

1. UNIVERSIDAD ANDRÉS BELLO.

**Facultad de Ingeniería.**

**Ingeniería Civil Informática.**

**ANÁLISIS DE SEÑALES PERIODICAS EN APLICACIONES WEB.**

**Memoria para optar al título de Ingeniero Civil en Informática.**

**Autor:**

**Marco Andres Bahamonde Sanhueza**

**Profesor Guía: Dr. Carlos Gómez Pantoja**

**Santiago de Chile,**

**2015.**

INDICE DE CONTENIDO

[1. Introducción 1](#_Toc440819594)

[1.1. Motivación 1](#_Toc440819595)

[1.2. Contribuciones de la tesis 2](#_Toc440819596)

[1.3. Contextualización 2](#_Toc440819597)

[2. Descripción del Problema 5](#_Toc440819598)

[2.1. Enunciado del Problema 5](#_Toc440819599)

[2.2. Objetivos 6](#_Toc440819600)

[2.2.1. Objetivo General 6](#_Toc440819601)

[2.2.2. Objetivos Específicos 6](#_Toc440819602)

[2.2.3. Resultados Esperados, Métricas y Criterios de Éxito 7](#_Toc440819603)

[2.3. Hipótesis 8](#_Toc440819604)

[2.4. Alcance 8](#_Toc440819605)

[3. Marco Teórico 8](#_Toc440819606)

[3.1. Señales 8](#_Toc440819607)

[3.1.1. Señal de tiempo continuo o analógico 8](#_Toc440819608)

[3.1.2. Señal digital o de tiempo discreto 9](#_Toc440819609)

[3.2. Tipos de señales en aplicaciones web. 11](#_Toc440819610)

[3.2.1. Periódica 11](#_Toc440819611)

[3.2.2. Permanentes 11](#_Toc440819612)

[3.2.3. Ráfaga 12](#_Toc440819613)

[3.2.4. Series Temporales 13](#_Toc440819614)

[3.2.5. Clasificación descriptiva de series 15](#_Toc440819615)

[3.3. Procesos Estocásticos 15](#_Toc440819616)

[3.3.1. Proceso estocástico estacionario 15](#_Toc440819617)

[3.4. Métodos para la eliminación de la componente Tendencia 17](#_Toc440819618)

[3.5. Medias Móviles 18](#_Toc440819619)

[3.6. Análisis Estadístico 19](#_Toc440819620)

[3.6.1. Variables Estadísticas 19](#_Toc440819621)

[3.6.1.1. Promedio o Media Aritmética 19](#_Toc440819622)

[3.6.1.2. Varianza 19](#_Toc440819623)

[3.6.1.3. Covarianza 20](#_Toc440819624)

[3.6.1.4. Correlación 20](#_Toc440819625)

[3.6.2. Auto correlación 20](#_Toc440819626)

[3.6.2.1. Función de Auto correlación Simple (AFC) 21](#_Toc440819627)

[3.6.2.2. Función de Auto correlación Parcial (PAFC) 22](#_Toc440819628)

[3.7. Análisis espectral 22](#_Toc440819629)

[3.7.1. Representación de Fourier de una sucesión 23](#_Toc440819630)

[3.7.2. Representación de una sucesión periódica 24](#_Toc440819631)

[3.7.3. Transformada Discreta de Fourier. 24](#_Toc440819632)

[3.7.4. Transformada Rápida de Fourier. 25](#_Toc440819633)

[3.7.5. Densidad Espectral (SPD) 26](#_Toc440819634)

[3.7.6. Periodograma 26](#_Toc440819635)

[4. Estado del Arte 28](#_Toc440819636)

[5. Métodos de identificación de Periodicidad. 30](#_Toc440819637)

[5.1. Método de Vlachos, Yu & Castelli. 30](#_Toc440819638)

[5.1.1. Descubriendo Periodos candidatos 31](#_Toc440819639)

[5.1.2. Validación de Periodos candidatos. 33](#_Toc440819640)

[5.1.3. Identificación del pico cercano 34](#_Toc440819641)

[5.2. Método de detección propuesto en la investigación (Método autoMM) 36](#_Toc440819642)

[5.2.1. Suavizado de la muestra, Medias Móviles 36](#_Toc440819643)

[5.2.2. Encontrando rezagos importantes con la función de Autocorrelación 38](#_Toc440819644)

[5.2.3. Diferenciación de rezagos, extrayendo periodos candidatos 40](#_Toc440819645)

[5.2.4. Filtrado de periodos candidatos, identificando periodos finales 40](#_Toc440819646)

[6. Muestras utilizadas en la investigación 41](#_Toc440819647)

[6.1. Organizaciones involucradas en las muestras utilizadas 41](#_Toc440819648)

[6.1.1. The Text REtrieval Conference (TREC) 41](#_Toc440819649)

[6.1.2. WEB TRACKS de TREC 42](#_Toc440819650)

[6.1.3. Google Trends 42](#_Toc440819651)

[6.2. Origen de las muestras utilizadas en la investigación 43](#_Toc440819652)

[6.3. Preparación de la data 43](#_Toc440819653)

[7. Aplicación de Método de Vlachos y Método autoMM. 44](#_Toc440819654)

[7.1. Consultas tipo Ráfaga. 44](#_Toc440819655)

[7.1.1. Consulta 1: “get rich” 44](#_Toc440819656)

[7.1.1.1. Método de Vlachos 45](#_Toc440819657)

[7.1.1.2. Método AutoMM 47](#_Toc440819658)

[7.1.2. Consulta 2: “google earth” 50](#_Toc440819659)

[7.1.2.1. Método de Vlachos 51](#_Toc440819660)

[7.1.2.2. Método autoMM 51](#_Toc440819661)

[7.1.3. Conclusión métodos consultas tipo Ráfaga 54](#_Toc440819662)

[7.2. Consultas tipo Estables. 55](#_Toc440819663)

[7.2.1. Consulta 1: “google images” 55](#_Toc440819664)

[7.2.1.1. Método de Vlachos 55](#_Toc440819665)

[7.2.1.2. Método autoMM 56](#_Toc440819666)

[7.2.2. Consulta 2: “cats” 60](#_Toc440819667)

[7.2.2.1. Método de Vlachos 61](#_Toc440819668)

[7.2.2.2. Método autoMM 62](#_Toc440819669)

[7.2.3. Conclusión métodos consultas tipo Estables. 65](#_Toc440819670)

[7.3. Consultas tipo Periódica 66](#_Toc440819671)

[7.3.1. Consulta 1: “mistletoe” 66](#_Toc440819672)

[7.3.1.1. Método de Vlachos 66](#_Toc440819673)

[7.3.1.2. Método autoMM 69](#_Toc440819674)

[7.3.2. Consulta 2: “farm and fleet” 72](#_Toc440819675)

[7.3.2.1. Método de Vlachos 73](#_Toc440819676)

[7.3.2.2. Método autoMM 75](#_Toc440819677)

[7.3.3. Conclusión métodos consultas tipo Periódica 78](#_Toc440819678)

[8. Análisis Comparativo 79](#_Toc440819679)

[8.1. Proporción de Acuerdo Observado y Coeficiente kappa de Cohen. 79](#_Toc440819680)

[8.1.1. Proporción de Acuerdo Observado 79](#_Toc440819681)

[8.1.2. Índice Kappa de Cohen 80](#_Toc440819682)

[8.2. Matriz de confusión. 82](#_Toc440819683)

[8.2.1. Métricas de la matriz de confusión. 83](#_Toc440819684)

[8.3. Preparando los casos de prueba para la Evaluación de Métodos 84](#_Toc440819685)

[8.4. Procedimiento para la Evaluación 85](#_Toc440819686)

[8.4.1. Resultados Proporción de Acuerdo Observado para Método AutoMM 86](#_Toc440819687)

[8.4.2. Resultado Índice de Proporción de Acuerdo para Método de Vlachos 89](#_Toc440819688)

[8.4.3. Resultados Índices de kappa de Cohen para Método AutoMM 90](#_Toc440819689)

[8.4.4. Resultados Índice de Kappa de Cohen para Método de Vlachos 93](#_Toc440819690)

[8.4.5. Resultados Matriz de Concordancia 93](#_Toc440819691)

[8.5. Análisis de resultados 95](#_Toc440819692)

[8.6. Métricas Objetivos Específicos 96](#_Toc440819693)

[8.6.1. Métrica Objetivo Especifico 1 96](#_Toc440819694)

[8.6.2. Métrica Objetivo Especifico 2 97](#_Toc440819695)

[8.6.3. Métrica Objetivo Especifico 3 98](#_Toc440819696)

[9. Conclusión 98](#_Toc440819697)

**ÍNDICE DE TABLAS**

[Tabla 1: Situación Actual, Resultados Esperados, Métricas y Criterios de Éxito. 7](#_Toc440819698)

[*Tabla 2: Tipos y nombres de las consultas utilizadas en la demostración.* 44](#_Toc440819699)

[*Tabla 3: Periodo candidato Periodograma con su rango de alcance, consulta "get rich"..* 46](#_Toc440819700)

[Tabla 4: Rezagos importantes función de autocorrelación, consulta "get rich".. 46](#_Toc440819701)

[*Tabla 5: Periodos candidatos por iteración de medias móviles consulta "get rich".* 49](#_Toc440819702)

[*Tabla 6: Candidatos por media móvil.* 54](#_Toc440819703)

[*Tabla 7: Periodos candidatos por iteración de medias móviles consulta "google images".* 59](#_Toc440819704)

[*Tabla 8: Periodos Candidatos por iteración de medias móviles consulta "cats".* 64](#_Toc440819705)

[*Tabla 9: Candidatos finales y su frecuencia de aparición consulta "cats".* 65](#_Toc440819706)

[*Tabla 10: Periodos candidatos del Periodograma y sus rangos de alcance consulta "mistletoe".* 67](#_Toc440819707)

[*Tabla 11: Rezagos importantes de la función de autocorrelación consulta "mistletoe"* 68](#_Toc440819708)

[*Tabla 12: Periodos candidatos por media móvil consulta "mistletoe".* 71](#_Toc440819709)

[*Tabla 13: Periodos candidatos finales y su frecuencia de aparición en las iteraciones de suavizado.* 71](#_Toc440819710)

[*Tabla 14: Tabla de periodos finales del método AutoMM.* 72](#_Toc440819711)

[*Tabla 15: Periodos candidatos Periodograma y sus rangos de alcance.* 73](#_Toc440819712)

[*Tabla 16: Rezagos importantes del resultado de la función autocorrelación de la consulta "farm and fleet".* 74](#_Toc440819713)

[Tabla 17: Rezagos importantes extraídos de la función de autocorrelación consulta "farm and fleet". 75](#_Toc440819714)

[*Tabla 18: Periodos candidatos por iteración de media móvil.* 77](#_Toc440819715)

[*Tabla 19: Periodos finales y su frecuencia de aparición en las iteraciones.* 77](#_Toc440819716)

[Tabla 20: Tabla Distribución de Acuerdos. 79](#_Toc440819717)

[Tabla 21: Métricas de medición del índice de kappa de cohen. 81](#_Toc440819718)

[Tabla 22: Clasificación índice kappa por Fleiss, Cicchetti & Sparrow.. 81](#_Toc440819719)

[Tabla 23: Tabla matriz de confusión para un clasificador.. 82](#_Toc440819720)

[Tabla 24: Conjuntos de tipos de consultas para evaluar la calidad de resultados los métodos VL y AutoMM. 84](#_Toc440819721)

[Tabla 25: Tabla de Resultados Índice de Acuerdo Observado por media móvil del 1 al 20. 86](#_Toc440819722)

[Tabla 26: Tabla de Resultados Índice de Acuerdo Observado por media móvil del 21 al 40. 87](#_Toc440819723)

[Tabla 27: Valor máximo y mínimo índice de acuerdo observado por conjunto de método Vlachos. 89](#_Toc440819724)

[Tabla 28: Tabla de Resultados Índice Kappa de Cohen de AutoMM por tipo de consulta, media móvil del 1 al 20. 90](#_Toc440819725)

[Tabla 29: Tabla de Resultados Índice Kappa de Cohen de Auto MM por tipo de consultas, medias móviles del 21 al 40. 91](#_Toc440819726)

[Tabla 30: Valor máximo y mínimo índice de acuerdo observado. 93](#_Toc440819727)

[Tabla 31: Tabla de resultados matriz de concordancia conjuntos PQ-SQ-AQ-OBQ. 94](#_Toc440819728)

[Tabla 32: Tabla de resultados matriz de concordancia conjuntos PQ-SQ. 94](#_Toc440819729)

[Tabla 33: Tabla de resultados matriz de confusión conjuntos PQ-AQ. 94](#_Toc440819730)

[Tabla 34: Tabla de resultados matriz de confusión PQ-OBQ. 94](#_Toc440819731)

[Tabla 35: Objetivo Especifico 1. 96](#_Toc440819732)

[Tabla 36: Objetivo Especifico 2. 97](#_Toc440819733)

[Tabla 37: Objetivo Especifico 3. 98](#_Toc440819734)

**ÍNDICE DE ILUSTRACIONES**

[Ilustración 1: Gráfico de una señal analógica o continua. 9](#_Toc440818973)

[Ilustración 2: Gráfico que representa una señal digital o discreta.. 9](#_Toc440818974)

[Ilustración 3: Conversión señal continua a discreta. 10](#_Toc440818975)

[Ilustración 4: Ejemplo de señal de carácter periódico, consulta “christmas present”. 11](#_Toc440818976)

[Ilustración 5: Consulta permanente "Java jdk". 12](#_Toc440818977)

[Ilustración 6: Ejemplo de Señales en Ráfaga, consulta “octopus paul”. 13](#_Toc440818978)

[Ilustración 7: Gráfico que representa las componentes de una serie por separado. 14](#_Toc440818979)

[Ilustración 8: Diagrama de la metodología propuesta por Vlachos. 31](#_Toc440818980)

[Ilustración 9: Extracción de periodos candidatos Periodograma. 32](#_Toc440818981)

[Ilustración 10: Proceso de identificación de periodos a través de método de Vlachos, Yu & Castelli de la consulta "Navidad". 34](#_Toc440818982)

[Ilustración 11: Diagrama de la metodología propuesta en la Investigación. 36](#_Toc440818983)

[Ilustración 12: Suavizado de la muestra y resultado de Autocorrelación. 37](#_Toc440818984)

[Ilustración 13: Extracción de rezagos de la función de autocorrelación. 39](#_Toc440818985)

[Ilustración 14: Eliminación de rezagos candidatos no importantes. 39](#_Toc440818986)

[*Ilustración 15: Serie de tiempo de la consulta "get rich" en semanas.* 45](#_Toc440818987)

[*Ilustración 16: Periodograma de la consulta "get rich".* 45](#_Toc440818988)

[*Ilustración 17: Gráfico de autocorrelación, intervalo de confianza y rezago candidato.* 46](#_Toc440818989)

[*Ilustración 18: Función de autocorrelación consulta "get rich" secuencia original.* 47](#_Toc440818990)

[*Ilustración 19: Gráfico de la muestra con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación.* 48](#_Toc440818991)

[*Ilustración 20: Gráfico de la muestra Media Móvil 4 y 5 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia* 49](#_Toc440818992)

[*Ilustración 21: Serie de tiempo de la consulta "google earth" en semanas.* 50](#_Toc440818993)

[*Ilustración 22: Gráfico Periodograma de la consulta "google earth".* 51](#_Toc440818994)

[*Ilustración 23: Gráfico de autocorrelación muestra original consulta "google earth".* 52](#_Toc440818995)

[*Ilustración 24: Gráfico de la serie de tiempo “Google Earth” con Media Móvil 2 (izquierda) y 3 (derecha) con su respectivo gráfico de autocorrelación.* 53](#_Toc440818996)

[*Ilustración 25: Gráfico de la serie de tiempo “Google Earth” con Media Móvil 4 (izquierda) y 5 (derecha) con su respectivo gráfico de autocorrelación.* 53](#_Toc440818997)

[*Ilustración 26: Serie de tiempo en semanas consulta "google images".* 55](#_Toc440818998)

[*Ilustración 27: Gráfico de Función de autocorrelación de la consulta "google images"* 56](#_Toc440818999)

[*Ilustración 28: Gráfico función de autocorrelación consulta "google images".* 56](#_Toc440819000)

[*Ilustración 29: Gráfico de la función de autocorrelación para la serie original "google images"* 57](#_Toc440819001)

[*Ilustración 30: Gráfico de la serie de tiempo “google images” con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación.* 58](#_Toc440819002)

[*Ilustración 31: Gráfico de la muestra con Media Móvil 4 y 5 con su respectivo gráfico de autocorrelación.* 59](#_Toc440819003)

[*Ilustración 32: Serie de tiempo de la consulta "cats" en semanas.* 60](#_Toc440819004)

[*Ilustración 33: Gráfico Periodograma de la consulta "cats".* 61](#_Toc440819005)

[*Ilustración 34: Gráfico función autocorrelación de la consulta "cats".* 61](#_Toc440819006)

[*Ilustración 35: Gráfico función autocorrelación serie original consulta "cats".* 62](#_Toc440819007)

[*Ilustración 36: Gráfico de la muestra con Media Móvil 2 (izquierda) y 3 (derecha) con su respectivo gráfico de autocorrelación consulta “cats”.* 63](#_Toc440819008)

[*Ilustración 37: Gráfico de la muestra con Media Móvil 4 y 5 con su respectivo gráfico de autocorrelación.* 64](#_Toc440819009)

[*Ilustración 38: Serie de tiempo en semanas de la consulta "mistletoe".* 66](#_Toc440819010)

[*Ilustración 39: Gráfico de la función de autocorrelación relacionado a la consulta "mistletoe".* 67](#_Toc440819011)

[*Ilustración 40: Gráfico función de autocorrelación consulta "mistletoe".* 68](#_Toc440819012)

[*Ilustración 41: Gráfico función autocorrelación de la serie de tiempo original consulta "mistletoe".* 69](#_Toc440819013)

[*Ilustración 42: Gráfico de la muestra con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación.* 70](#_Toc440819014)

[*Ilustración 43: Gráfico de la muestra con Media Móvil 4 y 5 con su respectivo gráfico de autocorrelación.* 70](#_Toc440819015)

[*Ilustración 44: Serie de tiempo original en semanas consulta "farm and fleet"* 72](#_Toc440819016)

[Ilustración 45: Periodograma consulta "farm and fleet". 73](#_Toc440819017)

[*Ilustración 46: Gráfico de autocorrelación de la serie de tiempo original consulta "farm and fleet".* 74](#_Toc440819018)

[*Ilustración 47: Gráfico de función de autocorrelación de la muestra original consulta "farm and fleet".* 75](#_Toc440819019)

[*Ilustración 48: Gráfico de la muestra con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación.* 76](#_Toc440819020)

[*Ilustración 49: Gráfico de la muestra con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación.* 76](#_Toc440819021)

[Ilustración 50: Gráfico de Resultados Índice de Acuerdo Observado por conjuntos de clasificación, Medias móviles del 1 al 40. 88](#_Toc440819022)

[Ilustración 51: Gráfico índices kappa de cohen por conjuntos de clasificación de consultas, iteración de medias móviles del 1 al 40. 92](#_Toc440819023)

**RESUMEN**

La presente investigación se enfoca en la determinación de un método efectivo para la identificación automatizada de señales periódicas y no periódicas dentro del contexto de la web.

