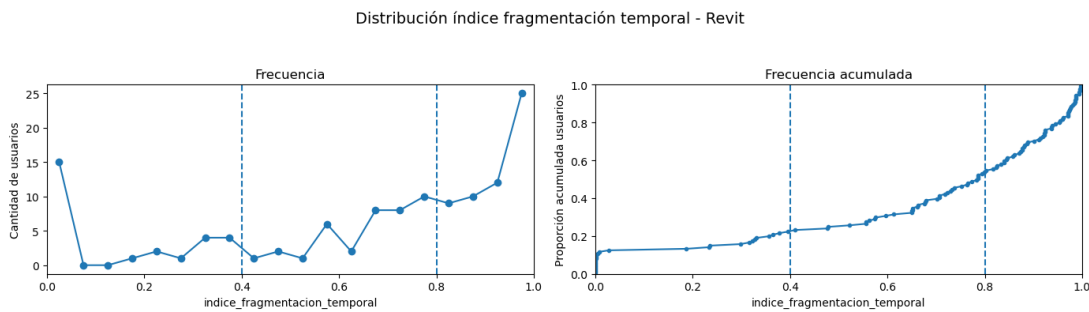


Figura 5.18: Distribución del indicador de fragmentación temporal en Revit.

En Revit, los valores de $I_{\text{Fragmentacin}}$ se distribuyen de forma más amplia. Se observan usuarios a lo largo de casi todo el rango del indicador, aunque la frecuencia aumenta nuevamente en los valores más altos.

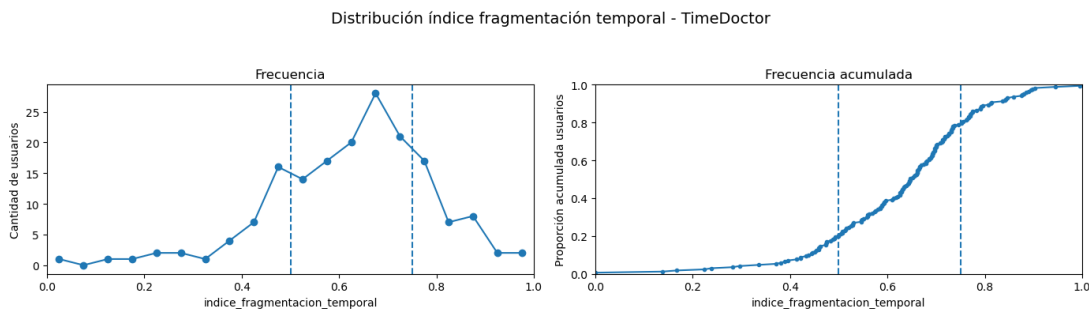


Figura 5.19: Distribución del indicador de fragmentación temporal en TimeDoctor.

En TimeDoctor, la mayor parte de los usuarios se concentra en valores intermedios y altos, principalmente entre 0,5 y 0,8. A diferencia de AutoCAD, la distribución presenta una concentración más clara alrededor de la zona central del indicador.

En conjunto, las tres plataformas presentan distribuciones diferentes de $I_{\text{Fragmentacin}}$, por lo que los puntos de corte se definieron de forma separada para cada una. En AutoCAD se utilizaron 0,7 y 0,85; en Revit, 0,4 y 0,8; y en TimeDoctor, 0,5 y 0,75. En cada caso, los valores inferiores al primer corte corresponden a *Baja fragmentación*, los valores entre ambos cortes a *Fragmentación media* y los valores superiores al segundo corte a *Alta fragmentación*.

5.5.2. Distribución de usuarios por categoría

Una vez definidos los rangos de $I_{\text{Fragmentacin}}$ para cada plataforma, los usuarios fueron clasificados en *Baja fragmentación*, *Fragmentación media* y *Alta fragmentación*. La Figura 5.20 muestra la cantidad de usuarios en cada categoría para AutoCAD, Revit y TimeDoctor.

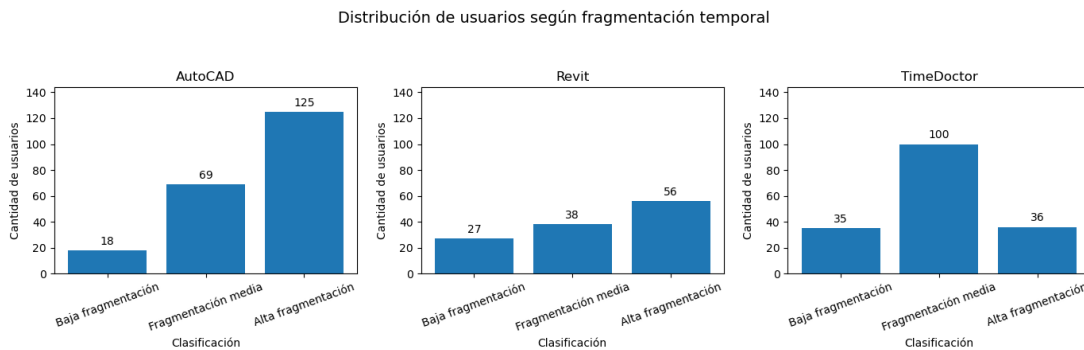


Figura 5.20: Distribución de usuarios según las categorías de fragmentación temporal.

En AutoCAD predomina la categoría *Alta fragmentación*, con 125 usuarios, seguida por *Fragmentación media* con 69 y *Baja fragmentación* con 18. En Revit también existe una mayor cantidad de usuarios en *Alta fragmentación*, aunque la distribución entre las tres categorías es más equilibrada.

En TimeDoctor el comportamiento es diferente, ya que la mayoría de los usuarios se concentra en *Fragmentación media*, con 100 usuarios, mientras que las categorías *Baja fragmentación* y *Alta fragmentación* presentan cantidades similares. Estos resultados son consistentes con las distribuciones observadas anteriormente y muestran que la fragmentación temporal presenta comportamientos distintos entre las tres plataformas.

5.5.3. Variabilidad del indicador

Para complementar el análisis gráfico y la distribución de usuarios por categoría, se analizaron los cuartiles y percentiles del indicador $I_{\text{Fragmentacin}}$ en cada plataforma. Los resultados se presentan en la Tabla 5.10.

Indicador	Plataforma	Q_1	Q_3	IQR	P_{10}	P_{90}	$P_{90} - P_{10}$
Fragmentación temporal	AutoCAD	0,8126	0,9340	0,1214	0,7624	0,9800	0,2176
Fragmentación temporal	Revit	0,6496	0,9365	0,2869	0,3310	0,9858	0,6548
Fragmentación temporal	TimeDoctor	0,5282	0,7310	0,2029	0,4462	0,8118	0,3657

Cuadro 5.10: Medidas descriptivas de dispersión del indicador de fragmentación temporal

En AutoCAD, el 50 % central de los usuarios presenta valores de $I_{\text{Fragmentación}}$ entre $Q_1 = 0,8126$ y $Q_3 = 0,9340$, con un $IQR = 0,1214$. Al considerar el 80 % central, los valores se extienden desde $P_{10} = 0,7624$ hasta $P_{90} = 0,9800$, con una diferencia de 0,2176. Esto muestra una distribución relativamente concentrada en valores altos del indicador.

En Revit, el 50 % central se encuentra entre $Q_1 = 0,6496$ y $Q_3 = 0,9365$, con el mayor IQR de las tres plataformas, igual a 0,2869. El 80 % central se extiende desde $P_{10} = 0,3310$ hasta $P_{90} = 0,9858$, con una diferencia de 0,6548. Estos valores muestran una alta variabilidad entre los usuarios, abarcando desde niveles bajos hasta niveles altos de fragmentación temporal.

En TimeDoctor, el 50 % central de los usuarios se encuentra entre $Q_1 = 0,5282$ y $Q_3 = 0,7310$, con un $IQR = 0,2029$. Al considerar el 80 % central, los valores van desde $P_{10} = 0,4462$ hasta $P_{90} = 0,8118$, con una diferencia de 0,3657. Esto muestra una variabilidad intermedia entre AutoCAD y Revit, con valores distribuidos principalmente en niveles medios del indicador.

La fragmentación temporal muestra diferencias más marcadas entre plataformas. Revit presenta la mayor variabilidad, TimeDoctor un comportamiento intermedio y AutoCAD concentra sus valores en un rango más reducido y principalmente alto.

5.5.4. Respuesta a la pregunta de investigación

Los resultados muestran que los usuarios presentan diferencias en sus niveles de fragmentación temporal. En las tres plataformas se observan valores distintos de $I_{\text{Fragmentación}}$, desde usuarios con una actividad más continua hasta otros con mayores niveles de interrupción.

Las medidas de dispersión respaldan estas diferencias. Revit presenta la mayor variabi-

lidad entre sus usuarios, TimeDoctor muestra un comportamiento intermedio y AutoCAD presenta una distribución más concentrada y principalmente en valores altos del indicador.

En fragmentación temporal también se observan diferencias entre usuarios, aunque su distribución varía más entre plataformas. Los resultados respaldan la hipótesis y las categorías permiten describir estos distintos niveles.

Capítulo 6

Relación entre indicadores

En este capítulo se analiza la relación entre los indicadores construidos y los perfiles generales de actividad obtenidos mediante GMM, cuyos parámetros fueron estimados mediante el algoritmo EM.

Para ello, primero se integran las variables construidas en una única tabla a nivel de usuario. Luego, se analiza la asociación entre las variables categóricas mediante la prueba Chi-cuadrado y el coeficiente V de Cramér. Posteriormente, se comparan los valores de los indicadores continuos entre los perfiles generales de actividad y estas diferencias se evalúan mediante la prueba de Kruskal–Wallis. En una primera etapa, estos análisis se realizan por separado para AutoCAD, Revit y TimeDoctor. Posteriormente, los p -values correspondientes a una misma hipótesis se combinan mediante el método de Fisher para evaluar la evidencia conjunta obtenida al considerar las tres plataformas.

