Universidad Católica de Santa María

Facultad de Ciencias e Ingenierías Físicas y Formales
Escuela Profesional de Ingeniería de Sistemas

"Minería de Procesos: Descubrimiento de Concept Drift en la Perspectiva de Usuarios"

Sistemas de Información: Minería de Procesos

Tesis presentada por los Bachilleres:

- Fernández Mendoza, Martha Lucía
- González Córdova, Mónica Lucía

Para obtener el Título Profesional de

INGENIERO DE SISTEMAS

Arequipa – Perú
2015

PRESENTACIÓN

El siguiente trabajo tiene el propósito de Descubrir el *Concept Drift* en la Minería de Procesos desde otra perspectiva de cambio, ya que las propuestas anteriores se enfocan en la perspectiva de flujo, acá se proponen características, calculadas mediante técnicas de sociometría, para encontrar el *Concept Drift* desde la perspectiva de usuarios.

AGRADECIMIENTOS

Agradecemos a nuestros formadores, a nuestras familias y amigos que nos han ayudado a llegar hasta el punto que nos encontramos.

Gracias por todo el proceso, nada sencillo, de transmitirnos conocimientos y la orientación que nos seguirán por el resto de nuestra vida profesional.

Dedicamos este trabajo a todas las personas que contribuyeron en él y a aquellas personas que nos brindaron su apoyo incondicional en todo momento, para poder *log*rar este objetivo, especialmente a nuestros padres, porque sin su apoyo, ánimos y paciencia no hubiéramos llegado a

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Las técnicas de minería de procesos permiten extraer información de *logs* de eventos, que algunas empresas, que utilizan sistemas de información, registran al ejecutar sus procesos de negocio.

Dentro del análisis de estos *logs* de eventos es posible encontrar cambios súbitos en el comportamiento normal de los procesos ejecutados. A estos cambios se les conoce como *Concept Drift*. Esta localización de cambios puede aplicarse a varías perspectivas de comportamiento como, usuarios, recursos, tiempo, o flujo.

El actual trabajo se centra en descubrir el *Concept Drift* desde la perspectiva de usuarios al aplicar las técnicas de sociometría para analizar los *log* de eventos. Se *log*raron encontrar cambios en esta perspectiva utilizando *logs* sintéticos, comparando su comportamiento mediante gráficos de valores acumulados.

Palabras Clave: Minería de Procesos, Log de Eventos, Concept Drift, Sociometría, Usuarios.

ABSTRACT

Process Mining techniques allow to extract information from event logs, some companies that use information systems, record this log events when their business processes are running.

In the analysis of these logs of events is posible to find sudden changes in the normal behavior of the processes executed. These changes are known as Concept Drift. This change discover can be applied to various perspectives of behavior such as users, resources, time, or flow.

The current work is focused on discovering Concept Drift from the perspective of users by applying sociometric techniques to analyze event logs. In the experiment, Concept Drifts were found in this new perspective using synthetic logs and contrasting behaviour by plotting acummulated data.

Key words: Process Mining, Event Logs, Concept Drift, Sociometric, Users.

INTRODUCCIÓN

El propósito de la Minería de Procesos es descubrir, monitorear y mejorar procesos reales al extraer conocimiento de *logs* de eventos como fuente de información, algunos de los cuales se encuentran disponibles en los sistemas de información que utilizan algunas empresas.

Debido a la gran cantidad de información que puede ser extraída de estos *logs* de eventos es que surge la necesidad de crear nuevas técnicas necesarias para extraer información importante.

Estas técnicas de análisis de información sobre los *logs* de eventos, complementan los enfoques existentes de BPM (*Business Process Management*). El ciclo de vida de BPM describe las diferentes fases del manejo de procesos de negocios en particular.

La técnica de *Concept Drift* se refiere a la detección de cambios en un proceso de negocio, es decir si se descubre una anomalía en el comportamiento del proceso en una ejecución es que ha ocurrido un cambio. Esta técnica es aplicada usualmente para descubrir cambios en el flujo del proceso, por esto el enfoque de este trabajo es descubrir estos cambios desde la perspectiva de usuarios.

La estructura de este trabajo de investigación es como sigue, en el capítulo I explicaremos el problema en el *Concept Drift*, la solución propuesta y los objetivos; en el capítulo II hablaremos sobre los fundamentos teóricos y el estado del arte; en el capítulo III se detallaran los alcances y limitaciones, los métodos de investigación y el plan de análisis de datos; en el capítulo IV detallaremos el modelo propuesto; en el

capítulo V analizaremos los resultados obtenidos mediante casos de estudio; terminamos con las conclusiones, recomendaciones y trabajos futuros.

CAPÍTULO I: PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

La idea de Minería de Procesos es descubrir, monitorear y mejorar procesos reales al extraer conocimiento de *logs* de eventos como fuente de información, algunos de los cuales se encuentran disponibles en los sistemas de información (van der Aalst, y otros, 2012)

Concept Drift es una técnica se encarga del análisis de los cambios que un proceso puede experimentar mientras está siendo analizado, por ejemplo, las influencias de temporadas y su periodicidad poseen un comportamiento distinto, o los cambios en un proceso que puede producir una nueva legislación. Entender estos cambios es de suma importancia para la gestión de procesos.

En la mayor parte de propuestas para solucionar el problema de *Concept Drift* se utilizan procesos bien definidos, aislados y estáticos. Sin embargo, los procesos son fluidos y se relacionan entre sí; haciendo necesarias técnicas para obtener resultados útiles y reales. Estas técnicas sólo se enfocan en localizar el cambio pero no sustentan el motivo del mismo.

Además que, en la mayoría de propuestas, solamente se encuentra el *Concept Drift* analizando los procesos según una única perspectiva de cambio, la cual se centra en el flujo de actividades. Sin embargo sería importante poder encontrar este cambio en otras perspectivas, como lo son tiempo, usuarios, recursos, etc. (van der Aalst W. M., 2011)

El objetivo de nuestra investigación es detectar el *Concept Drift* en la perspectiva de usuarios con la finalidad de proponer una nueva perspectiva para el análisis de *Concept Drift* en la minería de procesos.

1.2 Objetivos de la Investigación

1.2.1 General

Proponer características que permitan descubrir el *Concept Drift* en la perspectiva de usuarios.

1.2.2 Específicos

- Identificar la perspectiva de los puntos de cambio utilizando las características planteadas.
- Implementación del algoritmo como *plug-in* no publicado en ProM.
- Simulación del *plug-in* con *logs* sintéticos.

1.3 Preguntas de Investigación

- ¿Es posible encontrar un *Concept Drift* en la perspectiva de usuarios utilizando características de sociometría utilizando las pruebas de hipótesis en ventanas?
- ¿Un cambio en el comportamiento de los usuarios puede afectar el flujo de un proceso?
- ¿Existe una mejor granularidad al aplicar un tamaño de grupo de casos para analizar el *log*?
- ¿Cuál es la prueba de hipótesis más adecuada para encontrar el *Concept Drift* para las características de sociometría?
- ¿Existe una relación en los resultados de la perspectiva de flujo y los resultados de la perspectiva de usuarios?

1.4 Línea y Sub-línea de Investigación a la que corresponde el Problema

Sistemas de Información y Bases de Datos

- Minería de Procesos

1.5 Palabras Claves

Concept Drift, usuarios, sociometría, test de hipótesis

1.6 Solución Propuesta

1.6.1 Justificación

El *Concept Drift* es una línea de investigación emergente, extensa y relativamente novedosa en la minería de procesos, que aún no ha recibido la atención suficiente. Es considerado un reto importante en el "Process Minning Manifesto" (van der Aalst, y otros, 2012), un documento sobre los avances científicos con respecto al área.

Muchas soluciones han sido propuestas para poder detectar correctamente el *Concept Drift* en los procesos, a estas propuestas les hace falta entornos visuales en los que la información generada por estos algoritmos pueda ser explotada correctamente. Las propuestas y algoritmos desarrollados actualmente se enfocan únicamente en la perspectiva de flujo, la relación precede-antecede entre los procesos, dejando de lado otras perspectivas importantes como las de usuarios, recursos y datos que pueden ser importantes para explicar los cambios.

El poder detectar el *Concept Drift* en otras perspectivas ayudaría mucho para el análisis de los procesos por parte de las empresas. Así podrían actuar correctamente ante diferentes problemas y poder planificar mejor sus

actividades, por ejemplo, si se tiene conocimiento sobre el cambio inusual en el desarrollo de actividades por parte de un usuario en una etapa del proceso, se podría identificar el porqué de este cambio y tomar las medidas correspondientes para poder solucionarlo o quizás mejorar el proceso.

Es por esto que proponemos un algoritmo que sea capaz de identificar *Concept Drift* en la perspectiva de usuarios mediante características que consideramos relevantes, además de mostrar gráficas y tablas capaces de mostrar la información adecuadamente.

1.6.2 Descripción de la Solución

Se propone un algoritmo capaz de identificar el *Concept Drift* enfocado a las perspectivas de usuarios, adaptando el algoritmo propuesto por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014). Para poder identificar el *Concept Drift* en esta perspectiva se van a considerar tres características que consideramos importantes como lo son: La distribución de carga laboral entre los usuarios, la entropía usuario-actividad y la participación del usuario en el proceso.

El algoritmo tendrá como datos de entrada *logs* finitos y de datos sintéticos, estos *logs* serán procesados de manera *off-line*, ya que el algoritmo a adaptar los procesa de manera *off-line* y procesarlos de manera *on-line* sale del enfoque de esta propuesta.

1.7 ALCANCES Y LIMITACIONES

Alcances

- Detectar el *Concept Drift* en la perspectiva de cambio de usuarios.

- Se desarrollará una gráfica donde se mostraran los resultados correspondientes.
- Se modificará la interfaz del *plug-in* para que el usuario pueda ingresar y visualizar la nueva perspectiva de cambio.
- Se procesarán los datos de forma *off-line*.

Limitaciones

- Las limitaciones de resultados serán acordes a las limitaciones de los clasificadores elegidos.
- Se utilizará únicamente *logs* de datos sintéticos.
- Se trabajará únicamente con los tipos de *Concept Drift* que el algoritmo de referencia es capaz de encontrar.
- Se depende de las limitaciones del *plug-in* implementado en ProM.

1.8 APORTE

En la investigación se busca añadir al *plug-in* de ProM, implementado por (R.P. Jagadeesh Chandra Bose, 2010), una nueva perspectiva de cambio a nivel de usuarios mediante técnicas de sociometría.

1.9 TIPO Y NIVEL DE LA INVESTIGACIÓN

Tipo: Básica

Nivel: Exploratorio

1.10 POBLACIÓN Y MUESTRA O UNIVERSO

Utilización de dos *logs* sintéticos para demostrar la aplicación de la técnica.

1.11 Métodos, Técnicas e Instrumentos de Recolección de Datos

Los valores obtenidos después de aplicar el algoritmo en un *log* en las gráficas y tablas de salida de ambos algoritmos.

1.12 Plan de Análisis de los Datos

Comparación de los resultados y variaciones de las gráficas con el contenido de los *logs* para explicar la variación de ambos.

- Tamaño de Grupo: Comparación de resultados de las gráficas con diferentes tamaños de grupos.
- Tiempo de ejecución por característica: Tiempo de análisis de un mismo *log* con cada una de las características.
- Efectividad del Algoritmo: Cantidad de puntos de *Concept Drift* encontrados contra los esperados.

CAPÍTULO II: FUNDAMENTOS TEÓRICOS

2.1 ESTADO DEL ARTE (ANTECEDENTES DE LA INVESTIGACIÓN)

Las técnicas de minería de procesos son capaces de extraer información de los *logs* de eventos disponibles en los sistemas de información. Estas técnicas proveen nuevas formas de descubrir, monitorear y mejorar procesos. Existen dos principales enfoques para el creciente interés en la minería de procesos. Por un lado, más eventos son grabados en consecuencia se añade información detallada acerca de la historia de los procesos. Mientras que por otro lado, existe una necesidad de mejorar y apoyar los procesos de negocio en un ambiente competitivo.

Ante esta problemática los miembros y partidarios del grupo de trabajo de Minería de Procesos de la IEEE publicaron el "Process Minning Manifesto" (ProM, 2012). El cual es una "declaración pública de los principios e intenciones" para promover la investigación, el desarrollo, implementación, evolución y entendimiento de la minería de procesos.

En (van der Aalst, y otros, 2012), se presentan los principales retos de esta metodo*logía*; el cuarto reto hace referencia al *Concept Drift*. El término de *Concept Drift* es la situación en la cual un proceso cambia cuando está siendo analizado. Por ejemplo, en el inicio de un *log* de eventos existen dos actividades que en principio son concurrentes y luego pueden ser secuenciales. Estos cambios impactan en los procesos y es de vital importancia detectarlos y analizarlos. (Carmona & Gavaldà, 2012) Indica que el *Concept Drift* es una problemática importante para cualquier escenario de análisis que envuelva

datos ordenados temporalmente. En la última década se han realizado muy pocos trabajos que tratan con el reto de *Concept Drift* en minería de procesos.

Empezamos nuestra investigación informándonos acerca de las técnicas utilizadas hasta hoy en día para localizar el *Concept Drift* en minería de procesos. En (Maggi., 2013) se basa en el algoritmo de Bose (R.P. Jagadeesh Chandra Bose, 2010) en el cual, su algoritmo base, toma un conjunto de instancias calculando las características correspondientes, luego, se realiza una prueba de hipótesis que mostrará la diferencia entre ambos conjuntos. En la propuesta de (Maggi., 2013) se presentan cinco enfoques para detectar el *Concept Drift off-line*. En el primer enfoque, el tamaño del paso o la distancia entre las ventanas que se analizan, permite mejorar la velocidad de los algoritmos obviando pasos intermedios. Este tamaño debe ser tanteado de acuerdo al tiempo y la granularidad necesaria. En el segundo enfoque, algoritmo de "Detección de Punto de Cambio", se calcula el valor p, luego itera haciendo particiones a la población hasta ubicar el segmento donde ocurrió el *Concept Drift*; este algoritmo permite extraer los puntos de *Concept Drift* sin la necesidad de analizar el caso manualmente. En el tercer enfoque, algoritmo de "Ventanas Adaptativas" o ADWIN, se utiliza al primer algoritmo agregando un tamaño máximo y mínimo de ventana para encontrar el mejor valor de p; relaja la dependencia del primer algoritmo en el tamaño de la población fijo y reduce la cantidad de falsos positivos y falsos negativos. En el cuarto enfoque, algoritmo de "Captura de Variaciones Graduales con Poblaciones Continuas", se enfoca en los *Drift*s graduales; se utiliza una brecha entre ambas poblaciones, que disminuye el tamaño constantemente, seguido se hacen pruebas de hipótesis hasta que uno de los nuevos valores es mayor al antiguo. Finalmente, se presenta un último

algoritmo “Captura Dinámica de Múltiples Pedidos utilizando Períodos de Tiempo”, en esta variante la población mínima de ADWIN se convierte en el período mínimo, el tamaño máximo de población se convierte en el máximo período y el tamaño de paso se convierte el período de paso; la definición del tamaño de la población como períodos de tiempo en vez de cantidad de casos permite detectar un *Concept Drift* en nivel micro o macro en *logs* con dinámica en multi-orden (*multi-order dynamics*), donde los cambios de procesos pueden ser en distintos niveles de granularidad. Las pruebas se realizaron con la implementación de los algoritmos en *plug-in* de *Concept Drift* en ProM, un *framework* para minería de procesos, con data sintética (Registro de demanda del seguro médico de Bose), y data real. Se comparó los resultados ante diversos tipos de *Concept Drift*.

(Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014) Se basa en el trabajo anterior de Bose (R.P. Jagadeesh Chandra Bose, 2010) al igual que la investigación anterior. En esta propuesta se desarrolla un *framework* que utiliza “Pruebas de Hipótesis para la Detección de *Drift*” (*Hypothesis Tests for Drift Detection*) como las siguientes pruebas de hipótesis, Kolmogorov-Smirnov, Mann-Whitney U y Hottelling T2. Estas pruebas se utilizan para evaluar y comparar grupos o poblaciones de datos y así poder identificar el *Concept Drift*. Las principales desventajas de esta propuesta son las siguientes: las características definidas no son capaces de hallar todas las clases de *Concept Drift*, ya que son muy generales y es necesario el uso de características más especializadas. Por otro lado, cuando se evalúa un *log*, no necesariamente se va a hacer uso de las mismas características que se utilizaron para evaluar otros *logs*, es por esto que se necesita identificar qué características son necesarias

para cada *log* y así mejorar la complejidad computacional. En cuanto al *Concept Drift* recurrente, esta propuesta no se ha enfocado a este tipo de *Concept Drift*, sino solamente, al *Concept Drift* repentino y gradual. Las pruebas utilizadas en (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014) se realizaron con data real y sintética, a través de modificaciones en el *plug-in* de *Concept Drift* en ProM. Es importante recalcar que se compararon las pruebas de hipótesis utilizando diversos algoritmos, comparando rendimiento en el tiempo y el tamaño de la muestra.

Seguidamente, se revisó el trabajo de (Fei, Liquin, Guang Yun, & Xiaolei, 2013) donde se muestra otra mejora para el algoritmo de Bose (R.P. Jagadeesh Chandra Bose, 2010) con respecto al tiempo. Para *log*rarlo utiliza una rápida expansión y reducción del tamaño de la muestra para detectar el *Concept Drift*. Utiliza la prueba de hipótesis estadística de Kolmogorov-Smirnov. En sus pruebas se usaron cuatro modelos de procesos y se demostró que la complejidad de tiempo se redujo en un 10%.

Los trabajos mencionados anteriormente están enfocados a hallar el *Concept Drift* con *logs* finitos y procesos aislados, son propuestas que analizan datos de manera fuera de línea u *off-line*, sin embargo, surge la necesidad de poder encontrar el *Concept Drift* al momento en que ocurre, ya que permite un mejor análisis de los procesos de negocio y acerca la aplicación a la realidad. Es posible adaptar estos trabajos a análisis en tiempo real, ya que consideran las estampas de tiempo, sin embargo, existen otras propuestas dirigidas a encontrar el *Concept Drift on-line*, las propuestas relacionadas las veremos a continuación.

Los autores (Carmona & Gavaldà, 2012) presentan una técnica *on-line* para detectar el *Concept Drift* mediante el monitoreo secuencial de *logs* de un

sistema. Es una técnica multi-fase que utiliza la teoría de la interpretación abstracta, los autores se enfocan en el dominio abstracto de poliedros convexos, en el cual se genera una población de puntos para el aprendizaje mediante un vector, utilizando los primeros casos del *log*, con esta información se arma el poliedro. Seguidamente mediante la técnica de ventana adaptativa se van tomando los casos siguientes también como una colección de puntos; si un punto cae en el área del poliedro significa q no se ha detectado un cambio pero, si el punto cae fuera del área, entonces se ha encontrado un *Concept Drift*. Este algoritmo reacciona a los cambios en una forma *on-line*, casi en tiempo real. Sin embargo una de las desventajas de esta propuesta es el tamaño del vector para armar el poliedro, el cual puede aumentar su tamaño al analizar *logs* más amplios, lo cual genera más complejidad en tiempo y espacio computacional. Además existe la posibilidad de que ciertos puntos sean obviados del vector, ya sea porque se analizaron anteriormente o no se computaron en la etapa de aprendizaje, lo cual haría que ciertos tipos de *Concept Drift* no puedan ser detectados.

(Lin Feng, 2013) Presenta un *framework* que utiliza un estrategia de piscinas de clasificadores y el método de divergencia KL, que mide la similitud entre el bloque actual de datos y los bloques de las piscinas. Este *framework* mantiene el control de la cantidad de clasificadores a través de una estrategia de olvido. Se realizaron pruebas con data sintética y real. El algoritmo fue comparado con otros modelos de clasificación Knn y un árbol de decisión, demostrando una precisión mayor en un 15%. Una de sus ventajas es el manejo del *Drift* gradual y como trabajo a futuro se plantea concentrarse en la clasificación de flujo.

En (Sujatha & Anil Kumar, 2013), se presenta otra forma de lidiar con el *Concept Drift* utilizando estampas de tiempo, para *log*rarlo se divide el enfoque en cuatro algoritmos. El primer algoritmo se encarga de la clasificación en línea, en caso la instancia no se haya clasificado. El segundo algoritmo, buscará la clasificación para la instancia en la memoria secundaria. El tercer algoritmo, creará un clasificador si la instancia no ha sido clasificada en los clasificadores de la memoria secundaria. Finalmente un cuarto algoritmo se encarga de liberar la memoria principal. En las simulaciones de los algoritmos, se muestra que el tiempo para la clasificación aumenta linealmente a medida que se aumentan los registros. En las pruebas se compara con un árbol de clasificación y clasificadores *Naive Bayes.*

Identificamos que los enfoques para el *Concept Drift on-line* son pocos y sólo se centran en la perspectiva de flujo. Sin embargo, al igual en el *Concept Drift off-line*, las características y métodos utilizados no *log*ran encontrar todos los casos de *Concept Drift*, y muy pocos llegan a un nivel de granularidad más fino. Otro problema encontrado en ambos tipos de propuestas es que el *Concept Drift* sólo se señala, más no se explica la razón el cambio, esto es algo que sería de gran importancia.

En (Cheng-Jung Tsai a, 2007) se presenta un enfoque novedoso donde se minan las reglas o razones del *Concept Drift* mediante árboles de decisión. Esta propuesta utiliza como entrada un *stream* de *Concept Drift*, hallado con anterioridad mediante algún método, el cual los autores no detallan. Seguidamente se arma un árbol CDR, el cual devolverá las ramas del árbol con mayor apoyo y la confianza. En esta investigación se hicieron pruebas sólo con dos bloques de datos en un *data stream*, si se utilizaría un número mayor de

datos el árbol CDR se volvería más complejo y crecería en cuanto a tamaño, esto produciría que la complejidad creciera exponencialmente. Una mejora planteada para este trabajo es el uso de algoritmos de discretización.

(van der Aalst W. M., 2013) Propone la noción de *Process Cube* donde los eventos y modelos de procesos se encuentran organizados utilizando diferentes dimensiones, ya que existe la necesidad de visualizar la información acerca de los eventos desde diferentes ángulos. En contraste con la mayoría de propuestas que se enfocan en procesos en solitario y técnicas *off-line* esta propuesta se enfoca en múltiples procesos interrelacionados que pueden cambiar a través del tiempo. El autor hace referencia a los cubos de datos OLAP (*Online Analytical Processing*), definiendo las nociones de OLAP tales como *slicing*, *dicing*, *rolling up* y *drilling down* para manejar los datos sobre eventos, siendo utilizadas para comparar, combinar y dividir celdas de proceso en nivel de *log* y nivel de modelo. Las nociones de *Process Cube* están cercanamente relacionadas a los alcances de divide y vencerás en la minería de procesos, donde grandes *logs* pueden dividirse en *sub-logs* para mejorar el desempeño y la escalabilidad.

Uno de los desafíos que proponen (van der Aalst W. M., 2013) en cuanto a *Concept Drift* es acerca del tiempo, ya que en la dimensión de ventana de tiempo se puede asociar los casos completos si son *short-running cases*, en cambio en el caso de *long-running cases* se debe asociar eventos individuales a las ventanas de tiempo, ya que el proceso puede variar mientras la instancia está corriendo. Utilizando vectores de características escogidos cuidadosamente se podría analizar los *Drift*s usando ventanas adaptativas de tiempo.

La propuesta de (van der Aalst W. M., 2013), ha sido implementada para poder ser aplicada en varios tipos de análisis sobre minería de procesos. Esta propuesta

es muy atractiva para solucionar los diferentes problemas que hemos encontrado sobre *Concept Drift*, ya que el uso de cubos nos permite analizar de una forma más detallada los *logs*, haciendo posible llegar a un nivel de granularidad más fino, y se muestra la posibilidad de hallar el *Concept Drift* de acuerdo a las diferentes perspectivas de cambio relevantes para el negocio.

Proponemos un algoritmo capaz de encontrar el *Concept Drift* en diferentes perspectivas de cambio, de una manera *on-line*, para lo cual utilizaremos la propuesta de (R.P. Jagadeesh Chandra Bose, 2010), ya que nos permitiría obtener *sub-logs* más especializados, ya sea en un nivel de detalle más fino o en diferentes perspectivas de cambio. Además que sería posible explicar los cambios en los *logs* utilizados mediante *footprints* o árboles CDR.

El trabajo planteado resume los trabajos anteriores dando un enfoque más práctico y útil para la aplicación del *Concept Drift* en la minería de procesos. Al mismo tiempo debido a la amplitud del mismo, se presentaría en una fase experimental con datos sintéticos y la efectividad del mismo depende en gran medida de los algoritmos utilizados.

2.2 BASES TEÓRICAS

2.2.1 Minería de Procesos (*Process Minning*)

Según (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014), la minería de procesos o *Process Minning* es una disciplina de investigación relativamente joven que se encuentra por un lado, entre el aprendizaje de máquina y la minería de datos y el modelado de procesos y análisis por otro. Existen tres tipos de minería de procesos:

Descubrimiento: Toma un *log* de eventos y produce un modelo sin usar ninguna información a priori.

Conformidad: Un modelo de procesos existente es comparado con un *log* de eventos del mismo proceso. La revisión de conformidad es usada para verificar si la realidad grabada en el *log*, es conforme al modelo y vice versa.

Mejora: Consiste en extender y mejorar un modelo de proceso existente, usando la información acerca del proceso actual obtenido de un *log* de eventos.

En la *Tabla 1*, se muestra un fragmento de un *log* de eventos. Donde se ilustra su información típica.

Tabla 1 Ejemplo de un Log de Eventos

Id Caso	Id Evento	Propiedades				
		Marca de tiempo	Actividad	Recurso	Costo	...
1	35654423	30-12-2010:11.02	Registrar requerimiento	Pete	50	...
	35654424	31-12-2010:10.06	Examinar a fondo	Sue	400	...
	35654425	05-01-2011:15.12	Revisar ticket	Mike	100	...
	35654426	06-01-2011:11.18	Decidir	Sara	200	...
	35654427	07-01-2011:14.24	Rechazar requerimiento	Pete	200	...
2	35654483	30-12-2010:11.32	Registrar requerimiento	Mike	50	...
	35654485	30-12-2010:12.12	Revisar ticket	Mike	100	...
	35654487	30-12-2010:14.16	Examinar casualmente	Pete	400	...
	35654488	05-01-2011:11.22	Decidir	Sara	200	...
	35654489	08-01-2011:12.05	Pagar compensación	Ellen	200	...
3	35654421	30-12-2010:14.32	Registrar requerimiento	Pete	50	...
	35654422	30-12-2010:15.06	Examinar casualmente	Mike	400	...
	35654424	30-12-2010:16.34	Revisar ticket	Ellen	100	...
	35654425	06-01-2011:09.18	Decidir	Sara	200	...
	35654426	06-01-2011:12.18	Reinicializar requerimiento	Sara	200	...
	35654427	06-01-2011:13.06	Examinar a fondo	Sean	400	...
	35654430	08-01-2011:11.43	Revisar ticket	Pete	100	...
	35654431	09-01-2011:09.55	Decidir	Sara	200	...
	35654433	15-01-2011:10.45	Pagar compensación	Ellen	200	...
4	35654441	06-01-2011:15.02	Registrar requerimiento	Pete	50	...
	35654443	07-01-2011:12.06	Revisar ticket	Mike	100	...
	35654444	08-01-2011:14.43	Examinar a fondo	Sean	400	...
	35654445	09-01-2011:12.02	Decidir	Sara	200	...
	35654447	12-01-2011:15.44	Rechazar requerimiento	Ellen	200	...
...	...	...	...	...	...	...