La investigación enfrenta la problemática del escaso número de métodos automatizados para la identificación y clasificación de señales periódicas y no periódicas, la incertidumbre en la efectividad y fiabilidad de estos. Como también la falta de datasets o conjunto de prueba de señales periódicas y no periódicas formalmente clasificadas en publicaciones relacionadas a series de tiempo. Todo esto relacionado al contexto de las aplicaciones web.

La hipótesis planteada en la investigación es comprobar si “Es posible establecer un método automatizado para la detección de señales periódicas y no periódicas que sea efectivo y fiable”.

A continuación se presenta el desarrollo de la investigación, que comprende aspectos de estudio de herramientas, diseño e implementación de métodos automatizados en etapa experimental.

# Introducción

## Motivación

Las aplicaciones web son herramientas capaces de generar una gran cantidad de datos de distinta índole dependiendo del objetivo para el cual haya sido creada, ya sea dentro del contexto de una organización corporativa, donde la información está relacionada al ámbito laboral que lo representa, o bien, en el contexto del usuario casual, donde la información es variada y de ámbito general.

En la actualidad, la mayoría de las aplicaciones web están fuertemente ligadas a las redes sociales, donde la gente lee y/o publica información de distinta índole, lo cual permite la interacción entre las personas, facilitando el intercambio de información relevante a través del medio. Esta interacción entre aplicaciones web, redes sociales y personas, genera una gran cantidad de datos de distintos eventos relevantes de la sociedad, que quedan registrados como acontecimientos históricos a través del tiempo.

Estos datos, al quedar registrados, pueden ser tratados como señales a través del tiempo. Estas señales se pueden clasificar en señales en ráfaga, señales permanentes y señales periódicas. (Chen Z. , Yang, Ma, Lei, & Gao, 2011)

La detección de señales periódicas y el análisis de patrones repetitivos, tienen una amplia aplicabilidad en diversos contextos de la actividad humana. Específicamente en el contexto de la web, donde pueden utilizarse para determinar tópicos recurrentes a través del tiempo y además determinar su comportamiento futuro.

En esta investigación se realizará un estudio de las distintas herramientas y metodologías utilizadas para la identificación de señales periódicas con el objetivo de realizar un estudio de evaluación de métodos y determinar su efectividad y grado de fiabilidad.

## Contribuciones de la tesis

La investigación expone el uso de distintas herramientas enfocadas a la detección de periodos dentro de distintos tipos de series de tiempo, con esto, se implementaron dos métodos automatizados para la identificación de señales periódicas.

El primer método se enfoca en transformar la señal al dominio de las frecuencias con la idea de encontrar posibles periodos los cuales son refinados con la herramienta de autocorrelación, todo esto sin alterar la señal original. Y el segundo método se enfoca en alterar la muestra, realizando iteraciones de suavizado por medio de medias móviles, con la finalidad de minimizar las variaciones de esta, disminuyendo la componente de aleatoriedad para luego encontrar periodos utilizando la herramienta de autocorrelación, diferenciando los rezagos importantes de esta.

Con el objetivo de demostrar la efectividad de las distintas herramientas implementadas y también la fiabilidad en la detección automática de señales tipo periódica y no periódicas.

## Contextualización

Una señal es cualquier magnitud que se puede medir de alguna forma y que contiene información sobre el comportamiento o la naturaleza de algún fenómeno. Están presentes en la vida diaria; como por ejemplo, sirven para controlar la salud de las personas (electrocardiogramas, ecografías), mostrar variaciones de índices económicos (IPC, variación del Dólar, valor de la UF, etc.), estudiar el comportamiento de fenómenos naturales (pronóstico del tiempo, variaciones de temperatura, etc.), comunicar y también para almacenar y recuperar información de distinta índole.

Desde el punto de vista matemático una señal es una función de una o varias variables cuyo dominio de definición puede ser continuo o una cantidad finita de muestras u observaciones.

Algunas señales contienen observaciones que son registradas en instantes de tiempo sucesivos, por ejemplo, horas, días, meses, trimestres, semestres, etc., o bien registradas por algún equipo de forma continua. Estas observaciones se pueden utilizar para realizar previsiones de sucesos futuros, con el fin de realizar una planificación adecuada en un área de conocimiento en particular. Estas previsiones se basan en el análisis de sucesos pasados, los cuales son llamados análisis de series de tiempo.

Las series de tiempo son un conjunto de mediciones de cierto fenómeno o experimento registradas secuencialmente en el tiempo. Pueden ser denotadas por un valor en particular, el cual puede ser, la frecuencia de una acción o evento, el valor de una medición financiera o científica, etc, las cuales fueron registradas en un instante de tiempo en determinado.

En el contexto de la web, estas series de tiempo se pueden generar acorde al historial de frecuencias, ya sea para cada consulta en un motor de búsqueda o base de datos, acciones generadas por los usuarios en una aplicación web o cantidad de clics realizadas dentro de una página web. Estas acciones o consultas se representan como una frecuencia en una unidad de tiempo determinado (ya sea en días o meses) generando datos en base a una serie de tiempo.

La periodicidad es una interesante propiedad para muchos conjuntos de series de tiempo. Un periodo puede ser informalmente definido como un patrón repetido en sí mismo. Este patrón provee información útil acerca de la estructura inherente de una señal.

La importancia de la periodicidad puede ser medida o estimada, por el hecho de conceptos básicos como el número de horas en el día y número de días en el año, los cuales son motivo de la naturaleza periódica del movimiento de la tierra.

El patrón periódico es utilizado en distintas áreas como:

* En ciencias naturales, el análisis periódico y la estimación de la periodicidad es importante en el aspecto de las disciplinas como astronomía, meteorología y oceanología. Ya que, sugieren un potencial de anomalías y ayudan a entender la relación causal entre diferentes procesos. Por ejemplo, está bien establecido que la variabilidad solar afecta enormemente al cambio climático. (Friss-Cristensen & Lassen, 1991).
* En Medicina, donde se utilizan medidas biométricas que exhiben fuertes periodicidades, existe un gran interés en detectar anomalías en periodicidades. Disturbios en patrones periódicos pueden ser notados en muchas enfermedades. Por ejemplo, el síndrome de Touretes en los pacientes exhibe un aumento de la frecuencia de pestañeo normal de una persona (Tulena, y otros, 1997).

En industrias de manufactura, es una herramienta indispensable en máquinas autónomas. Puede ser utilizado para el examen de espectros de vibración causados por partes rotatorias. Donde, un cambio en la estructura periódica de las vibraciones de la máquina son un buen indicador de desgaste de máquinas.

Utilizando un modelado de series de tiempo, en el contexto web, es posible estimar tendencias y periodicidades para predecir valores futuros en la frecuencia de las consultas, acciones del usuario o demanda de transacciones en una página web enfocada al comercio. Se ha estimado que más del 7% del tráfico de búsqueda son consultas cíclicas relacionadas a eventos estacionales (Metzler, Jones, Peng, & Zhang, 2009).

Esta búsqueda del comportamiento es importante para los motores de búsqueda, ya que permite detectar y predecir consultas estacionales y responder a estas con resultados relevantes en el tiempo, por lo que puede ser utilizada en la mejora de las políticas de rastreo y ranking de los resultados de búsqueda. O en el caso de los sistemas operativos y sistemas de bases de datos, para detectar periodos de baja actividad para agendar tareas de mantenimiento.

# Descripción del Problema

## Enunciado del Problema

El análisis de eventos pasados permite realizar predicciones de eventos futuros con el objetivo de prever y evitar posibles problemas, como también planificar y sacar ventaja de este conocimiento.

En el caso de las aplicaciones web, el análisis de señales relacionadas al comportamiento de los usuarios puede considerarse como una oportunidad de mejora en el funcionamiento de las características de las aplicaciones o servicios web, con el fin de conocer la ocurrencia de eventos sucesivos o próximos.

Utilizando la información obtenida de las acciones de los usuarios, se puede predecir el comportamiento futuro de estos, lo cual se puede utilizar para mejorar los servicios de aplicaciones web. Por ejemplo, en los motores de búsqueda se pueden indexar las consultas de tipo periódica de tal manera de que la información de las respuestas a estas sean correspondientes a hechos que sucedieron en los eventos de alzas de frecuencias periódicas, o bien, en el evento próximo relacionado a la consulta.

También permiten prevenir posibles problemas a futuro con respecto a la sobrecarga de servidores teniendo en cuenta eventos anteriores de alzas abruptas periódicas, como por ejemplo, en la cantidad de conexiones simultáneas y extracción de información en un intervalo corto de tiempo.

Todo esto ayudará a mejorar la satisfacción de las necesidades y experiencias del usuario dentro de las aplicaciones, que es una de las prioridades en los servicios webs.

Dentro del estudio realizado, se hallaron varias publicaciones que exponen el uso de diversas herramientas para identificar patrones periódicos dentro de señales relacionadas al contexto de la web, pero existe un escaso número que se dedica a proponer y exponer métodos automatizados para la identificación y clasificación de señales tipo periódicas. Además, estos métodos propuestos están evaluados por eficiencia de cómputo pero no se tiene certeza de su efectividad y fiabilidad.

Otro aspecto importante es la falta de datasets o conjuntos de señales ya clasificados de manera formal en publicaciones relacionadas al contexto web, sin esto la dificultad en la evaluación de la implementación de métodos automatizados aumenta.

## Objetivos

## Objetivo General

Determinar un método efectivo para la identificación automatizada de señales periódicas y no periódicas en el contexto de la web

## Objetivos Específicos

1. Implementar y automatizar un conjunto de métodos para la detección de señales periódicas.[OE1]
2. Determinar un conjunto de casos de estudios para la evaluación experimental de dichos métodos.[OE2]
3. Evaluar y seleccionar un método efectivo para determinar señales periódicas.[OE3]

## Resultados Esperados, Métricas y Criterios de Éxito

Tabla 1: Situación Actual, Resultados Esperados, Métricas y Criterios de Éxito. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obj.Esp | Situación Actual | Resultados Esperados | Métricas | Criterios de Éxito |
| [OE1] | No se tienen métodos implementados ni tampoco herramientas de software que permitan automatizar la detección de señales periódicas. | Un conjunto de métodos implementados para realizar una etapa experimental para la detección de señales periódicas. | Numero de métodos implementados. | >=2 |
| [OE2] | No existe en la literatura un dataset consistente para evaluar señales periódicas. | Un dataset de series de tiempo en el contexto web para la evaluación experimental. Este dataset está compuesto por distintos tipos de señales. | Número de series de tiempo. | >=400 |
| [OE3] | No existe en la literatura referencia en términos de calidad para evaluar la efectividad de los métodos. | Un método con efectividad alta para la discriminación entre señales periódicas y no periódicas. | RAW | >=80% |
| Kappa de Cohen | >=60% |
| Precision, Recall, Specificity | >60% |

## Hipótesis

Es posible establecer un procedimiento automatizado que detecte señales periódicas y no periódicas que sea fiable.

## Alcance

El presente trabajo estará circunscrito al siguiente alcance:

* Sólo se analizarán señales relacionadas a aplicaciones web (consultas o acciones).
* Sólo se analizarán señales con tiempo secuencial separados de manera uniforme (semanas).

# Marco Teórico

## Señales

Se define una señal como una función matemática que depende de una o más variables independientes cuyo valor da información sobre el fenómeno físico al que está asociado.

Existen dos tipos de señales básicas, las ***señales en tiempo continuo o analógico*** y las ***señales de tiempo discreto o digitales*** (Morón, 2011).

## Señal de tiempo continuo o analógico

Una señal es una señal de tiempo continuo cuando la variable independiente es una variable continua. Estas señales están definidas para una “línea” continua de valores de esa variable; es decir, el valor puede ser especificado en cualquier instante de un intervalo de tiempo dado, ya sea mediante una expresión matemática o gráficamente por medio de una curva; en otras palabras, la variable independiente puede tomar cualquier valor real (Morón, 2011).

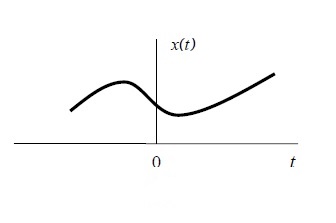


Ilustración : Gráfico de una señal analógica o continua. Fuente: Señales y Sistemas, Morón, 2011.

## Señal digital o de tiempo discreto

Si está definida en puntos de tiempo discretos, entonces es una señal en tiempo discreto, generalmente generada a través de un muestreo de la señal de tiempo continuo. Como una señal de tiempo discreto está definida solamente en tiempos discretos, como se aprecia en la ilustración 2, se puede identificar como una secuencia de números, denotada por o , donde, para nuestros propósitos, es un entero.

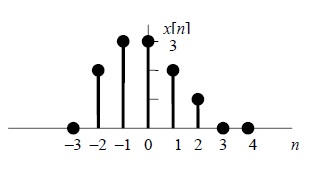


Ilustración : Gráfico que representa una señal digital o discreta. Fuente: Señales y Sistemas, Morón, 2011.

Una señal de tiempo discreto puede representar un fenómeno para el cual la variable independiente es siempre discreta. Por ejemplo, el promedio diario de los valores de la bolsa de valores es, por su naturaleza, una señal que evoluciona en puntos discretos en el tiempo (es decir, el cierre del día). En el caso de las aplicaciones web, la cantidad de consultas realizadas en un intervalo puntual, que puede ser promediado en días, semanas, meses, etc.

Como se aprecia en la ilustración 3, una señal de tiempo discreto, , también puede obtenerse mediante un muestreo de una señal de tiempo continuo para obtener los valores, a través de un dispositivo o procedimiento que convierta la información analógica a forma digital mediante redondeo llamado convertidor analógico-digital.

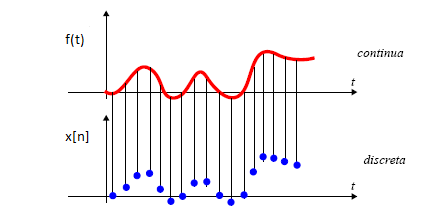


Ilustración : Conversión señal continua a discreta. Fuente: Elaboración Propia.

Una señal continúa con observaciones puede denotarse en su forma discreta en una forma abreviada como o .

Y a los valores de se les denomina muestras u observaciones; el intervalo de tiempo entre muestras se llama intervalo de muestreo. Cuando estos intervalos son iguales (muestra uniforme), entonces , donde la constante es el intervalo de muestreo.

En el caso de las señales de aplicaciones web, al ser de forma digital, al tener una cantidad discreta de valores, se interpretan como una señal discreta, la cual, al tener sus intervalos de muestreo uniformes, en este caso tiempo, se pueden interpretar como una serie de tiempo.

## Tipos de señales en aplicaciones web.

Dentro de las aplicaciones web las consultas se pueden clasificar de manera temporal con respecto a su frecuencia (Chen, Yang, Ma, Lei, & Gao, 2011). Estas se clasifican en:

* Periódicas.
* Permanentes.
* Ráfaga.

## Periódica

Son señales realizadas periódicamente y siguen un patrón idéntico o similar durante periodos sucesivos. Por ejemplo “Christmas Present” (Regalos de navidad en inglés) aumenta su frecuencia en un ciclo anual ya que la gente tiende a elegir sus regalos de navidad en Diciembre de cada año.