6.1. Asociación entre variables categóricas

En una primera etapa se analiza la relación entre las categorías asignadas a los indicadores y los perfiles generales de actividad. El objetivo es determinar si existe asociación entre estas clasificaciones o si representan aspectos diferentes de la actividad de los usuarios.

6.1.1. Prueba Chi-cuadrado de independencia

Para analizar la relación entre las variables categóricas, se construyó una tabla de contingencia para cada par de variables. En estas tablas, cada celda representa la cantidad

de usuarios que pertenece simultáneamente a una categoría de cada variable.

Se analizaron todas las combinaciones entre el perfil general de actividad y los cuatro indicadores categorizados, además de las relaciones entre los propios indicadores. Para cada comparación se aplicó la prueba Chi-cuadrado de independencia, descrita en la Sección 3.9, obteniendo el estadístico χ^2 , los grados de libertad ν y el correspondiente p -value. Se utilizó un nivel de significancia de $\alpha = 0,05$ para evaluar la existencia de asociación entre las variables.

Antes de interpretar los resultados, se revisaron las frecuencias esperadas de las tablas de contingencia. Se consideró adecuada la aproximación Chi-cuadrado cuando ninguna frecuencia esperada fue inferior a 1 y al menos el 80 % de las celdas presentó una frecuencia esperada mayor o igual a 5.

Los resultados obtenidos para AutoCAD, Revit y TimeDoctor se presentan en la Tabla 6.1. En todas las comparaciones se obtuvieron $\nu = 4$ grados de libertad.

Posteriormente, los p -values obtenidos para una misma comparación en las tres plataformas se combinaron mediante el método de Fisher, descrito en la Sección 3.12, con el objetivo de evaluar la evidencia conjunta de asociación considerando los tres conjuntos de datos.

Variables comparadas	AutoCAD		Revit		TimeDoctor	
	χ^2	p	χ^2	p	χ^2	p
Perfil general – Concentración horaria	24,3050	0,0001	14,9179	0,0049	17,2453	0,0017
Perfil general – Diversidad documental	16,4546	0,0025	28,1747	< 0,0001	41,6885	< 0,0001
Perfil general – Concentración documental	52,2759	< 0,0001	35,0644	< 0,0001	127,0484	< 0,0001
Perfil general – Fragmentación temporal	34,5952	< 0,0001	51,1863	< 0,0001	2,4188	0,6592
Concentración horaria – Diversidad documental	6,6994	0,1527	16,4300	0,0025	18,0070	0,0012
Concentración horaria – Concentración documental	22,4658	0,0002	27,0053	< 0,0001	22,3878	0,0002
Concentración horaria – Fragmentación temporal	6,4288	0,1693	7,8829	0,0960	0,6857	0,9531
Diversidad documental – Concentración documental	98,5991	< 0,0001	40,2233	< 0,0001	81,0568	< 0,0001
Diversidad documental – Fragmentación temporal	23,2804	0,0001	4,9744	0,2899	5,0745	0,2797
Concentración documental – Fragmentación temporal	40,1754	< 0,0001	28,7603	< 0,0001	4,0476	0,3996

Cuadro 6.1: Resultados de la prueba Chi-cuadrado entre las variables categóricas

Los resultados muestran que algunas relaciones presentan evidencia de asociación en las tres plataformas, mientras que otras varían según el conjunto de datos analizado. Para complementar esta comparación y obtener una medida conjunta de la evidencia observada,

se combinaron los p -values correspondientes a una misma relación mediante el método de Fisher.

Variables comparadas	X_F	p -value combinado	Resultado
Diversidad documental – Concentración documental	198,4908	$3,976 \times 10^{-40}$	Evidencia global significativa
Perfil general – Concentración documental	193,6137	$4,336 \times 10^{-39}$	Evidencia global significativa
Perfil general – Fragmentación temporal	74,2400	$5,503 \times 10^{-14}$	Evidencia global significativa
Perfil general – Diversidad documental	70,2779	$3,586 \times 10^{-13}$	Evidencia global significativa
Concentración documental – Fragmentación temporal	59,2067	$6,522 \times 10^{-11}$	Evidencia global significativa
Concentración horaria – Concentración documental	56,5001	$2,306 \times 10^{-10}$	Evidencia global significativa
Perfil general – Concentración horaria	42,5163	$1,454 \times 10^{-7}$	Evidencia global significativa
Concentración horaria – Diversidad documental	29,1487	$5,702 \times 10^{-5}$	Evidencia global significativa
Diversidad documental – Fragmentación temporal	23,2306	0,0007	Evidencia global significativa
Concentración horaria – Fragmentación temporal	8,3355	0,2145	Sin evidencia global significativa

Cuadro 6.2: Combinación de los p -values de las pruebas Chi-cuadrado mediante el método de Fisher

Los resultados por plataforma muestran que varias de las relaciones analizadas presentan p -values inferiores a 0,05, aunque este comportamiento no es igual en todos los casos. La relación entre diversidad y concentración documental destaca por presentar asociación significativa en AutoCAD, Revit y TimeDoctor, además de registrar algunos de los mayores valores de χ^2 . En cambio, la relación entre concentración horaria y fragmentación temporal no presenta evidencia de asociación en ninguna de las tres plataformas.

También se observan relaciones cuyo resultado depende de la plataforma. Por ejemplo, la asociación entre el perfil general y la fragmentación temporal es significativa en AutoCAD y Revit, pero no en TimeDoctor. Estas diferencias muestran que una misma relación puede presentar comportamientos distintos según el conjunto de datos analizado.

Al combinar mediante el método de Fisher los p -values correspondientes a una misma comparación, se obtuvo evidencia conjunta significativa en nueve de las diez relaciones analizadas. La única excepción fue la relación entre concentración horaria y fragmentación temporal, cuyo p -value combinado fue 0,2145. Por lo tanto, esta relación no presenta evidencia de asociación ni al analizar las plataformas por separado ni al considerar sus resultados de manera conjunta.

La significancia obtenida mediante Fisher debe interpretarse como evidencia conjunta a partir de los resultados de las tres plataformas y no como una indicación de que la asociación sea significativa individualmente en cada una de ellas. Esto se observa, por

ejemplo, en la relación entre el perfil general y la fragmentación temporal, que presenta un resultado combinado significativo a pesar de no ser significativa en TimeDoctor.

Sin embargo, tanto la prueba Chi-cuadrado como la combinación de sus p -values permiten evaluar la existencia de evidencia de asociación, pero no indican su intensidad. Por esta razón, el análisis se complementa a continuación mediante el coeficiente V de Cramér.

6.1.2. Coeficiente V de Cramér

Una vez evaluada la existencia de asociación mediante la prueba Chi-cuadrado, se utilizó el coeficiente V de Cramér, definido en la Sección 3.10, para cuantificar la intensidad de la asociación entre cada par de variables categóricas.

El coeficiente se calculó de manera independiente para AutoCAD, Revit y TimeDoctor. Con el objetivo de facilitar la comparación entre plataformas, la Tabla 6.3 presenta los valores obtenidos en cada una de ellas, junto con el promedio de V de Cramér para cada relación. Las relaciones se ordenaron de mayor a menor según este promedio.

Además, se incorpora el p -value combinado obtenido previamente mediante el método de Fisher. Este valor permite complementar la lectura de la tabla indicando si existe evidencia conjunta de asociación entre las variables considerando las tres plataformas, mientras que los valores de V de Cramér permiten comparar la intensidad de dicha asociación.

Variables comparadas	AutoCAD	Revit	TimeDoctor	Promedio V	p -value combinado
Diversidad documental – Concentración documental	0,4880	0,4276	0,4883	0,4680	$3,976 \times 10^{-40}$
Perfil general – Concentración documental	0,3553	0,3992	0,6113	0,4553	$4,336 \times 10^{-39}$
Perfil general – Diversidad documental	0,1994	0,3579	0,3502	0,3025	$3,586 \times 10^{-13}$
Perfil general – Fragmentación temporal	0,2891	0,4824	0,0843	0,2853	$5,503 \times 10^{-14}$
Concentración horaria – Concentración documental	0,2329	0,3504	0,2566	0,2800	$2,306 \times 10^{-10}$
Concentración documental – Fragmentación temporal	0,3115	0,3616	0,1091	0,2607	$6,522 \times 10^{-11}$
Perfil general – Concentración horaria	0,2423	0,2604	0,2252	0,2426	$1,454 \times 10^{-7}$
Concentración horaria – Diversidad documental	0,1272	0,2733	0,2301	0,2102	$5,702 \times 10^{-5}$
Diversidad documental – Fragmentación temporal	0,2371	0,1504	0,1222	0,1699	0,0007
Concentración horaria – Fragmentación temporal	0,1246	0,1893	0,0449	0,1196	0,2145

Cuadro 6.3: Valores de V de Cramér en AutoCAD, Revit y TimeDoctor, ordenados según su valor promedio

Los resultados muestran que las asociaciones de mayor intensidad promedio se concentran principalmente en las variables documentales. La relación entre diversidad y concentración documental presenta el mayor valor promedio de V de Cramér, con $V = 0,4680$, se-

guida por la relación entre el perfil general y la concentración documental, con $V = 0,4553$. En ambos casos, el p -value combinado obtenido mediante Fisher indica evidencia conjunta significativa de asociación en las tres plataformas.

La concentración documental también aparece relacionada con otras variables. Su asociación promedio con la concentración horaria alcanza $V = 0,2800$, mientras que con la fragmentación temporal alcanza $V = 0,2607$. Además, el perfil general presenta asociaciones con la diversidad documental $V = 0,3025$, la fragmentación temporal $V = 0,2853$ y la concentración horaria $V = 0,2426$. En todos estos casos, los p -values combinados son inferiores a 0,05.

En el extremo inferior se encuentran las relaciones que involucran la fragmentación temporal. La asociación entre diversidad documental y fragmentación temporal presenta un promedio de $V = 0,1699$, mientras que la relación entre concentración horaria y fragmentación temporal alcanza el menor valor de la tabla, con $V = 0,1196$. Esta última es además la única comparación cuyo p -value combinado no resulta significativo, con $p = 0,2145$.