Fuente: Process Minning. Discovery, Conformance and Enhancement of Business Processes

La *Tabla 1* muestra un *log* relacionado al manejo de requerimientos para una compensación, consideraremos cada una de las líneas de registro como un

evento. Usando esta imagen, podemos listar nuestros supuestos acerca de los *logs* de eventos.

- Un proceso consiste de casos.
- Un caso consiste de eventos, de forma tal que cada evento se relaciona a un sólo caso precisamente.
- Los eventos en un caso están ordenados.
- Los eventos pueden tener atributos. Por ejemplo, actividad, tiempo, costo y recursos.

2.2.2 *Concept Drift* en Minería de Procesos

El *Concept Drift*, en el aprendizaje de máquinas y en la minería de datos, se refiere a las situaciones de cambio entre los datos de entrada con respecto la variable objetivo. La cual se intenta predecir mediante un modelo, esta variable posee cambios a través del tiempo de forma imprevista. A comparación de la minería de datos y el aprendizaje automático donde el *Concept Drift* se enfoca en cambios simples como variables, en la minería de procesos trabaja con cambios en estructuras más complejas como modelos de procesos describiendo concurrencia, elecciones, bucles y cancelaciones.

Perspectivas de Cambio en *Concept Drift*

- *Perspectiva de Control de Flujo/Comportamiento:* Se refiere al comportamiento y la estructura en un modelo de procesos. Los cambios del control de flujos pueden ser clasificados en operaciones como inserción, borrado, substitución y la reordenación de los fragmentos del proceso.

- *Perspectiva de datos:* Se refiere a los cambios en la producción y consumo de datos y el efecto de los datos.

- *Perspectiva de recursos:* Se refiere a los cambios en los recursos, sus roles, la estructura organizacional y su influencia en la ejecución de un proceso.

Naturaleza de los *Drift*s

Se identifican cuatro clases de cambios:

- *Sudden Drift:* Corresponde a la substitución de un proceso M1 con un nuevo proceso M2. M1 deja de existir desde el momento de la sustitución

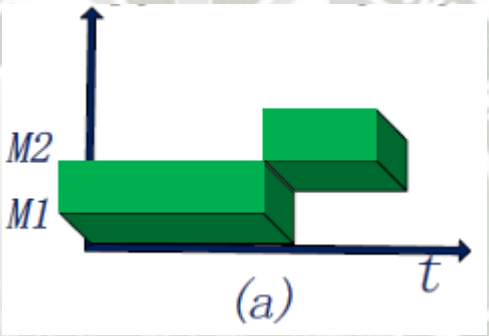

Ilustración 1 Sudden Drift
Fuente: An algorithm for Detecting Concept Drift Based on Ceontext in Process Mining

- *Gradual Drift:* A diferencia del *Sudden Drift*, ambos procesos coexisten por un tiempo, descontinuando M1 gradualmente.

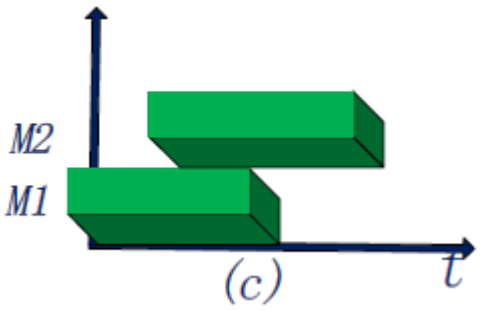

Ilustración 2 Gradual Drift
Fuente: An algorithm for Detecting Concept Drift Based on Ceontext in Process Mining

- *Recurring Drift:* Corresponde al escenario donde un set de procesos reaparece después de un tiempo (sustituido una y otra vez).

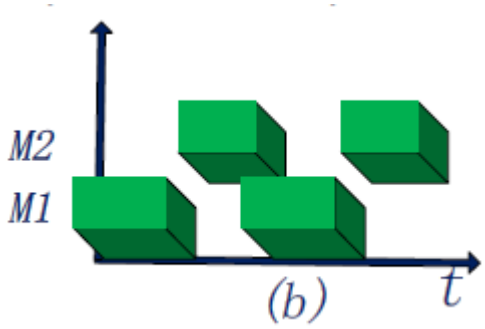

Ilustración 3 Recurring Drift
Fuente: An algorithm for Detecting Concept Drift Based on Ceontext in Process Mining

- *Incremental Drift:* Se refiere a un escenario donde la sustitución de un proceso es dado a través de pequeños cambios incrementales.

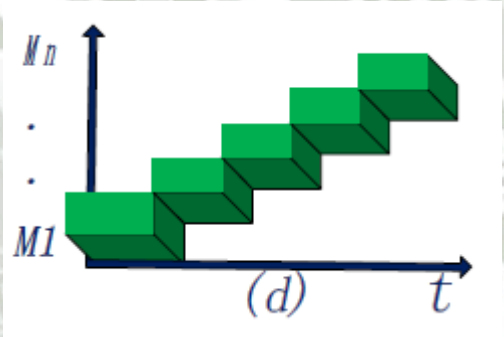

Ilustración 4 Incremental Drift
Fuente: An algorithm for Detecting Concept Drift Based on Ceontext in Process Mining

2.2.3 Análisis y Comparación de clasificadores informáticos aplicados al *Concept Drift*

Tabla 2 Tabla Comparativa de Clasificadores Informáticos

Variables	Clasificadores		
	Ventanas	**Piscina de clasificadores**	**Poliedros convexos**
Tiempo de ejecución	Moderado	Alto	Alto
Granularidad	Dependiente del tamaño de ventana	Alta	Media
Precisión	Alta, dependiendo del tamaño de Ventana	Media	Media o baja dependiendo del tamaño del *log*
Uso de recursos	Moderado	Alto	Alto, dependiendo del tamaño del *log*
Capacidad de adaptación Online	Si	Si	Si

Fuente: Elaboración Propia

La *Tabla 2* muestra una comparación sobre los clasificadores informáticos aplicados para la detección de *Concept Drift*. Uno de los clasificadores con mayor precisión y bajo tiempo de ejecución son las ventanas adaptativas, ya que estas pueden variar en tamaño se puede parametrizar la granularidad deseada, a mayor tamaño de ventana se tiene una granularidad más baja mientras que a menor tamaño de ventana se puede obtener un resultado mucho más a detalle.

2.2.4 *Plug-in* ProM *Concept Drift*

El trabajo propuesto por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014) señala los siguientes desafíos en *Concept Drift*:

a. Detectar el Punto de Cambio: Se refiere a localizar un *Concept Drift* en un *log* de procesos, seguidamente se procede a detectar los periodos de tiempo en los que ocurrió el *Drift*.

b. Localización y Caracterización del cambio: Se refiere a caracterizar la naturaleza del cambio (perspectivas de flujo, datos o recursos) e

identificar las regiones de cambio de un proceso. Implica encontrar el cambio exacto en el proceso.

c. Descubrimiento del proceso de cambio: Se refiere a poner el cambio en perspectiva para un análisis posterior, para esto se necesitan técnicas y herramientas.

Ésta propuesta se enfoca principalmente en dos puntos, primero en la detección y localización del cambio, y segundo en la caracterización del cambio. Todo esto lo trabaja de manera *off-line*.

La premisa básica del *Concept Drift* define que las características de los casos antes del punto de cambio difieren de las características de los casos después del punto de cambio. Basándose en ésta premisa los autores plantean su algoritmo en dos pasos: 1) Capturar las características de los casos, extracción de características y 2) Identificar cuando estas características cambian, detección del *Drift*. Las características pueden definirse por caso o a nivel de *sub-log*.

Los *logs* procesados para determinar las características de los casos pueden observarse como un *stream* de datos de valores de características donde pruebas estadísticas pueden utilizarse para determinar cambios.

Las características definidas pueden ser locales o globales. Las características globales están definidas sobre un *log* de eventos, mientras que las características locales están definidas a un nivel de caso. Para la perspectiva de cambio de flujo proponen dos variables globales: 1) *Relation Type Count* (RC) y 2*) Relation Entropy* (RE); y dos variables locales 1) *Window count* (WC) y 2) *J Measure* (J). Estas características están definidas como sigue:

1. *RC*: La RC con respecto a la relación de precede (antecede) es una función: $f_{RC}^{\mathcal{L}}: \mathcal{A} \to \mathbb{N}_0 \times \mathbb{N}_0 \times \mathbb{N}_0$, definida sobre el conjunto de actividades $\mathcal{A}$. $f_{RC}^{\mathcal{L}}$ de una actividad, $x \in \mathcal{A}$ con respecto a la relación precede (antecede) sobre un *log* de eventos $\mathcal{L}$ es la tripleta $\langle c_A, c_S, c_N \rangle$ dónde c_A, c_S y c_N son el número de actividades en que $\mathcal{A}$ siembre (always), a veces (sometimes) y nunca (never) precede (antecede) x, respectivamente en el *log* de eventos$\mathcal{L}$.
2. *RE*: La RE con respecto a la relación precede (antecede) es una función: $f_{RE}^{\mathcal{L}}: \mathcal{A} \to \mathbb{R}_0^+$, definida sobre el conjunto de actividades. $f_{RE}^{\mathcal{L}}$ de una actividad, $x \in \mathcal{A}$ con respecto a la relación de precede (antecede) es la entropía de la métrica RC. En otras palabras $f_{RE}^{\mathcal{L}}(x) = -p_A log_2(p_A) - p_S log_2(p_S) - p_N log_2(p_N)$ donde, $p_A = c_A/|\mathcal{A}|$ y $p_S = c_S/|\mathcal{A}|$ y $p_N = c_N/|\mathcal{A}|$, y $\langle c_A, c_S, c_N \rangle = f_{RC}^{\mathcal{L}}(x)$.
3. *WC*: Dado un tamaño de ventana $l \in \mathbb{N}$, la WC con respecto a la relación que precede (antecede) es una función, $f_{WC}^{l,t}: \mathcal{A} \times \mathcal{A} \to \mathbb{N}_0$, definido sobre un conjunto de pares de actividades. Dado un caso **t** y un tamaño de ventana l, sea $S^{l,t}(a)$ la bolsa de todas las subsecuencias $t(i, i+l-1)$, tal que $t(i) = a$. Sea $\mathcal{F}^{l,t}(a,b) = \left[s \in S^{l,t}(a) | \exists_{1<k\leq|s|}\, s(k) = b\right]$, p.e., la bolsa de subsecuencias en **t** empezando por a y precedido por b dentro de una ventana de tamaño l. La WC de la relación b precede a, $f_{WC}^{l,t}(a,b) = |\mathcal{F}^{l,t}(a,b)|$.
4. *J Measure*: Existe el hecho en el que se puede considerar la relación *b* precede a como una regla: si la actividad a ocurre, entonces la actividad *b* probablemente puede ocurrir. El *J Measure* con respecto a la relación precede (antecede) como una función $f_J^{l,t}: \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+$definida sobre

un conjunto de pares de actividades y una ventana de longitud $l \in \mathbb{N}$. Sea $p^t(a)$ y $p^t(b)$ las probabilidades de la ocurrencia de las actividades a y b, respectivamente en un caso **t**. Sea $p^{l,t}(a,b)$ la probabilidad de que b preceda a en una ventana de longitud l, p.e., $p^{l,t}(a,b) = |\mathcal{F}^{l,t}(a,b)|/|S^{l,t}(a)|$. Entonces el *J Measure* para una ventana de longitud l está definida como $f_J^{l,t}(a,b) = p^t(a)CE^{l,t}(a,b)$ donde $CE^{l,t}(a,b)$ es la entropía cruzada de a y b (b precede a dentro de la ventana de longitud l) y se encuentra definida como: $CE^{l,t}(a,b) = p^{l,t}(a,b)log_2\left(\frac{p^{l,t}(a,b)}{p^t(b)}\right) + (1-p^{l,t}(a,b))log_2\left(\frac{1-p^{l,t}(a,b)}{1-p^t(b)}\right)$

El *framework* propuesto por los autores identifica los siguientes pasos:

1. Extracción y selección de características: en este paso se definen las características de los casos en un *log* de eventos. Se han definido cuatro características que caracterizan la perspectiva de flujo de las instancias de procesos en un *log* de eventos.
2. Generación de Poblaciones: Un *log* de eventos puede ser transformado en un *stream* de datos basado en las características seleccionadas en el paso anterior. Este paso define las poblaciones ejemplo para estudiar los cambios es las características de los casos. Para definir las poblaciones los autores han considerado ventanas no sobrepuestas, continuas y de tamaño fijo.
3. Comparación de poblaciones: Una vez que se han generado las poblaciones de ejemplo, el siguiente paso es analizar estas poblaciones para encontrar cambios en las características. Utilizan métodos de hipótesis estadísticos para comparar las poblaciones.

4. Visualización Interactiva: El resultado de los estudios comparativos en las poblaciones pueden ser representadas intuitivamente a un análisis.
5. Analizar Cambios: Técnicas de visualización como el gráfico de los *Drift*s pueden asistir para identificar los puntos de cambio.

2.2.5 Minería organizacional y Redes Sociales

El análisis de Redes Sociales está relacionado a la sociometría que se refiere a los métodos para representar los datos en las relaciones interpersonales en un grafo o matriz.

En caso de utilizar una estructura de grafo los nodos en una red social corresponden a las entidades organizacionales, normalmente los recursos que se encuentran en el *log*, roles, grupos o departamentos. Los arcos corresponden a las relaciones que se dan entre las entidades donde el peso corresponde a la "importancia". Por ejemplo, si el nodo "y" es más importante que el nodo "x" y "z" porque tiene más tamaño. La relación entre "y" y "z" es más fuerte que la relación entre "x" y "z". La interpretación de la importancia depende en la red social.

Existe una gran variedad de métricas para analizar las redes sociales y para caracterizar el rol de los nodos individuales en un diagrama. Por ejemplo, si todos los nodos tienen una corta distancia a un nodo y todos los caminos geodésicos visitan el nodo. También existen diferentes métricas para la intuitiva noción de "centralidad".

El índice de Bavelas-Leavitt es un ejemplo bien conocido que se basa en los caminos geodésicos en el grafo. Sea i un nodo y sea $D_{j,k}$ la distancia geodésica del nodo j al nodo k. Es posible definirlo como:

$$BL(i) = \frac{\sum j,k\ D_{j,k}}{\sum j,k\ D_{j,i} + D_{i,k}}$$

Otras métricas son la "cercanía", donde se toma el valor 1 entre la suma de todas las distancias geodésicas a un nodo. La "intermediación", la cual es un radio basado en el número de caminos geodésicos visitando un nodo.

Claramente, los *logs* de eventos con $\#_{resource}(e)$ atributos proveen una excelente fuente de información para el análisis de redes sociales. Por ejemplo, basados en el *log* de eventos uno puede contar cuantas veces un trabajo ha sido delegado de un recurso a otro. Siguiendo el caso 1: $\langle a^{Pete}, b^{Sue}, d^{Mike}, e^{Sara}, h^{Pete}\rangle$ existe un paso de trabajo de Pete a Sue y Mike antes de terminar el trabajo de a. Nótese que b y d son concurrentes y por ende Sue nunca delega trabajo a Mike. Sin embargo Sue y Mike delegan trabajo a Sara, ya que la actividad e necesita el ingreso de b y d. Finalmente Sara delega el trabajo a Pete. Hay cinco delegaciones de trabajo $(a^{Pete}, b^{Sue}), (a^{Pete}, d^{Mike}), (b^{Sue}, e^{Sara}), (d^{Mike}, e^{Sara}), (e^{Sara}, h^{Pete})$

Sin embargo, es posible construir redes sociales a nivel de departamentos, equipos o roles. Asumimos que existen tres roles: Asistente, Experto y Manager. Permitimos que los roles sean descubiertos por la frecuencia de los patrones en el *log*. Más aún dicha información esta típicamente disponible en el sistema de información. Es posible contar el número de delegaciones a nivel de rol. Considerando el caso anterior, utilizando la información de los roles, se puede representar $\langle a^{Asisstant}, b^{Expert}, d^{Asistant}, e^{Manager}, h^{Asisstant}\rangle$. Podemos encontrar cinco delegaciones: una de Asistente a Experto, otra de Asistente a

Asistente y otra de Manager a Asistente. Con estos datos se arma la *Tabla 3* con el promedio de delegaciones por caso.

Tabla 3 Promedio de las Delegaciones por Caso

	Pete	Mike	Ellen	Sue	Sean	Sara
Pete	0.135	0.225	0.09	0.06	0.09	1.035
Mike	0.225	0.375	0.15	0.1	0.15	1.725
Ellen	0.09	0.15	0.06	0.04	0.06	0.69
Sue	0	0	0	0	0	0.46
Sean	0	0	0	0	0	0.69
Sara	0.885	1.475	0.59	0.26	0.39	1.3

Fuente: Elaboración Propia

Contar el número de delegaciones de trabajo es solo un método de construir una red social de un *log* de eventos. Existen varios tipos de redes sociales, por ejemplo uno puede contar cuantas veces dos recursos han estado sobre el mismo caso. Es posible medir la similitud de dos recursos, en la tabla siguiente cada fila en la matriz puede ser vista como el perfil de un recurso, este vector indica las características del recurso. Por ejemplo, Pete tiene el perfil $P_{Pete} = (0.30, 0.0, 0.345, 0.69, 0.0, 0.0, 0.135, 0.165)$, Mike tiene el perfil $P_{Mike} = (0.2, 0.0, 0.575, 1.15, 0.0, 0.0, 0.225, 0.275)$ y Sara tiene el perfil $P_{Sara} = (0.0, 0.0, 0.0, 0.0, 2.3, 1.3, 0.0, 0.0)$. Se puede observar claramente la similitud entre Pete y Mike mientras Pete y Sara no lo son. La distancia entre dos puntos se mide a través dela Distancia Minkowski, la distancia Hamming y el coeficiente de relación de Pearson.

Tabla 4 Perfiles Calculados de los Usuarios

	A	B	C	D	E	F	G	H
Pete	0.3	0	0.345	0.69	0	0	0.135	0.165
Mike	0.5	0	0.575	1.15	0	0	0.225	0.275
Ellen	0.2	0	0.23	0.46	0	0	0.09	0.11
Sue	0	0.46	0	0	0	0	0	0
Sean	0	0.69	0	0	0	0	0	0
Sara	0	0	0	0	2.3	1.3	0	0

Fuente: Elaboración Propia

Al mismo tiempo las técnicas de *clustering* como k-means y el *clustering* jerárquico aglomerativo se pueden utilizar para agrupar los recursos de acuerdo al perfil. Finalmente en el ejemplo observando la *Tabla 4* nos percatamos que Pete, Mike y Ellen son similares y tienen una fuerte relación social en la red basada en similitud. Igualmente Sue y Sean tiene una fuerte relación en la red social basada en similitud. Sara es claramente diferente comparada con los otros dos en otros grupos.

2.2.6 ProM *Framework*

ProM (Process Mining *framework*) es un *framework* de código abierto para algoritmos de minería de procesos. Provee una plataforma para usuarios y desarrolladores de algoritmos de minería de procesos para que sean fácil de utilizar y fáciles de extender.

ProM es un entorno de *plug-in*s para la minería de procesos la cual provee una base común para todo tipo de técnicas de minería de procesos, desde la importación, exportación y filtrado de *logs* de eventos para su análisis y la visualización de resultados.

La última versión disponible de ProM es la 6.4.1 “CopR+”

Ilustración 5 Pantalla de Inicio ProM 6.4.1
Fuente: ProM 6 Getting Started

Liberada últimamente la versión 6.5 "SilvR" publicado este mismo año.

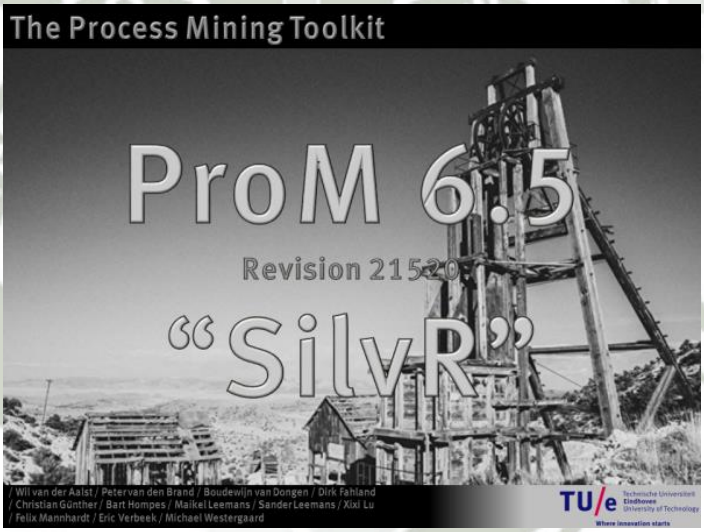

Ilustración 6 Pantalla de Inicio ProM 6.5
Fuente: ProM 6 Getting Started

Interfaz de Usuario

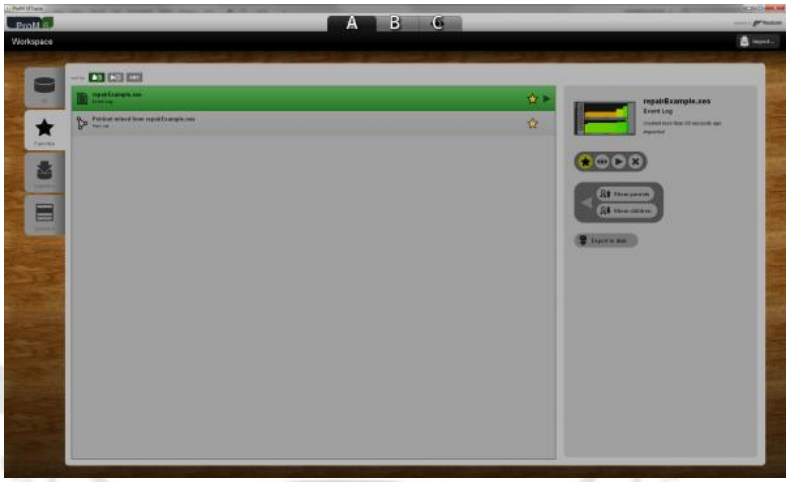

Ilustración 7 Interfaz de usuario
Fuente: ProM 6 Getting Started

En la *Ilustración 6* se muestra mediante letras cada objeto:

A) Vista "Workspace": En esta vista se muestra todos los recursos, como *logs* y redes Petri, que pueden haber sido importados al *framework* o que son un resultado de la ejecución de una acción u otros recursos.

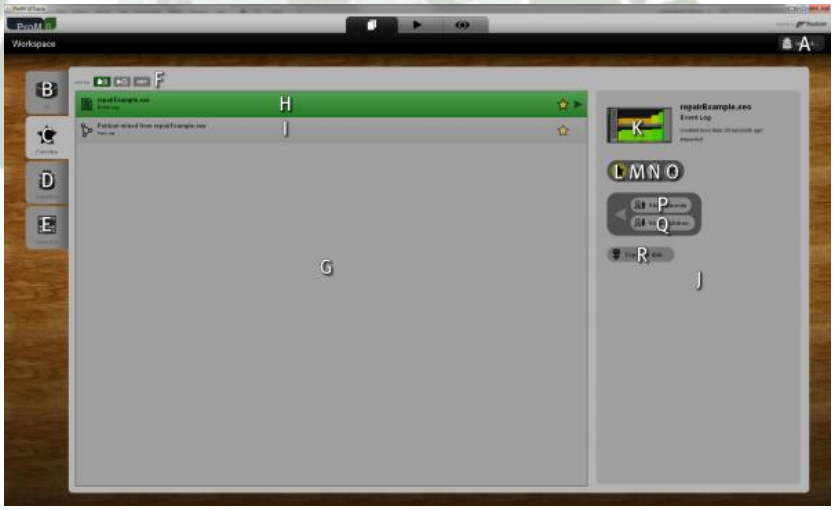

Ilustración 8 Vista "Workspace"
Fuente: ProM 6 Getting Started

En la *Ilustración 7* se muestra mediante letras cada objeto de la vista "Workspace", dónde:

a. El botón "Import…" importa un recurso de un archivo a ProM 6x.

b. La pestaña “All” muestra todos los recursos en el panel de recursos (G)
c. La pestaña “Favorites” muestra los recursos favoritos en el panel de recursos (G)
d. La pestaña “Imported” muestra todos los recursos importados en el panel de recursos (G)
e. La pestana “Selection” permite visualizar la selección de recursos (P y Q) en el panel de recursos (G)
f. Los botones de ordenado ordenan los recursos en (G), son ordenados según fecha de creación, último uso y por nombre.
g. El panel de recursos muestra los recursos apropiados en el orden apropiado.
h. Un recurso seleccionado.
i. Un recurso no seleccionado.
j. La vista de recurso, muestra los detalles del recurso seleccionado.
k. Una vista previa del recurso seleccionado, con detalles adicionales como el nombre, tipo y origen.
l. El botón “Favorites” cambia de estado si el recurso seleccionado es un favorito.
m. El botón “View” muestra la vista del recurso seleccionado.
n. El botón “Action” lo lleva a la vista de acción donde el recurso seleccionado es añadido como un input para la acción.
o. El botón “Remove” elimina el recurso seleccionado.
p. El botón “Show parents” lo lleva a la pestaña de “Selection” (E) mostrando los recursos padre del recurso seleccionado.

q. El botón “Show children” lo lleva a la pestaña de “Selection” (E) mostrando los recursos hijo del recurso seleccionado.

r. El botón “Export to disk” exporta un recurso de ProM a un archivo.

B) Vista “Action”: Muestra todas las acciones que pueden tomar ciertos recursos como inputs, resultar en cierto tipo de outputs y si igualan algún criterio de filtrado. Utilizando esta vista es fácil reconocer las acciones apropiadas para cada recurso.