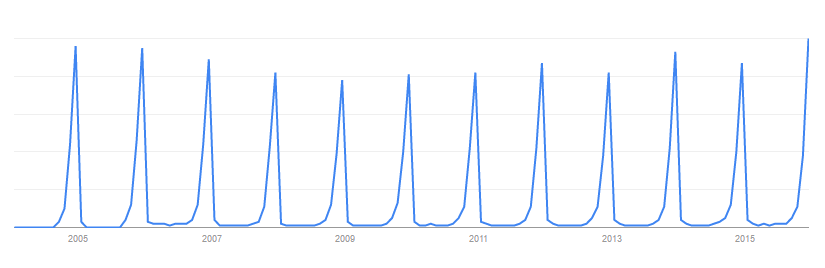


Ilustración : Ejemplo de señal de carácter periódico, consulta “christmas present”. Fuente: Google Trends.

## Permanentes

Representan la necesidad de información usual, frecuente y constante de los usuarios. Este comportamiento mantiene una frecuencia constante en el tiempo, puede oscilar pero de manera disminuida y no tiene alzas abruptas ni comportamientos periódicos. Por ejemplo la consulta “Java jdk”, cuyos resultados se muestra en la Ilustración 5, representa una señal permanente.

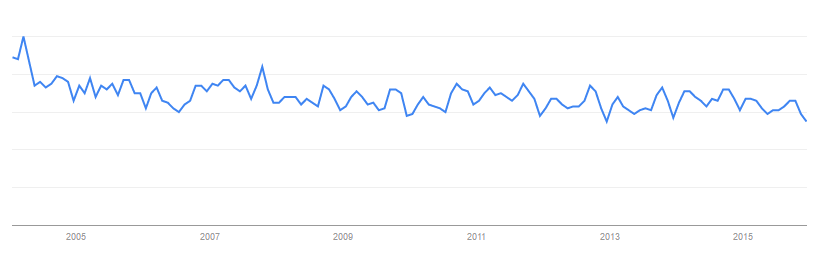


Ilustración : Consulta permanente "Java jdk". Fuente: Google Trends.

En la Ilustración 5, se aprecia el comportamiento de la consulta “java jdk”, la cual, mantiene una frecuencia constante y tiene oscilaciones que no son predominantes, como también, sus alzas no son abruptas y pueden representar la salida de una nueva versión de “Java jdk”.

## Ráfaga

Las señales en ráfaga se dan habitualmente en aplicaciones web relacionadas a redes sociales, como también en motores de búsqueda, donde las consultas se basan en la búsqueda de información acerca de un tema social.

Donde, se produce un alza abrupta en la frecuencia en un intervalo de tiempo específico y que en el pasado era relativamente raro de ser consultado, luego de llegar a su máxima frecuencia esta disminuye considerablemente llegando al número de consultas de los intervalos anteriores previo al alza, en motores de búsqueda estas señales son nombradas como “consultas en ráfaga”.



Ilustración : Ejemplo de Señales en Ráfaga, consulta “octopus paul”. Fuente: Google Trends.

En la ilustración 6 se aprecia una “señal de ráfaga”, ya que su comportamiento normal se ve interrumpido por un alza abrupta en la frecuencia de las consultas, para luego, descender y volver al comportamiento previo a dicho evento.

## Series Temporales

Una serie temporal, como se dijo anteriormente, es una secuencia de datos a través del tiempo ordenados cronológicamente y espaciados entre sí de manera uniforme, ya sea por horas, días, meses y años. Estos datos son denotados como con como el valor de la variable de en el instante (Villavicencio, s.f).

Permiten extraer información con la idea de realizar distintos tipos de análisis, principalmente para conocer su comportamiento en el pasado o realizar futuras predicciones.

Estas series temporales se componen de tres componentes (Molinero, 2004). Las cuales son:

* **Tendencia:** Se ve reflejada en el constante aumento o decrecimiento de los valores a lo largo del tiempo o el cambio a largo plazo de la media.
* **Estacionalidad:** Componente que representa un comportamiento reiterativo de valores en uno o más periodos.
* **Aleatoriedad o ruido:** Componente que no responde a ningún patrón de comportamiento y no es posible extraer información de ninguna índole. Es el resultado de factores fortuitos o aleatorios que no tienen ni una relación entre sí dentro de la serie de tiempo.

De las tres componentes nombradas, las dos primeras son determinísticas, mientras que la última es aleatoria. Así, la serie se puede denotar como:

Donde:

: Serie de tiempo original.

: Componente de la tendencia en la serie.

: Componente estacional.

: Componente aleatoria.

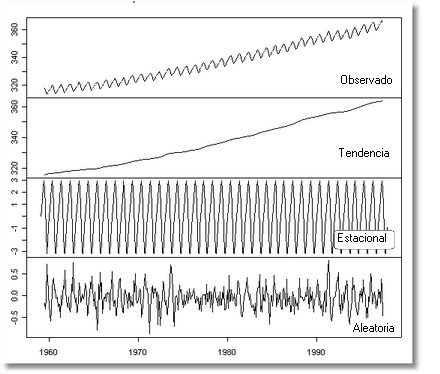


Ilustración : Gráfico que representa las componentes de una serie por separado. Fuente: Análisis de Series Temporales (2004), Luis M. Molinero.

## Clasificación descriptiva de series

La serie de tiempo se puede clasificar en Estacionaria o No estacionaria dependiendo de las siguientes condiciones (Villavicencio, s.f):

**Estacionaria:** Se dice estacionaria cuando es estable, es decir, cuando la media y variabilidad son constantes a lo largo del tiempo. Esto refleja que gráficamente los valores de la serie tienden a oscilar alrededor de una media constante y la variabilidad con respecto a esa media también prevalece constante en el tiempo.

**No estacionaria:** Son series donde la media y/o variabilidad cambian en el tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a lo largo del plazo, por lo que la serie no oscila alrededor de un valor constante.

## Procesos Estocásticos

Se describen como una secuencia de datos que evolucionan en el tiempo. Las series temporales se definen como un caso particular de los procesos estocásticos.

## Proceso estocástico estacionario

Un proceso estocástico se dice ser estacionario cuando su media y varianza son constantes en el tiempo y si el valor de covarianza entre dos periodos depende solamente de la distancia o rezago entre estos dos periodos de tiempo y no del tiempo en el cual se ha calculado la covarianza (Villavicencio, s.f) (Peralta, s.f).

Sea una serie de tiempo, para que sea estacionaria sigue las siguientes propiedades:

Estable en Media:

Estable en Varianza:

Estable en autocovarianza

Donde , la covarianza (o autocovarianza) al rezago , es la covarianza entre los valores de y que están separados por instancias.

Resumiendo, si una serie de tiempo es estacionaria, su media, su varianza y autocovarianza (en diferentes rezagos) permanecen iguales sin importar el momento en el cual se midan; es decir, son invariantes en el tiempo. Pero esto es casi imposible. Es por esto que se definen dos posibles procesos estacionarios:

**Proceso estacionario en sentido estricto:** Si las distribuciones marginales de todas las variables son idénticas y además la distribución finitodimensional de cualquier conjunto de variables solo depende de los retardos. Es decir, si

Para cualquier y , donde, denota la distribución conjunta de .

**Proceso estacionario en sentido débil:** Si la serie es estable en media y en autocovarianza se dice ser estacionaria en sentido débil.

## Métodos para la eliminación de la componente Tendencia

Existen distintos métodos para eliminar el comportamiento de la tendencia con el fin de “corregir” la media de la serie y así observar de mejor manera la estacionalidad en la serie.

**Diferenciación:** este método consiste en suponer que la tendencia evoluciona lentamente en el tiempo, de manera que en el instante *t* la tendencia debe estar próxima a la tendencia en el instante . De esta forma se resta el valor de la serie con el siguiente para obtener una serie resultante libre de tendencia, se utiliza principalmente para atenuar el decrecimiento o aumento en media de la serie temporal. Esta operación consiste en pasar la serie original a una serie :

De este modo, la serie resultante es la componente estacionaria junto con la componente aleatoria.

**Transformación Logarítmica**: este método requiere del método anterior para ser llevada a cabo. Consiste en pasar los valores a valores logarítmicos , luego hacer una diferenciación. Se utiliza principalmente cuando la muestra presenta variaciones del coeficiente de variancia elevados.

## Medias Móviles

Una media móvil es la media de un conjunto de instancias durante un determinado periodo de tiempo. Usualmente se utilizan para conocer la tendencia a través del tiempo y suavizar la serie ofreciendo una correlación suavizada entre los valores y el transcurso del tiempo.

Se denomina media móvil de longitud *n* a la siguiente ecuación:

Donde es el promedio del periodo y representa el valor en el periodo de *t*. La constante *n* determina el número de periodos o largo del periodo a promediar.

Cabe recalcar que la serie de medias móviles comienza después de los primeros *n* datos, ya que los valores de MA empiezan después de primer conjunto de datos.

Por ejemplo, si se tiene una serie de tiempo en meses:

* Enero=200
* Febrero=230
* Marzo=260
* Abril = 180

La media móvil de Abril se obtiene promediando los valores reales de Enero Febrero y Marzo: MA (Marzo) = (200+230+260)/3=230. El pronóstico de Abril se obtiene promediando Febrero, Marzo y Abril: MA (abril) = (230+260+180)/3=223, y así sucesivamente.

Si se acierta con la longitud de las medias móviles el método puede ser útil para descubrir la tendencia. También, si se tiene en cuenta las variaciones cíclicas que se repiten cada cierto periodo eliminando la tendencia de la serie, el cálculo de las medias móviles permitiría descubrir si la serie es estacionaria.

Como se ha dicho anteriormente, una condición determinante para que la serie sea estacionaria es que su media a lo largo de la serie sea constante. En este caso, si se tiene un periodo , la media en el periodo donde , debiese de ser constante para dichos periodos.

## Análisis Estadístico

## Variables Estadísticas

En este segmento se explicará, de forma resumida, acerca algunas variables estadísticas que son de suma importancia para el entendimiento de los procedimientos que se hablarán más adelante.

## Promedio o Media Aritmética

Se define como el cociente de la suma de los valores observados de la variable en estudio y el número de observaciones.

Donde,

: promedio; es el valor de la observación; *i* es el índice del valor y *N* es el número total de observaciones de la muestra.

## Varianza

Es el promedio de las desviaciones de las observaciones con respecto a su media. Es una medida de variabilidad absoluta:

Donde,

: es el valor de la observación; *i* es el índice del valor y *N* es el número total de observaciones de la muestra.

## Covarianza

La covarianza, es una medida del grado en que dos variables aleatorias se mueven en la misma dirección o en direcciones opuestas de una con respecto a la otra. Si dos variables aleatorias generalmente se mueven en la misma dirección se dirá que tienen una covarianza positiva. Si por el contrario, tienden a moverse en direcciones contrarias, se dirá que tienen covarianza negativa. (Weisstein, s.f)

La covarianza se mide como el valor que se espera de los productos de las desviaciones de dos variables aleatorias respecto a sus correspondientes medias. Una varianza es un caso especial de covarianza.

Dadas dos variables estadísticas *x* e *y*. La covarianza se define como:

Donde, *N* es el largo de las series; y valor de las observaciones y es el número del registro dentro de la serie.

## Correlación

La correlación estadística determina la relación o dependencia que existe entre dos variables. Esto permite determinar si los cambios en una de las variables influyen en los cambios de la otra.

## Auto correlación

En ocasiones los valores de las variables de las series de tiempo no son independientes entre sí, sino que tienen valores determinados dependientes de valores anteriores. La auto correlación es una herramienta que sirve para encontrar patrones repetitivos dentro de una señal y existen dos formas de medir estas dependencias.

## Función de Auto correlación Simple (AFC)

La función de auto correlación simple mide la correlación entre dos valores de la misma serie separadas por periodos.

Dentro de la secuencia temporal de *N* observaciones se puede formar parejas de observaciones contiguas y calcular el coeficiente de correlación de estas parejas. Este coeficiente se denomina “coeficiente de auto correlación de orden 1” y se puede denotar como . Análogamente se puede formar parejas con puntos separados por una distancia 2, es decir . y calcular un nuevo coeficiente con retraso o lag de orden 2. Si se preparan muchas parejas con puntos separados a una distancia *k*, calcularemos el coeficiente de orden *k*.

Para el coeficiente de autocorrelación, se puede calcular un error estándar y con ello se establece un intervalo de confianza para dicho coeficiente. Esto permite identificar los retardos o valores con mayor correlación entre sí.

La función de autocorrelación es el conjunto de coeficientes de autocorrelación comenzando desde hasta un máximo que no puede exceder la mitad de los valores observados. Es de gran importancia para estudiar la estacionalidad de la serie, ya que si se presenta el caso, los valores separados entre sí por intervalos iguales al periodo estacional deben estar correlacionados de alguna forma. Es decir, para un retardo igual al periodo estacional, el coeficiente de autocorrelación es significativamente distinto de 0 sobrepasando el intervalo de confianza (Barlett, 1955).

Esta función tiene las siguientes propiedades:

Donde es el valor de autocorrelación en el retraso o lag *k*. Si este valor no está dentro del intervalo descrito anteriormente, quiere decir que los valores entre el rezago *k* están fuertemente correlacionados indicando una posible estacionalidad de periodo k.

## Función de Auto correlación Parcial (PAFC)

La auto correlación parcial, mide la correlación entre dos variables separadas por *k* periodos, pero a diferencia de la función simple, esta calcula la correlación entre parejas de valores separados a una distancia *k* sin considerar la dependencia creada por las parejas o retardos intermedios, eliminando la correlación producida por retardos anteriores a k.

## Análisis espectral

Es una forma alternativa de tratar el análisis de series temporales, la cual consiste en transformar la función de serie de tiempo al dominio de la frecuencia. La idea de esto es encontrar posibles comportamientos que no son visibles dentro del dominio del tiempo, como la estacionalidad (POLLOCK, s.f).

El Análisis Armónico o de Fourier, afirma que cualquier función analítica arbitraria, o serie de tiempo en el caso de esta investigación, puede ser perfectamente aproximada a través de la sumatoria de funciones seno y cosenos, por lo que dicha transformación descompone la función o señal de serie de tiempo en series o una función de ondas armónicas o sinusoidales. La precisión de dicha aproximación depende del número de funciones realizadas en la suma.

Esta representación espectral, basadas en ondas armónicas capaces de representar cualquier fenómeno, es una herramienta adecuada para sacar a la luz componentes periódicas, ya que consiste en expresar la serie como una suma de sucesiones periódicas, concretamente trigonométricas, por lo tanto, permite centrar la atención en las frecuencias predominantes.

Una de las herramientas utilizadas para realizar la transformación es la Transforma Rápida de Fourier la cual se hablará más adelante.

## Representación de Fourier de una sucesión

Dada una sucesión finita , puede escribirse como una suma de senos y cosenos de la siguiente manera:

Donde son las frecuencias naturales. Cuando *T* es par, *n = T/2*. Entonces, , mientras que y .

Lo que da, para par:

Y si es impar entonces y

Los coeficientes y son la Amplitud de las funciones sinodales.

Por otro lado los coeficientes y se obtienen como:

;

De esta forma la sucesión se puede representar como un conjunto de pares o como un conjunto de pares . La primera corresponde a la serie en el dominio del tiempo y la segunda en el dominio de las frecuencias. Donde se dice que es la energía contenida en la frecuencia .

También la serie de Fourier se puede expresar en exponenciales complejas. Sustituyendo y por:

,

Con esto se obtiene:

La cual es la serie de Fourier, lo que dicho anteriormente, representa la serie original en una función de series trigonométricas.

## Representación de una sucesión periódica

Se dice que una sucesión de longitud es periódica de periodo , donde .

Es decir, , por lo tanto, el conjunto de valores se repite sucesivamente. Se puede decir que para simplificar , se puede escribir:

Puesto que todos los senos y cosenos que aparecen son periódicos de periodo , entonces el lado derecho de la ecuación es también periódico. Esto significa que la ecuación anterior no solo cumple que sino que también para . En la representación de Fourier de una sucesión periódica de periodo aparecen solamente las frecuencias .

## Transformada Discreta de Fourier.

La idea central de la transformada de Fourier es sencilla. El análisis espectral descompone una señal en sus frecuencias constitutivas y almacena la amplitud de cada componente en el dominio de la frecuencia. Por lo tanto una Transformación de Fourier es la representación de la señal en el dominio de la frecuencia la cual contiene la misma información que la función original (en el dominio del tiempo). El análisis de Fourier permite examinar la señal desde otro punto de vista.

* Para una señal periódica utilizando la transformada de Fourier, toma los coeficientes espectrales y interpretándolos como armónicos de la serie que generan una línea espectral.
* Para la transformación en señales no periódicas, la transformada genera un espectro continuo en la función de la frecuencia digital.

La Transformada de Discreta de Fourier es considerada como una representación de la serie de Fourier enfocada a señales digitales no periódicas de longitud finita.