La comparación de las asociaciones muestra que la concentración documental alcanza los valores más altos de V de Cramér, especialmente en su relación con la diversidad documental y con el perfil general. Sin embargo, los valores individuales de V de Cramér también muestran que la intensidad de algunas relaciones cambia entre plataformas, por lo que el promedio se utiliza únicamente como una medida descriptiva para facilitar su comparación.

6.1.3. Interpretación de las asociaciones

Los resultados muestran que las relaciones entre las variables no se comportan de la misma forma en las tres plataformas. Algunas asociaciones se mantienen de manera consistente, mientras que otras dependen en mayor medida del conjunto de datos analizado.

Las relaciones más consistentes se concentran en los indicadores documentales. La diversidad y la concentración documental presentan asociación en las tres plataformas, lo que resulta coherente con su construcción, ya que ambas se basan en la distribución del tiempo entre los documentos utilizados, aunque representan características diferentes de

dicha distribución. Además, ambos indicadores presentan asociaciones con el perfil general de actividad en AutoCAD, Revit y TimeDoctor.

La concentración horaria presenta un comportamiento menos uniforme. Algunas de sus relaciones se mantienen entre plataformas, como ocurre con la concentración documental, mientras que otras cambian según el conjunto de datos analizado. La fragmentación temporal muestra una variación aún mayor, con asociaciones presentes principalmente en AutoCAD y Revit y sin evidencia de asociación con las demás variables en TimeDoctor.

La evidencia obtenida hasta este punto destaca a los indicadores documentales como los más consistentemente relacionados con el perfil general y con las demás variables analizadas. La concentración horaria y, especialmente, la fragmentación temporal muestran un comportamiento más dependiente de la plataforma. A continuación, el análisis se complementa utilizando los valores continuos de los indicadores para evaluar si sus distribuciones difieren entre los perfiles generales de actividad.

6.2. Comparación de indicadores entre perfiles

En esta sección se comparan los valores de los indicadores continuos entre los perfiles generales de actividad obtenidos mediante GMM. Primero se revisan gráficamente las distribuciones de cada indicador, considerando su ubicación, dispersión y superposición entre los clusters. Posteriormente, estas diferencias se evalúan mediante la prueba de Kruskal–Wallis.

6.2.1. Concentración horaria en AutoCAD

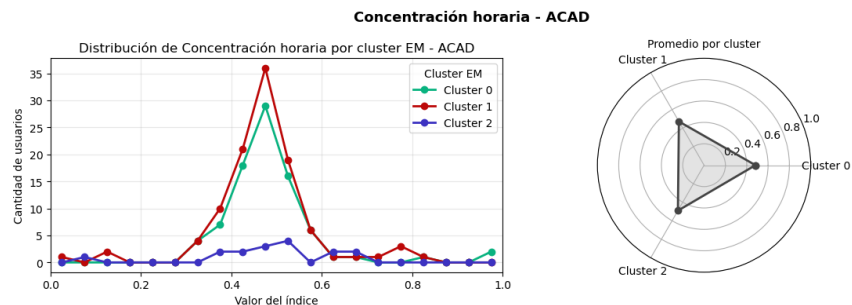


Figura 6.1: Comparación del indicador de concentración horaria entre clusters en AutoCAD

Los clusters 0 y 1 concentran la mayor parte de sus valores de P_{MM} entre 0,3 y 0,6, con una alta superposición entre sus distribuciones. El cluster 2 presenta una mayor dispersión y una menor cantidad de observaciones, aunque también coincide con los otros grupos en la zona central del indicador.

Los valores promedio son cercanos entre los tres clusters. En conjunto, el análisis gráfico muestra diferencias acotadas en la concentración horaria entre los perfiles de AutoCAD.

6.2.2. Concentración horaria en Revit

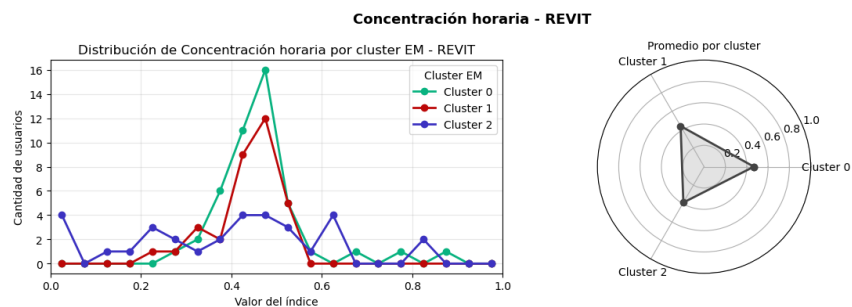


Figura 6.2: Comparación del indicador de concentración horaria entre clusters en Revit

Los clusters 0 y 1 concentran la mayor parte de sus valores de P_{MM} entre 0,3 y 0,6 y presentan una alta superposición. El cluster 2 muestra una distribución más dispersa, con

usuarios tanto en valores bajos como altos del indicador, aunque mantiene una superposición importante con los otros grupos.

Los valores promedio también son cercanos entre los tres clusters. Por lo tanto, las diferencias observadas gráficamente entre los perfiles de Revit son reducidas.

6.2.3. Concentración horaria en TimeDoctor

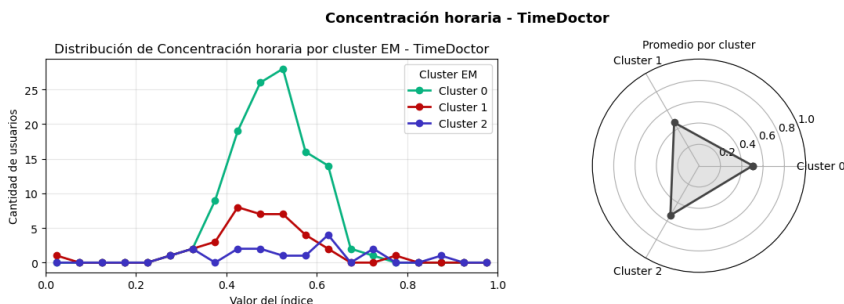


Figura 6.3: Comparación del indicador de concentración horaria entre clusters en TimeDoctor

El cluster 0 concentra la mayor parte de sus valores de P_{MM} entre 0,4 y 0,65. Los clusters 1 y 2 presentan distribuciones más dispersas y una menor concentración de usuarios.

La superposición entre los grupos es menor que en AutoCAD y Revit, aunque sigue presente en parte del rango del indicador. Los valores promedio son similares, por lo que las principales diferencias observadas se encuentran en la dispersión y forma de las distribuciones.

6.2.4. Comparación general de la concentración horaria

En las tres plataformas se observa una superposición importante de los valores de P_{MM} entre los perfiles generales de actividad. AutoCAD y Revit presentan las distribuciones más similares entre clusters, mientras que en TimeDoctor se observan mayores diferencias en su dispersión.

En términos descriptivos, la concentración horaria no muestra una separación clara entre los perfiles. Estas diferencias se evalúan posteriormente mediante la prueba de

Kruskal–Wallis para determinar si son estadísticamente significativas.

6.2.5. Diversidad documental en AutoCAD

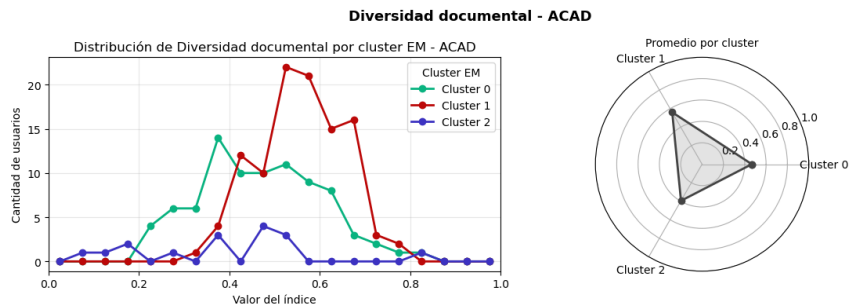


Figura 6.4: Comparación del indicador de diversidad documental entre clusters en AutoCAD

El cluster 1 concentra la mayor parte de sus valores de $H_{\text{norm}}(\mathbf{q})$ entre 0,5 y 0,65, mientras que el cluster 0 presenta una distribución más amplia y desplazada hacia valores menores. El cluster 2 contiene menos usuarios y presenta una mayor dispersión.

Los clusters 0 y 1 se superponen principalmente entre 0,45 y 0,60, aunque sus distribuciones muestran una separación mayor que la observada anteriormente para la concentración horaria. Los valores promedio también presentan diferencias, especialmente entre los clusters 0 y 1.

6.2.6. Diversidad documental en Revit

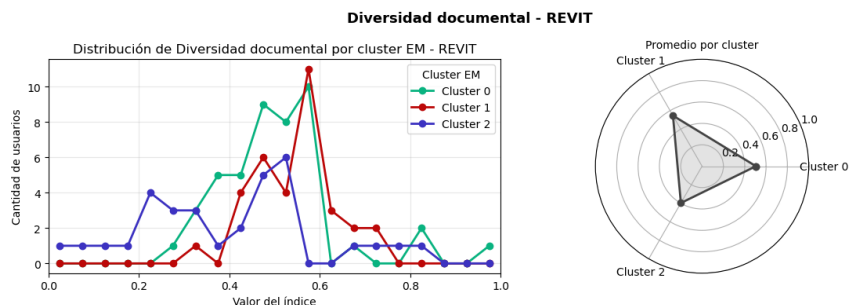


Figura 6.5: Comparación del indicador de diversidad documental entre clusters en Revit

Los clusters 0 y 1 concentran gran parte de sus valores de $H_{\text{norm}}(\mathbf{q})$ entre 0,4 y 0,6, generando una alta superposición entre ambos grupos. El cluster 2 presenta una distribución más dispersa y con una mayor presencia de valores bajos del indicador.

Aunque existe una superposición importante, se observan diferencias en la forma y dispersión de las distribuciones. Los valores promedio también presentan diferencias entre los clusters.