Ilustración 9 Interfaz de la Vista "Action"
Fuente: ProM 6 Getting Started

En la *Ilustración 8* se muestra mediante letras cada objeto de la vista “Action”, dónde:

a. El botón “Activity…” abre una vista de actividad, dónde se muestran las acciones que han sido ejecutadas y cuales siguen corriendo.

b. El panel de input muestra los recursos de entrada. Si se inicia una acción, entonces estos recursos serán utilizados como inputs de ésta acción.

c. Un recurso.

d. El botón “Workspace” lo lleva a la vista “Workspace” con el correspondiente recurso seleccionado.
e. El botón “Remove” elimina el recurso correspondiente del panel de recursos de entrada.
f. El contenedor de recursos permite añadir un recurso al panel de recursos de entrada.
g. El panel de acción muestra las acciones disponibles basados en el panel de recursos de entrada (B), el panel de salida (M), y el filtro de acciones (L)
h. El elemento “Action” muestra que la acción correspondiente es interactiva y va a requerir una configuración por parte del usuario.
i. El elemento “Action” muestra que la acción correspondiente es un batch, y no requiere una configuración por parte del usuario.
j. El botón “Interactive” cambia de estado según las acciones sean interactivas o no.
k. El botón “Batch” cambia de estado según las acciones sea de tipo batch o no.
l. El cambo “Search” permite filtrar los nombres de las acciones.
m. El panel de salida (output) muestra los tipos de output de la acción seleccionada.
n. El tipo de elemento muestra que un recurso de un tipo correspondiente está en el panel de salida correspondiente.
o. El contenedor de tipo permite añadir un tipo al panel de salida.

p. El botón “Reset” resetea la vista de acción. Limpia el panel de recursos de entrada, el panel de tipos de salida, y el filtro, además deselecciona todas las acciones del panel de acciones.

C) Vista “View”: Muestra un recurso, o un panorama de todos los recursos que existan para una vista.

Ilustración 10 Interfaz de la vista "View"
Fuente: ProM 6 Getting Started

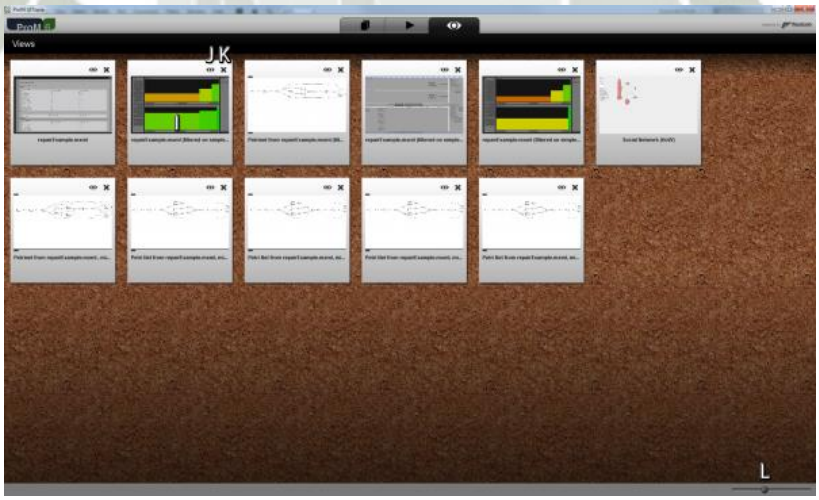

Ilustración 11 Panorama de los recursos.
Fuente: ProM 6 Getting Started

En la *Ilustración 9* se muestra mediante letras cada objeto de la vista “View”, dónde:

a. La lista desplegable permite seleccionar entre diferentes vistas en el mismo recurso.
b. El botón “Refresh” refresca la vista actual.
c. El botón “Print” imprime la vista actual.
d. El botón “Favorites” cambia de estado si el recurso actual es un favorito.
e. El botón “Action” permite cambiar a la vista acción con el recurso correspondiente añadido como un input para la acción.
f. El botón “Workspace” permite cambiar a la vista “Workspace” con el recurso correspondiente seleccionado.
g. El botón “View” permite cambiar a la vista de panorama, *Ilustración 10*.
h. El área principal muestra la vista en el correspondiente recurso.
i. Un slider muestra una vista previa del correspondiente recurso.
j. El botón “View” abre la vista para el correspondiente recurso.
k. El botón “Remove” elimina la vista del objeto correspondiente, no elimina el recurso sino la vista en el recurso.
l. El slider en la esquina inferior derecha, controla el tamaño de los objetos en el panel. (Group, 2010)

CAPÍTULO III: MODELO PROPUESTO

3.1 MODELO PROPUESTO

Se propone un algoritmo que complemente al *Plug-in* de *Concept Drift* de ProM para que este sea capaz de hallar el *Concept Drift* en la perspectiva de usuarios.

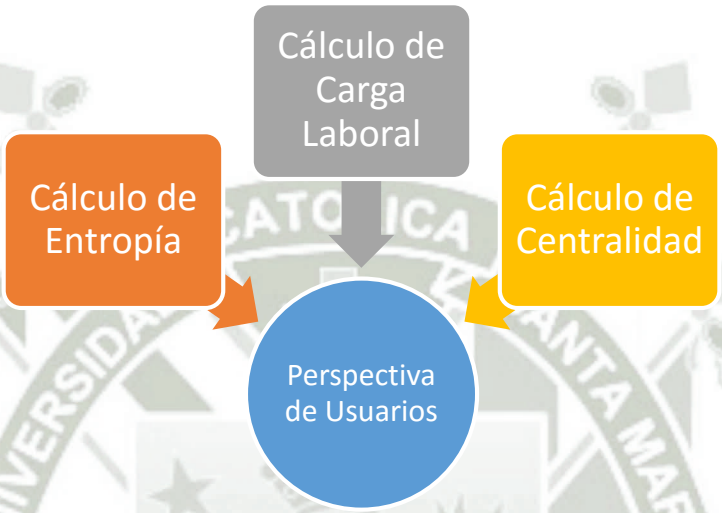

Ilustración 12 Modelo Propuesto
Fuente: Elaboración Propia

En la *Ilustración 11* se puede observar que el modelo propuesto para hallar el *Concept Drift* en la perspectiva de usuarios se compone de tres características. Como característica global tenemos el cálculo de la entropía, esta se aplica a todo el *log*. Como características específicas tenemos el cálculo de la carga laboral y el cálculo de la centralidad, las cuales a diferencia de la entropía son calculadas a nivel de grupos de casos y no de todo el *log*.

El modelo propuesto no considera los roles ni el nivel jerárquico a los que los usuarios pertenezcan, es decir se considera a los usuarios según la cantidad de actividades en las que estos participen sin clasificarlos ni distinguirlos por roles o jerarquía. Se decidió utilizar conceptos de minería organizacional y redes sociales.

3.1.1 Algoritmo *Concept Drift*

El proceso que realiza el *Plug-in* de *Concept Drift* de ProM, es el siguiente:

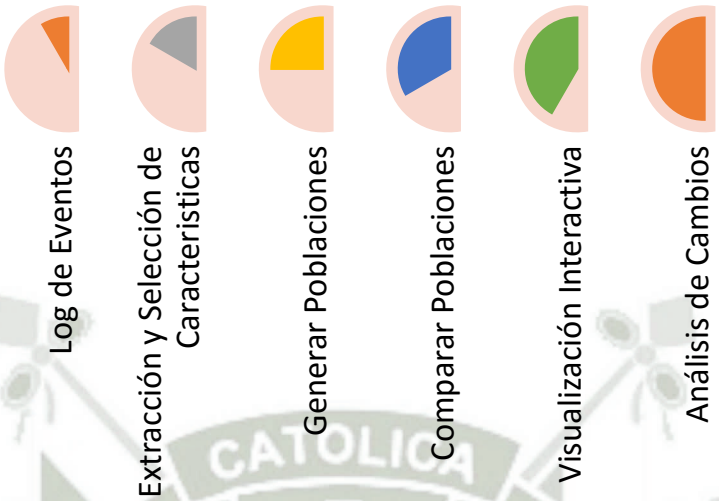

Ilustración 13 Proceso para analizar Concept Drift
Fuente: Elaboración Propia

El flujo de trabajo sigue el establecido en (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014), en la etapa de extracción y selección de características se utilizan métodos distintos para poder encontrar el *Concept Drift* en la perspectiva de usuarios.

El proceso de Extracción y Selección de Características, posee los siguientes pasos:

- Extracción de actividades del *log*
- Encriptación de Nombres de Actividades en el *log*
- Generación de Pares de Actividades, donde se calcula la entropía de la relación entre ambas, por caso.

3.1.2 Propuesta para hallar *Concept Drift* en la perspectiva de usuarios

Al analizar el algoritmo desarrollado por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014), se observó que las características usadas para encontrar el *Concept Drift* en la perspectiva de flujo podrían ser utilizadas de una

forma similar para encontrarlo en la perspectiva de usuario. Para esto se ha definido las siguientes características: 1) Relación de Entropía la cual se basa en la misma idea del cálculo de entropía en el algoritmo original, 2) Carga laboral, asignación de tareas por usuarios y 3) Centralidad, la cual refleja la participación de un usuario en el proceso.

3.1.2.1 Relación de Entropía

La entropía es la medida del desorden o la peculiaridad de ciertas combinaciones. Esta medida general se utiliza en el trabajo de Bose, y la reutilizamos debido a que nos permite darnos una idea general acerca del estado del *log*. Utilizamos la fórmula aplicada a los usuarios correspondientes:

$$f_{RE}^{L}(x) = -p_A \log_2(p_A) - p_S \log_2(p_S) - p_N \log_2(p_N)$$

Donde, p_A, p_B, p_C es la probabilidad que el usuario realice una tarea siempre, algunas veces o nunca.

A modo de ejemplo, en un *log* pequeño consideramos que se contaron la cantidad de veces que un usuario atendió la tarea en la *Tabla 5.*

Tabla 5 Cantidad de Veces que un Usuario atendió una Tarea

Actividad	Usuario				
	Indefinido	A	B	C	D
A	3	1	0	0	2
B	2	0	1	0	0
C	2	0	0	1	0
D	3	1	1	0	0
E	1	1	0	0	0
F	5	0	0	0	0

Fuente: Elaboración Propia

Consideramos que existen en total 6 casos que poseen todas las actividades y cuantas veces las atendieron los usuarios. Luego de obtener la información correspondiente armamos una segunda tabla, *Tabla 6.*

Tabla 6 Relación, Siempre (A), A Veces (S) y Nunca(N) que un usuario realiza una actividad

Actividad	Usuario				
	Indefinido	A	B	C	D
A	S	S	N	N	S
B	S	N	S	N	N
C	S	N	N	S	N
D	S	S	S	N	N
E	S	S	N	N	N
F	A	N	N	N	N

Fuente: Elaboración Propia

En la *Tabla 6* clasificamos la frecuencia correspondiente al número de actividades que cada uno de los usuarios atendió. En nuestro trabajo utilizamos rangos relacionados a la frecuencia de atención de casos. Finalmente contamos la probabilidad de siempre (A), a veces(S) y nunca(N) y aplicamos la fórmula de entropía que aplica Bose.

Aplicando la fórmula para el usuario indefinido se aplica de la siguiente forma:

$$C_A = 1, C_s = 5, C_N = 0$$

Luego se considera que el total de casos es 6. Y procedemos a calcular las probabilidades $P_A = 0.1667, P_s = 0.833, P_N = 0$y seguimos con el reemplazo en la fórmula:

$$f_{RE}^{L}(x) = -0.1667\, log_2(0.1667) - 0.833\, log_2(0.833) - 0 = 0.65$$

Tabla 7 Cantidad de S, A y N de los usuarios y cálculo de Entropía

Usuario	A	S	N	Entropía
Indefinido	1	5	0	0.65002242
A	0	3	3	1
B	0	2	4	0.91829583
C	0	1	5	0.65002242
D	0	1	5	0.65002242
F	1	0	0	0

Fuente: Elaboración Propia

En el resultado de entropía en la *Tabla 7* se observa que las actividades con un resultado más cercano a cero son los que han mantenido la misma cantidad de tareas en todo el *Log* analizado, a estos se les puede considerar como los más

estables, sin embargo los que mantienen un resultado más distanciado a cero se les puede considerar como menos estables y poco predecibles.

3.1.2.2 Carga laboral

Consideramos Carga Laboral a la cantidad de actividades que un usuario ejecuta en un determinado número de casos. Considerando la explicación básica de sociometría, armamos una lista con la cantidad de tareas atendidas por usuario lo cual llamaremos perfiles. Luego se procede a ponderarlas y compararlas con el resto de usuarios.

Para comparar los resultados de perfiles de usuario se utiliza la distancia de Minkowski, aplicando la siguiente fórmula:

$$distancia = \sum_{i=1}^{n} |x_i - y_i|^p \quad ^{1/p}$$

Donde:

n: Cantidad de actividades existentes.

p: Se considera como índice utilizado el 1.5.

$P = (x_1, x_2, x_3, \dots x_n)$ Es el perfil del usuario que se analiza

$Q = (y_1, y_2, y_3, \dots y_n)$ Es el perfil del usuario de comparación

Con la distancia Minkowski obtenida por pares de usuarios se procede a evaluar los resultamos mediante las pruebas de hipótesis ya establecidas en el *Plug-in* de *Concept Drift*. Se puede establecer que, mientras los perfiles de los pares de usuarios sigan manteniendo la misma distancia en el *Log* no se han producido cambios en cuanto a la carga laboral asignada, sin embargo si esta distancia entre dos usuarios varía se puede decir que uno de ellos tiene más o menos carga asignada en un determinado caso del *Log*, produciéndose así un *Concept Drift*.

Considerando el *log* anterior se arma la siguiente tabla, *Tabla 8* (mejora de la *Tabla 5* para fines didácticos), cada fila se considera como un perfil.

Tabla 8 Cantidad de Actividades Realizadas por Usuario

Usuario	Actividad					
	A	B	C	D	E	F
Indefinido	3	2	2	3	1	5
A	1	0	0	1	1	0
B	0	1	0	1	0	0
C	0	0	1	0	0	0
D	2	0	0	0	0	0

Fuente: Elaboración Propia

Para ejemplificar el cálculo utilizaremos el usuario indefinido, utilizamos el valor 1.5 como el valor por defecto de p. Consideramos que la tabla anterior es un momento. Obtenemos los valores con el resto de usuarios:

$$Indefinido\ vs\ A = \sqrt[1.5]{(2)^{1.5} + (2)^{1.5} + (2)^{1.5} + (2)^{1.5} + (0)^{1.5} + (5)^{1.5}}$$
$$= 7.97$$

$$Indefinido\ vs\ B = \sqrt[1.5]{(3)^{1.5} + (1)^{1.5} + (2)^{1.5} + (2)^{1.5} + (1)^{1.5} + (5)^{1.5}}$$
$$= 8.33$$

$$Indefinido\ vs\ C = \sqrt[1.5]{(3)^{1.5} + (2)^{1.5} + (1)^{1.5} + (3)^{1.5} + (1)^{1.5} + (5)^{1.5}}$$
$$= 8.87$$

$$Indefinido\ vs\ D = \sqrt[1.5]{(1)^{1.5} + (2)^{1.5} + (2)^{1.5} + (3)^{1.5} + (1)^{1.5} + (5)^{1.5}}$$
$$= 8.33$$

Por comparación el usuario indefinido tiene una mayor similitud con el usuario A. Al comparar la similitud entre usuario podemos definir el nivel de especialización que tiene un usuario dentro del flujo.

3.1.2.3 Centralidad

Consideramos el aspecto de interacción entre los nodos (Usuarios) y las actividades para el mismo utilizamos el concepto de centralidad de los grafos. Con los usuarios como nodos y la cantidad de veces que una actividad se ha derivado de un usuario a otro como arista, se procede a armar un grafo ponderado, donde el usuario al que más aristas se dirijan y tengan mayor peso será el usuario central, o el más importante del proceso, a esto le llamamos centralidad.

La centralidad se obtiene a través de la fórmula de Bavelas-Leavitt que se muestra continuación:

$$BL(i) = \frac{\sum j,k \ D_{j,k}}{\sum j,k \, D_{j,k} + D_{i,k}}$$

La fórmula se encarga de obtener todas las distancias geodésicas de y hacia el nodo i.

Para explicarlo se muestra un grafo de ejemplo. El siguiente grafo se arma con la cantidad de actividades que un usuario delega a otro, por ejemplo si una actividad empezó con el usuario B y luego lo toma el usuario A se considerará con un valor de 1 de B hacia A.

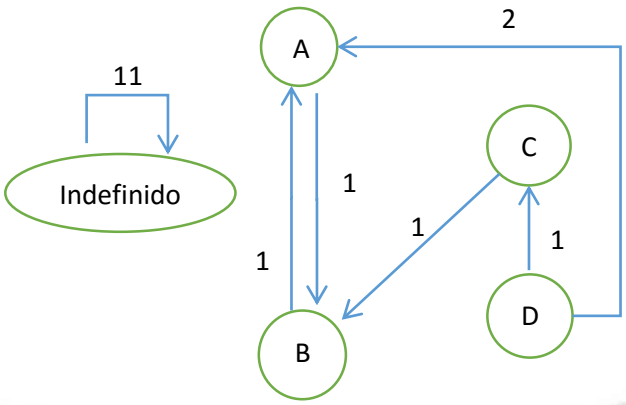

Ilustración 14 Grafo de Ejemplo
Elaboración Propia

Utilizando los montos del gráfico procederemos a obtener el valor de la fórmula de Bavelas-Leavitt para el usuario B.

$$BL(i) = \frac{(17)}{(1)+2} = 5.66667$$

En el caso del usuario indefinido, se considera el llamado recursivo dentro de la sumatoria de las distancias geodésicas. Debido a que la carga de trabajo siempre cae sobre sí mismo y se considera relevante.

Al mismo tiempo es posible utilizar la medida de la centralidad, si consideramos que la fórmula de centralidad de Bavelas indica que:

$$Centralidad = \frac{Total\ Distancia\ geodésica}{Distancia\ geodésica\ desde\ y\ hacia\ nodo}$$

Se entiende que a medida que un usuario tenga una mayor interacción en el flujo el resultado de la centralidad se acercará más a cero lo que lo pondría en un punto central y vital del flujo dirigiendo las tareas. En caso contrario mientras menos interacciones posean su resultado de centralidad tendrá un mayor valor.

3.1.2.4 Algoritmo Propuesto

En el algoritmo propuesto se define un tamaño de grupo de casos para analizar la muestra con las características anteriormente presentadas. Se utiliza este tamaño de grupo para obtener un mejor cálculo de la distancia de Minkowski y del Bavelas-Leavitt, ya que si se utilizan las actividades de un solo caso no se obtendría un resultado en el que se pueda observar un *Concept Drift* por que la muestra sería muy pequeña. Utilizando la referencia del *Concept Drift* de Bose seguimos los siguientes pasos:

- Generación de Listas de Actividades
- Generación de Listas de Usuarios
- Generación de Matrices Usuario por Actividad según Tamaño de Grupo
- Cálculo de características a través de dos clases
    - Generales: Entropía
    - Específicas: Obtención de la Distancia Minkowski y Bavelas-Leavitt
- Las listas obtenidas con los datos resueltos se envían a las pruebas de hipótesis ya definidas por Bose.
- Al finalizar la ejecución se grafican los resultados. Los resultados graficados para la propuesta son el acumulado, por caso o caso, de los valores devueltos por la prueba de hipótesis.

Esta solución tiene aplicaciones variadas pero dependen mucho del flujo al cual se le aplica. Por ejemplo, tenemos un flujo de trabajo de proyectos informáticos y un usuario con resultados de baja entropía, la carga laboral como distinta y finalmente con el valor de centralidad muy lejano de 1. Significaría que el usuario es realmente valioso en la empresa ya que es una persona que realiza todo tipo

de tareas pero cierra los circuitos de requerimiento. En cambio si se trata de un flujo de ventas, estas medidas o cualidades no serían beneficiosas ya que les conviene personas que empiecen con el flujo o promocionen las ventas, por ende tendría entropía alta, carga laboral muy parecida al resto y una centralidad más cercana al cero.

Algoritmo Entropía

Algoritmo de Obtención de entropía

Recibe: Matriz [actividades, usuarios] llena con los datos de todo el *log*.

Se obtienen los rangos de Nunca (Never), algunas veces (Sometimes) y siempre (Always).

```
PARA CADA Usuario
       PARA CADA Actividad
              Se obtiene la probabilidad
              Se clasifica de acuerdo al rango en A, S o N.
       FIN PARA

       Con los totales se calcula la probabilidad para Pa, Ps    y
Pn.
       fr(x) = -Pa log2(Pa) -Pb log2(Pb) -Pc  log2(Pc)
       Agregar f(x) en el vector de salida.
FIN PARA
```

Algoritmo Obtención de distancia de Minkowski y Bavelas-Leavitt

```
Algoritmo Obtención de distancia de Minkowski y Bavelas-Leavitt

Recibe: Log, tamaño de grupo

Se inicializan las variables, matrices, grafo y lista de
matrices.

PARA CADA Trace en el log
      Se obtiene los datos
      SI el contador es menor al tamaño de grupo
            AnalizaTrace(TraceActual)
      SINO
            Añadir a la lista de Matrices la matriz
      temporal
            Se inicializa la matriz temporal
            Añadir a la lista de Grafos el grafo
      temporal
            Se inicializa el grafo temporal
            Se inicializa contador
      FIN SI
FIN PARA
CalcularMinkowski()
CalcularBavelas()
```

```
AnalizoTrace
Recibe: Trace

PARA CADA Evento en el Trace
      Se obtiene el recurso actual
      Se obtienen las posiciones dentro de la matriz
      temporal y se lleva la cuenta.

      SI existe recurso anterior
            SI en el grafo existe relación recurso
      anterior a recurso actual
                  Se aumenta la ponderación de la
            arista
            SINO
                  Se agrega la relación en el grafo
            FIN SI

            Agregar al grafo temporal la arista con
      origen de Usuario_Anterior al
      Usuario_Actual con una ponderación de 1
      FIN SI
      Usuario_Anterior = Usuario_Actual
FIN PARA
```

```
CalcularMinkowski
Recibe: Lista de Matrices, usuario

PARA CADA usuario A
      PARA todos los usuarios B que no sea A
            PARA Cada Matriz en la Lista
                  PARA todas las actividades K
                        Auxiliar
                        = (matriz[k,a] – matriz[k,b])^p
                  FIN PARA
            FIN PARA
            Agregar resultado a la matriz de valores
      FIN PARA
FIN PARA
```

```
CalcularBavelas
Recibe: Lista de Grafos, usuario

Obtengo en la variable T el total de la distancia geodésica
PARA CADA Usuario A
      PARA CADA Grafo en la Lista
            PARA CADA Usuario B
                  SI es una actividad autoasignada
                        A += peso de la autoasignación
                        B += peso de la autoasignación
                  SINO
                        A += distancia dijkstra a B
                        B += distancia dijkstra Desde B
                  FIN SI
            FIN PARA
            Se agrega al arreglo para usuario
                  = T / A + B
      FIN PARA
```

3.1.3 Análisis de los datos mediante Ventanas

Un *log* de eventos es transformado en una secuencia de datos, esta secuencia es el resultado del cálculo de las características definidas previamente. Esta secuencia de datos ***D***, puede ser considerada como una serie de tiempo de m valores, como se muestra en la *Ilustración 14*.

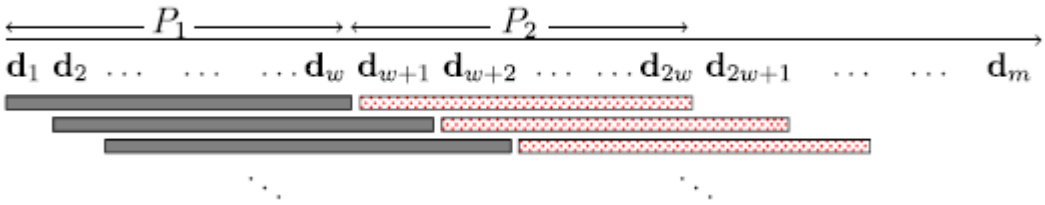

Ilustración 15 Idea básica para detección de Drifts utilizando Test de Hipótesis. El dataset de valores de características es considerado como una serie temporal para pruebas de hipótesis. P1 y P2 son dos poblaciones de tamaño w.

Fuente: Dealing with Concept Drifts in Process Minning (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014)

Cada valor ***d*** perteneciente al conjunto de datos ***D*** corresponde al valor de una característica para un caso. Para detectar *Concept Drift* se considera una serie de poblaciones sucesivas de valores de tamaño ***w***, el tamaño de ventana, e investigar si existe una diferencia significante entre dos sub-secuencias de poblaciones.

Una ventana movible de tamaño ***w*** es usada para generar las poblaciones, en la figura se muestra un escenario donde dos poblaciones $P_1 = d_1, d_2, \dots, d_w$ y $P_2 = d_{w+1}, d_{w+2}, \dots, d_w$ de tamaño ***w*** son consideradas. En la siguiente iteración la ventana recorrerá una posición teniendo como resultado las poblaciones correspondientes a $P_1 = d_2, d_3, \dots, d_{w+1}$ y $P_2 = d_{w+2}, d_{w+3}, \dots, d_{w+n}$.

Las dos poblaciones generadas se comparan mediante pruebas de hipótesis para cada par de actividad o par usuario actividad. El valor devuelto para cada par de actividad o par actividad usuario por la prueba de hipótesis se almacena para ser posteriormente graficado. Se considera la acumulación de grupos para permitir que las características de carga laboral y centralidad tengan mayor cantidad de datos para la aplicación.

Para graficar el ejemplo, consideramos:

- Tamaño de *Log*: 10 casos
- Tamaño de Grupo: 3
- Tamaño Ventana: 2
- Tamaño de Salto: 1
- Tamaño de Paso: 1

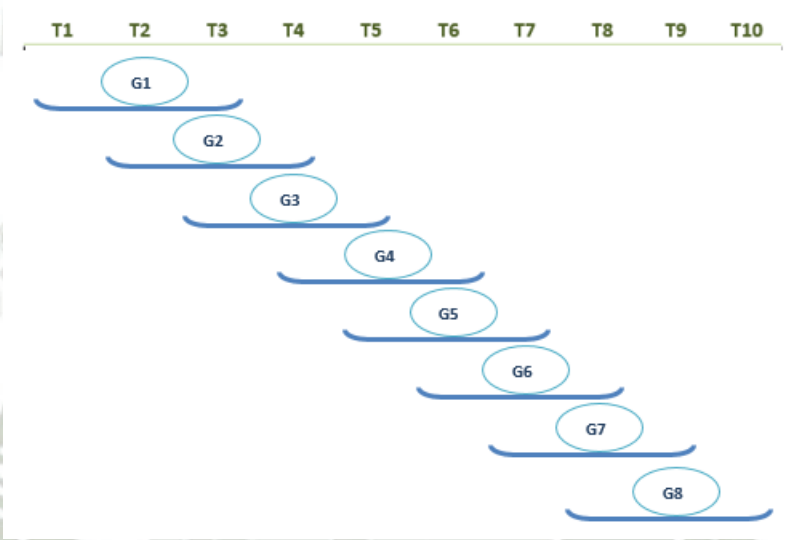

Ilustración 16 Agrupación de casos (Traces) por el algoritmo. Esta agrupación es necesaria para la aplicación de las técnicas de sociometría.
Fuente: Elaboración Propia.