La TDF de una señal definida en el rango se define como:

Y también existe la Transformada Inversa de Fourier la cual permite recuperar la serie original en el dominio del tiempo:

## Transformada Rápida de Fourier.

Es un algoritmo que realiza el cálculo para la Transformada Discreta de Fourier. Su importancia radica en el hecho que elimina una gran parte de cálculos repetitivos de la FDT, por lo tanto retorna una solución más rápida. (Vasco, s.f)

(Cooley & Tukey, 1965) demuestran que es posible dividir el cómputo de la Transformada discreta de Fourier en dos partes pequeñas. Desde la definición de la TDF se puede obtener:

Se separa la TFD en dos términos los cuales se ven similares, uno para números pares y otro para números impares. El truco viene en el uso de la simetría de los términos. Porque el rango de es , mientras que el rango de es . Por propiedades de simetría se puede apreciar que se necesita realizar la mitad del cómputo por sub problema, con una eficiencia algorítmica de y respectivamente. Mientras se minimiza la transformada de Fourier en valores pares, se puede reaplicar la aproximación de “Divide y Vencerás”, disminuyendo el costo de cómputo a la mitad cada vez que se haga hasta que no haya un beneficio. Esta aproximación recursiva tiene un valor de computo de eficiencia algorítmica en el mejor de los casos de .

## Densidad Espectral (SPD)

La densidad espectral representa e informa la estimación de la distribución de la potencia o energía distribuida de una señal sobre las distintas frecuencias de las que está formada, es decir, su espectro. En el caso de las señales temporales, se puede interpretar como la representación gráfica de la contribución de cada frecuencia de ondas más influyentes dentro de la señal.

Gracias a esto, se pueden identificar las frecuencias de onda que representan la señal estacionaria de la onda, o también, tienen una mayor influencia en la representación de serie temporal.

Con esto se obtienen las magnitudes de las ondas en cada frecuencia y se observar de manera apropiada las ondas más influyentes o predominantes de la señal dentro del espectro. También se le suele llamar “Periodograma”.

## Periodograma

El Periodograma es un estimador de densidad de potencia de la señal y puede ser descrito de la siguiente manera:

Donde es la magnitud de la Transformada Rápida de Fourier equivalente a la densidad espectral de la Potencia de la serie de tiempo y es el número de muestras de la serie.

Si se presta atención, se puede advertir que los valores de FTT no son más que los valores complejos de la salida de la FFT para cada frecuencia discreta, el módulo elevado al cuadrado de la componente real y su componente imaginaria.

La representación gráfica de esta herramienta en la serie representa la densidad espectral de la señal donde el eje representa el “poder” de la señal de dicha frecuencia y el eje representa la frecuencia.

El uso de esta herramienta consiste en determinar las frecuencias de las señales (onda) con mayor poder o influencia dentro de la serie transformada y determinar los periodos de la señal original.

En esta herramienta hay que establecer lo siguiente:

* No hay estacionalidad si no hay picos destacables.
* Cada pico destacable identifica un periodo que incluso puede ser un ciclo.
* A cada amplitud destacable le corresponde una frecuencia cuya inversa es el periodo estacional o el ciclo, con lo que el Periodograma identifica la longitud del periodo estacional y en su caso el ciclo.

# Estado del Arte

Esta sección corresponde a un resumen de los trabajos más importantes relacionados al estudio de la detección de señales con carácter periódico relacionados a la web.

**(Vlachos, Meek, Vagena, & Gunopulos, 2004)** utilizan logs de consultas del motor de búsqueda de MSN (portal de servicios web de Microsoft). Para cada pregunta o frase realizada en la consulta, se construyen series de tiempo en relación a la frecuencia de las palabras utilizadas en dichas consultas. El estudio se dedica a realizar un análisis de la herramienta de la Transformada Discreta de Fourier (TFD) enfocada al uso de los coeficientes de Fourier, la reconstrucción de la serie de tiempo a través de estas y también la relación de estos coeficientes en el Poder de Densidad Espectral para detectar periodicidades dentro de la serie de tiempo. Logrando determinar un umbral en la herramienta de Poder de Densidad Espectral que resalta las frecuencias que contienen mayor energía en la señal.

**(Vlachos, Yu, & Castelli, 2005)** presentan un método automatizado no parametrizado para detectar periodicidades en señales de series de tiempo sin alterar sus datos. Resumiendo, su método se centra en el uso del Periodograma, herramienta similar al Poder de Densidad Espectral de la señal, para obtener periodos candidatos. Luego, se establecen ciertas condiciones de evaluación con el resultado de la herramienta de autocorrelación de la serie de tiempo para validar los periodos. Los datos de pruebas utilizados en este estudio para probar el método son: 16 secuencias de logs de consulta MSN; y datos de MIT-BIH Arriytmia.

**(Parthasarathy, Mehta, & Srinivasan, 2006)** presentan un algoritmo el cual consiste en detectar la periodicidad en series de tiempos utilizando una combinación de análisis en el tiempo-frecuencias y autocorrelación. También consideran aspectos relacionados a la no-estacionalidad, mediante técnicas de ventanas móviles con filtros de Kalman. Ellos realizaron pruebas de su propuesta con 14 observaciones relacionados a nivel del agua y temperaturas diarias de un lago. Concluyen que siendo el periodo real de la muestra, en algunos casos las herramientas descubren como como periodo real. En esos casos ellos manualmente dividían la periodicidad por factores de 2 después de realizar un análisis visual de la señal. También el tamaño del filtro de ventanas móviles de Kalman dependía del número de observaciones de la muestra.

**(Jones & Diaz, 2007)** Se enfocaron en analizar las propiedades temporales de las consultas web y la calidad de sus resultados con respecto a su histograma. Identificaron 3 tipos de perfiles generales de consultas web: atemporales (no periódicas), temporalmente no ambigua (con una sola alza abrupta) y temporalmente ambigua (múltiples alzas abruptas)

**(Shokouhi, 2011)** analiza eventos estacionales de logs de búsqueda de Google. Estos eventos reflejan la forma estacional de picos de frecuencias en consultas relacionadas a estos eventos. El autor propone el uso de técnicas de descomposición de series de tiempo para la identificación y medición de la estacionalidad en las consultas. El historial de frecuencias en cada muestra puede ser descompuesta en 3 componentes (1) nivel; (2) tendencia; y estacionalidad (3). Shohouki declaro que para las consultas estacionales, la componente de estacionalidad es el factor más significativo e identificable teniendo un parecido con la distribución de las observaciones originales de la consulta.

**(Chen Z. , Yang, Ma, Lei, & Gao, 2011)** utilizan un algoritmo de clasificación diseñado para justificar la categoría de las consultas basada en la frecuencia grabada por los logs de consultas en la Web y presentan una taxonomía que agrupa las consultas de acuerdo a la distribución temporal presentada por las curvas de frecuencia de estas. La taxonomía temporal de consultas presentada clasifica las consultas en: Estables, no Estables, Periódicas, No periódicas, Ráfaga con un alza abrupta y Ráfaga con múltiples alzas abruptas.

**(Radinsky, y otros, 2012)** utilizan la autocorrelación en series de tiempo. Declarando a las consultas de carácter periódico con el coeficiente de correlación mayor a un umbral predefinido. Utilizando la estacionalidad de la muestra compuesta de 259 consultas de un log de búsqueda entre 2006 y 2011. Para cada consulta se forma una serie de tiempo diaria.

La mayoría de los trabajos relacionados a la detección de periodicidad en series de tiempo tienen en común como herramienta para detectar patrones periódicos la autocorrelación, la transformada de Fourier y el Poder de densidad espectral.

La autocorrelación describe la correlación entre valores de la misma serie de tiempo en diferentes tiempos. Y el poder de densidad espectral muestra las frecuencias de las ondas importantes que describen la representación de la serie de Fourier, que representan patrones repetitivos.

Entonces, el problema de minería de patrones periódicos se traduce a encontrar la correlación entre observaciones en dos tiempos diferentes en el dominio del tiempo. Y en el dominio de la frecuencia encontrar las ondas que tienen mayor influencia en la representación de la señal original dentro de la serie de Fourier.

# Métodos de identificación de Periodicidad.

En esta sección se describirán los métodos de trabajo para identificar señales de carácter periódico, el propuesto por **Vlachos, Yu & Castelli** y el definido en esta investigación, que denominaremos **Método autoMM.**

El método de Vlachos fue elegido ya que fue enfocado en la detección señales tipo periódica relacionados a consultas realizadas a motores de búsqueda.

El método AutoMM fue desarrollado con el objetivo de ver si con solo la herramienta de autocorrelación es posible determinar señales de tipo de periódica.

## Método de Vlachos, Yu & Castelli.

El método propuesto por Vlachos, Yu & Castelli en su publicación “On Periodicity Detection and Structural Periodic Similarity (2005)” consiste de dos etapas, donde en cada una de estas se utiliza la información que proporcionan distintas herramientas.

La primera etapa está encargada de identificar y extraer periodos candidatos de la muestra utilizando la herramienta de Periodograma.

La segunda etapa se dedica a la validación de estos periodos y rectificación de valores finales de los periodos posibles de la muestra utilizando la herramienta de autocorrelación.

En esta investigación, se utiliza esta metodología pero con algunas modificaciones las cuales no alteran la idea principal de esta.

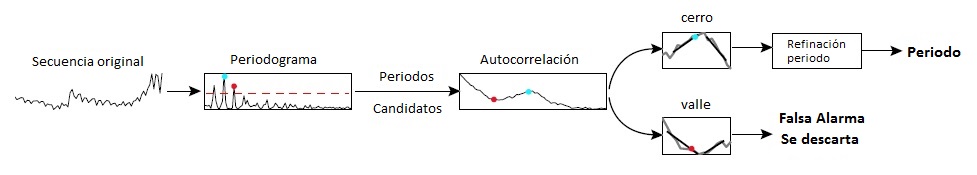


Ilustración : Diagrama de la metodología propuesta por Vlachos. Fuente: On Periodicity Detection and Structural Periodic Similarity (2005), Vlachos, Yu & Castelli.

## Descubriendo Periodos candidatos

Para la extracción de periodos candidatos a través del Periodograma, primero se determina un umbral apropiado para distinguir y diferenciar las frecuencias (o periodos en este caso) dominantes de la muestra. Si ninguna frecuencia de la secuencia excede el umbral establecido, se puede concluir que la muestra no es periódica.

Para especificar el umbral, primero se identifica cual es el valor del poder de energía o influencia de las frecuencias atribuidas a mecanismos aleatorios, esto es, cualquier muestra o proceso que no sea atribuido a un proceso aleatorio es considerado de interés dentro de la investigación.

Por ejemplo se examina una señal , donde la permutación de sus observaciones da como resultado una señal de . Esta nueva señal retendrá la estadística de primer orden de la muestra original (el promedio), pero no exhibirá ni un patrón o periodicidades gracias al proceso de la permutación.

Una vez hecho se obtiene una secuencia aleatoria a partir de la secuencia original la cual no tendrá un patrón definido y por esto el poder de energía máximo dentro de las frecuencias de la nueva secuencia aleatoria no debe de tener ninguna importancia. Por lo cual, una frecuencia que sobrepase el poder de energía definido por la muestra aleatoria se considera importante.

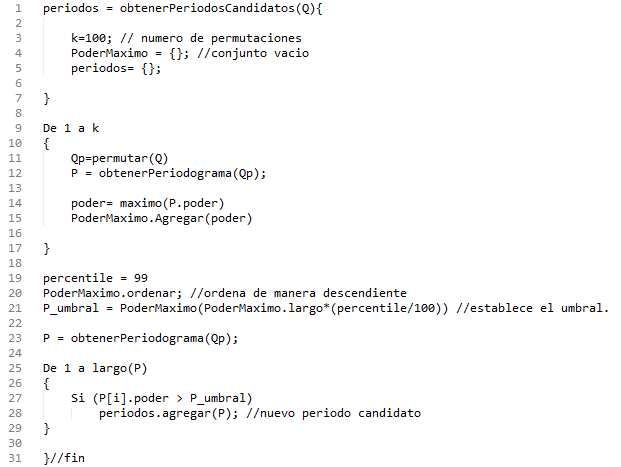


Ilustración : Extracción de periodos candidatos Periodograma. Fuente: On Periodicity Detection and Structural Periodic Similarity (2005), Vlachos, Yu & Castelli.

Si se quiere establecer un intervalo de confianza del 99% para determinar frecuencias importantes, el proceso de permutación de la muestra se repite 100 veces. En cada una de las iteraciones se guardan “las frecuencias importantes o máximas” de dichos procesos aleatorios con el fin de elegir el máximo valor de ellos, el cual establece un estimador suficiente de umbral que se está buscando.

Así, los periodos o frecuencias () que sobrepasen dicho umbral en el Periodograma, serán considerados como periodos candidatos.

Al extraer periodos candidatos de la muestra, hay que establecer un rango que abarca dicho periodo candidato dentro del Periodograma que define un alcance el cual es de gran importancia en la etapa de verificación de periodos.

Esto quiere decir que se considera los periodos inferior y superior como parte del rango considerado por el periodo candidato.

## Validación de Periodos candidatos.

Después de identificar los “picos” del Periodograma y extraer el conjunto de periodos candidatos de la muestra, se realiza una validación de estos con respecto a la información obtenida con la herramienta de autocorrelación.

Un indicio de un periodo candidato importante es cuando se da el caso en el cual dicho periodo cae cerca de un “cerro” dentro del gráfico de la función de autocorrelación. Si el periodo considerado cae en un “valle” o lejos de la punta de un “cerro”, este queda descartado de ser un periodo importante.

Para determinar el cerro cercano al periodo candidato, el valor en la punta del cerro próximo (el cual es un rezago importante dentro de la función de autocorrelación) debe de estar dentro del rango del periodo candidato establecido durante la última etapa del descubrimiento de periodos candidatos, con esto se puede refinar el valor del periodo candidato. Este paso es de suma importancia, ya que el valor correcto del periodo puede ser no descubierto por el Periodograma y puede ser refinado con el valor de la autocorrelación.

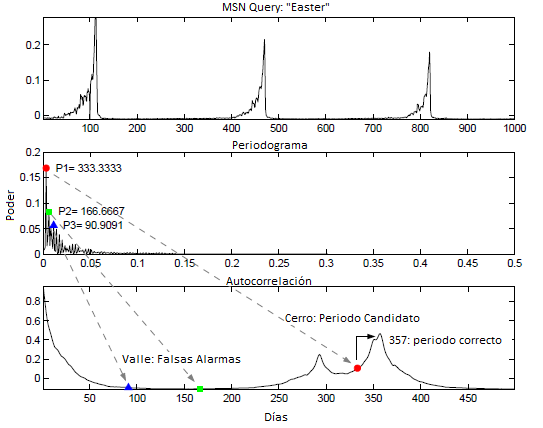


Ilustración : Proceso de identificación de periodos a través de método de Vlachos, Yu & Castelli de la consulta "Navidad". Fuente: On Periodicity Detection and Structural Periodic Similarity (2005), Vlachos, Yu & Castelli.

En la ilustración 10, se visualiza la demostración de la extracción de periodos candidatos del Periodograma, para luego ser verificados con la herramienta de autocorrelación, donde se aprecia la validación y refinamiento de los valores a través de la aproximación a la punta del cerro próximo.

## Identificación del pico cercano

Para identificar los valores máximos de los cerros dentro del gráfico de Autocorrelación, se realiza el cálculo de la segunda derivada sobre la función de la muestra, con el fin de encontrar el valor máximo de estos cerros, un paso que solo puede ser posible en funciones o señales continuas.

En esta investigación, al trabajar con señales del tipo discreta, no es posible realizar dicho paso, es por esto que se recurre a utilizar una alternativa que no altera la idea principal definida por Vlachos, Yu & Castelli. Esta alternativa consiste en utilizar un umbral para determinar periodos importantes dentro de la función de autocorrelación y no solamente que fuera un cerro.

Para identificar dichos valores importantes se utiliza un intervalo de confianza definido por (Barlett, 1955), el cual es el siguiente:

Donde es el coeficiente de autocorrelación en el rezago *k* y *N* es el largo de la muestra. Este intervalo establece un límite donde los valores del coeficiente de autocorrelación en un rezago *k* que sobrepasen dicho intervalo están fuertemente correlacionados indicando una alta correlación entre los valores separados por observaciones.

En este caso el valor del umbral está determinado por el límite superior del intervalo de confianza (sección 3.6.1), el cual si es superado establece que dicho valor esta correlacionado de manera positiva. Con esto, se guarda un conjunto de valores importantes, los que son “picos” que están por sobre el umbral.

Cuando se tiene certeza de que el periodo candidato cae en un cerro, vale decir, cerca de uno de estos valores importantes y no en un valle, se necesita descubrir si el valor “pico” del cerro cercano está dentro del rango del periodo candidato definido en la etapa anterior.

Si este valor “pico” importante está dentro del rango de un periodo candidato del Periodograma, se toma como periodo final el valor “pico” de la función de autocorrelación, en caso contrario, no se toma en cuenta.

En el caso de haber dos valores importantes dentro del rango del periodo candidato del Periodograma, se tomara como periodo final el valor del “pico” más cercano.

## Método de detección propuesto en la investigación (Método autoMM)

El método propuesto en la investigación consiste en realizar iteraciones de suavizado sobre la muestra a través de medias móviles, donde en cada una de estas iteraciones se utiliza la herramienta de autocorrelación simple, en conjunto con un intervalo de confianza que permite identificar y extraer los rezagos más importantes, para luego con un procedimiento de diferenciación entre dichos rezagos, se determina la distancia entre estos y se identifican posibles periodos en cada iteración.

Al finalizar las iteraciones se realiza un filtrado que consiste en la unión de candidatos finales contiguos y se consideran como periodos los que se repiten con mayor frecuencia en las iteraciones realizadas determinando la periodicidad de la muestra.