6.2.7. Diversidad documental en TimeDoctor

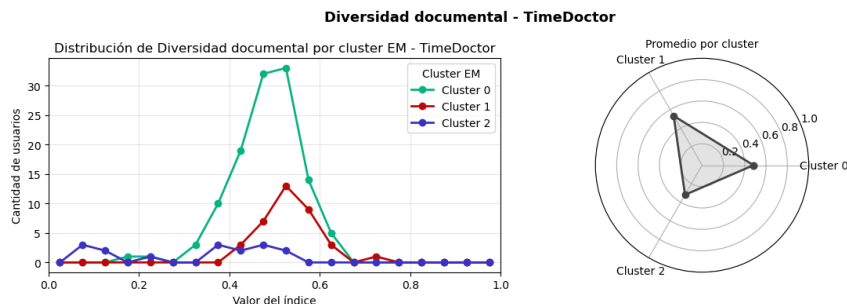


Figura 6.6: Comparación del indicador de diversidad documental entre clusters en TimeDoctor

El cluster 0 concentra la mayor parte de sus valores de $H_{\text{norm}}(\mathbf{q})$ entre 0,4 y 0,6. En cambio, los clusters 1 y 2 presentan distribuciones más dispersas y con una mayor presencia de valores bajos del indicador.

La superposición entre los grupos es menor que en Revit, principalmente por la mayor concentración del cluster 0 en la zona media del indicador. Los valores promedio también muestran diferencias entre los perfiles, en concordancia con las distribuciones observadas.

6.2.8. Comparación general de la diversidad documental

La diversidad documental presenta diferencias entre los perfiles generales de actividad, aunque estas no siguen el mismo comportamiento en las tres plataformas. En AutoCAD se observa una separación más clara entre los clusters 0 y 1, mientras que en Revit existe una mayor superposición. En TimeDoctor también se observan diferencias en la ubicación

y dispersión de las distribuciones.

En términos descriptivos, el indicador muestra una mayor separación entre algunos perfiles que la observada para la concentración horaria. Sin embargo, las distribuciones continúan presentando superposición, por lo que las diferencias observadas se evalúan posteriormente mediante la prueba de Kruskal–Wallis.

6.2.9. Concentración documental en AutoCAD

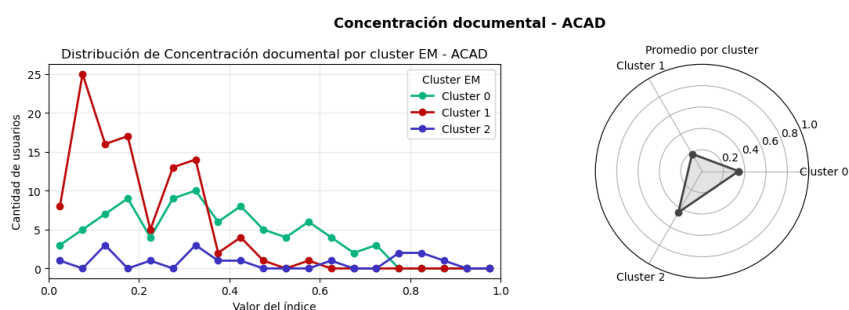


Figura 6.7: Comparación del indicador de concentración documental entre clusters en AutoCAD

El cluster 1 concentra la mayor parte de sus valores de $HHI(\mathbf{q})$ por debajo de 0,35. El cluster 0 presenta una distribución más amplia, con observaciones que se extienden hasta valores cercanos a 0,7.

El cluster 2 contiene menos usuarios y presenta una mayor presencia de valores altos del indicador, alcanzando valores cercanos a 0,8.

En términos descriptivos, se observa una diferencia entre los tres perfiles: el cluster 1 presenta menores niveles de concentración documental, el cluster 0 ocupa valores intermedios y el cluster 2 presenta los valores más altos. Los valores promedio muestran la misma tendencia.

6.2.10. Concentración documental en Revit

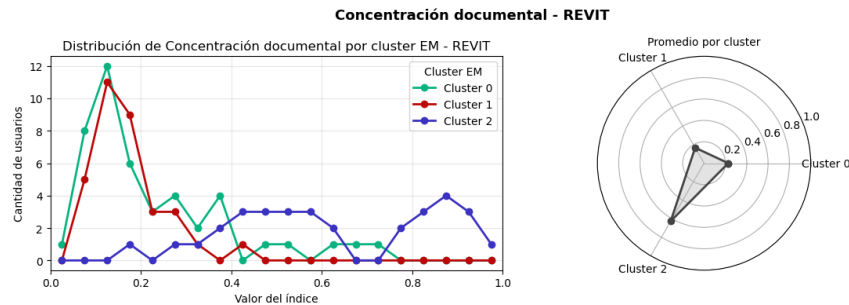


Figura 6.8: Comparación del indicador de concentración documental entre clusters en Revit

Los clusters 0 y 1 presentan una alta concentración de valores bajos de $HHI(q)$, con máximos cercanos a 0,1. Sin embargo, el cluster 0 se extiende hasta valores cercanos a 0,7, mientras que el cluster 1 presenta principalmente valores inferiores a 0,4.

El cluster 2 muestra una distribución distinta, con una mayor presencia de valores medios y altos y un aumento hacia valores cercanos a 0,9, donde los otros clusters presentan pocas observaciones.

Por lo tanto, los clusters 0 y 1 muestran una mayor superposición en los valores bajos, mientras que el cluster 2 presenta mayores niveles de concentración documental. Los valores promedio muestran el mismo comportamiento.

6.2.11. Concentración documental en TimeDoctor

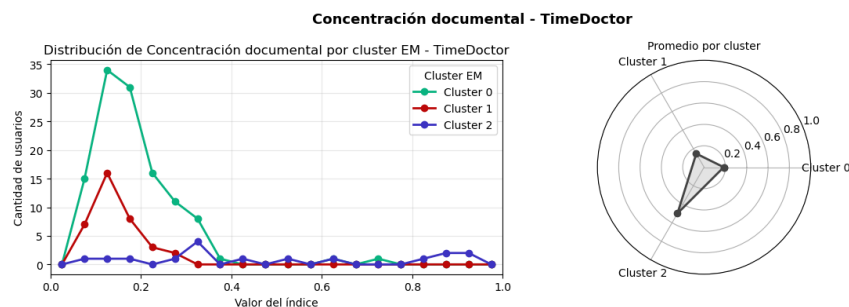


Figura 6.9: Comparación del indicador de concentración documental entre clusters en TimeDoctor

Los clusters 0 y 1 concentran la mayor parte de sus valores de $HHI(\mathbf{q})$ en la zona baja del indicador, con máximos cercanos a 0,1. El cluster 0 se extiende hacia valores algo mayores, mientras que el cluster 1 presenta principalmente valores inferiores a 0,3.

El cluster 2 presenta una distribución más amplia y alcanza valores cercanos a 0,9, diferenciándose de los otros grupos por su mayor presencia en niveles altos de concentración documental.

Los valores promedio mantienen esta tendencia, con valores mayores para el cluster 2 y menores para los clusters 0 y 1.

6.2.12. Comparación general de la concentración documental

En las tres plataformas se observan diferencias entre los perfiles en los valores de $HHI(\mathbf{q})$. Algunos clusters concentran sus observaciones en valores bajos, asociados a una distribución más equilibrada del tiempo entre los documentos, mientras que otros presentan valores mayores, asociados a una mayor concentración del tiempo en uno o pocos documentos.

La separación visual es especialmente clara para los perfiles que presentan los valores más altos de concentración documental, aunque continúa existiendo superposición entre algunos grupos. En términos descriptivos, este indicador muestra diferencias más marcadas entre los perfiles que las observadas para los otros indicadores. Posteriormente, la prueba de Kruskal–Wallis permitirá evaluar si estas diferencias son estadísticamente significativas.

6.2.13. Fragmentación temporal en AutoCAD

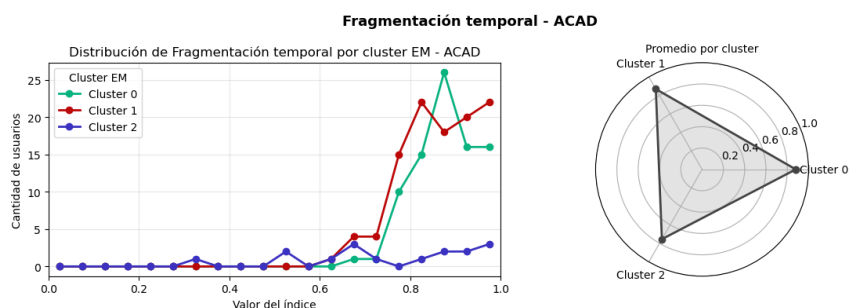


Figura 6.10: Comparación del indicador de fragmentación temporal entre clusters en AutoCAD

Los clusters 0 y 1 concentran la mayor parte de sus valores de $I_{\text{Fragmentación}}$ por sobre 0,7. El cluster 0 alcanza su mayor concentración cerca de 0,85, mientras que el cluster 1 mantiene presencia hasta los valores más altos del indicador.

El cluster 2 contiene menos usuarios y presenta una distribución más dispersa, principalmente entre valores intermedios y altos. Los valores promedio son altos y cercanos entre los tres grupos, siendo el cluster 0 el que presenta el mayor promedio.

6.2.14. Fragmentación temporal en Revit

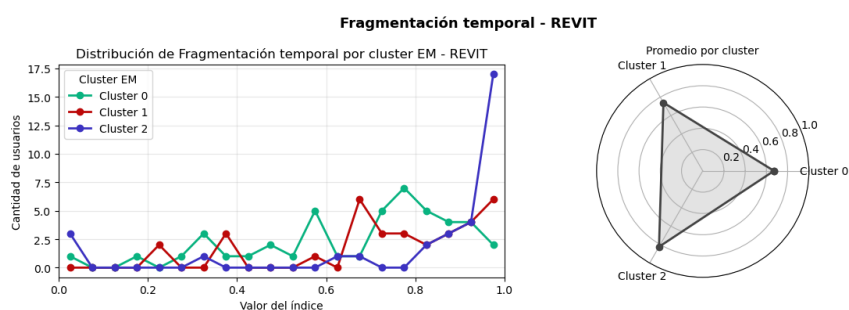


Figura 6.11: Comparación del indicador de fragmentación temporal entre clusters en Revit

Los tres clusters presentan valores de $I_{\text{Fragmentación}}$ distribuidos principalmente en niveles medios y altos. El cluster 0 concentra una mayor cantidad de usuarios cerca de 0,8, mientras que el cluster 1 presenta una distribución más extendida a lo largo del indicador.