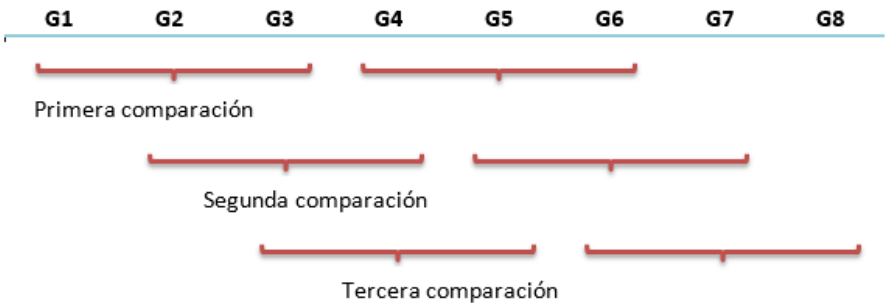

Ilustración 17 Evaluación por ventanas de los grupos de casos. Basado en el trabajo de Bose (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014).
Fuente: Elaboración Propia.

CAPÍTULO IV: ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS

4.1 CASO DE ESTUDIO

Para el análisis de datos utilizamos los gráficos proporcionados por el *plug-in* de ProM "Mine With Inductive Visual Miner" (Leemans, Fahland, & van der Aalst, 2014), como apoyo para mostrar el flujo de delegación de tareas entre los usuarios. En cuanto a los datos utilizados primero analizaremos si efectivamente las características propuestas son capaces de encontrar el *Concept Drift* en la perspectiva de usuarios mediante un *log* pequeño. Seguidamente analizaremos la granularidad y eficacia mediante un *log* de un tamaño más extenso. Finalmente se compararán los resultados de la perspectiva de flujo y de usuarios con el fin de determinar si existe alguna relación entre ambas.

4.1.1 Caso 1: *Log* Sintético Ejercicio 4

El *log* sintético, mostrado en el *Apéndice II*, está compuesto por 10 casos, los cuales tienen las siguientes actividades y usuarios mostrados en la *Tabla 9*.

Tabla 9 Usuarios y Actividades del Ejercicio 4

Usuarios	Actividades
A	Solicitud
B	Revisa_Sol
C	Aprueba_Sol
D	Genera_Datos
E	Cierra_Caso

Fuente: Elaboración Propia

En una ejecución normal del proceso, los usuarios ejecutan las tareas tal como se muestran en la *Tabla 9*, es decir, el usuario A ejecuta la actividad Solicitud, el usuario B la actividad Revisa_Sol y así sucesivamente.

El flujo de usuarios que grafica el *plug-in* de (Leemans, Fahland, & van der Aalst, 2014), corresponde a la *Ilustración 17*, en dicho gráfico se puede observar los diferentes caminos que suceden en el *log*. Mientras más coloreado sea el camino es que tiene un mayor uso o es el más común.

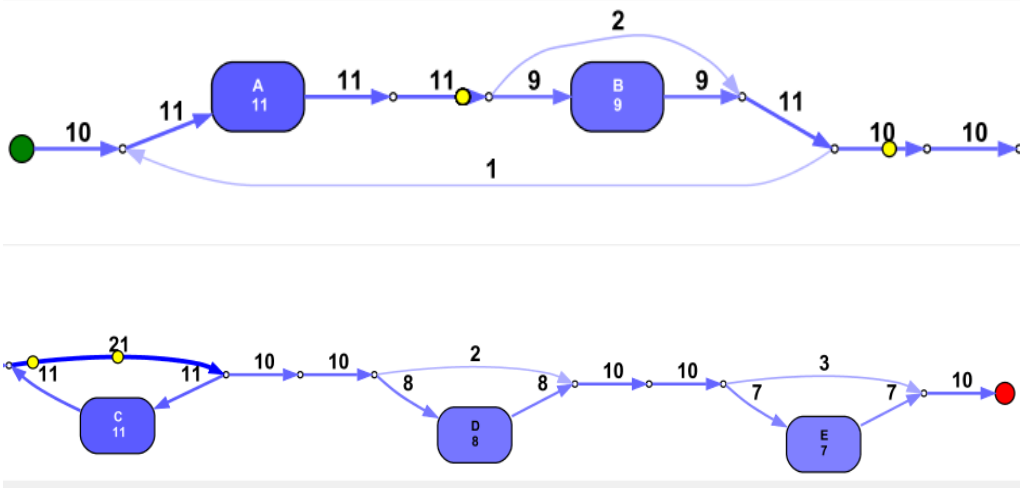

Ilustración 18 - Flujo de Usuarios Ejercicio 4
Fuente: Elaboración Propia mediante plug-in de ProM "Mine with Inductive visual Miner"

Tal como se aprecia en la *Ilustración 17*, el camino más usado es de A -> B-> C-> D-> E, sin embargo existen otros caminos donde se obvian usuarios, como es el caso de B, donde dos veces se omite a este usuario, D donde se omite dos veces a este usuario y E donde se omite 3 veces a este usuario. También se observa la cantidad de veces que un usuario es participa en el *log*, por ejemplo el usuario A, participa 11 veces mientras que el usuario E tiene una participación de 7. Este *plug-in* muestra los datos generales de todo el flujo como resumen gráfico, para hacer más comprensible el análisis.

Al mismo tiempo es posible observar más a detalle el flujo de cada caso. Como se ve en la *Ilustración 18*.

Ilustración 19 - Flujo de usuarios por casos
Fuente: Elaboración Propia mediante plug-in de ProM "Mine with Inductive visual Miner"

Los resultados a ser analizados, utilizando la perspectiva de usuarios se encuentran parametrizados como se muestra a continuación:

- Tamaño de Grupo: 3
- Tamaño de Ventana: 1
- Tamaño de población: 2
- Tipo de Prueba de hipótesis: Kolmogorov-Smirnov y Mann – Whitney

4.1.1.1 Cálculo de la Entropía

Como se describió en el *Capítulo IV*, la entropía se calcula en base a Usuario/Actividad. La *Tabla 10* muestra el conteo de actividades realizadas por cada usuario.

Tabla 10 Cantidad de veces que un usuario realiza una actividad

Usuario	Solicitud	Revisa Sol	Aprueba_Sol	Genera_Datos	Cierra_Caso
A	11	0	0	0	0
B	0	0	0	8	1
C	0	1	4	0	6
D	0	2	6	0	0
E	0	7	0	0	0

Fuente: Elaboración propia

Luego se realiza el análisis de la matriz A-S-N (A: Siempre, S: A veces, N: Nunca) en la *Tabla 11*. Se obtiene que los usuarios más predecibles o estables son los que realizan una sola actividad durante todo el *log*, esto se denota por la letra A (Siempre), sin embargo los usuarios B, C y D son los que realizan distintas actividades pero en casos muy específicos.

Tabla 11 Entropía Calculada Para el Ejercicio 4

Usuario	Solicitud	Revisa Sol	Aprueba_ Sol	Genera_ Datos	Cierra_ Caso	Entropía
A	A	N	N	N	N	0.72192809
B	N	N	N	S	S	0.97095059
C	N	S	S	N	S	0. 97095059
D	N	S	S	N	N	0.97095059
E	N	A	N	N	N	0.72192809

Fuente: Elaboración propia

Adicionalmente en la *Tabla 11* se muestra la entropía por usuario, teniendo como resultado que los usuarios más predecibles son los usuarios A, y E, estos usuarios también son considerados los más estables en el *log*. Los usuarios más difíciles de predecir o los que más cambian en cuanto a la frecuencia de actividades, son los usuarios B, C y D.

4.1.1.2 Cálculo distancia Minkowski

KOLMOGOROV-SMIRNOV

Aplicando la prueba de hipótesis Kolmogorov-Smirnov, se puede observar en la *Ilustración 19*, que en los índices 3 y 4, que engloban a los casos 3, 4, 5 y 4, 5, 6; respectivamente, ocurre un cambio.

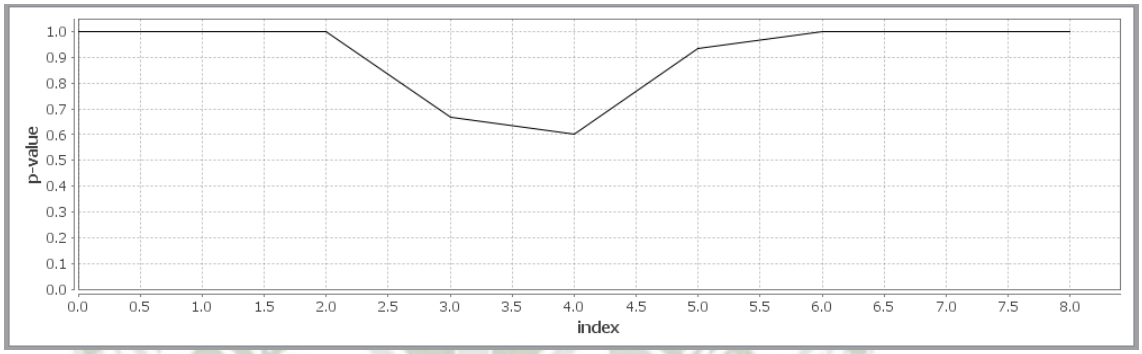

Ilustración 20 - Gráfica Distancia Minkowski con Test de Hipótesis Kolmogorov-Smirnov
Fuente: Elaboración Propia

De acuerdo al análisis del *log*, este resultado sería correcto, ya que en los casos 3, 4, 5, 6 el flujo de delegación de actividades entre usuarios cambia. En el grupo 3 el usuario B deja de participar en un caso, mientras que en el grupo 4 los usuarios C y D dejan de participar en un caso cada uno, como se observa en la *Tabla 12*. Esto origina que la carga laboral de estos usuarios varíe con respecto a los otros casos y se produzca un *Concept Drift*.

Tabla 12 - Resultados distancia Minkowski con Test de Hipótesis Kolmogorov-Smirnov

Índice	Probabilidad	Grupo	Casos/Grupo
0	1.0	0	*Case*7.0/*Case*1.0/*Case*2.0/
1	1.0	1	*Case*1.0/*Case*2.0/*Case*3.0/
2	1.0	2	*Case*2.0/*Case*3.0/*Case*4.0/
3	0.6666667	3	*Case*3.0/*Case*4.0/*Case*5.0/
4	0.6	4	*Case*4.0/*Case*5.0/*Case*6.0/
5	0.93333334	5	*Case*5.0/*Case*6.0/*Case*8.0/
6	1.0	6	*Case*6.0/*Case*8.0/*Case*9.0/
7	1.0	7	*Case*8.0/*Case*9.0/*Case*10.0/
8	1.0	8	*Case*9.0/*Case*10.0/null/

Fuente: Elaboración Propia

Su ejecución se demoró 5 milisegundos en obtener los resultados con el uso de kolmogorov smirnov al ejecutar el análisis.

MANN-WHITTNEY

Aplicando la prueba de hipótesis Mann-Whittney se observa que el cambio empieza a darse desde el índice 2 al índice 5 como se muestra en la *Ilustración 20*, que engloban los casos 2 al 8 (sin incluir el caso 7.0).

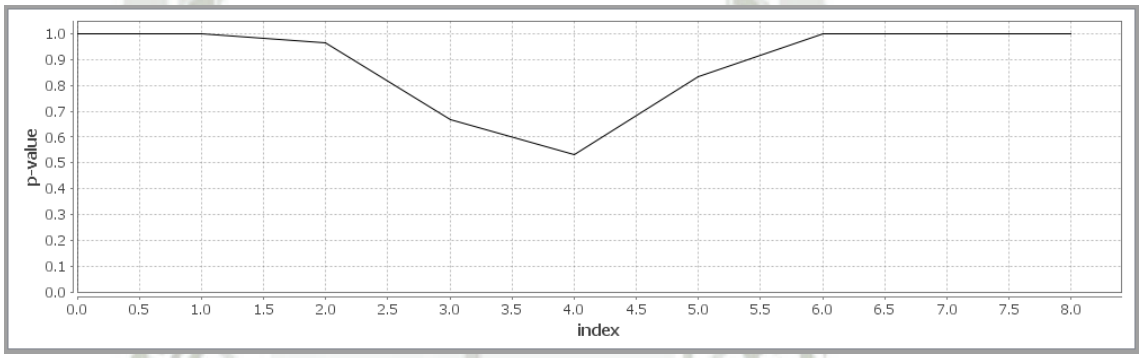

Ilustración 21 Gráfica Distancia Minkowski con Test de hipótesis Mann-Whittney
Fuente: Elaboración Propia

Los resultados mostrados en la gráfica se resumen en la *Tabla 13*.

Tabla 13 Resultados distancia Minkowski Test de hipótesis Mann-Whittney

Índice	Probabilidad	Grupo	Casos/Grupo
0	1.0	0	*Case*7.0/*Case*1.0/*Case*2.0/
1	1.0	1	*Case*1.0/*Case*2.0/*Case*3.0/
2	0.96666664	2	*Case*2.0/*Case*3.0/*Case*4.0/
3	0.6666667	3	*Case*3.0/*Case*4.0/*Case*5.0/
4	0.53333336	4	*Case*4.0/*Case*5.0/*Case*6.0/
5	0.8333333	5	*Case*5.0/*Case*6.0/*Case*8.0/
6	1.0	6	*Case*6.0/*Case*8.0/*Case*9.0/
7	1.0	7	*Case*8.0/*Case*9.0/*Case*10.0/
8	1.0	8	*Case*9.0/*Case*10.0/null/

Fuente: Elaboración Propia

Se puede observar que este test es más sensible a los cambios ya que todos los casos en la zona del cambio no siguen un patrón, sino que cada uno es diferente al anterior por una pequeña diferencia, esto se puede observar en la *Tabla 14.*

Tabla 14 Flujo de usuarios Casos 2 al 9

Caso	Flujo Usuarios
2.0	A, B, C, D, E
3.0	A, C, C, D
4.0	A, B, C, D, E
5.0	A, B, C, E
6.0	A, B, D ,E
8.0	A, B, C, C
9.0	A, B, C, D, E

Fuente: Elaboración Propia

CARGA DE LABORAL – CÁLCULO DISTANCIA MINKOWSKI

Para realizar un análisis más detallado mostramos a continuación la distancia Minkowski por cada usuario.

Tabla 15 DIstancia Minkowski para usuario A

Grupo de tiempo	A-B	A-C	A-D	A-E	Resultado
0	3	2.44726	2.44726	2	9.89452
1	2	3.37351	2.44726	2	9.82077
2	2	3.37351	2.44726	2	9.82077
3	2	3.1748	1.5874	2	8.7622
4	2.44726	1.5874	2	3	9.03466
5	2.44726	2.44726	1	2	7.89452
6	2.44726	2.08008	2	2	8.52734
7	3	2.85674	2	2	9.85674
8	2	2	2	2	8
					81.61152

Fuente: Elaboración Propia

Tabla 16 DIstancia Minkowski para usuario B

Grupo de tiempo	B-A	B-C	B-D	B-E	Resultado
0	3	4.33462	4.33462	4.00819	15.67743
1	2	4.33462	3.53872	3.1748	13.04814
2	2	4.33462	3.53872	3.1748	13.04814
3	2	4.16017	2.85674	3.1748	12.19171
4	2.44726	2.44726	3.53872	4.33462	12.76786
5	2.44726	3.88478	2.85674	3.53872	12.7275
6	2.44726	2.85674	3.53872	3.53872	12.38144
7	3	4.64919	4.00819	4.00819	15.66557
8	2	3.1748	3.1748	3.1748	11.5244
					119.03219

Fuente: Elaboración Propia

Tabla 17 DIstancia Minkowski para usuario C

Grupo de tiempo	C-A	C-B	C-D	C-E	Resultado
0	2.44726	4.33462	2.85674	3.53872	13.17734
1	3.37351	4.33462	3.72735	4.33462	15.7701
2	3.37351	4.33462	3.72735	4.33462	15.7701
3	3.1748	4.16017	2.85674	4.16017	14.35188
4	1.5874	2.44726	1.5874	3.72735	9.34941
5	2.44726	3.88478	1.5874	2.44726	10.3667
6	2.08008	2.85674	2.08008	2.08008	9.09698
7	2.85674	4.64919	2.85674	2.85674	13.21941
8	2	3.1748	3.1748	3.1748	11.5244
					112.62632

Fuente: Elaboración Propia

Tabla 18 DIstancia Minkowski para usuario D

Grupo de tiempo	D-A	D-B	D-C	D-E	Resultado
0	2.44726	4.33462	2.85674	2.44726	12.08588
1	2.44726	3.53872	3.72735	2.44726	12.16059
2	2.44726	3.53872	3.72735	2.44726	12.16059
3	1.5874	2.85674	2.85674	1.5874	8.88828
4	2	3.53872	1.5874	4.00819	11.13431
5	1	2.85674	1.5874	2.44726	7.8914
6	2	3.53872	2.08008	3.1748	10.7936
7	2	4.00819	2.85674	3.1748	12.03973
8	2	3.1748	3.1748	3.1748	11.5244
					98.67878

Fuente: Elaboración Propia

Tabla 19 DIstancia Minkowski para usuario E

Grupo de tiempo	E-A	E-B	E-C	E-D	Resultado
0	2	4.00819	3.53872	2.44726	11.99417
1	2	3.1748	4.33462	2.44726	11.95668
2	2	3.1748	4.33462	2.44726	11.95668
3	2	3.1748	4.16017	1.5874	10.92237
4	3	4.33462	3.72735	4.00819	15.07016
5	2	3.53872	2.44726	2.44726	10.43324
6	2	3.53872	2.08008	3.1748	10.7936
7	2	4.00819	2.85674	3.1748	12.03973
8	2	3.1748	3.1748	3.1748	11.5244
					106.69103

Fuente: Elaboración Propia

El usuario más difícil de clasificar es el usuario A, se puede inferir que su trabajo dentro del flujo es el menos parecido al del resto de usuarios por lo cual se entiende que es el usuario que cuenta con la carga laboral más especializada o exclusiva.

Analizando la *Ilustración 19* y las tablas anteriores podemos decir que el primer *Drift* se encuentra en los grupos 3 y 4. En la *Tabla 18* la distancia del usuario D varía en éstos dos grupos, de un valor de 8.88 a 11.13. Podemos inferir que es un cambio significativo, el usuario D cuenta con una distancia muy diferente con respecto a todos los usuarios ya que este realiza otra tarea o deja de participar.

4.1.1.3 Cálculo Bavelas-Leavitt

KOLMOGOROV-SMIRNOV

Como se aprecia en la *Ilustración 21* el cambio se produce en los grupos 3 y 4 como en los resultados de la característica anterior. Los resultados del gráfico se detallan en la *Tabla 20.*

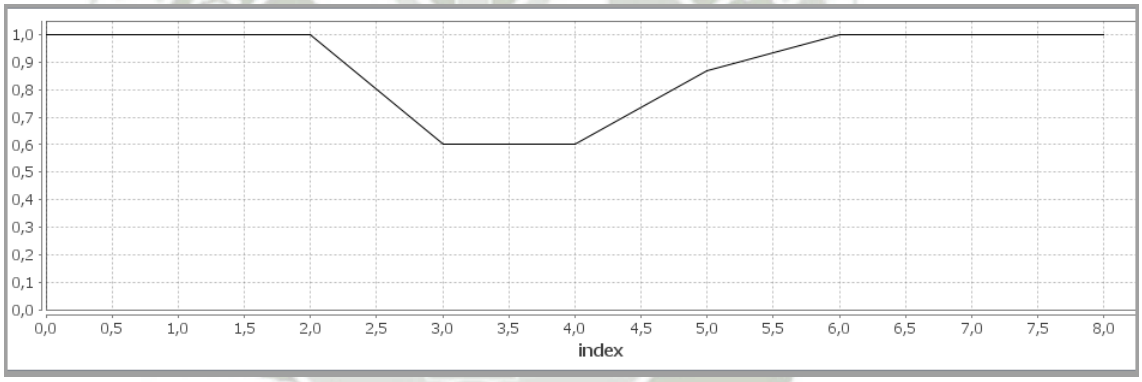

Ilustración 22 Grafica Bavelas Leavitt con Test de Hipótesis Kolmogorov-Smirnov
Fuente: Elaboración Propia

Analizando los resultados obtenidos con la *Tabla 14*, se encuentra que en el grupo 3 compuesto por los casos 3, 4 y 5, el usuario más central es C y el usuario E tiene muy baja participación. Mientras tanto en el grupo 4 compuesto por los casos 4, 5, 6 se obvia a los usuarios D y C en dos iteraciones respectivamente y los demás usuarios mantienen la misma centralidad.

Tabla 20 Datos Bavelas-Leavitt con Test de Hipótesis Kolmogorov-Smirnov

Índice	Probabilidad	Grupo	Casos/Grupo
0	1.0	0	*Case*7.0/*Case*1.0/*Case*2.0/
1	1.0	1	*Case*1.0/*Case*2.0/*Case*3.0/
2	1.0	2	*Case*2.0/*Case*3.0/*Case*4.0/
3	0.6	3	*Case*3.0/*Case*4.0/*Case*5.0/
4	0.6	4	*Case*4.0/*Case*5.0/*Case*6.0/
5	0.866666667	5	*Case*5.0/*Case*6.0/*Case*8.0/
6	1.0	6	*Case*6.0/*Case*8.0/*Case*9.0/
7	1.0	7	*Case*8.0/*Case*9.0/*Case*10.0/
8	1.0	8	*Case*9.0/*Case*10.0/null/

Fuente: Elaboración Propia

Su ejecución se demoró 3682 milisegundos en obtener los resultados con el uso de la prueba de hipótesis Kolmogorov Smirnov al ejecutar el análisis.

MANN-WHITTNEY

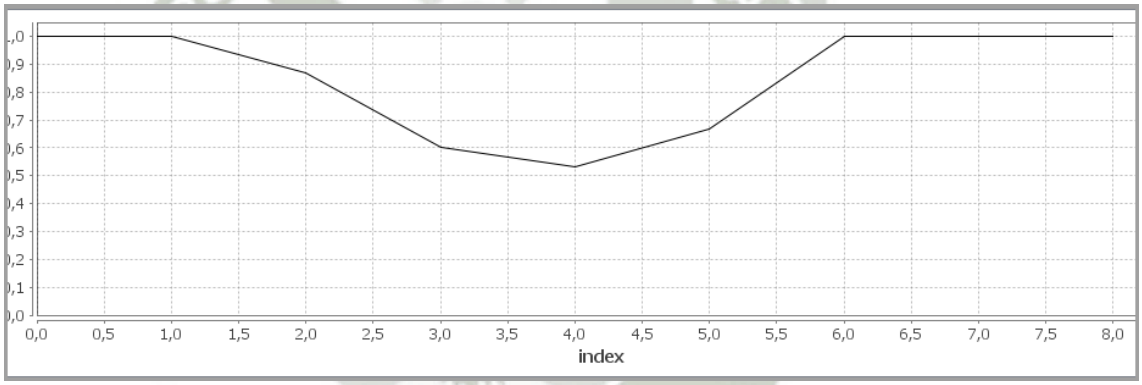

Ilustración 23 Gráfica Bavelas Leavitt con Test de Hipótesis Mann-Whittney
Fuente: Elaboración propia

Esta prueba de hipótesis detecta los *Concept Drift*s entre el grupo 2 al grupo 6, estos grupos se encuentran compuestos por los casos del 2 al 8, excepto el caso 7. Analizando los resultados obtenidos con la *Tabla 21*, la centralidad de los usuarios B y C cambia desde el grupo 2. En el grupo 3 el cambio más significativo es del usuario E, en el grupo 3 y 4 los usuarios C y E, en el grupo 5 existe un ligero cambio en la centralidad de casi todos los usuarios y en el grupo 6 en adelante la

centralidad se estabiliza. Ya que esta prueba de hipótesis es más sensible, es capaz de detectarlos.

Tabla 21 Datos Bavelas-Leavitt con Test de Hipótesis Mann-Whittney

Índice	Probabilidad	Grupo	Casos/Grupo
0	1.0	0	*Case*7.0/*Case*1.0/*Case*2.0/
1	1.0	1	*Case*1.0/*Case*2.0/*Case*3.0/
2	0.86666667	2	*Case*2.0/*Case*3.0/*Case*4.0/
3	0.6	3	*Case*3.0/*Case*4.0/*Case*5.0/
4	0.53333336	4	*Case*4.0/*Case*5.0/*Case*6.0/
5	0.66666667	5	*Case*5.0/*Case*6.0/*Case*8.0/
6	1.0	6	*Case*6.0/*Case*8.0/*Case*9.0/
7	1.0	7	*Case*8.0/*Case*9.0/*Case*10.0/
8	1.0	8	*Case*9.0/*Case*10.0/null/

Fuente: Elaboración propia

La centralidad calculada por la prueba de hipótesis Kolmogorov-Smirnov, encuentra un cambio en los grupos de casos 3 y 4 al igual que en el cálculo de la distancia Minkowski (Carga Laboral por usuario), además la prueba de hipótesis Mann-Whittney sigue siendo más sensible a los cambios en el *log* ya que detecta los *Drift*s desde el grupo 2 hasta el grupo 6.

La ejecución del algoritmo con este caso de prueba al utilizar la prueba de hipótesis de Kolmogorov-Smirnov para el cálculo de la distancia Minkowski demoró 4 milisegundos y para el cálculo de Bavelas – Leavitt demoró 1 milisegundo. En el caso de la prueba de hipótesis de Mann – Whitney tuvo una duración de 6 milisegundos para el cálculo de la distancia Minkowski y 1 milisegundo para el cálculo de Bavelas – Leavitt.

CENTRALIDAD POR USUARIOS – CÁLCULO BAVELAS-LEAVITT

Con la finalidad de analizar la centralidad de cada usuario y sustentar el *Concept Drift*, si consideramos que la fórmula de centralidad de Bavelas y la premisa de análisis del título de *Centralidad (página 40)*, se entiende que un usuario que

participa bastante en el flujo tendrá un índice más cercano a cero, mientras que no posea un coeficiente mayor. En la *Tabla 22* se muestran las centralidades por usuarios y su sumatoria.