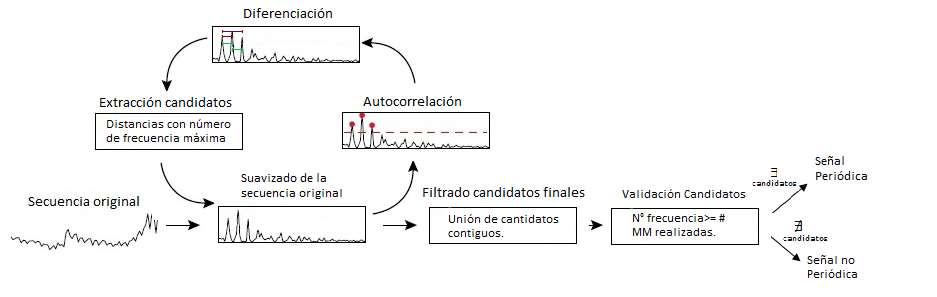


Ilustración : Diagrama de la metodología propuesta en la Investigación. Fuente: Elaboración Propia.

## Suavizado de la muestra, Medias Móviles

Cada iteración comienza con el uso de medias móviles, herramienta dedicada a tomar un conjunto de observaciones de manera consecutiva generando una serie de promedios con el fin de suavizar la muestra.

En cada iteración se aumenta en 1 el número de observaciones tomado por las medias móviles, aumentando el suavizado de la muestra.

Esto es de suma importancia ya que también suaviza el resultado de la función de autocorrelación, eliminando “picos” o rezagos candidatos que no son importantes dentro del gráfico de autocorrelación y dejando en evidencia en cada iteración los que son importantes.

Hay que tener en cuenta que a medida que se va aumentando el número de medias móviles, la serie de tiempo comienza desde la observacion del valor de las medias móviles, por ejemplo, en el caso de una serie de tiempo en semanas y el valor de medias móviles es 4, se comienza desde la semana 4 promediando el conjunto de las últimas 4 semanas (1, 2, 3 y 4), y así sucesivamente. Aunque se “pierdan” las primeras observaciones de la muestra, la estructura de esta no se pierde.

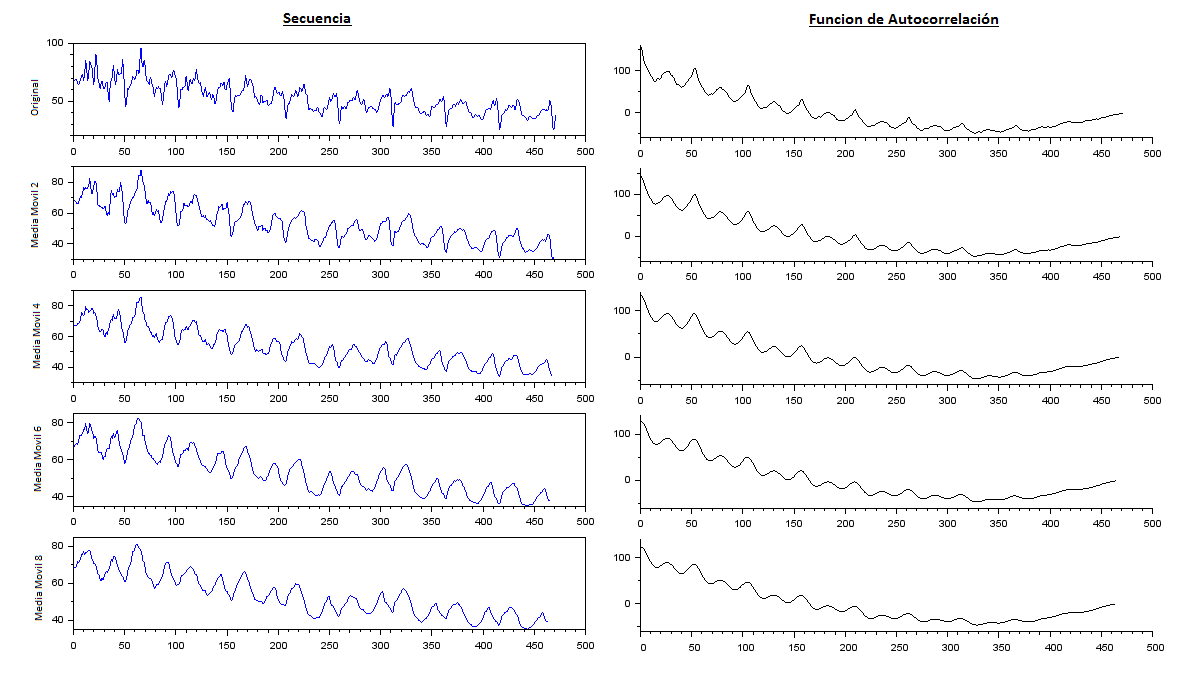


Ilustración : Suavizado de la muestra y resultado de Autocorrelación. Fuente: Elaboración Propia.

En la ilustración 12, se puede apreciar como la herramienta de media móvil va modificando la muestra original, suavizando el resultado obtenido de la función de autocorrelación.

## Encontrando rezagos importantes con la función de Autocorrelación

Luego de suavizar la muestra se utiliza la función de autocorrelación simple, herramienta que determina la correlación entre dos observaciones separadas por observaciones para identificar los rezagos importantes dentro de la muestra. Para ello se utiliza un intervalo de confianza definido por (Barlett, 1955), el cual es el siguiente:

Donde es el coeficiente de autocorrelación en el rezago *k* y *N* es el largo de la muestra. Este intervalo establece un límite donde los valores del coeficiente de autocorrelación en un rezago *k* que sobrepasen dicho intervalo están fuertemente correlacionados indicando una fuerte correlación entre los valores separados por k observaciones.

Es por esto que dentro del gráfico de autocorrelación los valores que estén en un “pico” de un cerro, el cual sobrepase dicho intervalo de confianza, representan un valor importante y será almacenado.

En la ilustración 13, se aprecia el proceso de extracción de los “picos” o rezagos importantes que están por sobre del intervalo de confianza (línea denotada por color verde) dentro de la función de autocorrelación.

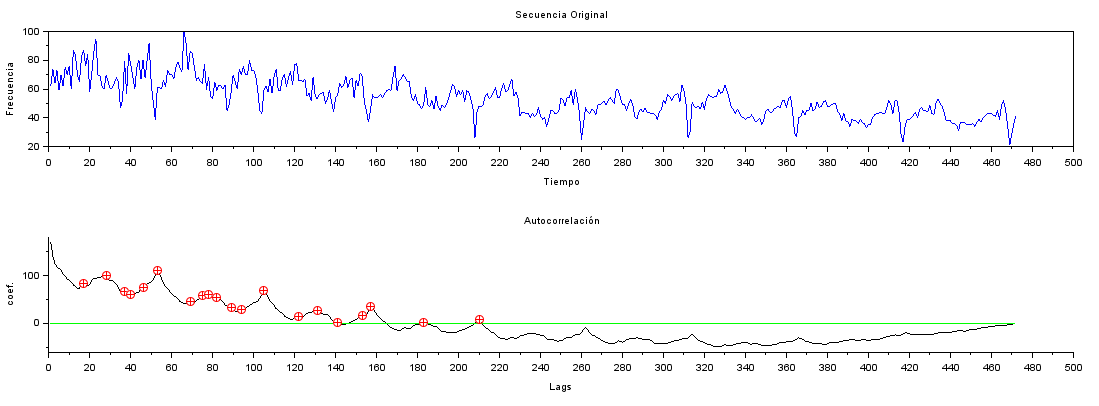


Ilustración : Extracción de rezagos de la función de autocorrelación. Fuente: Elaboración Propia.

También se puede apreciar que se consideran valores entre los cerros, es por esto que se recurre a suavizar la muestra con medias móviles eliminando falsos candidatos.

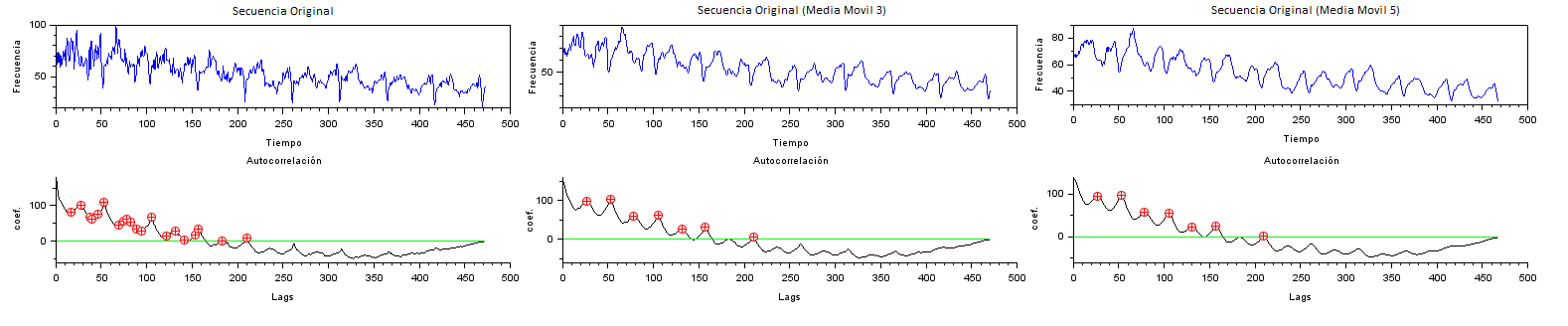


Ilustración : Eliminación de rezagos candidatos no importantes. Fuente: Elaboración Propia.

En la ilustración 14, los candidatos no importantes son eliminados gracias al suavizado realizado con la herramienta de medias móviles.

## Diferenciación de rezagos, extrayendo periodos candidatos

Teniendo guardados los rezagos importantes obtenidos por la función de autocorrelación, se realiza a un proceso de diferenciación de estos valores, donde se resta cada rezago importante con todos los demás, y se guardan los valores absolutos de las diferencias (valores positivos), obteniendo la distancia entre rezagos que representan posibles periodos dentro de la muestra.

Una vez almacenado estos posibles periodos, se realiza una revisión del número de las frecuencias de los valores extraídos y se guardan los que tienen el valor máximo de frecuencia por cada iteración de media móvil.

## Filtrado de periodos candidatos, identificando periodos finales

Una vez terminadas las iteraciones de medias móviles obteniendo conjuntos de posibles periodos en cada una de ellas, se realiza un proceso de unión de periodos candidatos contiguos y una validación de los candidatos finales para identificar la existencia de periodos y así determinar si la muestra es de carácter periódico.

Primero se reúnen todos los periodos candidatos de las iteraciones en un vector y se ordenan de menor a mayor. Una vez hecho esto, se realiza una unión de los valores contiguos, debido que se puede dar el caso donde en una iteración el periodo candidato tiene un valor de 52 y en otra 53. Esto sucede debido que suavizando la muestra, el valor del periodo puede variar levemente.

Por esta razón se realiza una unión de dichos valores contiguos, almacenando este conjunto en un vector auxiliar, para después extraer el valor promedio de estos y la frecuencia de las observaciones involucradas, con el fin de generar un nuevo periodo candidato final.

Una vez generado el conjunto de candidatos finales, se realiza un filtrado de estos, donde se eligen los que tienen una frecuencia superior o igual al número de iteraciones realizadas, o sea, el número de medias móviles realizadas durante el proceso de reconocimiento.

# Muestras utilizadas en la investigación

En esta sección se explicará la extracción y origen de la data utilizada para realizar las pruebas de precisión a los métodos nombrados anteriormente.

## Organizaciones involucradas en las muestras utilizadas

A continuación se explicará de forma breve alguna de las organizaciones relacionadas al origen de las muestras extraídas.

## The Text REtrieval Conference (TREC)

The Text REtrieval Conference (TREC), es una serie continua de talleres, co-auspiciado por el Instituto Nacional de Estándares y Tecnologías (NIST) y el Departamento de Defensa de Estados Unidos. Tiene por propósito dar apoyo y fomentar la investigación dentro de la comunidad de recuperación de información, proveyendo una infraestructura necesaria para la evaluación a gran escala de metodologías de extracción de información (Text Retrieval Context, s.f.). En particular, algunos de los objetivos del seminario de TREC son:

* Reforzar la investigación en la extracción de información basada en grandes colecciones de pruebas.
* Incrementar la comunicación entre la industria, academias y gobiernos al crear un foro abierto para intercambiar ideas relacionadas a la investigación.
* Para agilizar la transferencia de tecnologías de laboratorios de investigación en productos comerciales demostrando mejoras substanciales en metodologías de extracción en problemas del mundo real.
* Incrementar la disponibilidad de técnicas de evaluación apropiadas de uso por la industria y academia, incluyendo el desarrollo de nuevas técnicas de evaluación mejores aplicadas en sistemas actuales.

## WEB TRACKS de TREC

Es un seminario de TREC que consiste de un conjunto de trayectorias (líneas de acción) o áreas de enfoque, en las cuales se definen las tareas particulares para cada área.

Estas trayectorias tienen distintos propósitos. Primero, actúan como incubadoras para nuevas áreas de investigación: el comienzo de una trayectoria usualmente define cual es el problema, y esta trayectoria crea una infraestructura necesaria (colección de pruebas, evaluaciones, metodologías, etc.) para dar soporte a la investigación en dichas tareas.

Estas trayectorias también demuestran la robustez de tecnologías de extracción, en las cuales dichas técnicas pueden ser apropiadas para una variedad de tareas. Finalmente las trayectorias hacen a TREC más atractivo para una comunidad más amplia, proporcionando tareas que responden a los intereses de investigación de más grupos.

## Google Trends

Google Trends es una herramienta de Google, que permite mostrar el comportamiento de términos de búsqueda en Google o Youtube más populares del pasado reciente, de manera gráfica (Google Trends, s.f.).

La información proporcionada en gráficos representa la frecuencia de la búsqueda de un término en particular en los motores de búsqueda de Google y Youtube, el cual representa la frecuencia del término ingresado en una serie de tiempo.

La herramienta permite filtrar en intervalos de tiempo y de manera geográfica, como también comparar de manera visual las series de tiempo de varios términos dentro del gráfico.

## Origen de las muestras utilizadas en la investigación

Una agrupación de la Universidad de Shandong University of Science and Technology, extrajo un listado de consultas desde las Trayectorias de TREC eligiendo un número aproximado de 15000 consultas de manera aleatoria del listado, e ingresaron cada consulta en Google Trends para descargar los archivos de series de tiempo.

Ellos dedicaron a 4 asesores a anotar manualmente el patrón de estas consultas utilizando su propia taxonomía:

**SQ:** Consulta tipo Estable o Permanente.

**OBQ:** Consulta Ráfaga de una sola vez

**AQ:** Consulta tipo Ráfaga Aperiódica de múltiples veces.

**PQ:** Consulta tipo Ráfaga Periódica.

**Desconocido:** Los asesores no saben cómo clasificar la Consulta.

Las series de tiempo que ellos clasificaron contienen series de tiempo que van desde la fecha enero 2004 y abril del año 2013.

## Preparación de la data

En la presente investigación se descargaron de Google Trends cerca de 13000 consultas, elegidas desde el listado clasificado por la agrupación de profesores de la Universidad de Shangdong, con la finalidad de poner a prueba la precisión de los métodos nombrados anteriormente.

También, se desarrollaron distintas soluciones para realizar un filtrado de las muestras descargadas con el objetivo de recrear la data clasificada por la agrupación de profesores en la investigación, ya que la data que ellos clasificaron va desde los inicios del 2004 hasta abril del 2013.

# Aplicación de Método de Vlachos y Método autoMM.

En esta sección se aplicarán los métodos explicados anteriormente (Vlachos y autoMM), para cada tipo de señales con el fin de conocer el comportamiento de ambos métodos. Para estos efectos se utilizarán los datos de las consultas extraídas desde Google Trends y clasificadas por la agrupación de profesores de la Universidad de Shangdong.

Las consultas utilizadas en la demostración son:

*Tabla 2: Tipos y nombres de las consultas utilizadas en la demostración. Fuente: Elaboración propia.*

|  |  |
| --- | --- |
| Tipo de Señales | Nombre consulta |
| Ráfaga | “get rich”  “google earth” |
| Estables | “google images”  “cats” |
| Periódica | “mistletoe”  “farm and fleet” |

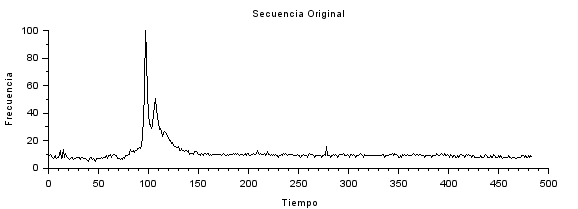
## Consultas tipo Ráfaga.

## Consulta 1: “get rich”

*Serie de tiempo en Semanas:* Del 2004-01-04 al 2013-04-06

*Cantidad de observaciones (semanas):* 483

En la ilustración 15, se puede apreciar un evento de alza abrupta en las frecuencias de la consulta en la semana número 100 de la serie de tiempo. Luego de dicho evento, el número de frecuencias decae siguiendo el comportamiento normal previo al evento de alza abrupta.