El cluster 2 también presenta una distribución dispersa, pero muestra una mayor presencia de valores cercanos a 1. Los valores promedio son altos en los tres grupos, siendo el cluster 2 el que presenta el mayor promedio.

6.2.15. Fragmentación temporal en TimeDoctor

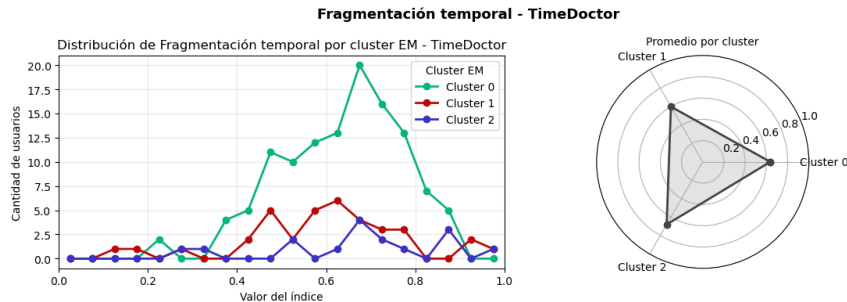


Figura 6.12: Comparación del indicador de fragmentación temporal entre clusters en TimeDoctor

El cluster 0 concentra gran parte de sus valores de $I_{\text{Fragmentación}}$ en niveles medios y altos, con una mayor presencia cerca de 0,7. El cluster 1 presenta un comportamiento similar, aunque con una distribución más dispersa.

El cluster 2 contiene menos usuarios y presenta una distribución más irregular, también concentrada principalmente en valores medios y altos. En esta plataforma existe una mayor superposición entre los tres grupos y sus valores promedio son cercanos.

6.2.16. Comparación general de la fragmentación temporal

La fragmentación temporal presenta comportamientos diferentes entre plataformas. En AutoCAD los tres perfiles se concentran principalmente en valores altos, mientras que en Revit se observan mayores diferencias en la forma y ubicación de sus distribuciones. En TimeDoctor, en cambio, existe una mayor superposición entre los grupos.

En términos descriptivos, el indicador no muestra una separación uniforme entre los perfiles de las tres plataformas. Las diferencias observadas se evalúan posteriormente mediante la prueba de Kruskal–Wallis para determinar si son estadísticamente significativas.

6.3. Resultados de la prueba de Kruskal–Wallis

En esta sección se evalúa si la distribución de cada indicador continuo presenta diferencias entre los perfiles generales de actividad. Para ello, se aplicó la prueba de Kruskal–Wallis, descrita en la Sección 3.11, considerando inicialmente los tres clusters de cada plataforma en conjunto.

La prueba global se realizó de manera independiente para AutoCAD, Revit y TimeDoctor. Además, para cada indicador se combinaron los p -values obtenidos en las tres plataformas mediante el método de Fisher, descrito en la Sección 3.12, con el objetivo de obtener una medida conjunta de la evidencia observada.

Cuando la prueba global de Kruskal–Wallis mostró diferencias significativas en las tres plataformas para un mismo indicador, el análisis se complementó con comparaciones por pares de clusters, considerando los pares 0–1, 0–2 y 1–2. Este criterio permite realizar las comparaciones bajo una misma estructura en AutoCAD, Revit y TimeDoctor y posteriormente combinar sus p -values mediante el método de Fisher.

Para complementar las comparaciones por pares entre plataformas, los p -values correspondientes a una misma comparación se combinaron también mediante el método de Fisher. De esta forma, el análisis considera tanto los resultados obtenidos individualmente en cada plataforma como la evidencia conjunta observada para cada comparación.

La Tabla 6.4 resume los resultados de la prueba global de Kruskal–Wallis para los cuatro indicadores en AutoCAD, Revit y TimeDoctor. Además, presenta el promedio del estadístico H_{KW} y el p -value combinado mediante el método de Fisher, con el fin de comparar los resultados individuales y la evidencia conjunta entre plataformas.

Indicador	AutoCAD		Revit		TimeDoctor		Estadístico	p -value
	H_{KW}	p -value	H_{KW}	p -value	H_{KW}	p -value	X_F	Fisher
Concentración horaria	1,7330	0,4204	2,0608	0,3569	3,8121	0,1487	7,6059	0,2684
Diversidad documental	35,5922	$1,867 \times 10^{-8}$	16,1649	0,0003	30,3348	$2,588 \times 10^{-7}$	82,0919	$1,320 \times 10^{-15}$
Concentración documental	40,0636	$1,997 \times 10^{-9}$	55,8675	$7,388 \times 10^{-13}$	27,7019	$9,652 \times 10^{-7}$	123,6330	$2,810 \times 10^{-24}$
Fragmentación temporal	5,4404	0,0659	19,5364	$5,724 \times 10^{-5}$	3,5076	0,1731	28,4844	$7,614 \times 10^{-5}$

Cuadro 6.4: Resultados de la prueba de Kruskal–Wallis y combinación de p -values mediante el método de Fisher

Los resultados muestran que la diversidad documental y la concentración documental son los indicadores que presentan el comportamiento más consistente entre plataformas. En AutoCAD, Revit y TimeDoctor, ambos presentan p -values inferiores a 0,05, por lo que existe evidencia de diferencias entre los perfiles generales de actividad. Este resultado también se mantiene al considerar conjuntamente las tres plataformas mediante el método de Fisher, obteniéndose $X_F = 82,0919$ para la diversidad documental y $X_F = 123,6330$ para la concentración documental.

La concentración horaria presenta un comportamiento distinto. En ninguna de las tres plataformas se encontraron diferencias estadísticamente significativas entre los perfiles. Este resultado se mantiene al considerar conjuntamente los tres p -values mediante Fisher, por lo que no se obtiene evidencia suficiente para afirmar que la distribución de este indicador difiere entre los perfiles generales de actividad.

La fragmentación temporal presenta mayores diferencias entre plataformas. En Revit se observaron diferencias significativas entre los perfiles, mientras que en AutoCAD y TimeDoctor no se rechazó la hipótesis nula. Sin embargo, la combinación de los tres p -values mediante Fisher resulta significativa, indicando que existe evidencia conjunta de diferencias al considerar los resultados de las tres plataformas. Este resultado no implica que la fragmentación presente diferencias significativas en cada plataforma por separado, sino que resume la evidencia obtenida en los tres análisis.

Las pruebas sobre los valores continuos refuerzan el resultado anterior: diversidad y concentración documental presentan las diferencias más consistentes entre los perfiles generales de actividad. La concentración horaria no muestra diferencias entre perfiles, mientras que la fragmentación temporal presenta un comportamiento menos uniforme entre plataformas.

A partir de estos resultados, las comparaciones por pares se realizan únicamente para los indicadores cuya prueba global de Kruskal–Wallis resultó significativa en AutoCAD, Revit y TimeDoctor. Bajo este criterio, las comparaciones posteriores se realizan para la diversidad documental y la concentración documental.

6.3.1. Comparaciones por pares entre clusters

Para los indicadores que presentaron diferencias significativas en la prueba global de Kruskal–Wallis, se realizaron comparaciones por pares entre los clusters 0–1, 0–2 y 1–2, aplicando nuevamente la prueba con $G = 2$. Esto permite identificar entre qué grupos se encuentran las diferencias observadas.

Como las etiquetas de los perfiles obtenidos mediante GMM fueron homologadas previamente entre las tres plataformas, las comparaciones 0–1, 0–2 y 1–2 corresponden a los mismos pares de perfiles en AutoCAD, Revit y TimeDoctor.

Las comparaciones se realizaron por separado en AutoCAD, Revit y TimeDoctor. Además, para cada par se combinaron los p -values obtenidos en las tres plataformas mediante el método de Fisher, utilizando un nivel de significancia de $\alpha = 0,05$.

6.3.2. Concentración horaria

La concentración horaria no presentó diferencias estadísticamente significativas entre los tres clusters en AutoCAD, Revit ni TimeDoctor. Por esta razón, no se realizaron comparaciones posteriores por pares para este indicador, ya que la prueba global de Kruskal–Wallis no entregó evidencia suficiente de diferencias entre los perfiles.

6.3.3. Diversidad documental

Dado que la prueba global de Kruskal–Wallis mostró diferencias significativas en la diversidad documental para AutoCAD, Revit y TimeDoctor, se realizaron comparaciones por pares entre los clusters 0–1, 0–2 y 1–2 en cada plataforma.

Posteriormente, los p -values obtenidos para cada par se combinaron mediante el método de Fisher, con el objetivo de evaluar la evidencia conjunta de diferencias entre los grupos. Los resultados se presentan en la Tabla [6.5](#).

Cuadro 6.5: Comparaciones por pares para el indicador de diversidad documental

Par	p AutoCAD	p Revit	p TimeDoctor	X_F	p -value Fisher	Resultado
0–1	$8,1 \times 10^{-5}$	0,4403	0,000432	35,9769	$2,785 \times 10^{-6}$	Significativo
0–2	0,3467	0,000750	$6,3 \times 10^{-5}$	35,8542	$2,942 \times 10^{-6}$	Significativo
1–2	0,9565	0,001764	$4,2 \times 10^{-5}$	32,9250	$1,084 \times 10^{-5}$	Significativo

Al analizar las plataformas por separado, las diferencias entre clusters no se presentan de la misma forma. En AutoCAD, solo la comparación entre los clusters 0 y 1 resulta significativa, mientras que en Revit las diferencias se observan entre los pares 0–2 y 1–2. En TimeDoctor, en cambio, las tres comparaciones presentan diferencias significativas.