Tabla 22 Centralidad por usuarios para el ejercicio 4

Grupo de Casos	A	B	C	D	E	Sumatoria de Centralidad
0	2.4	2	2	2.4	6	14.8
1	3.33333333	2.5	1.42857143	2	5	14.26190476
2	3.33333333	2.5	**1.42857143**	2	**5**	**14.26190476**
3	3	2.25	**1.28571429**	3	**4.5**	**14.03571429**
4	**3.33333333**	**1.66666667**	2.5	2.5	**3.333333333**	**13.33333333**
5	2.66666667	**1.33333333**	2	**4**	**4**	**14**
6	3	1.5	2.25	2.25	4.5	13.5
7	3.33333333	1.66666667	1.66666667	2.5	5	14.16666667
8	4	2	2	2	4	14

Fuente: Elaboración Propia

Utilizando el *plug-in* graficador anterior "Mine with Inductive visual Miner" vamos a verificar la zona de cambio que corresponde a los grupos del 2 al 6. Para demostrar los cambios en la centralidad de los usuarios y la concordancia con la *Tabla 22* anteriormente calculada.

Como podemos ver tanto en la *Ilustración 24* como en los resultados del Caso 2, se verifica que el usuario C es el que posee mayor interacción dentro del flujo. Mientras que el usuario A, que inicializa las tareas, posee un mayor índice y finalmente el usuario E, que finaliza las tareas, posee la mayor ponderación con respecto a la centralidad. Al mismo tiempo podemos ver que a partir de la *Ilustración 25* el valor de centralidad para el usuario D incrementa en un punto debido a que existen tareas que ya no son tomadas por el mismo. Luego en la *Ilustración 26* sucede el mismo comportamiento con el usuario C y el usuario B

se re-inserta. En la *Ilustración 27* se verifica se bifurcan las tareas entre los usuarios C y D. Para finalmente en la *Ilustración 28* se re estabiliza y se iguala al comportamiento de la *Ilustración 25*.

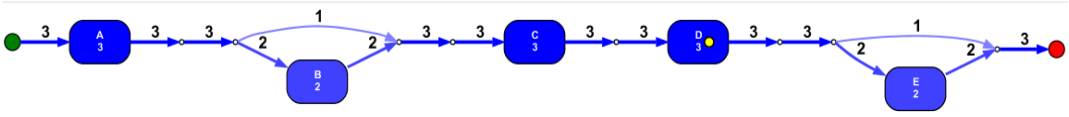

Ilustración 24 Ejecución de "Mine with Inductive Visual Miner" para el Caso 2
Fuente: Elaboración Propia

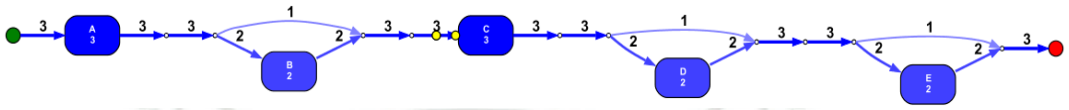

Ilustración 25 Ejecución de "Mine with Inductive Visual Miner" para el Caso 3
Fuente: Elaboración Propia

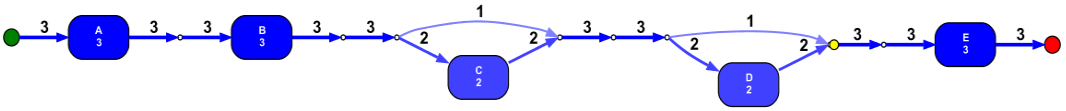

Ilustración 26 Ejecución de "Mine with Inductive Visual Miner" para el Caso 4
Fuente: Elaboración Propia

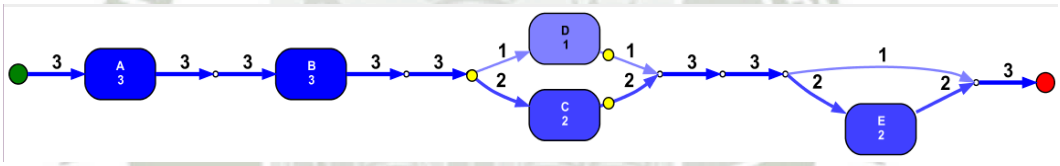

Ilustración 27 Ejecución de "Mine with Inductive Visual Miner" para el Caso 5
Fuente: Elaboración Propia

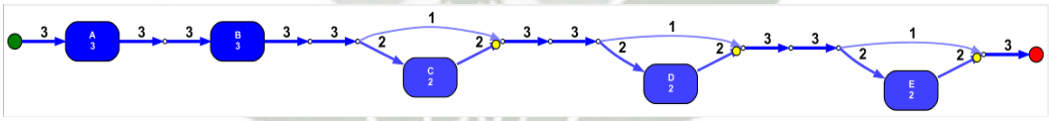

Ilustración 28 Ejecución de "Mine with Inductive Visual Miner" para el Caso 6
Fuente: Elaboración Propia

Se puede concluir que tanto los usuarios iniciales y finales poseerán una centralidad baja lo cual se considera normal, pero no se aplicaría el mismo criterio para los usuarios que participan al medio del flujo.

4.1.2 Caso 2: *Log* Sintético Ejercicio 5

Se ha tomado como referencia el Caso 5 de los *logs* de ejemplo que se encuentran en la página web de ProM (ProM, 2015). El *log* sintético está compuesto por 100 casos, los cuales tienen las siguientes actividades y usuarios mostrados en la *Tabla 23.*

Tabla 23 Usuarios y Actividades Ejercicio 5

Usuarios		**Actividades**	
Mike	Carol	Invite Reviewers	Time-out X
Pete	Pam	Time-out 1	Accept
__INVALID__	Sam	Time-out 2	Reject
John	Mary	Time-out 3	Decide
Sara		Get review 1	Invite Additional reviewer
Anne		Get review 2	Get Review X
Wil		Get review 3	Collect Reviews

Fuente: Elaboración Propia

El flujo de los usuarios se muestra en la *Ilustración 29*, se puede observar que los usuarios con mayor participación son Mike y Anne ya que los caminos más utilizados van a estos dos usuarios.

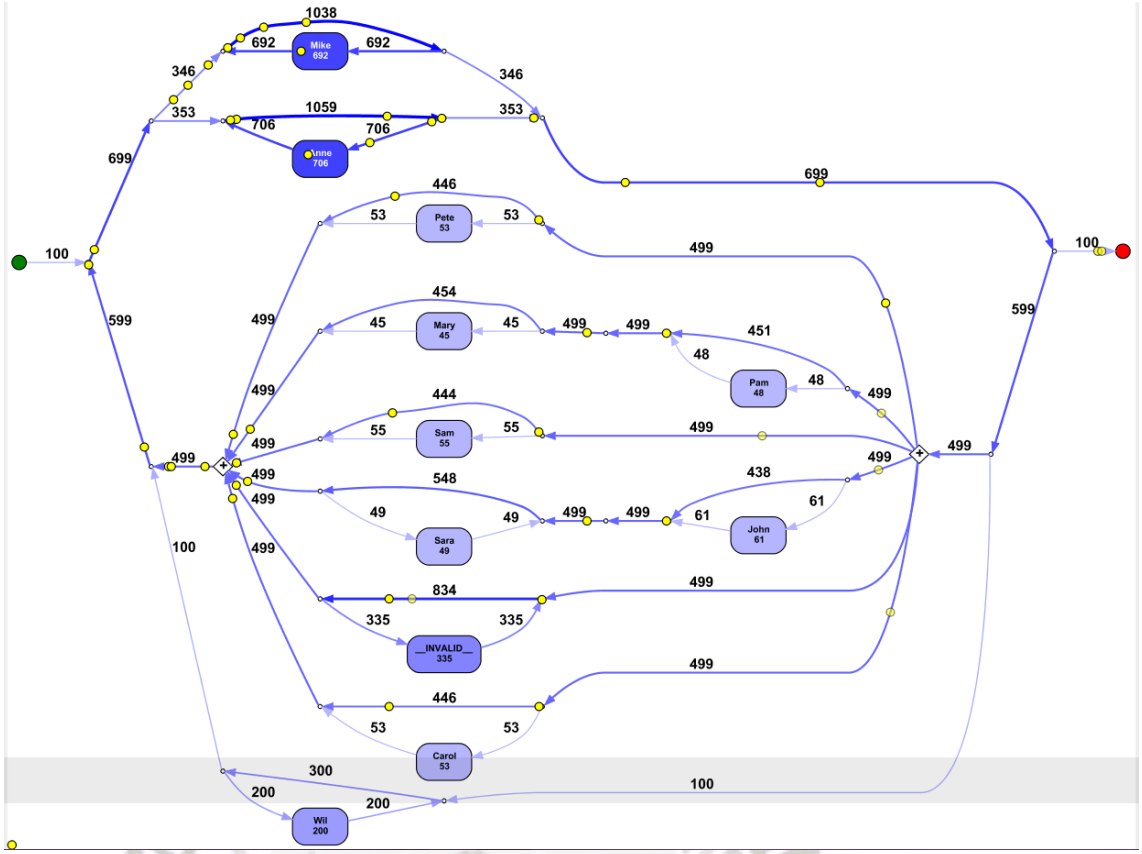

Ilustración 29 Flujo de Usuarios Ejercicio 5.
Fuente: Elaboración Propia mediante plug-in de ProM "Mine with Inductive visual Miner"

En este caso vamos a analizar los resultados obtenidos para la perspectiva de usuario variando el tamaño de grupo. En el Caso 1: *Log* Sintético Ejercicio 4 nos enfocamos a analizar si efectivamente se encuentran los puntos de cambio, ahora se analizarán los puntos de cambio encontrados dependiendo del tamaño de grupo que se defina.

El cálculo de la entropía no cambia si se toma otro tamaño de grupo, ya que la entropía se toma de forma global.

CÁLCULO ENTROPÍA

Tabla 24 Cálculo Entropía Ejercicio 5

Usuario	Entropía
Mike	0.9402859586706309
Pete	0.863120568566631
__INVALID__	0.863120568566631
John	0.863120568566631
Sara	0.863120568566631
Anne	0.9402859586706309
Wil	0.37123232664087563
Carol	0.863120568566631
Pam	0.863120568566631
Sam	0.863120568566631
Mary	0.863120568566631

Fuente: Elaboración Propia.

Como se puede ver en la *Tabla 24*, el usuario con más estabilidad en el *log*, es Wil, ya que este usuario generalmente hace una actividad; sin embargo los usuarios con un valor de entropía más elevado son Anne y Mike, ya que estos dos usuarios suelen hacer diferentes actividades.

4.1.2.1 Tamaño de Grupo: 1

A continuación se presenta el cálculo con un tamaño de grupo de 1, tanto para Minkowski en la *Ilustración 30* como para Bavelas en la *Ilustración 31,* es decir se analiza caso contra caso. Se encuentran muchos cambios desde el índice 20 al índice 79 en ambos cálculos. Se tienen cambios más significativos en el índice 70, en el caso de Bavelas-Leavitt, mientras que en la distancia Minkowski, hay un cambio más significativo en el índice 30.

CÁLCULO BAVELAS-LEAVITT

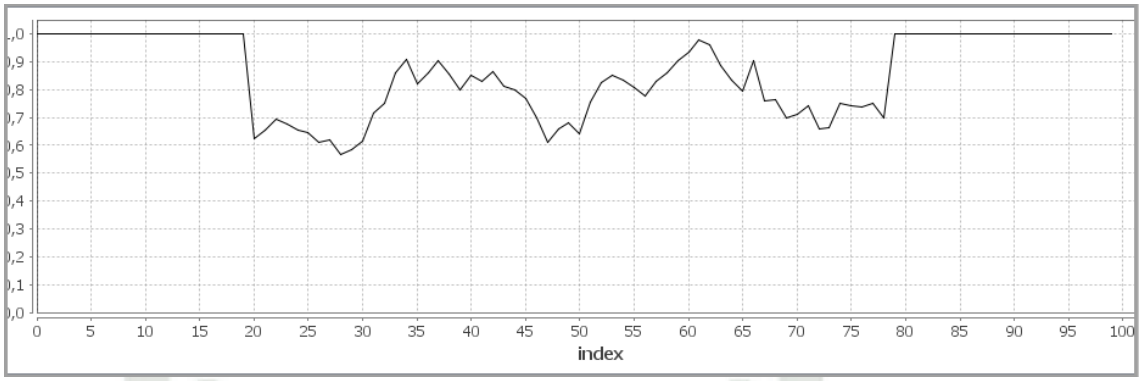

Ilustración 30 Resultados Acumulados de la prueba de hipótesis aplicada a la Bavelas-Leavitt con un tamaño de grupo de 1

DISTANCIA MINKOWSKI

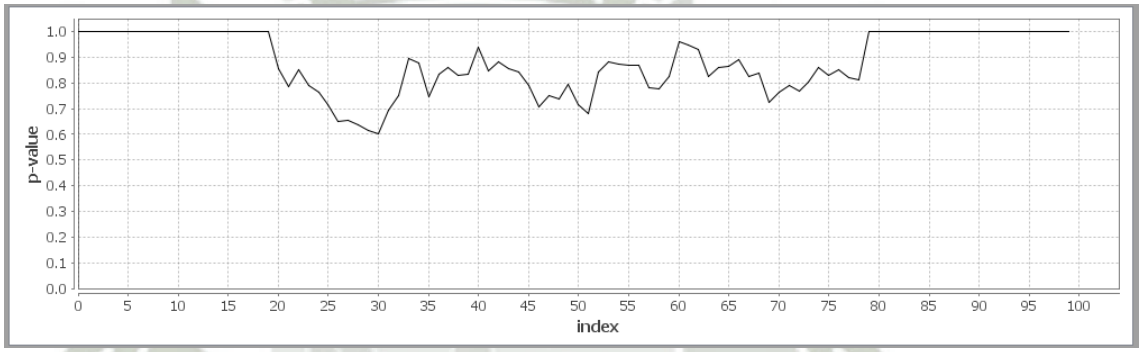

Ilustración 31 Resultados Acumulados de la prueba de hipótesis aplicada a la distancia Minkowski con un tamaño de grupo de 1
Fuente: Elaboración Propia

Podemos inferir entonces que la carga laboral no tiene muchas variaciones. Sin embargo en la centralidad de usuarios si se da un cambio importante en el índice 70, ya que el resultado es mucho más cercano a 0 como se observa en la *Tabla 25*.

Tabla 25 Resultados para un tamaño de grupo de 1

Distancia Minkowski			Bavelas-Leavitt		
Índice	**Probabilidad**	**Grupo de Casos**	**Índice**	**Probabilidad**	**Grupo de Casos**
0	1.0	0	**0**	1.0	0
1	1.0	1	**1**	1.0	1
2	1.0	2	**2**	1.0	2
3	1.0	3	**3**	1.0	3
4	1.0	4	**4**	1.0	4
5	1.0	5	**5**	1.0	5
6	1.0	6	**6**	1.0	6
7	1.0	7	**7**	1.0	7
8	1.0	8	**8**	1.0	8

9	1.0	9	**9**	1.0	9
10	1.0	10	**10**	1.0	10
11	1.0	11	**11**	1.0	11
12	1.0	12	**12**	1.0	12
13	1.0	13	**13**	1.0	13
14	1.0	14	**14**	1.0	14
15	1.0	15	**15**	1.0	15
16	1.0	16	**16**	1.0	16
17	1.0	17	**17**	1.0	17
18	1.0	18	**18**	1.0	18
19	1.0	19	**19**	1.0	19
20	0.8547981	20	**20**	0.6228841	20
21	0.78702724	21	**21**	0.654263	21
22	0.85121644	22	**22**	0.69549435	22
23	0.78918976	23	**23**	0.677957	23
24	0.7636675	24	**24**	0.65426373	24
25	0.7171854	25	**25**	0.6464955	25
26	0.6489955	26	**26**	0.6083356	26
27	0.6538522	27	**27**	0.61811143	27
28	0.63744915	28	**28**	0.5669332	28
29	0.6146887	29	**29**	0.58391535	29
30	0.6005361	30	**30**	0.6130615	30
31	0.69494134	31	**31**	0.71539414	31
32	0.7490344	32	**32**	0.74920774	32
33	0.8935988	33	**33**	0.8601497	33
34	0.87513703	34	**34**	0.908264	34
35	0.74645644	35	**35**	0.82028574	35
36	0.83200836	36	**36**	0.86088	36
37	0.8577053	37	**37**	0.9013811	37
38	0.8306121	38	**38**	0.8563594	38
39	0.8351839	39	**39**	0.7974909	39
40	0.94009525	40	**40**	0.8517352	40
41	0.8465535	41	**41**	0.82787234	41
42	0.8809509	42	**42**	0.86304545	42
43	0.8534419	43	**43**	0.8118665	43
44	0.8410556	44	**44**	0.7996564	44
45	0.78933334	45	**45**	0.7675745	45
46	0.70832866	46	**46**	0.6987552	46
47	0.750476	47	**47**	0.60940367	47
48	0.73775136	48	**48**	0.6588746	48
49	0.7934861	49	**49**	0.6782496	49
50	0.71412176	50	**50**	0.63875705	50
51	0.6811132	51	**51**	0.75425595	51
52	0.8414551	52	**52**	0.8230713	52
53	0.88102126	53	**53**	0.8507574	53
54	0.8717901	54	**54**	0.833395	54
55	0.8687353	55	**55**	0.8089708	55
56	0.86719656	56	**56**	0.7753293	56
57	0.78029484	57	**57**	0.8270628	57
58	0.7773802	58	**58**	0.8576124	58
59	0.8253026	59	**59**	0.9044457	59
60	0.96163243	60	**60**	0.93582857	60
61	0.9462052	61	**61**	0.97858834	61
62	0.9283209	62	**62**	0.9617819	62

63	0.82572025	63	**63**	0.885376	63
64	0.86167926	64	**64**	0.8326647	64
65	0.86549234	65	**65**	0.793696	65
66	0.89210176	66	**66**	0.90221435	66
67	0.82399464	67	**67**	0.7585228	67
68	0.83781296	68	**68**	0.7628684	68
69	0.7233179	69	**69**	0.6978433	69
70	0.7615596	70	**70**	0.71116406	70
71	0.789934	71	**71**	0.7428744	71
72	0.769036	72	**72**	0.66031265	72
73	0.8038551	73	**73**	0.66377616	73
74	0.8605752	74	**74**	0.7521899	74
75	0.8304228	75	**75**	0.74360186	75
76	0.8502311	76	**76**	0.73744524	76
77	0.81903297	77	**77**	0.7483086	77
78	0.8124099	78	**78**	0.69965434	78
79	1.0	79	**79**	1.0	79
80	1.0	80	**80**	1.0	80
81	1.0	81	**81**	1.0	81
82	1.0	82	**82**	1.0	82
83	1.0	83	**83**	1.0	83
84	1.0	84	**84**	1.0	84
85	1.0	85	**85**	1.0	85
86	1.0	86	**86**	1.0	86
87	1.0	87	**87**	1.0	87
88	1.0	88	**88**	1.0	88
89	1.0	89	**89**	1.0	89
90	1.0	90	**90**	1.0	90
91	1.0	91	**91**	1.0	91
92	1.0	92	**92**	1.0	92
93	1.0	93	**93**	1.0	93
94	1.0	94	**94**	1.0	94
95	1.0	95	**95**	1.0	95
96	1.0	96	**96**	1.0	96
97	1.0	97	**97**	1.0	97
98	1.0	98	**98**	1.0	98
99	1.0	99	**99**	1.0	99

Fuente: Elaboración Propia

Al momento de la ejecución con el tamaño 1 se obtuvieron resultados de 79 milisegundos a la ejecución con Kolmogorov Smirnoff al momento de calcular la carga laboral y 3 milisegundos para la obtención de la centralidad.

4.1.2.2 Tamaño de Grupo: 3

Se ha optado por mostrar los resultados considerando las pruebas de hipótesis de Kolmogorov-Smirnov y Mann-Whittney, podemos ver en ambas gráficas *Ilustración 32* e *Ilustración 33* poseen la misma forma. Sin embargo con la prueba

de hipótesis Kolmogorov-Smirnov los picos son menos fluidos o con mayores diferencias a comparación del segundo.

DISTANCIA MINKOWSKI

Kolmogorov-Smirnov

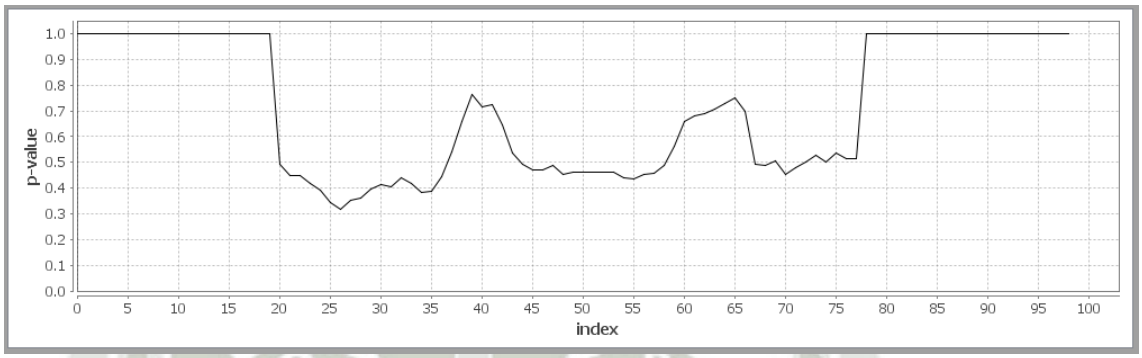

Ilustración 32 Resultados Acumulados de la prueba de hipótesis aplicada a la distancia Minkowski con un tamaño de grupo de 3 y el Test Kolmogorov-Smirnov
Fuente: Elaboración Propia

Mann-Whittney

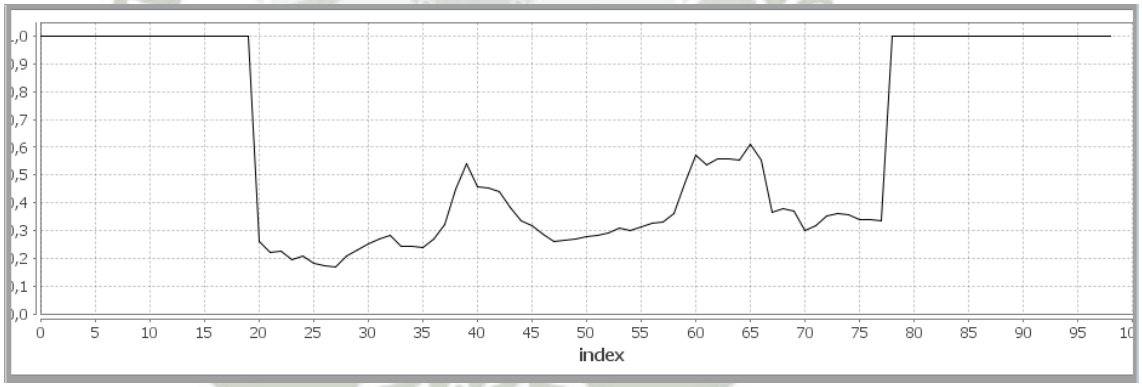

Ilustración 33 Resultados Acumulados de la prueba de hipótesis aplicada a la distancia Minkowski con un tamaño de grupo de 3 y el Test Mann Whitney
Fuente: Elaboración Propia

Seguidamente se prueba el cálculo de Bavelas para ambas pruebas en la *Ilustración 34* e *Ilustración 35*. Lo cual demuestra que la prueba de hipótesis Mann-Whittney es más exacta y mostrando una gráfica más fluida. Sin embargo, se comprobó que el tiempo de procesamiento del test de Kolmogorov-Smirnov

es mucho menor que el de Mann-Whittney, mientras la primera demoró un par de minutos, la segunda tiene un tiempo aproximado de una hora.