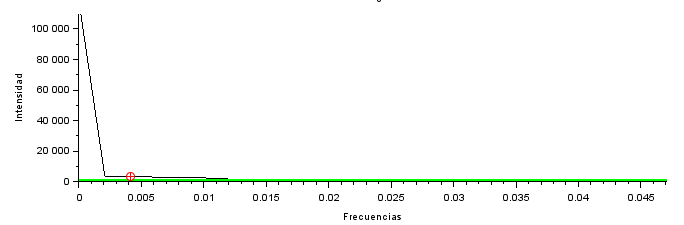


*Ilustración 15: Serie de tiempo de la consulta "get rich" en semanas. Fuente: Elaboración propia.*

## Método de Vlachos

**Etapa 1 - Descubriendo Periodos Candidatos (Periodograma)**

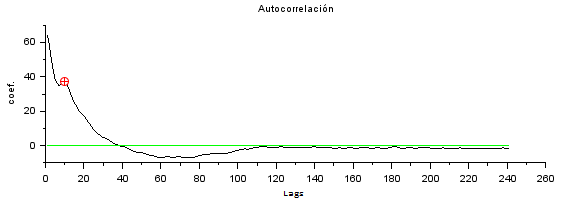
En el Periodograma de la Ilustración 16, se descubre un solo “pico” que está por sobre el umbral de poder definido explicado en la sección 5.1.1. El periodo candidato tiene valor un valor de 241,5 con un rango de alcance de 200 a 363 explicado en la sección 5.1.1.



*Ilustración 16: Periodograma de la consulta "get rich". Fuente: Elaboración Propia.*

**Etapa 2 - Validación de Periodos Candidatos (Autocorrelación)**

En el gráfico de función de autocorrelación de la Ilustración 17, se aprecia un rezago importante que sobrepasa el intervalo de confianza de (Barlett, 1955). El valor de dicho rezago es de 10.



*Ilustración 17: Gráfico de autocorrelación, intervalo de confianza y rezago candidato. Fuente: Elaboración Propia.*

**Etapa 3 - Validación de picos cercanos**

Luego de extraer el periodo candidato del Periodograma y el rezago importante de la función de autocorrelación, se verifica si dicho rezago está dentro del rango definido anteriormente en el Periodograma para determinar si es un candidato final a periodo.

*Tabla 3: Periodo candidato Periodograma con su rango de alcance, consulta "get rich". Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Periodo candidato** | **Rango periodo candidato** |
| 241,5 | [200 – 363] |

Tabla : Rezagos importantes función de autocorrelación, consulta "get rich". Fuente: Elaboración Propia.

|  |  |
| --- | --- |
| **Rezagos importantes:** | 10 |

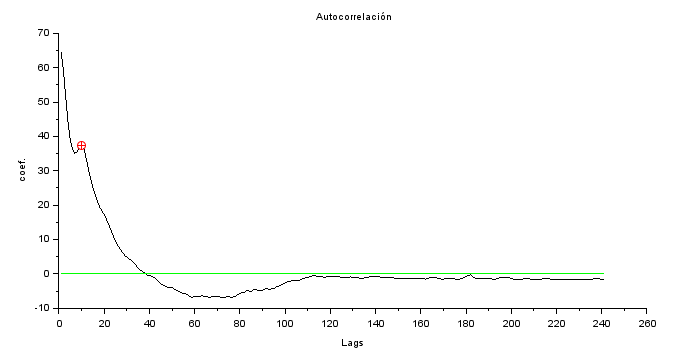
Como el rezago 10 no está dentro del rango de periodo candidato 241,5 - [200 – 365] se dice que la señal **no es periódica**.

## Método AutoMM

**Etapa 1 - Encontrando periodos candidatos.**

El método comienza aplicando la función de autocorrelación sobre la muestra original, extrayendo rezagos importantes que sobrepasan el intervalo de confianza de Bartlett.

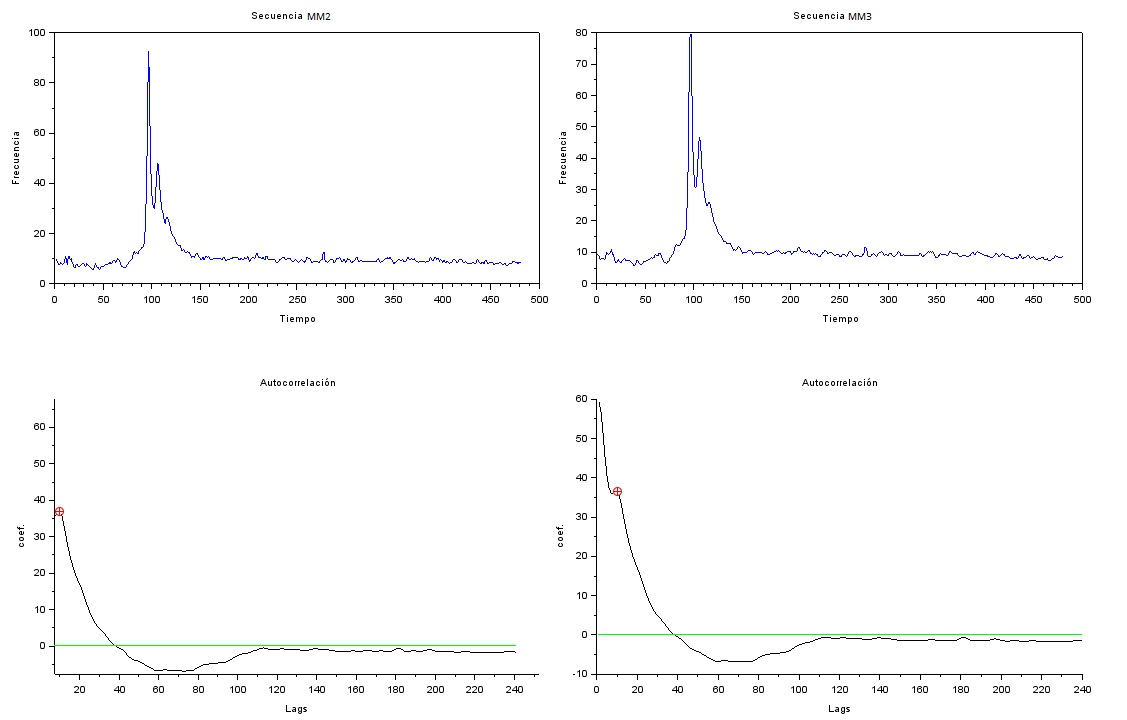
En el gráfico de autocorrelación de la muestra original (ilustración 18) se extrae solo un rezago importante. Como este método extrae periodos candidatos realizando un proceso de diferenciación entre todos los rezagos importantes de la función de autocorrelación, en este caso, no es posible realizar dicho proceso por lo que no existen periodos candidatos en la primera iteración.



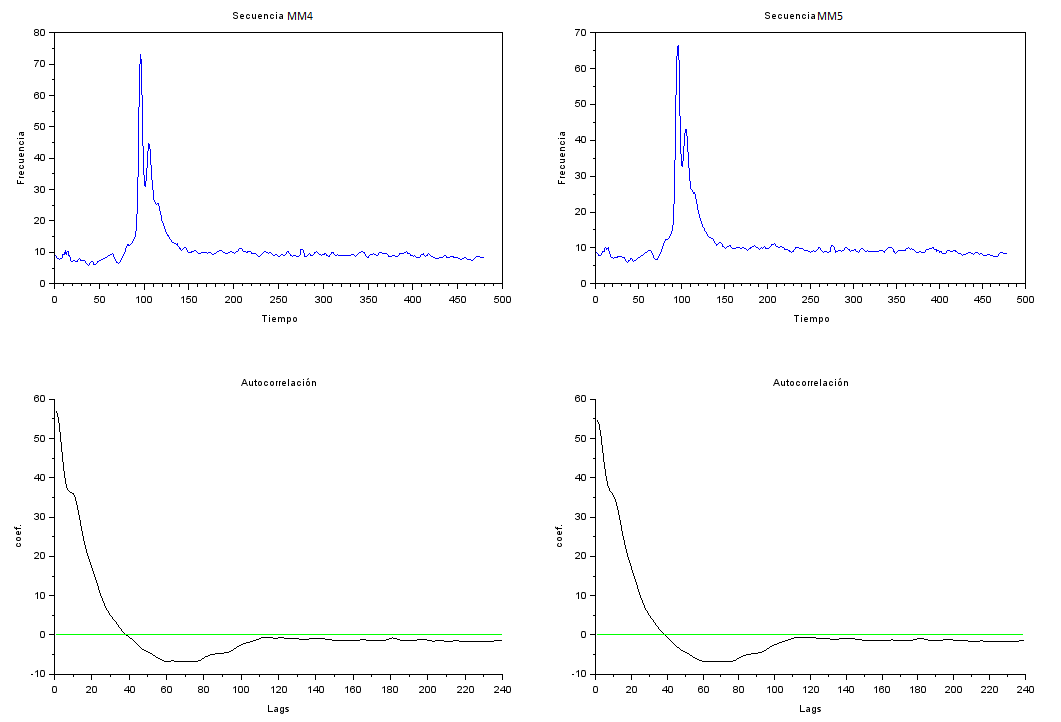
*Ilustración 18: Función de autocorrelación consulta "get rich" secuencia original. Fuente: Elaboración Propia.*

En las ilustraciones 19 y 20 se aprecia como el proceso de suavizado de la muestra, a través de las medias móviles, también suaviza el resultado de la función de autocorrelación, eliminando el rezago que se encontraba en la primera iteración, lo que significa que dicho rezago no era del todo importante en la búsqueda de periodos.

Al no haber rezagos en las siguientes iteraciones de suavizado, no se presentan periodos candidatos.



*Ilustración 19: Gráfico de la muestra con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*



*Ilustración 20: Gráfico de la muestra Media Móvil 4 y 5 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*

*Tabla 5: Periodos candidatos por iteración de medias móviles consulta "get rich". Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Medias Móviles** | **Periodos Candidatos** |
| 1 | - |
| 2 | - |
| 3 | - |
| 4 | - |
| 5 | - |

**Etapa 2 – Encontrando periodos finales**

Como no se hallaron periodos candidatos en ninguna de las iteraciones de suavizado. La consulta “get rich” se declara como **No Periódica**.

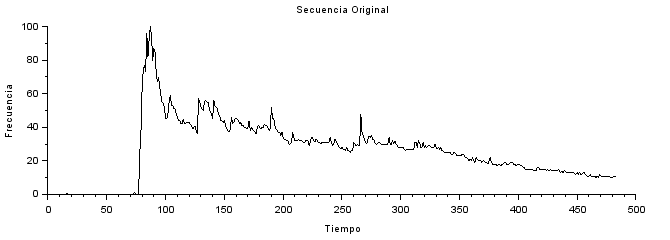
Como conclusión se establece que ambos métodos identifican adecuadamente el tipo de señal.

## Consulta 2: “google earth”

*Serie de tiempo en Semanas:* 2004-01-04 al 2013-04-06

*Cantidad de observaciones (semanas):* 483

En la Ilustración 22 a se aprecia un alza abrupta en la frecuencias de la consulta “google earth” en la semana 76 (del 2005-08-29 al 2005-09-03), que representa el lanzamiento de la aplicación Google Earth 3.0, que luego de llegar a un valor máximo decae para volver a un estado estable.

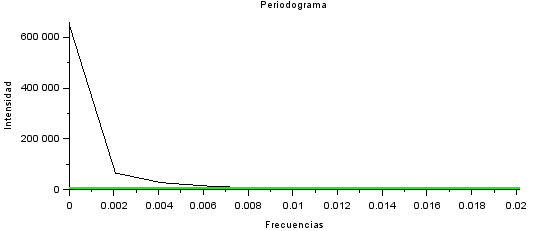


*Ilustración 21: Serie de tiempo de la consulta "google earth" en semanas. Fuente: Elaboración Propia.*

## Método de Vlachos

**Etapa 1 - Descubriendo Periodos Candidatos (Periodograma)**

En la Ilustración 22, el Periodograma no encuentra periodos candidatos, ya que no se aprecian “picos” importantes que sobrepasan el umbral definido dentro Periodograma. Con este resultado se puede declarar que la muestra es de carácter **No Periódico.**

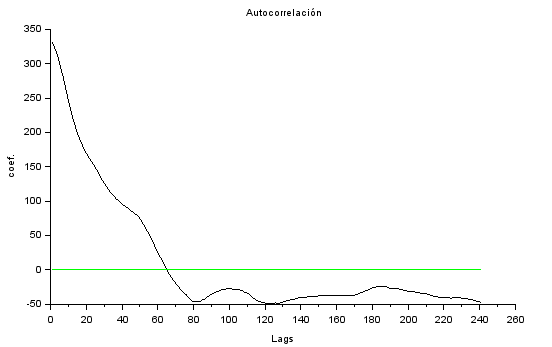


*Ilustración 22: Gráfico Periodograma de la consulta "google earth". Fuente: Elaboración Propia.*

## Método autoMM

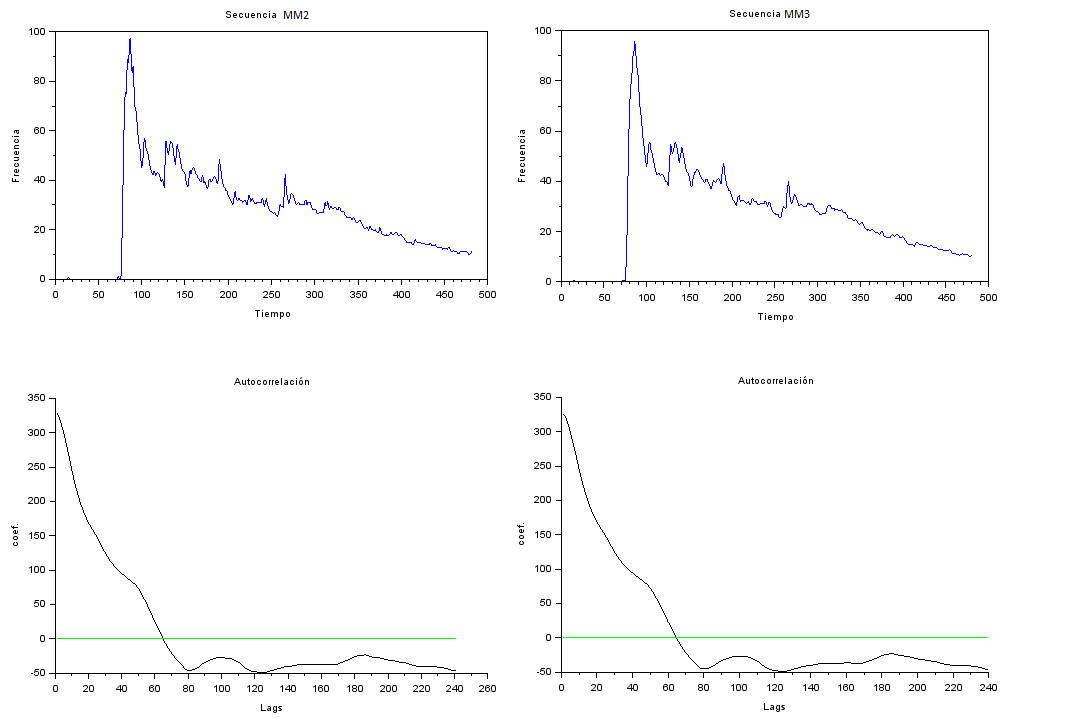
**Etapa 1 - Encontrando periodos candidatos.**

La información proporcionada por la función de autocorrelación nos dice que no se encuentran rezagos importantes, por lo tanto no existen periodos candidatos en la primera iteración.

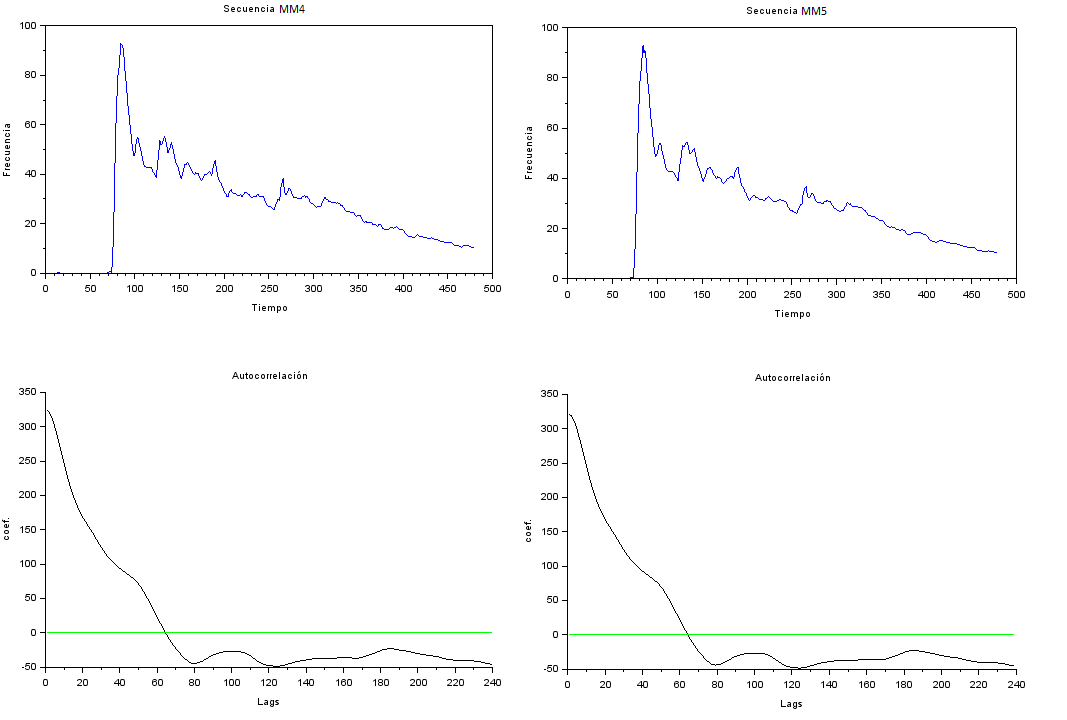


*Ilustración 23: Gráfico de autocorrelación muestra original consulta "google earth". Fuente: Elaboración Propia.*

Al iterar el suavizado de la muestra, se aprecia en las ilustraciones 24 y 25, las funciones de autocorrelación siguen sin extraer rezagos importantes. Con este resultado no se puede realizar el proceso de diferenciación para extraer periodos candidatos, por lo que se finaliza sin periodos candidatos en las iteraciones de suavizado.