Al combinar mediante Fisher los p -values correspondientes a cada par, las tres comparaciones presentan evidencia conjunta significativa. Se obtuvieron p -values combinados de $2,785 \times 10^{-6}$ para el par 0–1, $2,942 \times 10^{-6}$ para 0–2 y $1,084 \times 10^{-5}$ para 1–2.

Estos resultados indican que, aunque las diferencias observadas por separado dependen de la plataforma, al considerar conjuntamente la evidencia de AutoCAD, Revit y TimeDoctor se obtienen diferencias significativas para los tres pares de clusters. Esto refuerza el resultado de la prueba global, mostrando que la diversidad documental permite distinguir entre los perfiles generales de actividad.

6.3.4. Concentración documental

Dado que la prueba global de Kruskal–Wallis mostró diferencias significativas en la concentración documental para AutoCAD, Revit y TimeDoctor, se realizaron comparaciones por pares entre los clusters 0–1, 0–2 y 1–2 en cada plataforma.

Posteriormente, los p -values correspondientes a cada par se combinaron mediante el método de Fisher, con el objetivo de evaluar la evidencia conjunta de diferencias entre los grupos. Los resultados se presentan en la Tabla [6.6](#).

Cuadro 6.6: Comparaciones por pares para el indicador de concentración documental

Par	p AutoCAD	p Revit	p TimeDoctor	X_F	p -value Fisher	Resultado
0-1	$3,007 \times 10^{-12}$	$5,303 \times 10^{-8}$	$1,343 \times 10^{-12}$	141,2372	$5,493 \times 10^{-28}$	Significativo
0-2	0,08711	0,000597	$6,3 \times 10^{-5}$	39,0731	$6,925 \times 10^{-7}$	Significativo
1-2	0,09613	0,000228	$4,2 \times 10^{-5}$	41,6121	$2,193 \times 10^{-7}$	Significativo

Al analizar las plataformas por separado, las diferencias entre clusters no se presentan de la misma forma. En AutoCAD, solo la comparación entre los clusters 0 y 1 resulta significativa, mientras que los pares 0-2 y 1-2 no presentan diferencias significativas. En Revit y TimeDoctor, en cambio, las tres comparaciones resultan significativas.

Al combinar mediante Fisher los p -values correspondientes a cada par, las tres comparaciones presentan evidencia conjunta significativa. Se obtuvieron p -values combinados de $5,493 \times 10^{-28}$ para el par 0-1, $6,925 \times 10^{-7}$ para 0-2 y $2,193 \times 10^{-7}$ para 1-2.

En conjunto, los resultados muestran que la concentración documental presenta diferencias entre los perfiles generales de actividad. Aunque estas diferencias varían al analizar cada plataforma por separado, la combinación de los resultados mediante Fisher muestra evidencia significativa para los tres pares de clusters.

6.3.5. Fragmentación temporal

La fragmentación temporal presentó diferencias estadísticamente significativas entre los tres clusters únicamente en Revit. En AutoCAD y TimeDoctor, la prueba global de Kruskal-Wallis no entregó evidencia suficiente de diferencias entre los perfiles.

Dado que la prueba global no resultó significativa en las tres plataformas, y de acuerdo con el criterio definido para las comparaciones posteriores, no se realizaron comparaciones por pares para este indicador.

6.4. Propuesta de caracterización de los perfiles

A partir de los resultados obtenidos en las secciones anteriores, se propone una forma complementaria de caracterizar los perfiles generales de actividad mediante métricas

cuantitativas. El objetivo no es definir una clasificación única ni definitiva, sino mostrar que los clusters pueden interpretarse y ordenarse según distintos criterios, dependiendo del aspecto del comportamiento que se quiera analizar. Para ello, se consideran dos métricas: un puntaje documental y una medida de distancia respecto del comportamiento general.

6.4.1. Puntaje documental

El puntaje documental se construye a partir de los indicadores de diversidad documental, $H_{\text{norm}}(\mathbf{q}_i)$, y concentración documental, $HHI(\mathbf{q}_i)$, donde $\mathbf{q}_i$ representa la distribución del tiempo del usuario i entre los documentos utilizados.

La construcción del puntaje considera que ambos indicadores representan comportamientos opuestos. Una mayor diversidad documental indica una distribución más uniforme del tiempo entre los documentos utilizados, mientras que una mayor concentración documental indica que una mayor proporción del tiempo se concentra en uno o pocos documentos. Por esta razón, el puntaje aumenta con $H_{\text{norm}}(\mathbf{q}_i)$ y disminuye con $HHI(\mathbf{q}_i)$.

Para cada usuario i , la diversidad documental y la concentración documental se estandarizaron dentro de cada plataforma para que ambos indicadores fueran comparables. Luego, el puntaje documental se obtuvo combinando la diversidad en sentido positivo y la concentración en sentido negativo:

$$S_i = \frac{z(H_{\text{norm}}(\mathbf{q}_i)) - z(HHI(\mathbf{q}_i))}{2},$$

donde $z(\cdot)$ representa el valor estandarizado del indicador correspondiente.

De esta forma, valores mayores de S_i representan una tendencia hacia una mayor diversidad y una menor concentración documental, mientras que valores menores representan una mayor concentración y una menor diversidad.

El puntaje S_i puede tomar valores positivos o negativos y no posee un rango fijo común entre plataformas. Su interpretación se realiza de manera relativa: valores mayores representan una tendencia hacia una mayor diversidad y menor concentración documental, mientras que valores menores representan una mayor concentración y menor diversidad.

Posteriormente, para caracterizar cada perfil se obtiene el promedio de los puntajes de los usuarios pertenecientes al cluster C_k :

$$S_k = \frac{1}{n_k} \sum_{i \in C_k} S_i,$$

donde n_k corresponde al número de usuarios pertenecientes al cluster C_k .

La interpretación de S_k se realiza de manera relativa dentro de cada plataforma. Un valor mayor indica un perfil asociado a una mayor diversidad y una menor concentración documental, mientras que un valor menor indica una mayor concentración y una menor diversidad. Los valores intermedios representan perfiles ubicados entre ambos comportamientos.

6.4.2. Resultados del puntaje documental

La Tabla 6.7 presenta los resultados del puntaje documental para los perfiles obtenidos en AutoCAD, Revit y TimeDoctor. Además del valor promedio S_k , se incluyen la diversidad y concentración documental media de cada cluster y los cuartiles del puntaje individual, con el objetivo de observar cómo se distribuyen los usuarios dentro de cada perfil.

Cuadro 6.7: Resultados del puntaje documental por plataforma y cluster.

Plataforma	Cluster	n_k	Diversidad media	Concentración media	S_k	Q_1	Mediana	Q_3
AutoCAD	0	85	0,4667	0,3429	-0,3498	-0,9383	-0,4232	0,4007
AutoCAD	1	106	0,5603	0,1868	0,4146	-0,0615	0,5610	0,8737
AutoCAD	2	16	0,3950	0,4448	-0,8887	-2,5322	-0,7089	0,0655
Revit	0	45	0,5051	0,2270	0,2559	-0,1869	0,3892	0,6147
Revit	1	33	0,5458	0,1672	0,5079	0,2721	0,5421	0,7373
Revit	2	32	0,3973	0,6260	-0,8835	-1,7505	-0,8107	-0,2215
TimeDoctor	0	118	0,4826	0,1846	0,0916	-0,2361	0,2127	0,5046
TimeDoctor	1	36	0,5328	0,1462	0,4605	0,1475	0,5130	0,7035
TimeDoctor	2	16	0,3197	0,4986	-1,7120	-3,9791	-1,0585	-0,5102

En AutoCAD, el cluster 1 presenta el mayor puntaje documental, con $S_1 = 0,4146$, acompañado de la mayor diversidad media (0,5603) y la menor concentración media (0,1868). En el extremo contrario, el cluster 2 presenta el menor puntaje, $S_2 = -0,8887$, junto con la menor diversidad y la mayor concentración documental. El cluster 0 queda entre ambos, con $S_0 = -0,3498$, por lo que representa un comportamiento documental intermedio.

En Revit se observa el mismo orden. El cluster 1 alcanza el mayor puntaje, $S_1 = 0,5079$, mientras que el cluster 2 presenta el menor, con $S_2 = -0,8835$. El cluster 0 se ubica en una posición intermedia, con $S_0 = 0,2559$. En esta plataforma, además, los cuartiles muestran una separación más marcada entre los clusters 1 y 2, ya que el primer cuartil del cluster 1 es positivo ($Q_1 = 0,2721$), mientras que incluso el tercer cuartil del cluster 2 permanece en valores negativos ($Q_3 = -0,2215$).

En TimeDoctor, el cluster 1 vuelve a presentar el mayor puntaje documental, con $S_1 = 0,4605$, y el cluster 2 el menor, con $S_2 = -1,7120$. El cluster 0 presenta un valor cercano al centro de la escala, con $S_0 = 0,0916$. El cluster 2 muestra además una distribución claramente desplazada hacia valores negativos, con una mediana de $-1,0585$ y un tercer cuartil de $-0,5102$, reflejando una mayor concentración documental entre los usuarios que lo componen.

En las tres plataformas, el puntaje propuesto produce el mismo orden relativo entre los perfiles: el cluster 1 se orienta hacia una mayor diversidad documental, el cluster 0 presenta un comportamiento intermedio y el cluster 2 se orienta hacia una mayor concentración documental. Este resultado permite utilizar S_k como una medida resumida para caracterizar los perfiles según la forma en que distribuyen su tiempo entre los documentos utilizados, manteniendo además S_i como una medida aplicable a cada usuario de forma individual.