BAVELAS-LEAVITT

Kolmogorov-Smirnov

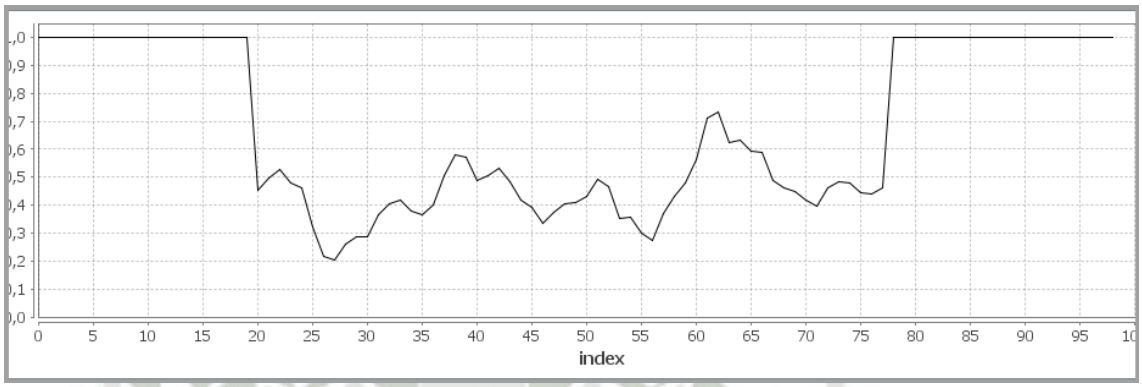

Ilustración 34 Resultados Acumulados de la prueba de hipótesis aplicada a la Bavelas-Leavitt con un tamaño de grupo de 3 y el test Test Kolmogorov-Smirnov
Fuente: Elaboración Propia

Mann-Whittney

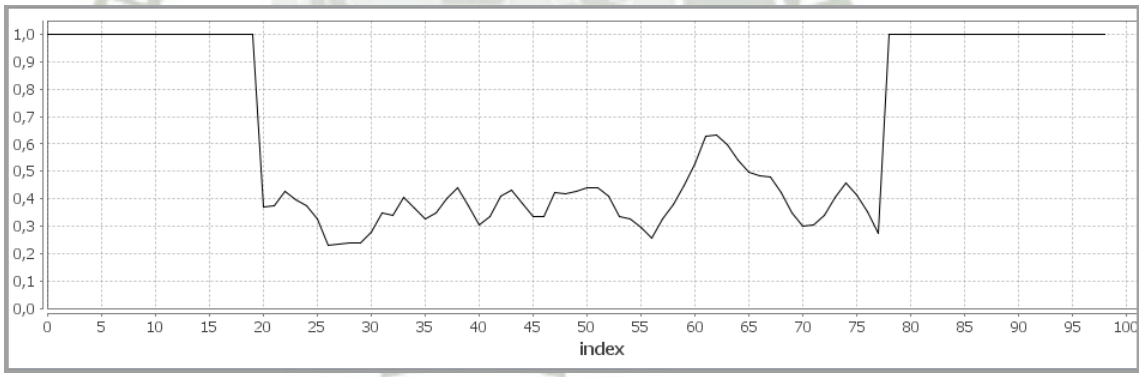

Ilustración 35 Resultados Acumulados de la prueba de hipótesis aplicada a la Bavelas-Leavitt con un tamaño de grupo de 3 y el test Test Mann-Whittney
Fuente: Elaboración Propia

Tabla 26 Resultados para un tamaño de grupo de 3

Bavelas-Leavitt			Distancia Minkowski		
Índice	**Probabilidad**	**Grupo de Casos**	**Índice**	**Probabilidad**	**Grupo de Casos**
0	1.0	0	**0**	1.0	0
1	1.0	1	**1**	1.0	1
2	1.0	2	**2**	1.0	2
3	1.0	3	**3**	1.0	3
4	1.0	4	**4**	1.0	4
5	1.0	5	**5**	1.0	5
6	1.0	6	**6**	1.0	6

7	1.0	7	**7**	1.0	7
8	1.0	8	**8**	1.0	8
9	1.0	9	**9**	1.0	9
10	1.0	10	**10**	1.0	10
11	1.0	11	**11**	1.0	11
12	1.0	12	**12**	1.0	12
13	1.0	13	**13**	1.0	13
14	1.0	14	**14**	1.0	14
15	1.0	15	**15**	1.0	15
16	1.0	16	**16**	1.0	16
17	1.0	17	**17**	1.0	17
18	1.0	18	**18**	1.0	18
19	1.0	19	**19**	1.0	19
20	0.4532798	20	**20**	0.493653	20
21	0.49840513	21	**21**	0.4484254	21
22	0.5255196	22	**22**	0.44849023	22
23	0.47907335	23	**23**	0.41632447	23
24	0.46095553	24	**24**	0.38989004	24
25	0.32033947	25	**25**	0.34178168	25
26	0.21611331	26	**26**	0.31742913	26
27	0.2032764	27	**27**	0.35435027	27
28	0.26206282	28	**28**	0.36280674	28
29	0.28873348	29	**29**	0.39421183	29
30	0.28718767	30	**30**	0.41280094	30
31	0.3632981	31	**31**	0.40313482	31
32	0.4040189	32	**32**	0.441285	32
33	0.4163502	33	**33**	0.41565627	33
34	0.380428	34	**34**	0.38295227	34
35	0.3666509	35	**35**	0.386037	35
36	0.40118855	36	**36**	0.44228747	36
37	0.5049717	37	**37**	0.5393821	37
38	0.5790357	38	**38**	0.6575277	38
39	0.5693899	39	**39**	0.76151973	39
40	0.487184	40	**40**	0.7131324	40
41	0.5065324	41	**41**	0.72587657	41
42	0.5303913	42	**42**	0.64592654	42
43	0.48193926	43	**43**	0.53451425	43
44	0.4193064	44	**44**	0.49210867	44
45	0.39368123	45	**45**	0.47153732	45
46	0.33421904	46	**46**	0.46975335	46
47	0.37247303	47	**47**	0.48916146	47
48	0.4047208	48	**48**	0.45200402	48
49	0.4071993	49	**49**	0.46373308	49
50	0.42934316	50	**50**	0.46216175	50
51	0.49113768	51	**51**	0.46322373	51
52	0.4670732	52	**52**	0.4628536	52
53	0.35421368	53	**53**	0.46341166	53
54	0.35671836	54	**54**	0.43839434	54
55	0.29967147	55	**55**	0.4363825	55
56	0.27446273	56	**56**	0.45219147	56
57	0.36930293	57	**57**	0.4574284	57
58	0.4306047	58	**58**	0.48618573	58
59	0.48073882	59	**59**	0.56218195	59
60	0.5639464	60	**60**	0.6566471	60

61	0.7101604	61	**61**	0.6812086	61
62	0.7313945	62	**62**	0.689747	62
63	0.6228491	63	**63**	0.70801985	63
64	0.6328277	64	**64**	0.727561	64
65	0.5908281	65	**65**	0.7503449	65
66	0.5900192	66	**66**	0.6995487	66
67	0.4894657	67	**67**	0.49361974	67
68	0.46218583	68	**68**	0.48608568	68
69	0.4476155	69	**69**	0.50724894	69
70	0.41617623	70	**70**	0.45241255	70
71	0.3957517	71	**71**	0.4793377	71
72	0.4606514	72	**72**	0.5003285	72
73	0.48210227	73	**73**	0.5257551	73
74	0.48117304	74	**74**	0.49945593	74
75	0.4454493	75	**75**	0.5359803	75
76	0.4410264	76	**76**	0.5126558	76
77	0.45957112	77	**77**	0.51190966	77
78	1.0	78	**78**	1.0	78
79	1.0	79	**79**	1.0	79
80	1.0	80	**80**	1.0	80
81	1.0	81	**81**	1.0	81
82	1.0	82	**82**	1.0	82
83	1.0	83	**83**	1.0	83
84	1.0	84	**84**	1.0	84
85	1.0	85	**85**	1.0	85
86	1.0	86	**86**	1.0	86
87	1.0	87	**87**	1.0	87
88	1.0	88	**88**	1.0	88
89	1.0	89	**89**	1.0	89
90	1.0	90	**90**	1.0	90
91	1.0	91	**91**	1.0	91
92	1.0	92	**92**	1.0	92
93	1.0	93	**93**	1.0	93
94	1.0	94	**94**	1.0	94
95	1.0	95	**95**	1.0	95
96	1.0	96	**96**	1.0	96
97	1.0	97	**97**	1.0	97
98	1.0	98	**98**	1.0	98

Fuente: Elaboración Propia

En la *Tabla 27* se puede observar los tiempos de ejecución para las pruebas de hipótesis Kolmogorov Smirnov y Mann – Whittney para un tamaño de gurpo de 3.

Tabla 27 - Tiempo de ejecución de las pruebas de hipótesis

	Kolmogorov - Smirnoff	**Mann - Whittney**
Distancia Minkowski	79 milisegundos	1288316 milisegundos
Bavelas-Leavitt	4 milisegundos	140298 milisegundos

Fuente: Elaboración Propia

4.1.2.3 Tamaño de Grupo: 7

En comparación con las gráficas con tamaño de grupo 1 o 3, al utilizar un tamaño de grupo de 7 se observan zonas de cambio mucho más amplias, mostrando solo cambios más drásticos en el *log*.

DISTANCIA MINKOWSKI

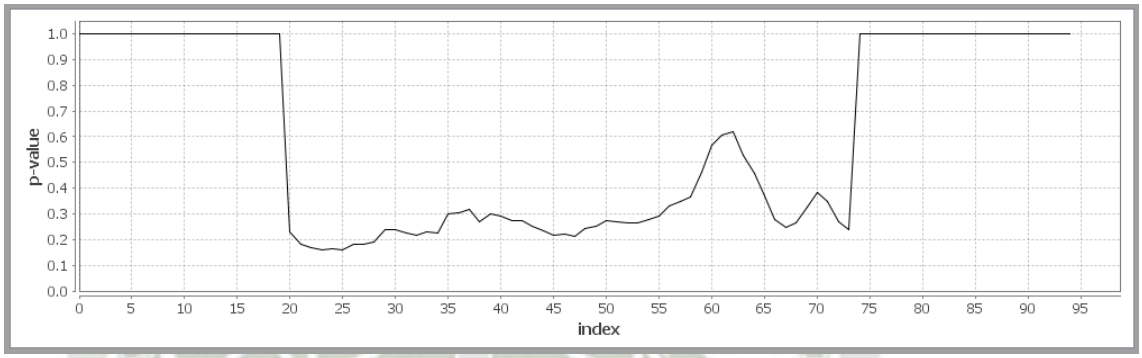

Ilustración 36 Resultados Acumulados de la prueba de hipótesis aplicada a la distancia Minkowski con un tamaño de grupo de 7
Fuente: Elaboración Propia

BAVELAS-LEAVITT

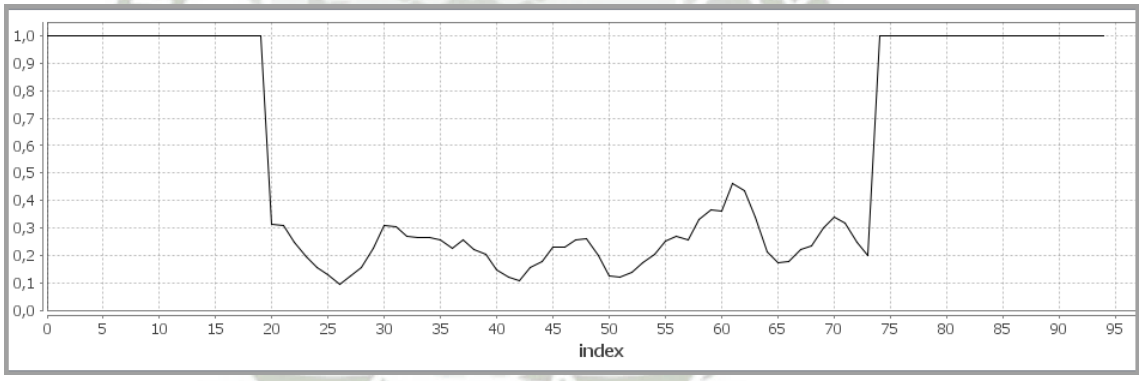

Ilustración 37 Resultados Acumulados de la prueba de hipótesis aplicada a la Bavelas-Leavitt con un tamaño de grupo de 7
Fuente: Elaboración Propia

Podemos decir que para el caso de la distancia Minkowski en la *Ilustración 36* existe una zona de cambio desde el índice 20 hasta el 60, que mantienen las mismas características, sin embargo entre el índice 60 al 68 se produce un cambio más significativo. Al comparar ambas ilustraciones se encuentran las diferencias entre la carga laboral y la centralidad. La ejecución demoró 99

milisegundos para obtener el cálculo de Minkowski y 5 milisegundos para encontrar el cálculo de Bavelas.

Tabla 28 Resultados para un tamaño de grupo de 7

Bavelas-Leavitt			Distancia Minkowski		
Índice	**Probabilidad**	**Grupo de Casos**	**Índice**	**Probabilidad**	**Grupo de Casos**
0	1.0	0	**0**	1.0	0
1	1.0	1	**1**	1.0	1
2	1.0	2	**2**	1.0	2
3	1.0	3	**3**	1.0	3
4	1.0	4	**4**	1.0	4
5	1.0	5	**5**	1.0	5
6	1.0	6	**6**	1.0	6
7	1.0	7	**7**	1.0	7
8	1.0	8	**8**	1.0	8
9	1.0	9	**9**	1.0	9
10	1.0	10	**10**	1.0	10
11	1.0	11	**11**	1.0	11
12	1.0	12	**12**	1.0	12
13	1.0	13	**13**	1.0	13
14	1.0	14	**14**	1.0	14
15	1.0	15	**15**	1.0	15
16	1.0	16	**16**	1.0	16
17	1.0	17	**17**	1.0	17
18	1.0	18	**18**	1.0	18
19	1.0	19	**19**	1.0	19
20	0.31217748	20	**20**	0.2317534	20
21	0.30871198	21	**21**	0.18093105	21
22	0.24766499	22	**22**	0.16730006	22
23	0.19293706	23	**23**	0.1605428	23
24	0.15473086	24	**24**	0.1628843	24
25	0.12915792	25	**25**	0.15834437	25
26	0.09577276	26	**26**	0.18229368	26
27	0.12643883	27	**27**	0.18302374	27
28	0.15734468	28	**28**	0.19159599	28
29	0.22561434	29	**29**	0.23811178	29
30	0.31014994	30	**30**	0.24041618	30
31	0.3052461	31	**31**	0.2258783	31
32	0.26689512	32	**32**	0.21704803	32
33	0.26345548	33	**33**	0.23126742	33
34	0.26262122	34	**34**	0.22741325	34
35	0.2569705	35	**35**	0.29996794	35
36	0.22610061	36	**36**	0.304481	36
37	0.25583616	37	**37**	0.3165299	37
38	0.22197403	38	**38**	0.27025336	38
39	0.20411544	39	**39**	0.29979688	39
40	0.14867897	40	**40**	0.28930885	40
41	0.12081195	41	**41**	0.27347177	41
42	0.10883855	42	**42**	0.2714136	42
43	0.15352552	43	**43**	0.2537237	43
44	0.17553921	44	**44**	0.23437035	44
45	0.2316	45	**45**	0.21767877	45
46	0.22977662	46	**46**	0.21882507	46

47	0.25455302	47	**47**	0.2141105	47
48	0.26190975	48	**48**	0.24157812	48
49	0.20023222	49	**49**	0.25104454	49
50	0.12569742	50	**50**	0.27553606	50
51	0.12076713	51	**51**	0.27064073	51
52	0.13784268	52	**52**	0.26284173	52
53	0.1718684	53	**53**	0.26529837	53
54	0.20493867	54	**54**	0.27804306	54
55	0.25312385	55	**55**	0.29310003	55
56	0.26858294	56	**56**	0.33026066	56
57	0.25400126	57	**57**	0.3482213	57
58	0.32905573	58	**58**	0.3661046	58
59	0.36512873	59	**59**	0.46109262	59
60	0.35994473	60	**60**	0.56675124	60
61	0.46132147	61	**61**	0.60614604	61
62	0.43448997	62	**62**	0.6203112	62
63	0.3389946	63	**63**	0.5281144	63
64	0.21251456	64	**64**	0.45622724	64
65	0.17416672	65	**65**	0.36983037	65
66	0.17879085	66	**66**	0.27937722	66
67	0.21875712	67	**67**	0.2460779	67
68	0.23332204	68	**68**	0.26477686	68
69	0.3013923	69	**69**	0.32622474	69
70	0.33958623	70	**70**	0.3808352	70
71	0.31631064	71	**71**	0.34994242	71
72	0.24802409	72	**72**	0.26903203	72
73	0.19893484	73	**73**	0.2378158	73
74	1.0	74	**74**	1.0	74
75	1.0	75	**75**	1.0	75
76	1.0	76	**76**	1.0	76
77	1.0	77	**77**	1.0	77
78	1.0	78	**78**	1.0	78
79	1.0	79	**79**	1.0	79
80	1.0	80	**80**	1.0	80
81	1.0	81	**81**	1.0	81
82	1.0	82	**82**	1.0	82
83	1.0	83	**83**	1.0	83
84	1.0	84	**84**	1.0	84
85	1.0	85	**85**	1.0	85
86	1.0	86	**86**	1.0	86
87	1.0	87	**87**	1.0	87
88	1.0	88	**88**	1.0	88
89	1.0	89	**89**	1.0	89
90	1.0	90	**90**	1.0	90
91	1.0	91	**91**	1.0	91
92	1.0	92	**92**	1.0	92
93	1.0	93	**93**	1.0	93
94	1.0	94	**94**	1.0	94

Fuente: Elaboración Propia

4.1.2.4 Tamaño de Grupo: 14

DISTANCIA MINKOWSKI

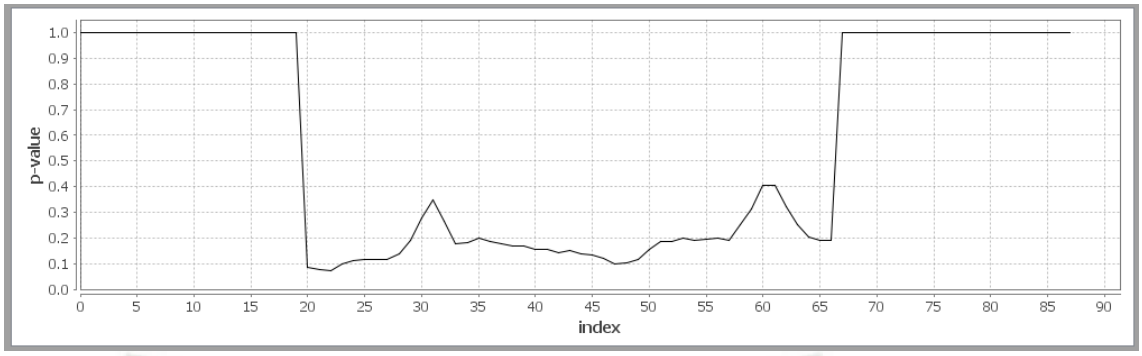

Ilustración 38 Resultados Acumulados de la prueba de hipótesis aplicada a la distancia Minkowski con un tamaño de grupo de 14
Fuente: Elaboración Propia

BAVELAS-LEAVITT

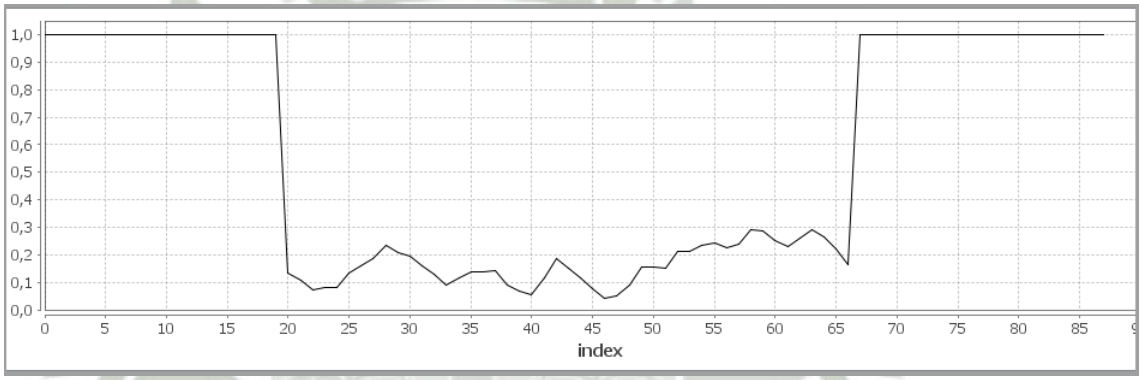

Ilustración 39 Resultados Acumulados de la prueba de hipótesis aplicada a la Bavelas-Leavitt con un tamaño de grupo de 14
Fuente: Elaboración Propia

Se puede observar en el cálculo de la distancia Minkowski en la *Ilustración 38* que existen dos zonas de cambio más importantes entre los índices 20 al 27 y 33 al 60. En cambio el cálculo de Bavelas-Leavitt en la *Ilustración 39* sigue con un patrón similar que en los otros cálculos de tamaño de grupo, cada vez resaltando menos los cambios en el *log*. Para este caso se obtuvo una ejecución de 74 milisegundos para la obtención de Minkowski y 3 milisegundos para la obtención de datos de Bavelas.

Tabla 29 Resultados para un tamaño de grupo de 14

Bavelas-Leavitt			Distancia Minkowski		
Índice	**Probabilidad**	**Grupo de Casos**	**Índice**	**Probabilidad**	**Grupo de Casos**
0	1.0	0	**0**	1.0	0
1	1.0	1	**1**	1.0	1
2	1.0	2	**2**	1.0	2
3	1.0	3	**3**	1.0	3
4	1.0	4	**4**	1.0	4
5	1.0	5	**5**	1.0	5
6	1.0	6	**6**	1.0	6
7	1.0	7	**7**	1.0	7
8	1.0	8	**8**	1.0	8
9	1.0	9	**9**	1.0	9
10	1.0	10	**10**	1.0	10
11	1.0	11	**11**	1.0	11
12	1.0	12	**12**	1.0	12
13	1.0	13	**13**	1.0	13
14	1.0	14	**14**	1.0	14
15	1.0	15	**15**	1.0	15
16	1.0	16	**16**	1.0	16
17	1.0	17	**17**	1.0	17
18	1.0	18	**18**	1.0	18
19	1.0	19	**19**	1.0	19
20	0.13281794	20	**20**	0.0852627	20
21	0.105694875	21	**21**	0.07485298	21
22	0.07243004	22	**22**	0.07435652	22
23	0.0819443	23	**23**	0.09953274	23
24	0.082097284	24	**24**	0.112993546	24
25	0.131653	25	**25**	0.11682783	25
26	0.1599078	26	**26**	0.116895854	26
27	0.18457979	27	**27**	0.115453124	27
28	0.23189576	28	**28**	0.1378763	28
29	0.2061397	29	**29**	0.19015476	29
30	0.19606476	30	**30**	0.2791056	30
31	0.161779	31	**31**	0.3462838	31
32	0.12870896	32	**32**	0.26528263	32
33	0.0900305	33	**33**	0.17628226	33
34	0.117488764	34	**34**	0.18237488	34
35	0.13925622	35	**35**	0.19703582	35
36	0.13937403	36	**36**	0.18528782	36
37	0.14163925	37	**37**	0.17530449	37
38	0.08970373	38	**38**	0.16720906	38
39	0.0696851	39	**39**	0.16681531	39
40	0.05685338	40	**40**	0.15587322	40
41	0.11450992	41	**41**	0.15450501	41
42	0.18759805	42	**42**	0.14200841	42
43	0.15290749	43	**43**	0.15190463	43
44	0.1141263	44	**44**	0.13953713	44
45	0.07783742	45	**45**	0.13523641	45
46	0.043279953	46	**46**	0.119062774	46
47	0.052416734	47	**47**	0.09692567	47
48	0.090360194	48	**48**	0.10113998	48
49	0.15737903	49	**49**	0.116435595	49
50	0.15465046	50	**50**	0.1537346	50

51	0.15237315	51	**51**	0.18401027	51
52	0.21072653	52	**52**	0.18539919	52
53	0.21147886	53	**53**	0.19936863	53
54	0.23617259	54	**54**	0.18849427	54
55	0.24444114	55	**55**	0.19294275	55
56	0.22466254	56	**56**	0.19872557	56
57	0.23684472	57	**57**	0.18852083	57
58	0.2896131	58	**58**	0.25208524	58
59	0.28851545	59	**59**	0.31195796	59
60	0.25271332	60	**60**	0.40378088	60
61	0.23088202	61	**61**	0.40641072	61
62	0.26196715	62	**62**	0.3214732	62
63	0.28911063	63	**63**	0.25148097	63
64	0.2632598	64	**64**	0.20396073	64
65	0.2212486	65	**65**	0.19041443	65
66	0.16421582	66	**66**	0.18864888	66
67	1.0	67	**67**	1.0	67
68	1.0	68	**68**	1.0	68
69	1.0	69	**69**	1.0	69
70	1.0	70	**70**	1.0	70
71	1.0	71	**71**	1.0	71
72	1.0	72	**72**	1.0	72
73	1.0	73	**73**	1.0	73
74	1.0	74	**74**	1.0	74
75	1.0	75	**75**	1.0	75
76	1.0	76	**76**	1.0	76
77	1.0	77	**77**	1.0	77
78	1.0	78	**78**	1.0	78
79	1.0	79	**79**	1.0	79
80	1.0	80	**80**	1.0	80
81	1.0	81	**81**	1.0	81
82	1.0	82	**82**	1.0	82
83	1.0	83	**83**	1.0	83
84	1.0	84	**84**	1.0	84
85	1.0	85	**85**	1.0	85
86	1.0	86	**86**	1.0	86
87	1.0	87	**87**	1.0	87

Fuente: Elaboración Propia

4.1.2.5 Tamaño de Grupo: 30

DISTANCIA MINKOWSKI

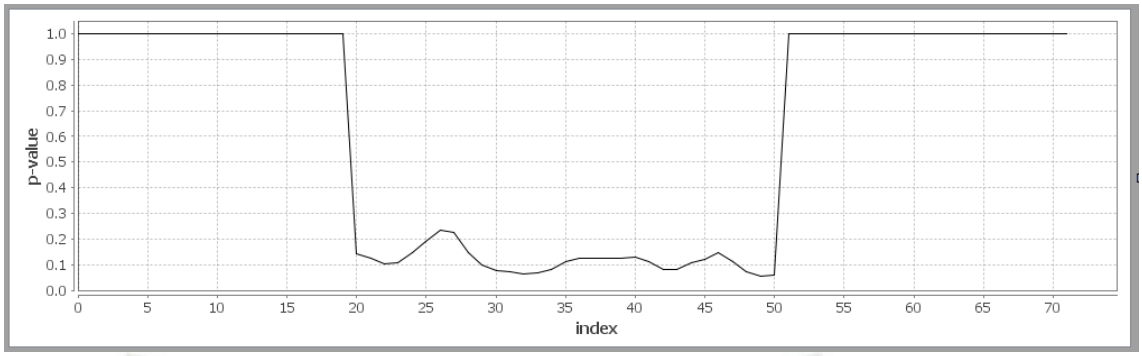

Ilustración 40 Resultados Acumulados de la prueba de hipótesis aplicada a la distancia Minkowski con un tamaño de grupo de 30
Fuente: Elaboración Propia

BAVELAS-LEAVITT

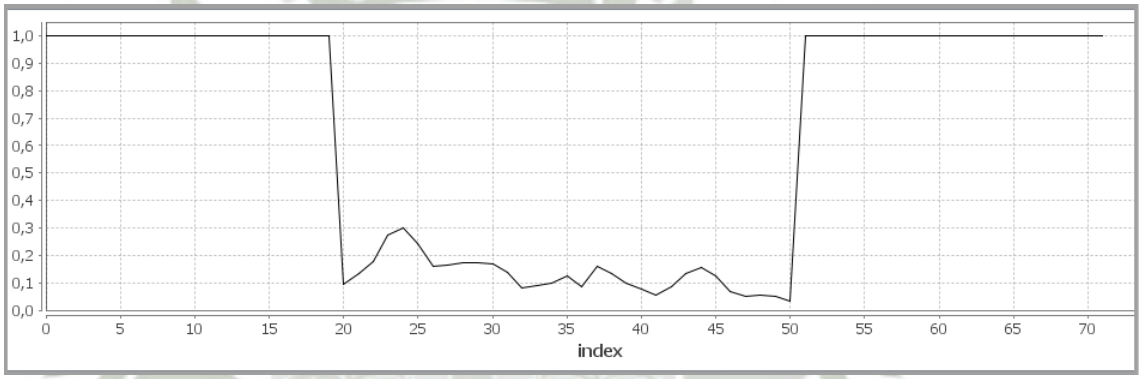

Ilustración 41 Resultados Acumulados de la prueba de hipótesis aplicada a la Bavelas-Leavitt con un tamaño de grupo de 30
Fuente: Elaboración Propia

Con un tamaño de grupo de 30 casos, que es más de un tercio del *log* comparados, se observan zonas de cambio mucho más amplias. En el cálculo de la distancia Minkowski en la *Ilustración 40* sólo se puede observar una zona de cambio importante, mientras que en Bavelas Leavitt en la *Ilustración 41*, se muestran zonas de cambio un poco más delimitadas que en los casos anteriores.

Con este tamaño de grupo podemos considerar que es una prueba de muy baja granularidad, ya que nos está indicando que efectivamente ocurre un cambio pero en un rango de 30 casos, los cuales pueden ser muy diferentes entre ellos. Poder entender cuáles han sido las razones de dichos cambios y en qué casos en

específico han ocurrido es más difícil. Para este cálculo demoró 53 milisegundos en ejecutar Minkowski y 3 para obtener Bavelas.