****

*Ilustración 24: Gráfico de la serie de tiempo “Google Earth” con Media Móvil 2 (izquierda) y 3 (derecha) con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*

****

*Ilustración 25: Gráfico de la serie de tiempo “Google Earth” con Media Móvil 4 (izquierda) y 5 (derecha) con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*

*Tabla 6: Candidatos por media móvil. Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| Medias Móviles | Periodos Candidatos |
| 1 | - |
| 2 | - |
| 3 | - |
| 4 | - |
| 5 | - |

**Etapa 2 – Encontrando periodos finales**

Al no haber periodos candidatos en ninguna de las iteraciones se concluye que la consulta “google earth” es de carácter **No Periódico.**

## Conclusión métodos consultas tipo Ráfaga

Como conclusión se establece que ambos métodos identifican adecuadamente el tipo de señal.

El método de Vlachos a pesar de encontrar un periodo candidato en el Periodograma de la consulta “get rich”, los rezagos importantes encontrados en la autocorrelación no está dentro del rango definido en el periodo candidato, por lo que se define como no periódica.

En la consulta “google earth” el Periodograma al no encontrar periodos candidatos se puede declarar no periódica, sin tener que realizar el cálculo de la función de autocorrelación.

El método de AutoMM en la consulta “get rich” encontró un rezago importante en la primera iteración, pero al suavizar la muestra, este rezago fue eliminado por lo que no era importante. Además, al no haber candidatos finales en ambas consultas, estas se declararon como no periódicas.

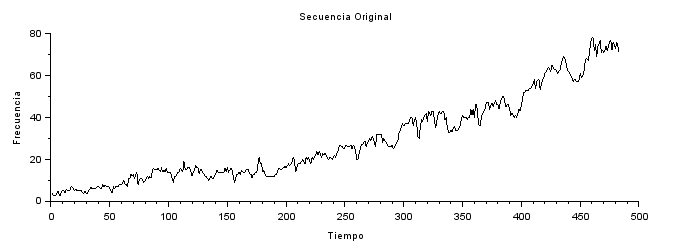
## Consultas tipo Estables.

## Consulta 1: “google images”

*Serie de tiempo en Semanas:* 2004-01-04 al 2013-04-06

*Cantidad de observaciones (semanas):* 483

En la Ilustración 26 se aprecia el comportamiento estable de la consulta “google images”, donde no se puede apreciar un alza abrupta o conducta reiterativa de las frecuencias dentro de la serie de tiempo, solo una tendencia leve en aumento de las consultas a medida que pasa el tiempo.

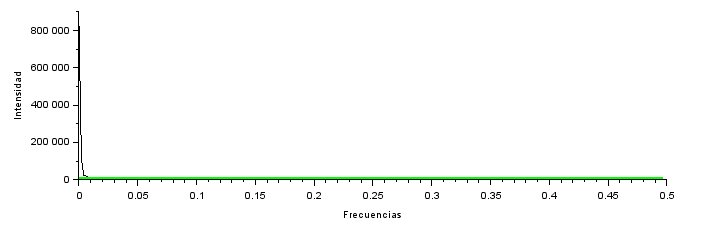


*Ilustración 26: Serie de tiempo en semanas consulta "google images". Fuente: Elaboración Propia.*

## Método de Vlachos

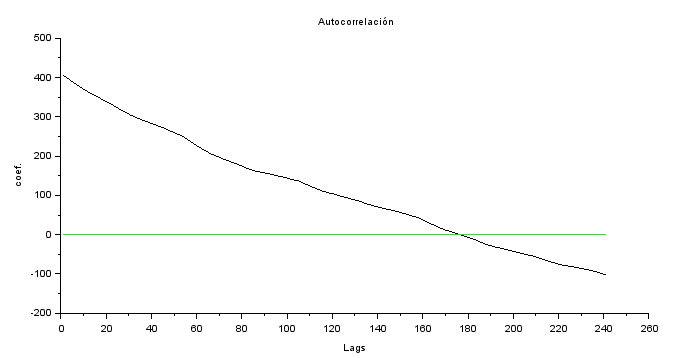
**Etapa 1 - Descubriendo Periodos Candidatos (Periodograma)**

En la ilustración 28 no se aprecia ni un solo cerro o “pico” que sobresalga del límite generado por el proceso aleatorio de la muestra, al no haber periodos candidatos, se establece que la consulta es de carácter **no periódico**.



*Ilustración 27: Gráfico de Función de autocorrelación de la consulta "google images". Fuente: Elaboración Propia.*

En la ilustración 28 se puede observar el comportamiento de los valores de la función de autocorrelación en series de tiempo de carácter estable. La consulta “google images” al no tener periodos candidatos extraídos en el paso anterior se declara **no periódica**.

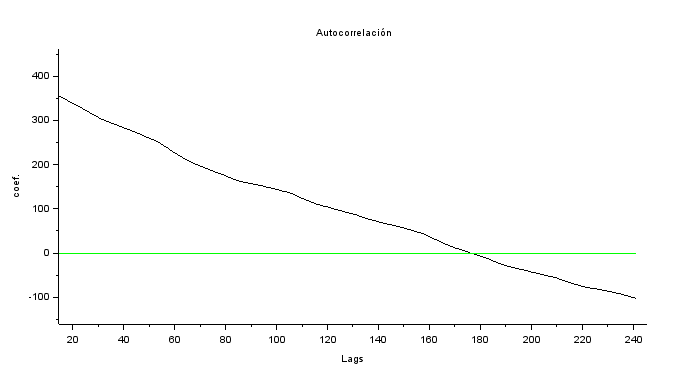


*Ilustración 28: Gráfico función de autocorrelación consulta "google images". Fuente: Elaboración Propia.*

## Método autoMM

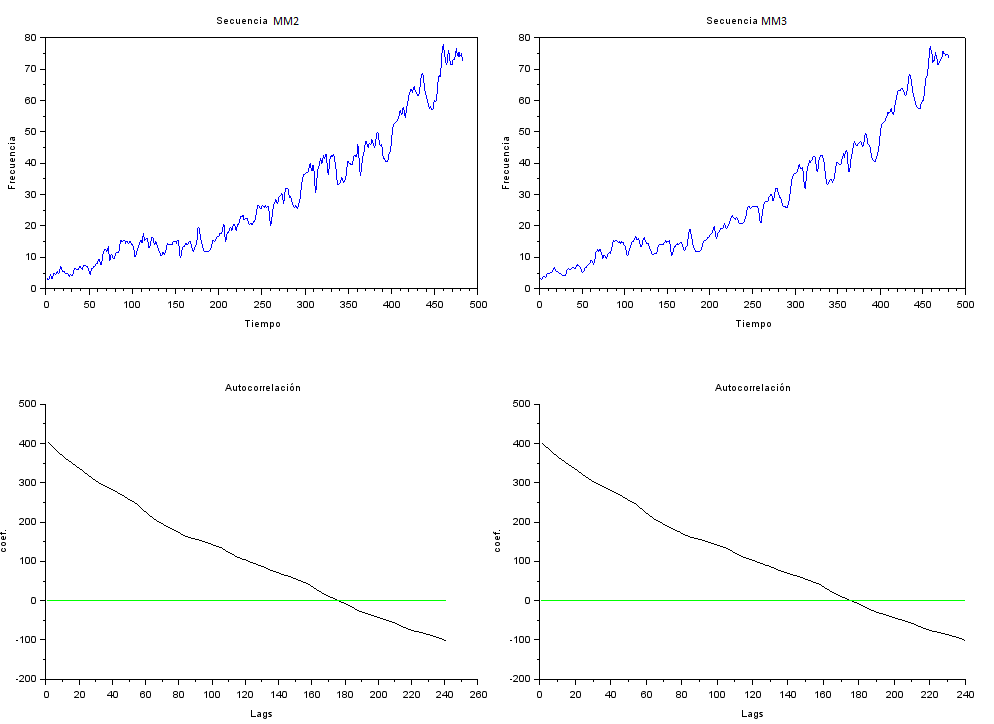
**Etapa 1 - Encontrando periodos candidatos.**

En el gráfico de la función de autocorrelación de la serie original (Ilustración 29), no se encuentran “picos” o cerros que sobresalgan por sobre el límite establecido por Barlett, por lo que no se puede realizar el proceso de diferenciación de rezagos y extraer periodos candidatos en la primera iteración.

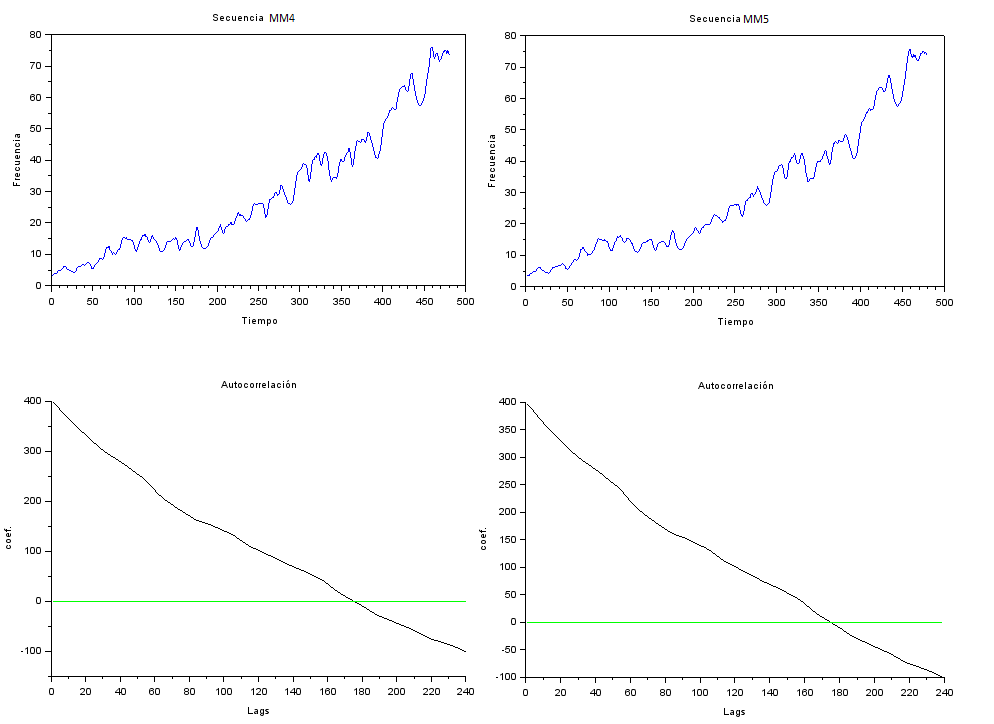


*Ilustración 29: Gráfico de la función de autocorrelación para la serie original "google images" Fuente: Elaboración Propia.*

A medida que se va iterando el suavizado de la muestra, como se aprecia en las ilustraciones 30 y 31, el gráfico de la función autocorrelación no varía ya que desde la primera iteración no se hallaban cerros o “picos” en el gráfico de autocorrelación de la primera iteración (ilustración 29).



*Ilustración 30: Gráfico de la serie de tiempo “google images” con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*



*Ilustración 31: Gráfico de la muestra con Media Móvil 4 y 5 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*

*Tabla 7: Periodos candidatos por iteración de medias móviles consulta "google images". Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Medias Móviles** | **Periodos Candidatos** |
| 1 | - |
| 2 | - |
| 3 | - |
| 4 | - |
| 5 | - |

**Etapa 2 – Encontrando periodos finales**

Al no haber periodos candidatos en las iteraciones de medias móviles la consulta se declara como **No Periódica**.

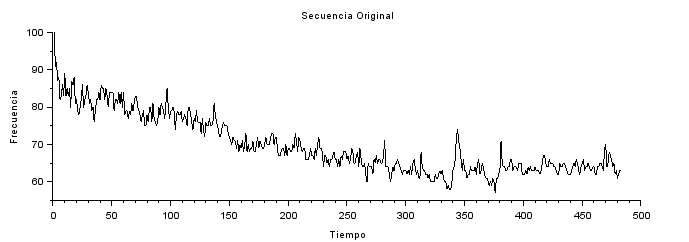
Como conclusión se establece que ambos métodos identifican adecuadamente el tipo de señal.

## Consulta 2: “cats”

*Serie de tiempo en Semanas:* 2004-01-04 al 2013-04-06

*Cantidad de observaciones (semanas):* 483

La consulta “cats” representa un comportamiento estable disminuyendo a través del tiempo con escasas alzas abruptas ni eventos reiterativos.

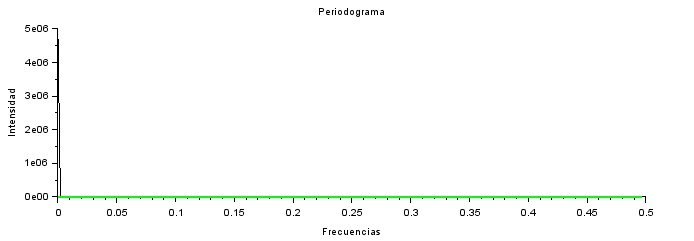


*Ilustración 32: Serie de tiempo de la consulta "cats" en semanas. Fuente: Elaboración Propia.*

## Método de Vlachos

**Etapa 1 - Descubriendo Periodos Candidatos (Periodograma)**

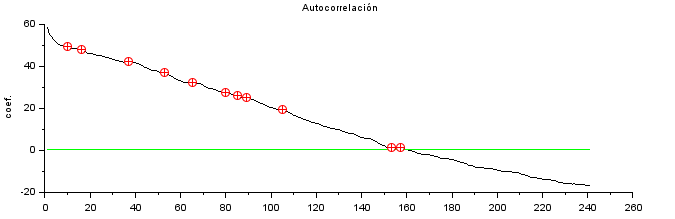
En la ilustración 33, no se aprecian picos que sobresalgan del umbral definido por el proceso aleatorio de la muestra, por lo que no existen periodos candidatos.



*Ilustración 33: Gráfico Periodograma de la consulta "cats". Fuente: Elaboración Propia.*

**Etapa 2 - Validación de Periodos Candidatos (Autocorrelación)**

En la ilustración 34, se aprecia el comportamiento de la función de autocorrelación, donde al haber pequeños cerros o “picos” a lo largo de los rezagos, el método los detecta como rezagos importantes. Estos son 10, 16, 37, 53, 65, 80, 85, 89, 105, 153, 157.



*Ilustración 34: Gráfico función autocorrelación de la consulta "cats". Fuente: Elaboración Propia.*

Pero al no haber periodos candidatos importantes por parte del Periodograma, la consulta “cats” se declara de carácter **No Periódico.**

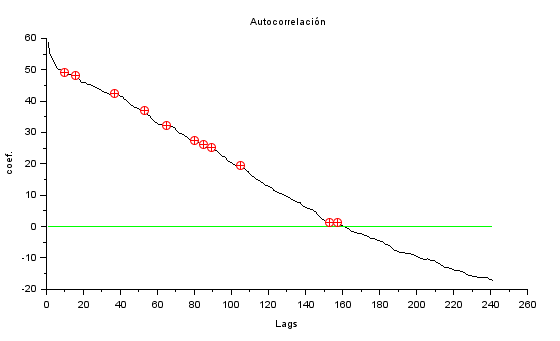
## Método autoMM

**Etapa 1 - Encontrando periodos candidatos.**

El gráfico de autocorrelación en la serie de tiempo original demuestra una tendencia a la caída constante, un comportamiento de la función de autocorrelación con las muestras del tipo estable.

Al haber pequeños “picos” a medida que caen los valores de la función de autocorrelación, estos al estar por encima del intervalo de confianza de (Barlett, 1955), se identifican como rezagos importantes.