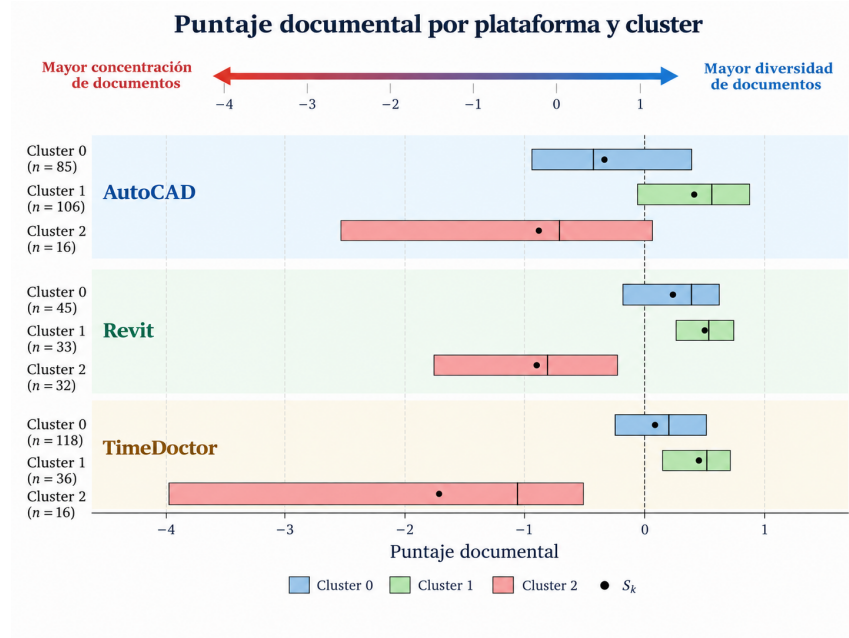


Figura 6.13: Distribución del puntaje documental por cluster en AutoCAD, Revit y TimeDoctor

La figura 6.13 presenta de forma visual la distribución del puntaje documental en cada cluster y plataforma, permitiendo corroborar las conclusiones obtenidas a partir de la Tabla 6.7. En las tres plataformas se observa el mismo orden relativo entre los perfiles: el cluster 1 se ubica hacia valores asociados a una mayor diversidad documental, el cluster 0 ocupa una posición intermedia y el cluster 2 se concentra hacia valores asociados a una mayor concentración documental. De este modo, la representación gráfica confirma que el puntaje documental resume de manera consistente las diferencias entre perfiles.

6.4.3. Distancia respecto del comportamiento general

La segunda métrica propuesta busca representar qué tan alejado se encuentra un usuario, o un perfil, respecto del comportamiento general observado en cada plataforma. Para ello se utilizan los mismos atributos considerados en la construcción de los perfiles generales de actividad mediante GMM.

Sea $\mathbf{x}_i$ el vector estandarizado de atributos del usuario i y sea $\bar{\mathbf{x}}$ el vector promedio de todos los usuarios de la plataforma. Para cada usuario se calcula la distancia euclidiana respecto de este promedio como

$$d_i = \|\mathbf{x}_i - \bar{\mathbf{x}}\|_2.$$

De esta forma, un valor pequeño de d_i indica que el usuario presenta un comportamiento más cercano al comportamiento general de la plataforma, mientras que un valor mayor indica que su combinación de atributos se encuentra más alejada de este.

Para extender esta medida a los perfiles, se calcula primero el vector promedio de los usuarios pertenecientes al cluster C_k :

$$\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i \in C_k} \mathbf{x}_i,$$

donde n_k corresponde al número de usuarios pertenecientes al cluster. Luego, la distancia del perfil respecto del comportamiento general se define como

$$D_k = \|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}\|_2.$$

La interpretación de D_k se realiza dentro de cada plataforma. Valores menores representan perfiles más cercanos al comportamiento general de los usuarios, mientras que valores mayores corresponden a perfiles más diferenciados respecto de este comportamiento. Esta métrica no indica si un perfil es mejor o peor que otro, sino únicamente el grado de separación respecto del comportamiento general observado.

6.4.4. Resultados de la distancia respecto del comportamiento general

La Tabla 6.8 presenta los resultados de la distancia respecto del comportamiento general para los perfiles obtenidos en AutoCAD, Revit y TimeDoctor. Se incluye la distancia D_k del centro de cada cluster respecto del centro global de la plataforma y, de manera complementaria, la distribución de las distancias individuales d_i de los usuarios que pertenecen a cada perfil.

Cuadro 6.8: Distancia respecto del comportamiento general por plataforma y cluster.

Plataforma	Cluster	n_k	D_k	Distancia media	Q_1	Mediana	Q_3
AutoCAD	0	85	0,6937	2,1560	1,5157	1,9774	2,4456
AutoCAD	1	106	0,8359	1,8212	1,3123	1,7126	2,1118
AutoCAD	2	16	3,2847	5,9247	2,0267	5,8213	7,2597
Revit	0	45	1,3765	2,9137	2,0042	2,5378	3,4776
Revit	1	33	0,7967	1,6295	1,3140	1,5606	1,9911
Revit	2	32	2,0477	2,9960	1,9176	2,2175	2,8292
TimeDoctor	0	118	0,2950	2,1319	1,3087	1,8044	2,4962
TimeDoctor	1	36	1,1419	2,4576	1,4142	2,1605	2,9107
TimeDoctor	2	16	3,1878	4,9862	2,0944	4,5181	6,1961

En AutoCAD, el cluster 0 presenta la menor distancia respecto del comportamiento general, con $D_0 = 0,6937$, seguido por el cluster 1 con $D_1 = 0,8359$. Ambos perfiles se encuentran relativamente próximos al centro global de la plataforma. En cambio, el cluster 2 alcanza una distancia considerablemente mayor, con $D_2 = 3,2847$, por lo que corresponde al perfil más diferenciado respecto del comportamiento general.

Las distancias individuales muestran además que los usuarios del cluster 2 presentan una mayor separación respecto del centro global. Su distancia mediana es 5,8213, frente a 1,9774 y 1,7126 en los clusters 0 y 1, respectivamente. Esto muestra que la mayor distancia del cluster 2 no se limita únicamente a la posición de su centro, sino que también se observa entre los usuarios que lo componen.

En Revit, el cluster 1 es el más cercano al comportamiento general, con $D_1 = 0,7967$. El cluster 0 ocupa una posición intermedia, con $D_0 = 1,3765$, mientras que el cluster 2 presenta la mayor distancia, con $D_2 = 2,0477$. Las distancias individuales siguen una tendencia similar, aunque los clusters 0 y 2 presentan valores más próximos entre sí, con medianas de 2,5378 y 2,2175, respectivamente.

En TimeDoctor se observa una separación más clara entre los perfiles. El cluster 0 presenta una distancia muy reducida respecto del centro global, con $D_0 = 0,2950$, mientras que el cluster 1 alcanza $D_1 = 1,1419$. El cluster 2 vuelve a ser el perfil más alejado, con $D_2 = 3,1878$. Esta diferencia también se observa en las distancias individuales, cuya mediana aumenta desde 1,8044 en el cluster 0 hasta 4,5181 en el cluster 2.

En conjunto, la métrica permite ordenar los perfiles según su grado de separación respecto del comportamiento general de cada plataforma. El cluster 2 corresponde al perfil más diferenciado en las tres plataformas, mientras que los perfiles más cercanos al centro global son el cluster 0 en AutoCAD y TimeDoctor, y el cluster 1 en Revit.

A continuación, se muestra la distribución de las distancias individuales de los usuarios para cada cluster y plataforma, con el objetivo de complementar los valores resumidos en la Tabla 6.8.

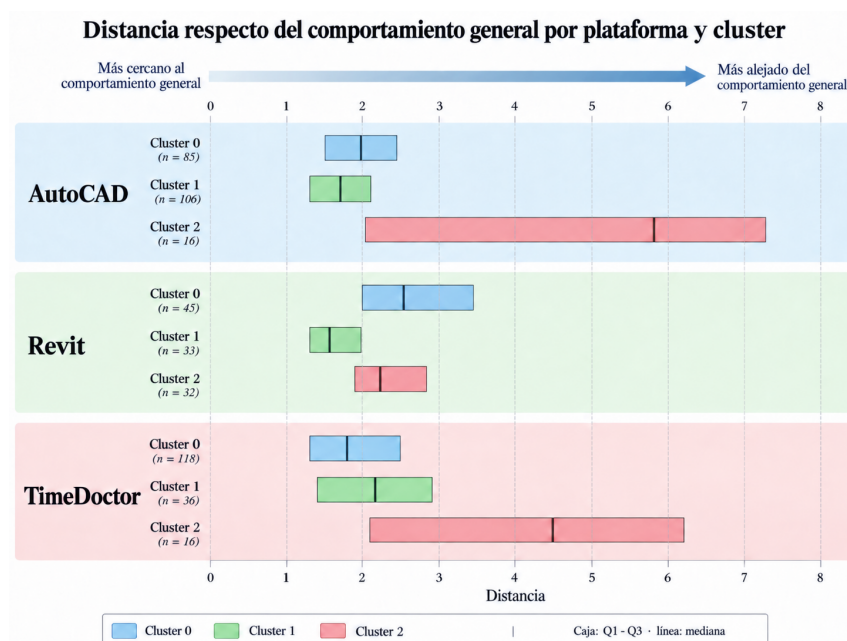


Figura 6.14: Distribución de la distancia por cluster en AutoCAD, Revit y TimeDoctor

La figura 6.14 complementa los resultados de la Tabla 6.8 al mostrar cómo se distribuyen las distancias individuales dentro de cada cluster. En AutoCAD y TimeDoctor se observa que el cluster 2 presenta una distribución desplazada hacia valores más altos y con mayor dispersión. En Revit, en cambio, las distribuciones de los clusters 0 y 2 presentan un mayor grado de superposición. Estas diferencias permiten observar la variabilidad interna de los perfiles, información que complementa la caracterización realizada mediante D_k .

Capítulo 7

Discusión

7.1. Interpretación general de los resultados

Los resultados muestran que es posible identificar diferencias entre los usuarios y representar distintos aspectos de su actividad a partir de los registros disponibles. Sin embargo, no todas las variables construidas se relacionan de la misma forma con los perfiles generales ni presentan el mismo comportamiento en AutoCAD, Revit y TimeDoctor.