Tabla 30 Resultados para un tamaño de grupo de 30

Bavelas-Leavitt			Distancia Minkowski		
Índice	**Probabilidad**	**Grupo de Casos**	**Índice**	**Probabilidad**	**Grupo de Casos**
0	1.0	0	**0**	1.0	0
1	1.0	1	**1**	1.0	1
2	1.0	2	**2**	1.0	2
3	1.0	3	**3**	1.0	3
4	1.0	4	**4**	1.0	4
5	1.0	5	**5**	1.0	5
6	1.0	6	**6**	1.0	6
7	1.0	7	**7**	1.0	7
8	1.0	8	**8**	1.0	8
9	1.0	9	**9**	1.0	9
10	1.0	10	**10**	1.0	10
11	1.0	11	**11**	1.0	11
12	1.0	12	**12**	1.0	12
13	1.0	13	**13**	1.0	13
14	1.0	14	**14**	1.0	14
15	1.0	15	**15**	1.0	15
16	1.0	16	**16**	1.0	16
17	1.0	17	**17**	1.0	17
18	1.0	18	**18**	1.0	18
19	1.0	19	**19**	1.0	19
20	0.09592491	20	**20**	0.14013115	20
21	0.13272075	21	**21**	0.1265607	21
22	0.17748475	22	**22**	0.10437372	22
23	0.27372712	23	**23**	0.10909561	23
24	0.30174974	24	**24**	0.14757952	24
25	0.2418907	25	**25**	0.1883534	25
26	0.1609474	26	**26**	0.23236577	26
27	0.16416034	27	**27**	0.22596733	27
28	0.1709183	28	**28**	0.14582956	28
29	0.1718998	29	**29**	0.09645997	29
30	0.17036128	30	**30**	0.07529418	30
31	0.13738619	31	**31**	0.07107519	31
32	0.08109978	32	**32**	0.063444704	32
33	0.091320984	33	**33**	0.069997266	33
34	0.09772994	34	**34**	0.08117297	34
35	0.123201326	35	**35**	0.11182448	35
36	0.0848121	36	**36**	0.12639995	36
37	0.15847875	37	**37**	0.12309147	37
38	0.13485594	38	**38**	0.123976566	38
39	0.09888414	39	**39**	0.12586112	39
40	0.07754237	40	**40**	0.12849121	40
41	0.0558149	41	**41**	0.11230574	41
42	0.085964635	42	**42**	0.08273295	42
43	0.13375263	43	**43**	0.08251178	43
44	0.15409091	44	**44**	0.106410205	44
45	0.12543763	45	**45**	0.12238564	45
46	0.06785152	46	**46**	0.14603944	46

47	0.050942887	47	**47**	0.113104045	47
48	0.052644283	48	**48**	0.07165536	48
49	0.052036993	49	**49**	0.05443849	49
50	0.033623587	50	**50**	0.058827553	50
51	1.0	51	**51**	1.0	51
52	1.0	52	**52**	1.0	52
53	1.0	53	**53**	1.0	53
54	1.0	54	**54**	1.0	54
55	1.0	55	**55**	1.0	55
56	1.0	56	**56**	1.0	56
57	1.0	57	**57**	1.0	57
58	1.0	58	**58**	1.0	58
59	1.0	59	**59**	1.0	59
60	1.0	60	**60**	1.0	60
61	1.0	61	**61**	1.0	61
62	1.0	62	**62**	1.0	62
63	1.0	63	**63**	1.0	63
64	1.0	64	**64**	1.0	64
65	1.0	65	**65**	1.0	65
66	1.0	66	**66**	1.0	66
67	1.0	67	**67**	1.0	67
68	1.0	68	**68**	1.0	68
69	1.0	69	**69**	1.0	69
70	1.0	70	**70**	1.0	70
71	1.0	71	**71**	1.0	71

Fuente: Elaboración Propia

Al utilizar un tamaño de grupo entre 3 a 14, se observan zonas de cambio más amplias con un nivel de granularidad elevado, en comparación a un grupo de 30. Consideramos que es mejor utilizar un tamaño de grupo mayor a 1, ya que para las propuestas de sociometría tiene más sentido realizarlas comparando el comportamiento de un grupo de casos, ya que individualmente no se puede considerar con certeza la carga laboral normal de un usuario o la centralidad de éste.

4.1.3 Comparación de resultados: Perspectiva de Flujo vs. Perspectiva de Usuarios

El algoritmo planteado por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014), como hemos indicado anteriormente, encuentra el *Concept Drift* desde la perspectiva de flujo. Ahora se procederá a evaluar los resultados

de ambas perspectivas, de usuarios y de flujo, con el fin de encontrar si existe alguna relación entre las ellas.

Se utilizó para estas pruebas el *log* del Caso 2, mostrado anteriormente, parametrizado como sigue:

- Tamaño de Grupo: 1
- Tamaño de Ventana: 10
- Tamaño de Población: 20
- Prueba de hipótesis: Kolmogorov-Smirnov

Se ha tomado como tamaño de grupo 1 para tener los resultados caso por caso, tal como lo hace en la perspectiva de flujo.

Procederemos a comparar los gráficos correspondientes a la perspectiva de flujo contra la perspectiva de usuarios.

PERSPECTIVA DE FLUJO

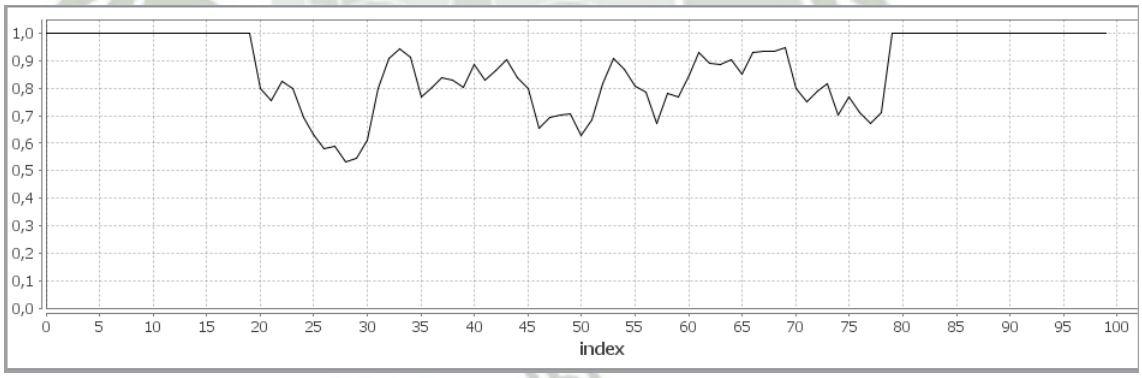

Ilustración 42 Gráfica Concept Drift Perspectiva de Flujo
Fuente: Elaboración Propia

PERSPECTIVA DE USUARIOS

Distancia Minkowski

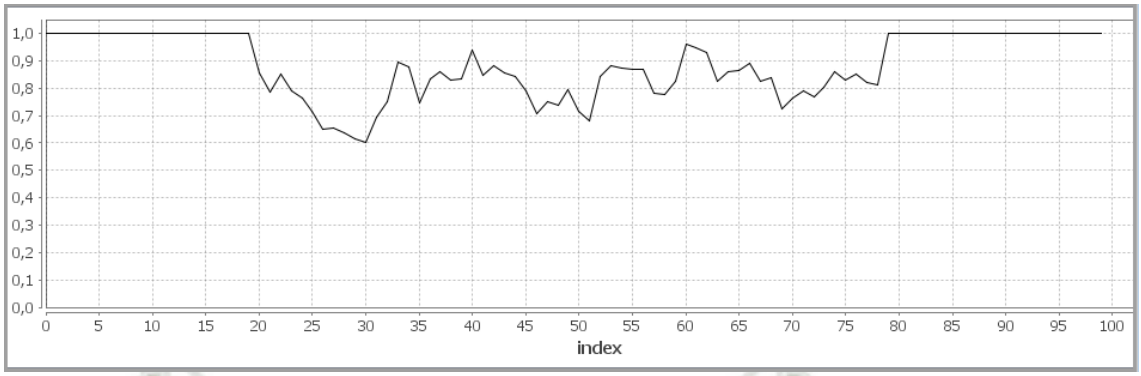

Ilustración 43 Gráfica Concept Drift Perspectiva de Usuarios – Distancia Minkowski
Fuente: Elaboración Propia

Bavelas – Leavitt

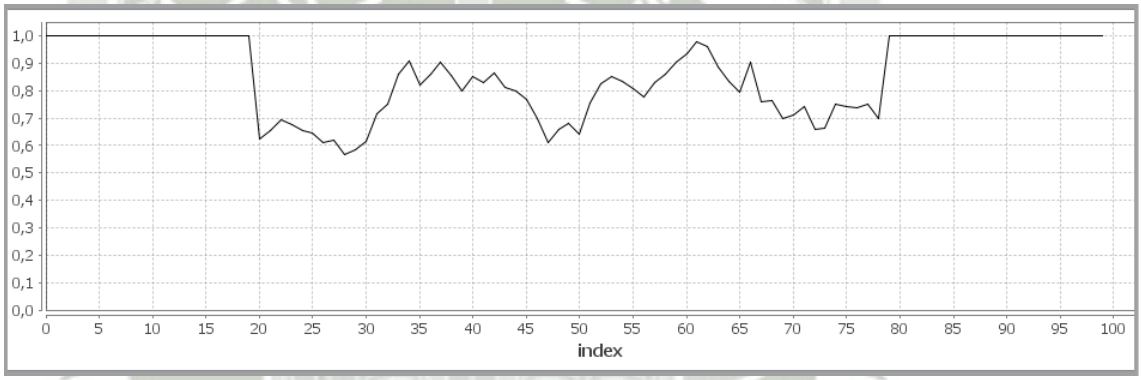

Ilustración 44 Gráfica Concept Drift Perspectiva de Usuarios – Bavelas – Leavitt
Fuente: Elaboración Propia

Se observa que las gráfica de la perspectiva de flujo, *Ilustración 42*, y la gráfica de la perspectiva de usuarios calculado mediante la distancia Minkowski, *Ilustración 43*, tienen una forma muy similar. Si comparamos los datos obtenidos de una zona de cambio, por ejemplo entre los índices 20 al 34 resumidos en la *Tabla 31*.

Tabla 31 Concept Drift en la perspectiva de flujo y de usuarios entre los índices 20 a 34

Índice	Perspectiva de Flujo		Perspectiva de Usuarios			
	Probabilidad	Caso	Probabilidad Distancia Minkowski	Probabilidad Bavelas – Leavitt	Grupo	Casos x Grupo
20	0.7988995	18	0.8547981	0.7974909	20	18/
21	0.75521356	22	0.78702724	0.8112334	21	22/
22	0.8248124	23	0.85121644	0.9013811	22	23/
23	0.7965057	24	0.78918976	0.8349273	23	24/
24	0.6937219	25	0.7636675	0.78221613	24	25/
25	**0.6330283**	26	**0.7171854**	**0.7882697**	25	26/
26	**0.5786897**	27	**0.6489955**	**0.77362806**	26	27/
27	**0.58731425**	28	**0.6538522**	**0.72769386**	27	28/
28	**0.5299962**	29	**0.63744915**	**0.60081565**	28	29/
29	0.54392684	30	0.6146887	0.5746941	29	30/
30	0.6083116	31	0.6005361	0.6331236	30	31/
31	0.79684806	33	0.69494134	0.63501173	31	33/
32	**0.90790987**	34	**0.7490344**	**0.7062625**	32	34/
33	**0.9426971**	32	**0.8935988**	**0.7431444**	33	32/
34	**0.91420203**	38	**0.87513703**	**0.7132218**	34	38/

Fuente: Elaboración propia

Se muestran los valores de probabilidad son similares y comprenden los mismos casos. También se encuentran los índices con probabilidad más cercanas a uno y más lejanas a 1 resaltados en negrita. Los casos comprendidos entre los índices 25 al 28 se puede observar que en la perspectiva de flujo existe un cambio más significativo que en la perspectiva de usuarios, al observar los datos de la *Tabla 32* notamos que el flujo de actividades de estos cuatro casos son muy diferentes entre todos, por ejemplo el caso 26 itera muchas más veces las actividades "Invite Addional Reviewer" y "Get Review X", además que no itera las actividades Get Review 1, 2 y 3, mientras que el caso 29 se cuenta con menos iteraciones de actividades e itera por lo menos una vez todas las actividades.

Tabla 32 Flujo de Actividades Casos 26, 27, 28 y 29

	Caso 26	Caso 27	Caso 28	Caso 29
1	invite reviewers	invite reviewers	invite reviewers	invite reviewers
2	Collect Reviews	Get Review 1	Get Review 3	get review 2
3	Decide	Get Review 2	Collect Reviews	get review 1
4	invite additional reviewer	Get Review 3	Decide	get review 3
5	invite additional reviewer	Collect Reviews	invite additional reviewer	Collect Reviews
6	get review X	Decide	invite additional reviewer	Decide
7	invite additional reviewer	invite additional reviewer	get review X	invite additional reviewer
8	invite additional reviewer	invite additional reviewer	invite additional reviewer	invite additional reviewer
9	invite additional reviewer	get review X	invite additional reviewer	get review X
10	get review X	invite additional reviewer	get review X	accept
11	invite additional reviewer	get review X	invite additional reviewer	
12	get review X	accept	invite additional reviewer	
13	invite additional reviewer		Get Review X	
14	get review X		Reject	
15	accept			

Fuente: Elaboración propia

En cambio la perspectiva de usuarios, al observar la *Tabla 33*, se puede notar que la participación y la carga laboral de los usuarios es muy variante, por ejemplo Pete sólo ejecuta una actividad en el caso 26 y no participa en los demás casos. Además la carga laboral de Mike es menor para los casos 28 y 29, también se puede observar que las tareas que ejecutan Sam, Mary, Carol y Jhon, no son constantes. Estos cambios en la cantidad de tareas hace que la distancia Minkowski calculada entre cada usuario sea diferente para cada caso en una forma significativa.

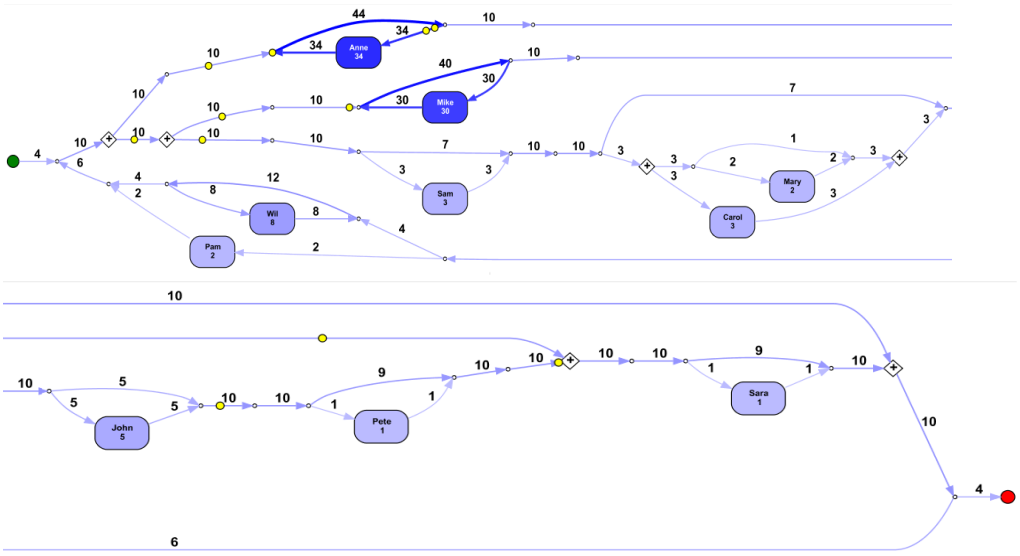

Ilustración 45 Gráfica resumen para Perspectivas de Usuarios. Casos 26, 27 ,28 y29

Fuente: Elaboración propia

En la *Ilustración 45*, podemos encontrar los usuarios con mayor participación y sus posiciones dentro del mismo. Podemos encontrar una clara sobrecarga de tareas con Anne y Mike, ambos inicializan y terminan los procesos. Mientras que los otros usuarios poseen poca carga.

Tabla 33 Actividades por Usuarios Casos 26, 27, 28 y 29

Usuario	Actividades			
	Caso 26	**Caso 27**	**Caso 28**	**Caso 29**
Anne	Invite reviewers Invite additional reviewer	invite reviewers invite additional reviewer	invite reviewers Collect reviews invite additional reviewer Reject	Collect reviews Invite additional reviewer
Sam	*	Get Review 1	Get Review X	Get Review X
Mary	*	Get Review 2	Get Review X	*
Carol	*	Get Review 3	Get Review X	Get Review 2
Mike	Collect Reviews Invite additional reviewer Accept	Collect Reviews Invite additional reviewer Accept	Invite additional reviewer	Invite reviewers Accept
Wil	Decide	Decide	Decide	Decide
Pam	Get Review X	Get Review X	*	*
jhon	Get Review X	Get Review X	Get Review 3	Get Review 1
Sara	*	*	*	Get Review 3
Pete	Get Review X	*	*	*

Fuente: Elaboración propia

En los casos comprendidos entre los índices 32 al 34 se puede observar que en la perspectiva de flujo el cambio encontrado es muy cercano a 1, esto significa que el proceso vuelve a un estado "normal", en cambio en la perspectiva de usuarios aún se sigue detectando un pequeño cambio.

En la *Tabla 34,* se puede observar que el flujo de actividades en los casos 32 y 38 es muy similar, salvo por unos casos que no son iterados, pero los cambios no son de la magnitud de los vistos anteriormente, es por esto que los valores son más cercanos a 1, ya que el proceso va normalizándose.

Tabla 34 Flujo de Actividades Casos 34, 32 y 38

	Caso 34	Caso 32	Caso 38
1	invite reviewers	invite reviewers	invite reviewers
2	get review 1	get review 1	Collect Reviews
3	get review 3	Collect Reviews	Decide
4	get review 2	Decide	invite additional reviewer
5	Collect Reviews	invite additional reviewer	invite additional reviewer
6	Decide	get review X	invite additional reviewer
7	Reject	invite additional reviewer	invite additional reviewer
8		invite additional reviewer	get review X
9		invite additional reviewer	invite additional reviewer
10		invite additional reviewer	get review X
11		get review X	accept
12		invite additional reviewer	
15		get review X	
16		accept	

Fuente: Elaboración propia

En la perspectiva de usuarios encontramos que al igual que en la perspectiva de flujo las actividades que realizan los usuarios y la participación de los mismos es similar. La mayoría de usuarios mantienen su carga laboral y los usuarios que no realizan tareas, no las realizan en todos los casos. Aquí el cambio más significativo sería la carga laboral de Anne, ya que esta cambia en los tres casos.

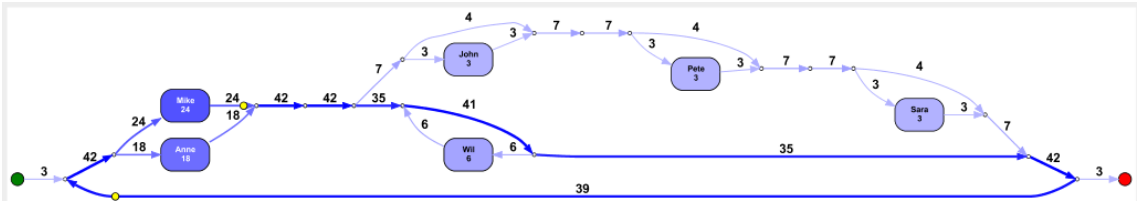

Ilustración 46 Gráfica resumen para Perspectivas de Usuarios. Casos 32, 34 y 38
Fuente: Elaboración propia

Como podemos ver en la perspectiva de usuarios el flujo es más resumido y estable. Ahora sólo encontramos tres caminos diferenciales y dos más utilizados.

Tabla 35 Actividades por Usuarios Casos34, 32 y 38

Usuario	Actividades		
	Caso 34	Caso 32	Caso 38
Anne	Invite reviewers Reject	Collect Reviews Accept	Collect Reviews Invite additional reviewer Accept
Sam	*	*	*
Mary	*	*	*
Carol	*	*	*
Mike	Collect Reviews	Invite reviewers Invite additional reviewer	Invite reviewers Invite additional reviewer
Wil	Decide	Decide	Decide
Pam	*	*	*
jhon	Get Review 1	Get Review X	*
Sara	Get Review 2	get review X	get review X
Pete	Get Review 3	get review X	get review X

Fuente: Elaboración propia

Se ha observado que los cambios que se detectan mediante la perspectiva de flujo y la perspectiva de usuarios, son encontrados en los mismos puntos. Sin embargo ambas perspectivas no necesariamente se encuentran directamente relacionadas, ya que puede mantenerse el mismo flujo de actividades pero los usuarios que las ejecutan pueden variar, y esto no supone que se haya realizado un cambio en el flujo, sino sólo un cambio a nivel de usuarios. Además que la perspectiva de flujo encuentra los cambios de precede/antecede entre actividades y la perspectiva de usuarios calcula en cambio mediante la cantidad de actividades realizadas y la participación de un usuario por caso.

CONCLUSIONES

1. Los resultados obtenidos muestran que es posible la detección del cambio en la perspectiva de usuarios mediante las características planteadas. La característica de entropía es capaz de distinguir un cambio a nivel global, indicando los usuarios con menos estabilidad en su comportamiento. La distancia Minkowski y Bavelas-Leavitt son capaces de encontrar cambios a nivel de Grupos de Casos, los cuales dependiendo de su tamaño pueden aportar mayor o menor granularidad. Estas dos últimas características encuentran el *Concept Drift* a nivel de promedio.

2. No necesariamente la perspectiva de flujo y de usuarios van a encontrar los cambios en los mismos puntos, ya que si bien el flujo de las actividades no cambia, los usuarios que la ejecutan si pueden cambiar, esto daría como resultado un *Concept Drift* sólo en la perspectiva de usuarios más no en la perspectiva de flujo y viceversa.

3. Se ha observado que es posible obtener resultados a mayor granularidad al minimizar o extender el tamaño de grupo. Se observó que mientras más pequeño sea el grupo de casos se obtiene un resultado más preciso.

4. Al utilizar la prueba de hipótesis Mann-Whittney se observó que este es más sensible para detectar los *Concept Drift* en un *log*, por esto creemos que es conveniente utilizarlo en casos de *logs* pequeños, para tener un mejor análisis, ya que al ser pocos datos con la prueba de Kolmogorov-Smirnov hay cambios no detectados.

5. Mediante esta investigación podemos aportar al campo de la minería de procesos en *Concept Drift*, la posibilidad de utilizar la perspectiva de usuarios. Esta perspectiva, en nuestra opinión, es altamente ventajosa ya que nos permite

identificar los cambios cuantificando el desempeño y la participación de los usuarios en los flujos de trabajo.

RECOMENDACIONES Y TRABAJOS FUTUROS

- Estudiar y aplicar otros conceptos más específicos de la sociometría. Por ejemplo, el concepto de cercanía de la red social. Con un mayor análisis de los comportamientos de los usuarios en el flujo es posible tener una visión más clara y completa sobre la perspectiva de usuarios.

- Utilizando perfiles de usuarios y niveles jerárquicos es posible obtener mayor información dentro del análisis en la perspectiva de usuarios. En la actual investigación se ha trabajado de forma horizontal, es factible modelar y analizar los datos de usuario aplicando jerarquías de cargos.

- Aplicación de las técnicas de localización de *Concept Drift* con *logs* en tiempo real. Tanto el algoritmo original como el trabajo expuesto son adaptables para trabajar con *logs* utilizando tramas de datos. La aplicación de algoritmos de análisis en tiempo real permiten otorgar, sustentar y tomar acción ante los cambios de comportamiento del flujo.

- Verificar los resultados utilizando *Concept Drift* del tipo gradual, recurrente e incremental en la perspectiva de usuarios.

- Evaluación y análisis exhaustivo de los datos y gráficos mostrados para hallar similitudes entre *logs*. Es posible generalizar o especificar resultados obtenidas a través del análisis y comparación de los *logs*. Realizar comparaciones a nivel de ciertas actividades o ciertos usuarios para análisis más específicos.

- Implementar una interfaz donde se pueda mostrar los datos y trabajar con ellos de una forma más cómoda. En el caso de Bavelas-Leavitt, poder especificar la

relación de usuarios contra usuario y en la Distancia Minkowski especificar los usuarios contra actividad, de forma gráfica y detallada.

REFERENCIAS

Carmona, J., & Gavaldà, R. (2012). Online Techniques for Dealing with *Concept Drift* in Process Mining.

Cheng-Jung Tsai a, *. C.-I.-P. (2007). v . *www.sciencedirect.com*, 15.

Chien-I, L., Cheng-Jung, T., Jhe-Hao, W., & Wei-Pang, Y. (2007). A Decision Tree-Based Approach to Mining the Rules of *Concept Drift*. *IEEE*.

Fei, D., Liquin, Z., Guang Yun, N., & Xiaolei, X. (2013). An Algorithim for Detecting *Concept Drift* Based On Context in Process Mining. *CPS*.

Group, P. M. (2010). *http://www.promtools.org*. Obtenido de http://www.promtools.org/doku.php?id=gettingstarted:start

Jagadeesh Chandra Bose, R. P., van der Aalst, W. M., Zliobaitè, I., & Pechenizkiy, M. (2014). Dealing With *Concept Drift*s in Process Mining. *IEEE*.

Leemans, S., Fahland, D., & van der Aalst, W. (2014). *Process and Deviantion Exploration with Inductive Visual Miner.*

Lin Feng, F. C. (2013). A Concept Similarity Based *Data stream* Classification Model. *Lin Feng, Feng Chen, Yuan Yao - 2013*, 9.

Maggi., R. J. (2013). Efficient Algorithms for Discovering *Concept Drift* in Business Processes. *2013 Fourth International Conference on Digital Manufacturing & Automation*, 8.

ProM. (21 de 08 de 2012). *Technische Universitet Eindhoven University of Technology*. Recuperado el 22 de 06 de 2014, de http://www.win.tue.nl/ieeetfpm/doku.php?id=shared:process_mining_manifesto

ProM. (25 de 05 de 2015). *ProM Tools*. Obtenido de ProM Tools: http://www.promtools.org/doku.php

R.P. Jagadeesh Chandra Bose, W. M. (2010). Handling *Concept Drift* in *Process Minning*. 8.

Sujatha, D., & Anil Kumar, M. (2013). Handling of Recurrence *Concept Drift* in *Data stream* Using Timestamp of Auziliary Learning Model. *IJITR*.

van der Aalst, W. M. (2011). *Process Mining.* Springer.

van der Aalst, W. M. (2013). Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining.

van der Aalst, W., Adriansyah, A., Alves de Medeiros, A. K., Arcieri, F., Baider, T., Blicke, T., . . . Castellanos, M. (2012). Proces Mining Manifesto. *IEEE*.

Westergaard, M. (13 de 09 de 2012). *Michael Westergaard's Home Page*. Obtenido de ProM Tutorial: https://westergaard.eu/category/prom-tutorial/

APÉNDICE I: GLOSARIO DE TÉRMINOS

Árbol KD (árbol k-dimensional). Estructura de datos que divide el espacio organizando puntos en un espacio euclídeo de k dimensiones. Emplea sólo planos perpendiculares a uno de los ejes del sistema de coordenadas. Todos los nodos de un árbol kd, desde el nodo raíz hasta los nodos hoja, almacenan un punto.

Divergencia Kullback-Leibler (divergencia de la información). Medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad P y Q. Mide el número esperado de extra bits requeridos en muestras de código de P cuando se usa un código basado en Q, en lugar de un código basado en P. Generalmente P representa la "verdadera" distribución de los datos, observaciones, o cualquier distribución teórica. La medida Q generalmente representa una teoría, modelo, descripción o aproximación de P.

Distribución conjunta. Dados dos eventos aleatorios X e Y, es la distribución de probabilidad de la intersección de eventos de X e Y, esto es, de los eventos X eY ocurriendo de forma simultánea.

Bootstrapping (Bootstrap). Método de re muestreo. Se utiliza para aproximar la distribución en el muestreo de un estadístico, aproxima el sesgo o la varianza de un análisis estadístico, así como para construir intervalos de confianza o realizar contrastes de hipótesis sobre parámetros de interés.

Support vector machines (SVMs, support vector networks) Modelos de aprendizaje supervisado asociado a algoritmos de aprendizaje que analizan datos y reconocen patrones, son usados para el análisis de clasificación y regresiones. Un algoritmo de entrenamiento de SVM arma un modelo que asigna los nuevos ejemplos a una categoría u otra, a través de un clasificador binario lineal no probabilístico.

Envolvente Convexa Se define la envolvente convexa, envoltura convexa o cápsula convexa de un conjunto de puntos X de dimensión n como la intersección de todos los conjuntos convexos que contienen a X.1 En geometría computacional existen numerosos algoritmos para calcular la envolvente convexa de un conjunto finito de puntos, con diversos grados de complejidad computacional. La complejidad del algoritmo de resolución se suele estimar en función del número n de puntos de entrada, y el número h de puntos de la correspondiente envolvente convexa.

APÉNDICE II: *LOG* SINTÉTICO EJERCICIO 4


```
<log xes.version="1.0" xes.features="nested-attributes" openxes.version="1.0RC7" xmlns="http://www.xes-standard.org/">
	<classifier name="Event Name" keys="concept:name"/>
	<classifier name="Resource" keys="org:resource"/>
	<string key="source" value="Rapid Synthesizer"/>
	<string key="concept:name" value="excercise4.mxml"/>
	<string key="lifecycle:model" value="standard"/>
	<trace>
		<string key="concept:name" value="Case1.0"/>
		<event>
			<string key="org:resource" value="A"/>
			<date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
			<string key="concept:name" value="Solicitud"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="B"/>
			<date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
			<string key="concept:name" value="Revisa_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="C"/>
			<date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
			<string key="concept:name" value="Aprueba_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="D"/>
			<date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
			<string key="concept:name" value="Genera_Datos"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="E"/>
			<date key="time:timestamp" value="2009-06-04T17:25:03.471+02:00"/>
			<string key="concept:name" value="Cierra_Caso"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
	</trace>
	<trace>
		<string key="concept:name" value="Case2.0"/>
		<event>
			<string key="org:resource" value="A"/>
			<date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
			<string key="concept:name" value="Solicitud"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="B"/>
			<date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
			<string key="concept:name" value="Revisa_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="C"/>
			<date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
```

```
                <string key="concept:name" value="Aprueba_Sol"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="D"/>
                <date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
                <string key="concept:name" value="Genera_Datos"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="E"/>
                <date key="time:timestamp" value="2009-06-04T17:25:03.471+02:00"/>
                <string key="concept:name" value="Cierra_Caso"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
</trace>
<trace>
        <string key="concept:name" value="Case3.0"/>
        <event>
                <string key="org:resource" value="A"/>
                <date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
                <string key="concept:name" value="Solicitud"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="C"/>
                <date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
                <string key="concept:name" value="Aprueba_Sol"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="C"/>
                <date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
                <string key="concept:name" value="Genera_Datos"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="D"/>
                <date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
                <string key="concept:name" value="Cierra_Caso"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
</trace>
<trace>
        <string key="concept:name" value="Case4.0"/>
        <event>
                <string key="org:resource" value="A"/>
                <date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
                <string key="concept:name" value="Solicitud"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="B"/>
                <date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
                <string key="concept:name" value="Revisa_Sol"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="C"/>
                <date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
                <string key="concept:name" value="Aprueba_Sol"/>
```

```
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="D"/>
                <date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
                <string key="concept:name" value="Genera_Datos"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="E"/>
                <date key="time:timestamp" value="2009-06-04T17:25:03.471+02:00"/>
                <string key="concept:name" value="Cierra_Caso"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
</trace>
<trace>
        <string key="concept:name" value="Case5.0"/>
        <event>
                <string key="org:resource" value="A"/>
                <date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
                <string key="concept:name" value="Solicitud"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="B"/>
                <date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
                <string key="concept:name" value="Revisa_Sol"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="C"/>
                <date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
                <string key="concept:name" value="Genera_Datos"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="E"/>
                <date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
                <string key="concept:name" value="Cierra_Caso"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
</trace>
<trace>
        <string key="concept:name" value="Case6.0"/>
        <event>
                <string key="org:resource" value="A"/>
                <date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
                <string key="concept:name" value="Solictud"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="B"/>
                <date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
                <string key="concept:name" value="Aprueba_Sol"/>
                <string key="lifecycle:transition" value="complete"/>
        </event>
        <event>
                <string key="org:resource" value="D"/>
                <date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
                <string key="concept:name" value="Genera_Datos"/>
                <string key="lifecycle:transition" value="complete"/>
```

```
		</event>
		<event>
			<string key="org:resource" value="E"/>
			<date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
			<string key="concept:name" value="Cierra_Caso"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
</trace>
<trace>
		<string key="concept:name" value="Case7.0"/>
		<event>
			<string key="org:resource" value="A"/>
			<date key="time:timestamp" value="2009-06-04T17:21:03.346+02:00"/>
			<string key="concept:name" value="Solicitud"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="B"/>
			<date key="time:timestamp" value="2009-06-04T17:22:03.346+02:00"/>
			<string key="concept:name" value="Revisa_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="A"/>
			<date key="time:timestamp" value="2009-06-04T17:23:03.346+02:00"/>
			<string key="concept:name" value="Solicitud"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="C"/>
			<date key="time:timestamp" value="2009-06-04T17:24:03.346+02:00"/>
			<string key="concept:name" value="Genera_Datos"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="D"/>
			<date key="time:timestamp" value="2009-06-04T17:25:03.346+02:00"/>
			<string key="concept:name" value="Cierra_Caso"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
</trace>
<trace>
		<string key="concept:name" value="Case8.0"/>
		<event>
			<string key="org:resource" value="A"/>
			<date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
			<string key="concept:name" value="Solicitud"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="B"/>
			<date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
			<string key="concept:name" value="Revisa_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="C"/>
			<date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
			<string key="concept:name" value="Genera_Datos"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
```

```
		<event>
			<string key="org:resource" value="C"/>
			<date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
			<string key="concept:name" value="Cierra_Caso"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
</trace>
<trace>
		<string key="concept:name" value="Case9.0"/>
		<event>
			<string key="org:resource" value="A"/>
			<date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
			<string key="concept:name" value="Solicitud"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="B"/>
			<date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
			<string key="concept:name" value="Revisa_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="C"/>
			<date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
			<string key="concept:name" value="Aprueba_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="D"/>
			<date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
			<string key="concept:name" value="Genera_Datos"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="E"/>
			<date key="time:timestamp" value="2009-06-04T17:25:03.471+02:00"/>
			<string key="concept:name" value="Cierra_Caso"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
</trace>
<trace>
		<string key="concept:name" value="Case10.0"/>
		<event>
			<string key="org:resource" value="A"/>
			<date key="time:timestamp" value="2009-06-04T17:21:03.471+02:00"/>
			<string key="concept:name" value="Solicitud"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="B"/>
			<date key="time:timestamp" value="2009-06-04T17:22:03.471+02:00"/>
			<string key="concept:name" value="Revisa_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
			<string key="org:resource" value="C"/>
			<date key="time:timestamp" value="2009-06-04T17:23:03.471+02:00"/>
			<string key="concept:name" value="Aprueba_Sol"/>
			<string key="lifecycle:transition" value="complete"/>
		</event>
		<event>
```

```
                        <string key="org:resource" value="D"/>
                        <date key="time:timestamp" value="2009-06-04T17:24:03.471+02:00"/>
                        <string key="concept:name" value="Genera_Datos"/>
                        <string key="lifecycle:transition" value="complete"/>
                </event>
                <event>
                        <string key="org:resource" value="E"/>
                        <date key="time:timestamp" value="2009-06-04T17:25:03.471+02:00"/>
                        <string key="concept:name" value="Cierra_Caso"/>
                        <string key="lifecycle:transition" value="complete"/>
                </event>
        </trace>
</log>
```

APÉNDICE III: TRABAJANDO CON PROM

Entorno de desarrollo

Se inicia instalando Java y una versión reciente de Eclipse. Hay que estar seguro de instalar la versión de subversión de Eclipse (1,2) o algún otro para el sistema operativo.

Existe una versión para principiantes GettingStarted ProM package de la página https://svn.win.tue.nl/repos/prom/Packages/GettingStarted/Trunk/.

Revisión Básica de ProM

Lo primero es diferenciar que ProM no es un editor. Es un modelo transformador y visor. En un nivel básico, necesitamos preocuparnos acerca de tres tipos de elementos: Objetos, *plug-in*s y visualizadores. Por ejemplos aquellos que están familiarizados con el modelo vista controlador como patrón de diseño, los objetos constituyen el modelo de diseño y los visualizadores las vistas. Como ProM no es un editor, no posee controlador.

Los *Plug-in*s en ProM tienen que vivir en un paquete, se crea confusión cuando no se realiza de esa forma. Es una buena idea utilizar un paquete como `org.processmining.plugins.<name>`, donde <name> es un versión del nombre del paquete en minúsculas. Como ejemplo utilizaremos el conocido HelloWorld como nombre de paquete `org.processmining.plugins.helloworld`.

Objects

Los objetos son la idea básica de ProM. Son utilizados como objetos regulares y son creados como objetos normales en Java. Una característica importante en ProM es que serializa automáticamente los datos en el disco. No requeriere que los objetos implementen la interfaz serealizable, pero tienes que garantizar un par de puntos. El principal es que los objetos no tengan referencias a objetos no serializables (unserializable) u objetos en tiempo real, o que los objetos no tengan referencias a hilos u objetos de interfaz de usuario.

Los objetos nunca deben modificarse una vez creados. Ellos no tienen que ser estrictamente inmutables. Una forma de garantizarlo es hacer que todos los métodos

mutator de paquetes sean sólo visibles, o sea tener solo get. Notese que los objetos no debe usar el standard Observable de Java, ya que no es necesario para la serialización. Luego de lo antes mencionado no existen restricciones, pero observando que el objeto principal debe ser publico y no utilizar parámetros genéricos. Se muestra un ejemplo de un simple modelo de objeto:

```
package org.processmining.plugins.helloworld;

public class Person {
  private Name name;
  private int age

  public Person() {
    age = 0;
  }

  public Name getName() {
    return name;
  }

  setName(Name name) {
    this.name = name;
  }

  public int getAge() {
    return age;
  }

  setAge(int age) {
    this.age = age;
  }
}

class Name {
  private String first, last;

  public Name(String first, String last) {
    this.first = first;
    this.last = last;
  }

  public String getFirst() {
    return first;
  }

  public String getLast() {
    return last;
  }
}
```

Plug-ins

Los *plug-ins* son un mecanismo básico de ProM. La principal idea de un plugin es que se transforme un modelo en otro. Notese que los objetos originales nunca deben ser cambiados y los nuevos objetos no deben ser alterados a medida que un plugin termina.

Los plugins deben ser básicamente solo métodos. Se pueden crear *plug-ings* de muchas maneras. Basicamente un *plug-in* es un numero de parámetros (incluidos objectos) y una configuración. La configuración describe como exactamente un plugin debe operar. Un *plug-in* expone un método para que ProM le llame, este métodos necesita una anotación en Java (Java Notation). Se presenta un *Plug-in* permitiendo que dos personas posean un hijo:

```
package org.processmining.plugins.helloworld;
import org.processmining.framework.plugin.annotations.*;
import org.processmining.framework.plugin.*;
import org.processmining.contexts.uitopia.*;
import org.processmining.contexts.uitopia.annotations.*;

@Plugin(name = "Procreate",
        parameterLabels   =   {   "Father",   "Mother",   "Procreation
Configuration" },
        returnLabels = { "Child" },
        returnTypes = { Person.class })
public class ProcreatePlugin {
  @UITopiaVariant(affiliation = "University of Life",
                  author = "Britney J. Spears",
                  email = "britney@westergaard.eu",
                  uiLabel = UITopiaVariant.USEPLUGIN)
  @PluginVariant(requiredParameterLabels = { 0, 1, 2 })
```

La variante UITopia especifica información extra expuesta en la interfaz de usuario de ProM, incluye el nombre, afiliación y correo del autor. También especifica una etiqueta utilizada para mostrar el *Plug-in* en la interfaz de usuario.

Interacción con la Interfaz de usuario de ProM

Todos los *plug-ins* poseen un PluginContext (o un UIPluginContext) como primer parámetro. Esto da acceso a muchas funcionalidades para comunicar estados a usuarios e interactuar con ProM y el usuario.

Los PluginContexts vienen en diferentes tipos. El básico es PluginContext, el cual no hace suposiciones sobre el ambiente de ProM. El UIPliginConext asume un entorno gráfico y debe ser utilizado solo si es requerido. Esta es la razón por la cual se dividen los *plug-in*s en dos variantes: una haciendo la computación sin asumir la presencia de una interfaz gráfica y otro para poblar la configuración de la interfaz de usuario.

El PluginContext, nos da acceso a *log*eos, reportes de progreso, y descripciones de los resultados, Ademas se tiene acceso a todos los *plug-in*s, objetos y conecciones del espacio de trabajo de ProM. Si se tiene un contexto gráfico (UIPluginContext), también se tienen acceso al método ShowConfiguration.

APÉNDICE IV: INTERFAZ DEL PLUGIN

PLUG-IN MODIFICADO EN PROM

En esta propuesta se ha modificado el *plug-in* de ProM creado por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014), los cambios más importantes son el agregado del parámetro para el tamaño de grupo y las gráficas y tablas de las características planteadas en el **¡Error! No se encuentra el origen de la referencia.**

Para poder ejecutar el *plug-in* primeramente iniciamos el ProM, se importa el *log* que se desea analizar y seleccionamos en el panel de acciones "*Concept Drift* Modificado", más adelante se detallará como utilizar el *plug-in* modificado.

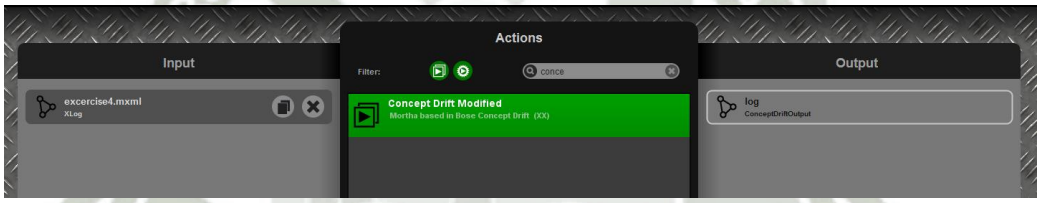

Ilustración 47 Plug-in modificado
Fuente: Elaboración propia

INGRESO DE PARÁMETROS

En la ventana "Feature Scope Configuration Step", se seleccionan las actividades que se desean analizar, en este caso se seleccionan todas las actividades ya que para nuestras pruebas y el enfoque que hemos dado en la perspectiva de usuarios se prefiere el uso de todas las actividades.

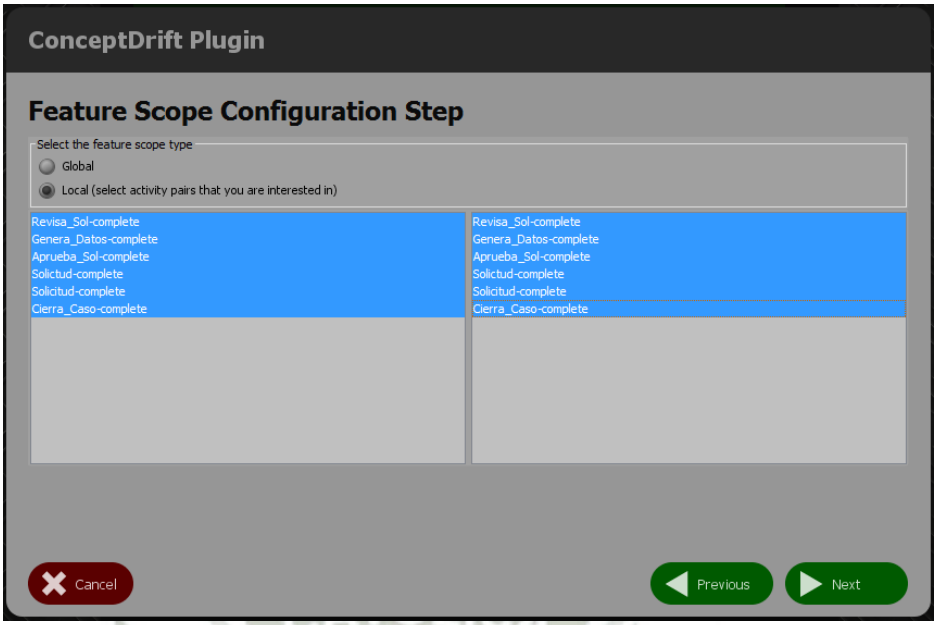

Ilustración 48 Selección de Actividades
Fuente: Elaboración propia

En la ventana "Feature Configuration Step" es donde se configura como será el análisis de las características para las perspectivas de flujo y de usuarios. Para la perspectiva de usuarios se ha añadido el campo de texto tamaño de grupo, que es el valor que definirá el tamaño de los grupos de casos.

Además de ingresar el tamaño de grupo, es importante seleccionar el tamaño de ventana.

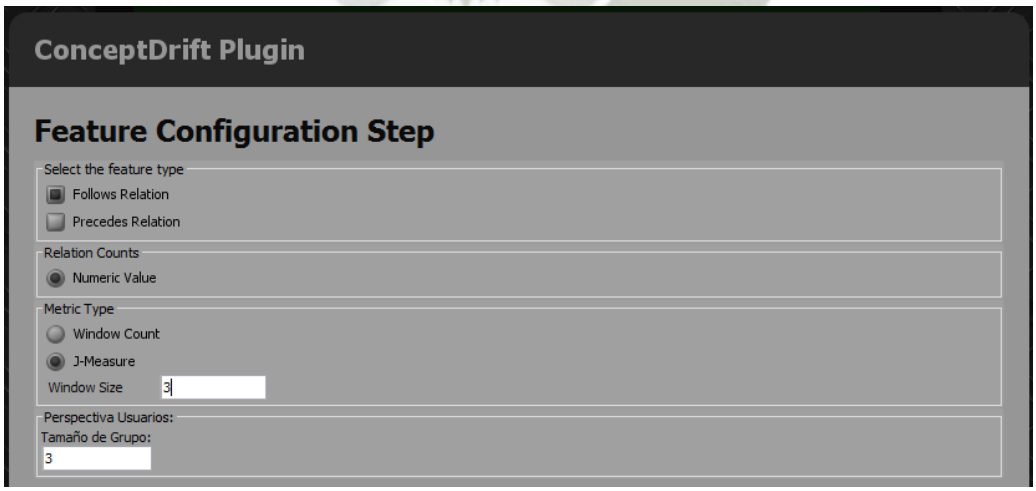

Ilustración 49 Configuración de características
Fuente: Elaboración propia

Seguidamente en la ventana "*Concept Drift* Learning Algorithm Configuration Step" se selecciona la prueba de hipótesis que se quiere usar y opciones de selección de población, entre las más importantes se encuentra el tamaño de población y el tamaño de paso.

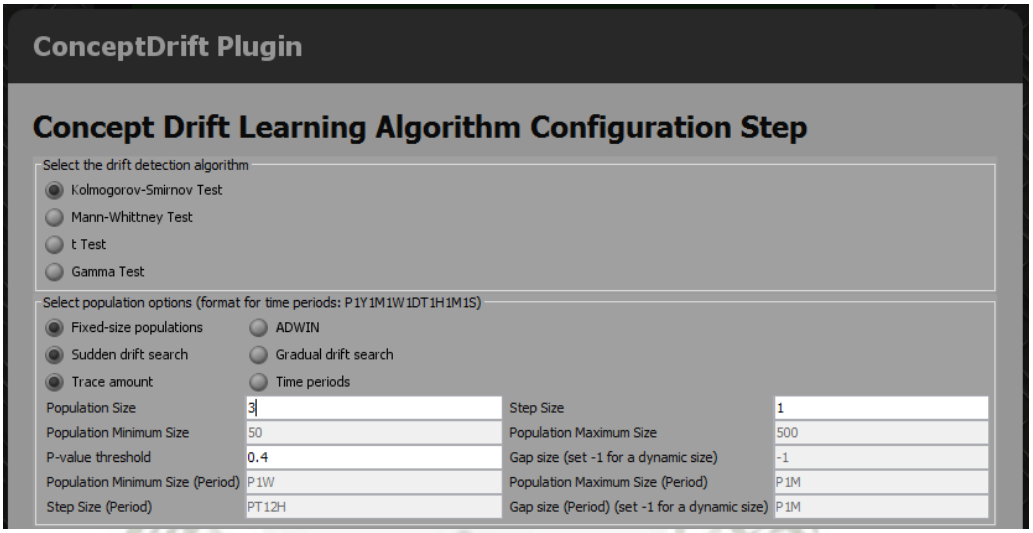

Ilustración 50 Configuración del Algoritmo Concept Drift

Fuente: Elaboración Propia

VISUALIZACIÓN DE DATOS

En este panel se muestran los resultados obtenidos en el lado derecho, la gráfica en la parte superior y las tablas de datos en la parte inferior; mediante los botones en el lado izquierdo.

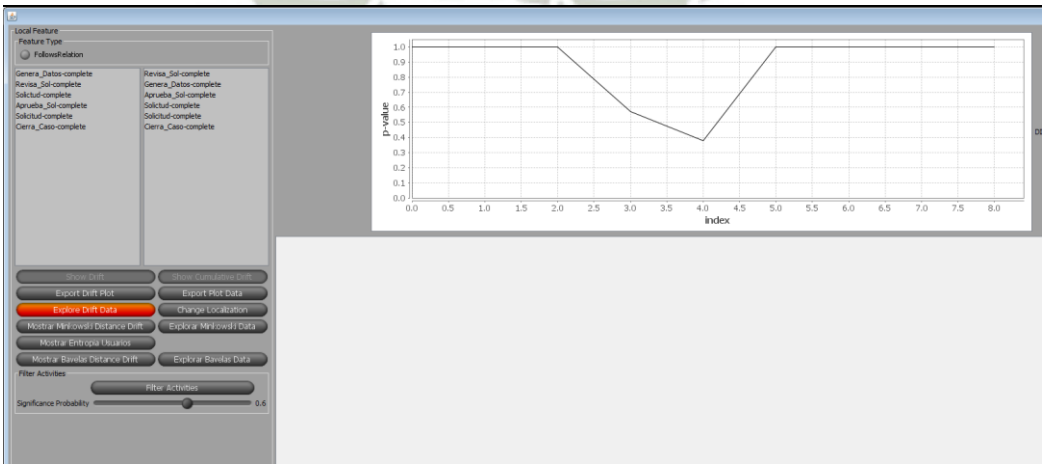

Ilustración 51 Panel de Resultados

Fuente: Elaboración Propia

Se han agregado los siguientes botones:

- Mostrar Minkowski Distance *Drift*, el cual muestra la gráfica de los valores acumulados de la prueba de hipótesis seleccionada para la característica de Carga laboral o Distancia Minkowski; el botón "Explorar Minkowski Data" muestra en una tabla dichos datos.

- Mostrar Bavelas Distance *Drift*, el cual muestra la gráfica de los valores acumulados de la prueba de hipótesis seleccionada para la característica de Centralidad o Distancia Bavelas-Leavitt; el botón "Explorar Bavelas Data" muestra en una tabla dichos datos.

- Mostrar Entropía Usuarios, muestra una tabla con la entropía de cada usuario.

Ilustración 52 Juego de Botones Plug-in modificado
Fuente: Elaboración propia

APÉNDICE V: GLOSARIO

- ADWIN: Algoritmo de ventanas adaptativas.
- Clustering: Un algoritmo de agrupamiento, procedimiento de agrupación de una serie de vectores de acuerdo con un criterio.
- Concept Drift: Cambio de comportamiento.
- Datastream: flujo de datos.
- Dataset: conjunto de datos.
- Dicing: operación común aplicada en cubos OLAP.
- Drilling down: operación común aplicada en cubos OLAP.
- Footprints: Rastro de datos en la minería de procesos.
- Framework: Marco de trabajo.
- Gradual Drift: Tipo de concept drift en la minería de procesos, gradual.
- Incremental Drift: Tipo de concept drift en la minería de procesos, incremental.
- J-Measure: Carácterística utilizada por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014) en el plug-in Concept Drfit en ProM.
- Logs: Conjunto de registros.
- Long-running: Ejecución larga de un proceso.
- Naive Bayes: Clasificador bayesiano ingenuo.
- Off-line: Fuera de línea, en tiempo no real.
- On-line: En línea o en tiempo real.

- Online Analytical Processing:
- Plug In: Complemento del framework, en estos casos de ProM.
- Process Cube: Cubo de procesos.
- Process Mining: Minería de procesos.
- Recurring Drift: Tipo de concept drift en la minería de procesos, recurrente.
- Relation Entropy: Carácterística utilizada por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014) en el plug-in Concept Drfit en ProM.
- Relation Type Count: Carácterística utilizada por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014) en el plug-in Concept Drfit en ProM.
- Rolling up: Operación común aplicada en cubos OLAP.
- Short-running: Ejecución corta de un proceso.
- Slicing: operación común aplicada en cubos OLAP.
- Stream: Flujo, generalmente aplicado a datos.
- Sub-logs: Subconjunto de registros de un log.
- Sudden Drift: Tipo de concept drift en la minería de procesos, repentino.
- Window Count: Carácterística utilizada por (Jagadeesh Chandra Bose, van der Aalst, Zliobaitè, & Pechenizkiy, 2014) en el plug-in Concept Drfit en ProM.
- Workspace: Espacio de trabajo.