La principal diferencia aparece en los indicadores documentales. Tanto la diversidad como la concentración documental presentan una relación consistente con los perfiles generales. Esto se observa al trabajar con sus categorías y también con sus valores continuos. Aunque las diferencias específicas entre clusters cambian de una plataforma a otra, ambos indicadores permiten distinguir perfiles en AutoCAD, Revit y TimeDoctor.

Esta relación también ayuda a entender por qué diversidad y concentración documental aparecen fuertemente asociadas entre sí. Ambas se construyen a partir de la forma en que el tiempo se distribuye entre los documentos utilizados, pero observan esa distribución desde perspectivas diferentes. La diversidad representa qué tan uniforme es la distribución del tiempo, mientras que la concentración representa cuánto de ese tiempo se acumula en determinados documentos.

A partir de este comportamiento se propuso el puntaje documental, que resume ambos indicadores en una sola medida. Esta medida permite ordenar los perfiles según un comportamiento documental más distribuido o más concentrado y entrega una forma más

directa de interpretar una de las principales diferencias encontradas entre los clusters.

La distancia respecto del comportamiento general complementa esta caracterización desde otra perspectiva. En este caso no se observa una dimensión particular del comportamiento, sino qué tan alejada se encuentra la combinación de características de cada perfil respecto del promedio de su plataforma. En las tres plataformas, el cluster 2 aparece como el grupo más diferenciado respecto del comportamiento general.

La concentración horaria entrega una lectura distinta. Sus categorías presentan asociación con el perfil general, pero al analizar directamente los valores del indicador no se observan diferencias significativas entre los clusters. Esto muestra que las categorías permiten describir ciertas diferencias entre usuarios, pero que estas no generan una separación clara entre los perfiles generales.

La fragmentación temporal presenta el comportamiento menos estable entre plataformas. Solo en Revit se observaron diferencias significativas entre los perfiles al trabajar con los valores continuos. Sin embargo, al combinar los resultados de las tres plataformas mediante Fisher se obtiene evidencia conjunta significativa. Esto no significa que la diferencia esté presente individualmente en las tres plataformas, sino que la evidencia acumulada permite rechazar la hipótesis nula al considerar los resultados en conjunto.

En general, los resultados muestran que las variables construidas representan dimensiones diferentes de la actividad. Algunas permiten distinguir claramente los perfiles, mientras que otras aportan principalmente una descripción adicional del comportamiento de los usuarios. Por esta razón, la utilidad de cada indicador depende de la información que se quiera representar y no únicamente de que existan diferencias entre usuarios.

7.2. Discusión de las preguntas e hipótesis planteadas

Para el perfil general de actividad, los resultados respaldan la hipótesis planteada. Las métricas agregadas permitieron identificar grupos de usuarios con características diferentes en AutoCAD, Revit y TimeDoctor.

En la concentración horaria también se observaron diferencias entre los usuarios en la

forma de distribuir su actividad durante el día, por lo que la hipótesis asociada al indicador se encuentra respaldada. Sin embargo, estas diferencias no se reflejan de manera clara en los perfiles generales.

La diversidad y la concentración documental permitieron identificar distintas formas de distribuir el tiempo entre los documentos utilizados. Ambas hipótesis se encuentran respaldadas y, además, estos fueron los indicadores que mostraron la relación más consistente con los perfiles generales en las tres plataformas.

La fragmentación temporal también permitió distinguir distintos niveles de continuidad e interrupción entre los usuarios, respaldando la hipótesis planteada. Su relación con los perfiles, en cambio, fue menos uniforme y dependió de la plataforma analizada.

En síntesis, las cinco variables permitieron responder las preguntas planteadas. Los cuatro indicadores distinguen características diferentes entre los usuarios, pero no todos presentan la misma capacidad para diferenciar los perfiles generales.

7.3. Limitaciones del estudio

Una primera limitación corresponde al tipo de información disponible. El análisis se basa únicamente en los registros generados por AutoCAD, Revit y TimeDoctor, por lo que estos datos representan la actividad registrada por las plataformas y no necesariamente todas las acciones realizadas por los usuarios durante su trabajo.

Tampoco se incorporaron variables de contexto como el cargo, el tipo de proyecto o la función desempeñada. Esto permite identificar diferencias entre usuarios y perfiles, pero no determinar directamente qué factores explican esas diferencias.

Los períodos de monitoreo tampoco son equivalentes. AutoCAD y Revit cuentan con períodos de observación considerablemente más extensos que TimeDoctor. Por esta razón, la comparación entre plataformas se centra en los patrones y distribuciones observadas y no en comparar directamente la magnitud total de la actividad registrada.

Además, los perfiles generales fueron obtenidos de manera independiente en cada pla-

taforma. Sus etiquetas fueron homologadas posteriormente según las características observadas en cada grupo, lo que permite compararlos bajo una misma numeración, pero no significa que provengan de un único modelo estimado sobre las tres plataformas.

Finalmente, los resultados corresponden al período y a la organización estudiada. Su uso en otros contextos requeriría volver a evaluar el comportamiento de los indicadores y de los perfiles obtenidos.

7.4. Uso responsable de los indicadores

Los indicadores contruidos buscan describir patrones de uso observados en AutoCAD, Revit y TimeDoctor. No fueron diseñados para medir productividad, desempeño individual ni calidad del trabajo realizado.

Por esta razón, sus resultados deben interpretarse de manera agregada y considerando el contexto en que fueron generados. Los perfiles representan distintas formas de uso de las plataformas y no categorías de mejor o peor desempeño laboral.

Capítulo 8

Conclusiones

8.1. Conclusiones generales

A partir de los registros de AutoCAD, Revit y TimeDoctor fue posible construir variables que representan distintas dimensiones de la actividad de los usuarios y utilizarlas para identificar y caracterizar patrones de uso.

De esta forma, la pregunta de investigación puede responderse afirmativamente. Los registros originales, que por sí solos representan eventos individuales, pudieron transformarse en información agregada que permite comparar usuarios, identificar perfiles y estudiar qué características los diferencian.

Un resultado importante del trabajo es que no todas las variables construidas aportan la misma información. Algunas permiten distinguir con mayor claridad los perfiles generales, mientras que otras describen dimensiones del comportamiento que varían entre usuarios sin necesariamente separar esos perfiles.

8.2. Principales hallazgos

El primer hallazgo fue que las métricas agregadas permiten identificar perfiles generales de actividad. Para representarlos se seleccionó GMM, ya que permitió obtener grupos con una distribución de usuarios más adecuada para su interpretación que la obtenida mediante K-Means.

Entre los cuatro indicadores, la diversidad y la concentración documental fueron los que mostraron el comportamiento más consistente respecto de los perfiles generales en las tres plataformas. Esto indica que la forma en que el tiempo se distribuye entre los documentos utilizados constituye una de las principales dimensiones para diferenciar los perfiles identificados.

La concentración horaria y la fragmentación temporal también permitieron describir diferencias entre usuarios, pero su relación con los perfiles fue menos consistente. Esto refuerza la idea de que un indicador puede aportar información sobre una dimensión de la actividad sin necesariamente diferenciar los perfiles generales.

8.3. Aportes del trabajo

El principal aporte de este trabajo es proponer una estructura para transformar registros operacionales en variables interpretables a nivel de usuario. Esta estructura considera la preparación de los datos, la construcción de variables, el análisis de sus distribuciones, la identificación de perfiles y el estudio de las relaciones entre las distintas dimensiones representadas.

Otro aporte es que las variables no se consideraron útiles únicamente por poder ser calculadas. Su comportamiento fue evaluado primero entre los usuarios y posteriormente en relación con los perfiles generales y con las demás variables. Esto permite distinguir qué información aporta realmente cada indicador.

Finalmente, se propusieron dos medidas complementarias para interpretar los perfiles. El puntaje documental resume la relación entre diversidad y concentración documental, mientras que la distancia respecto del comportamiento general permite representar qué tan diferente es un usuario o grupo respecto del promedio de su plataforma.

8.4. Trabajo futuro

Como continuación de este trabajo, se podrían incorporar nuevas variables que permitan representar otras dimensiones de la actividad y evaluar si los patrones observados se mantienen en nuevos períodos de información.

La misma estructura de análisis también podría aplicarse a otras fuentes de datos de la organización. Un ejemplo es el análisis de horas hombre por proyecto, donde la construcción de métricas y la identificación de grupos podrían utilizarse para comparar proyectos con características similares y apoyar futuras estimaciones.

Finalmente, el análisis podría repetirse periódicamente a medida que se generen nuevos registros. Esto permitiría observar si los perfiles identificados y las relaciones entre los indicadores se mantienen o cambian con el tiempo.

Bibliografía

- [1] Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2), 118–144.
- [2] Laudon, K. C., & Laudon, J. P. (2020). *Management Information Systems: Managing the Digital Firm* (16th ed.). Pearson.
- [3] van der Aalst, W. (2016). *Process Mining: Data Science in Action*. Springer.
- [4] Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.
- [5] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.
- [6] Autodesk. (2026). *AutoCAD: software CAD 2D y 3D*. Disponible en: <https://www.autodesk.com/es/products/autocad/overview>
- [7] Autodesk. (2026). *Revit: software BIM para diseño y modelamiento*. Disponible en: <https://www.autodesk.com/es/products/revit/overview>
- [8] Time Doctor. (2026). *Workforce analytics for distributed teams*. Disponible en: <https://www.timedoctor.com/>
- [9] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- [10] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- [11] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer.

-
- [12] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1–38.
- [13] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- [14] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423; 27(4), 623–656.
- [15] Hirschman, A. O. (1945). *National Power and the Structure of Foreign Trade*. University of California Press.
- [16] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
- [17] Agresti, A. (2019). *An Introduction to Categorical Data Analysis* (3rd ed.). Wiley.
- [18] Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.
- [19] Heard, N. A. y Rubin-Delanchy, P. (2018). Choosing between methods of combining p -values. *Biometrika*, 105(1), 239–246.