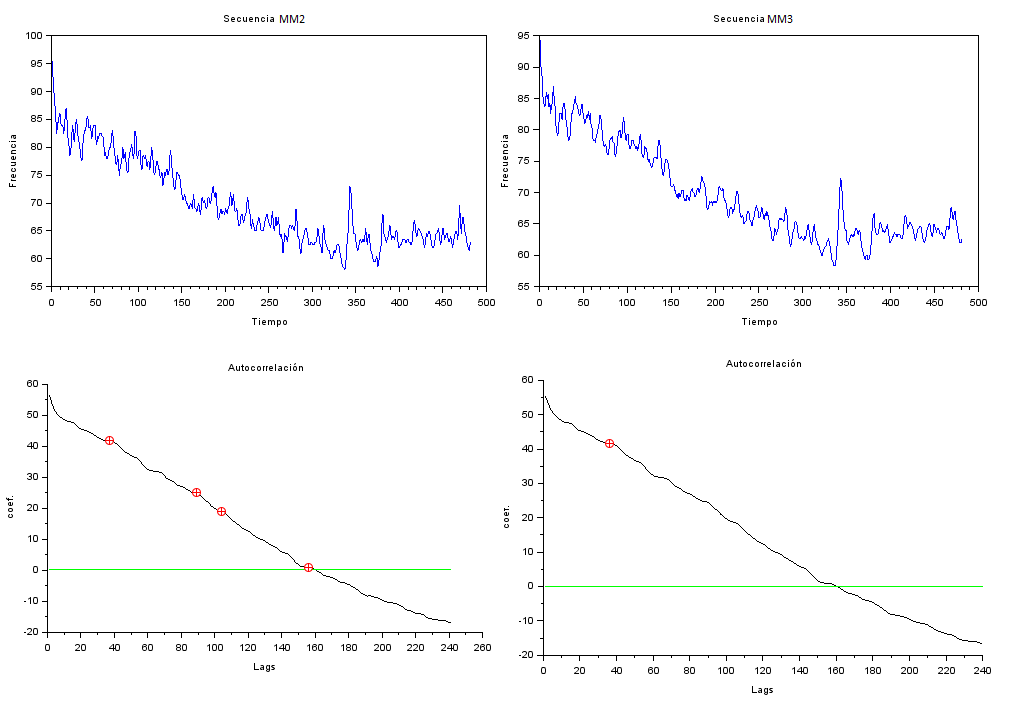
Se realiza el proceso de diferenciación de rezagos, dando como resultados la extracción de periodos candidatos extraídos con valores 52 y 68.



*Ilustración 35: Gráfico función autocorrelación serie original consulta "cats". Fuente: Elaboración Propia.*

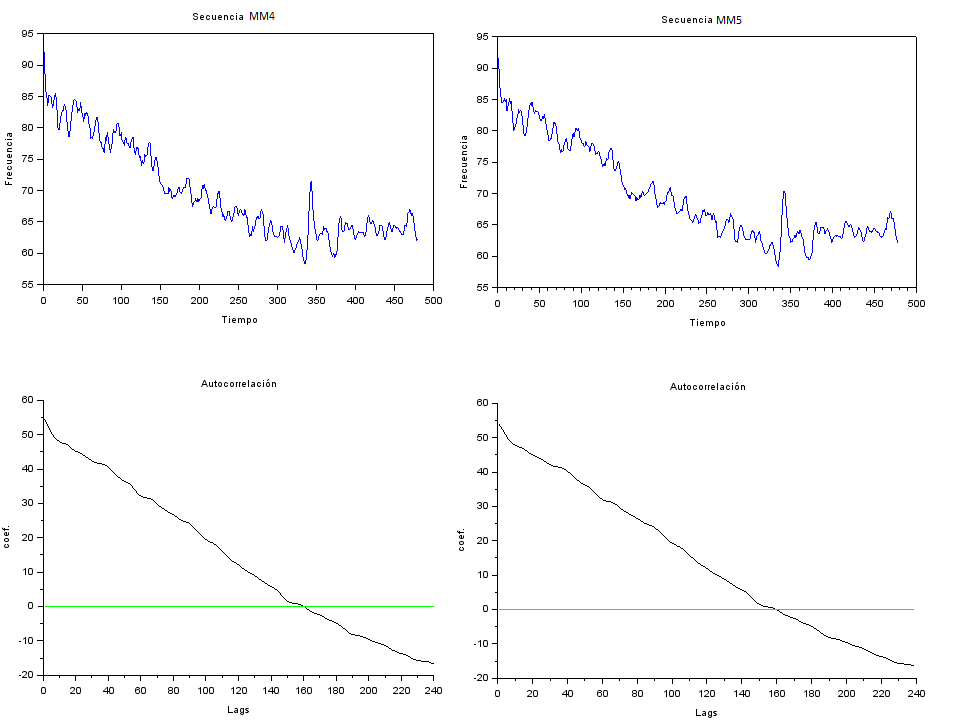
A medida que se va suavizando la muestra, como se aprecia en la ilustración 36, los rezagos que no son importantes se van eliminando y su cantidad va disminuyendo.

En la iteración de media móvil 2 se extrae como periodo importante los valores 52 y 67, y en la iteración de media móvil 3 solo se encuentra un rezago importante. Esta, al no tener más de dos rezagos importantes no se pueden extraer periodos candidatos por medio de diferencias de rezagos, por lo que en dicha iteración no hay candidatos.

****

*Ilustración 36: Gráfico de la muestra con Media Móvil 2 (izquierda) y 3 (derecha) con su respectivo gráfico de autocorrelación consulta “cats”. Fuente: Elaboración Propia.*

Debido a que en las iteraciones de medias móviles 4 y 5 el suavizado ha eliminado los rezagos que no eran importantes. Al no tener rezagos con los cuales efectuar el proceso de diferenciación, no existen periodos candidatos.

****

*Ilustración 37: Gráfico de la muestra con Media Móvil 4 y 5 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*

*Tabla 8: Periodos Candidatos por iteración de medias móviles consulta "cats". Fuente: Elaboración Propia.*

|  |  |  |
| --- | --- | --- |
| **Medias Móviles** | **Candidatos** | **Frecuencia** |
| **1** | **52 y 68** | **6** |
| **2** | **52 y 67** | **4** |
| **3** | **-** | **-** |
| **4** | **-** | **-** |
| **5** | **-** | **-** |

**Etapa 2 – Encontrando periodos finales**

Al finalizar la extracción de periodos candidatos, se hace una revisión general de estos y se observa que existen periodos candidatos contiguos 67 y 68 como se aprecia en la tabla 8.

En este caso se procede a realizar una unión de dichos periodos para generar uno con valor 68 con una frecuencia de 2, que representa el número de candidatos que generan dicho candidato final.

*Tabla 9: Candidatos finales y su frecuencia de aparición consulta "cats". Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Candidatos Finales** | **Frecuencia Total** |
| **52** | **2** |
| **68** | **2** |

Como este experimento tiene un total de 5 iteraciones y las frecuencias del conjunto de los candidatos finales no igualan o exceden el valor de la cantidad de iteraciones realizadas, se declara la muestra cómo **No Periódica.**

## Conclusión métodos consultas tipo Estables.

Como conclusión se establece que ambos métodos identifican adecuadamente el tipo de señal.

El método de Vlachos al no encontrar periodos candidatos en el Periodograma, no era necesario realizar el paso de la función de autocorrelación. Clasificando la consulta como no periódica.

El método de autoMM a pesar de encontrar varios rezagos y periodos candidatos en la primera iteración, a medida que se iteraba el suavizado de la muestra, se eliminaban los rezagos que no eran importantes y se reducía la cantidad de periodos candidatos hasta eliminarlos completamente. Ya en el paso de identificación de periodos finales, al ver que la frecuencia dentro del conjunto de periodos finales no excedían o igualaban el valor de la cantidad de iteraciones de suavizado, se concluía que dichas consultas no eran periódicas.

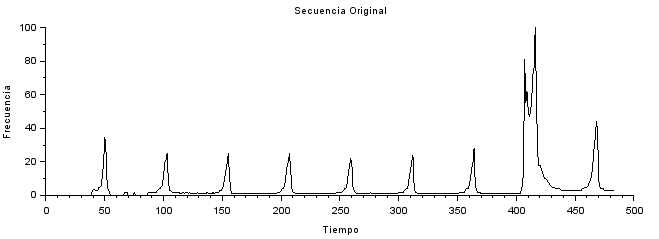
## Consultas tipo Periódica

## Consulta 1: “mistletoe”

*Serie de tiempo en Semanas:* 2004-01-04 al 2013-04-06

*Cantidad de observaciones (semanas):* 483

En la ilustración 38, se pueden apreciar a simple vista alzas abruptas reiterativas cada 50 semanas aproximadamente, que corresponden a la semana del día 25 de Diciembre donde se realiza el evento de “Navidad”.

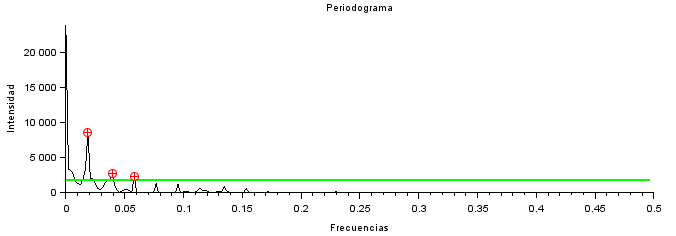


*Ilustración 38: Serie de tiempo en semanas de la consulta "mistletoe". Fuente: Elaboración Propia.*

## Método de Vlachos

**Etapa 1 - Descubriendo periodos candidatos (Periodograma)**

En la ilustración, se muestran “picos” que están por encima del umbral definido en el Periodograma. Estos representan periodos candidatos los cuales se extraen con sus respectivos rangos de alcance para ser utilizados en el proceso de validación de periodos candidatos.



*Ilustración 39: Gráfico de la función de autocorrelación relacionado a la consulta "mistletoe". Fuente: Elaboración Propia.*

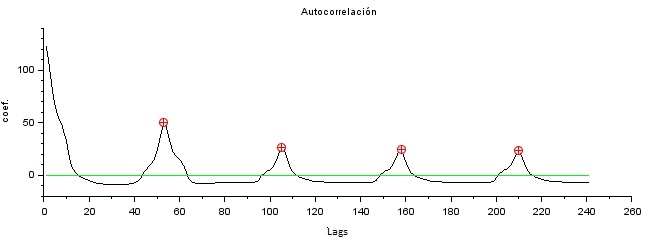
Los periodos candidatos extraídos son los siguientes:

*Tabla 10: Periodos candidatos del Periodograma y sus rangos de alcance consulta "mistletoe". Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Periodos Candidatos** | **Rango** |
| 53,7 | [50 – 58] |
| 25,4 | [24 – 27] |
| 17,3 | [16 – 19] |

**Etapa 2 - Validación de Periodos Candidatos (Autocorrelación)**

Una vez extraídos los periodos candidatos se aplica la función de autocorrelación sobre la serie de tiempo, donde se extraen los valores de los “picos” de los cerros que sobrepasen el intervalo de confianza de (Barlett, 1955), que representan rezagos importantes.



*Ilustración 40: Gráfico función de autocorrelación consulta "mistletoe". Fuente: Elaboración Propia.*

Los rezagos extraídos son 53, 103, 158 y 210. Se puede decir que son múltiplos de 53, que es el 1er valor de regazo extraído, lo que da indicios de un comportamiento periódico en la consulta “mistletoe”.

*Tabla 11: Rezagos importantes de la función de autocorrelación consulta "mistletoe" Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Rezagos Importantes:** | 53, 103, 158, 210 |

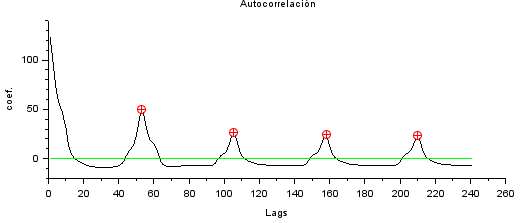
**Etapa 3 - Validación de picos cercanos**

Dentro de los rezagos importantes detectados por la función de autocorrelación (tabla 10) se puede decir que el único rezago que cae dentro de uno de los rangos de los periodos candidatos (tabla 9) es 53. Por lo que se puede concluir que la consulta “mistletoe” es de carácter Periódico con periodo de 53 semanas.

## Método autoMM

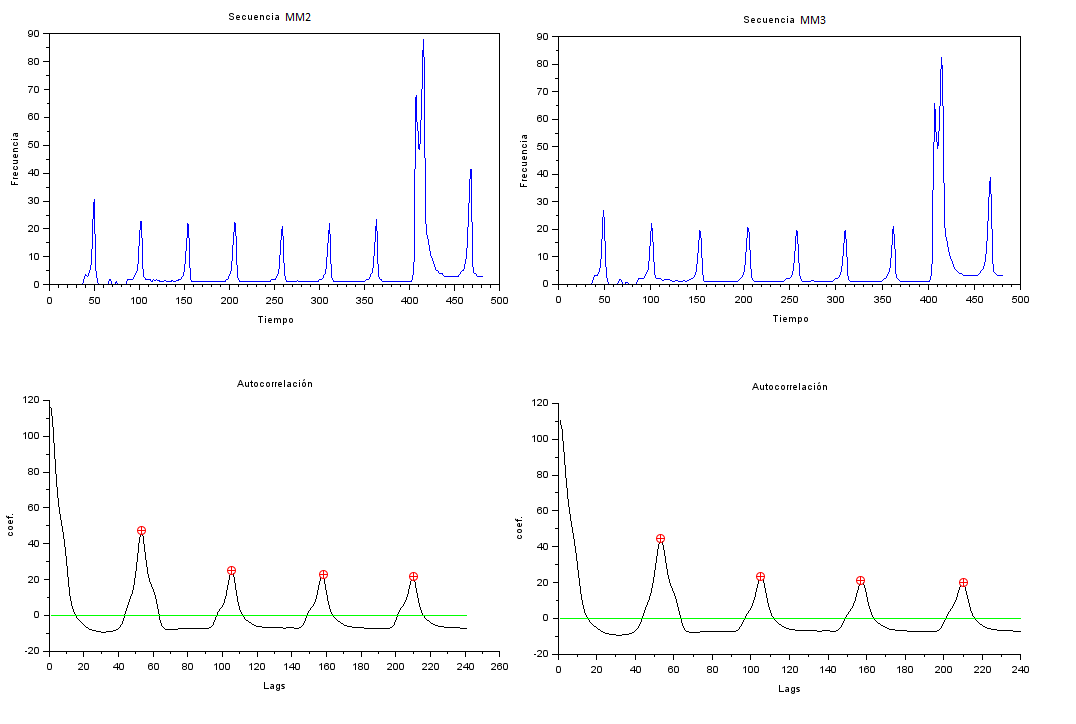
**Etapa 1 - Encontrando periodos candidatos.**

En la ilustración 41, se extraen los rezagos importantes con valores 53, 103, 158 y 210. Una vez extraídos estos rezagos se realiza un proceso de diferencias entre estos con la objetivo de extraer periodos candidatos de la serie de tiempo. Los periodos candidatos extraídos en la 1ra iteración son: 52 y 104.

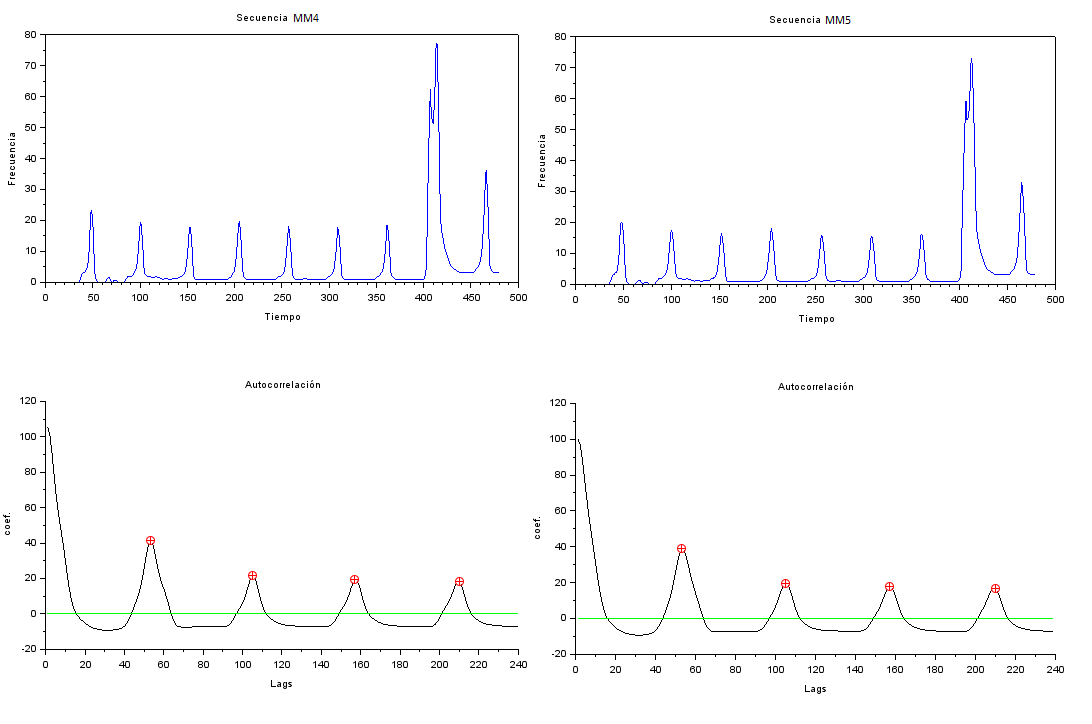


*Ilustración 41: Gráfico función autocorrelación de la serie de tiempo original consulta "mistletoe". Fuente: Elaboración Propia.*

En las iteraciones de suavizado, en las ilustraciones 42 e 43, se aprecia que dichos “picos” o rezagos importantes persisten, al igual que los periodos candidatos extraídos en cada iteración.



*Ilustración 42: Gráfico de la muestra con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*



*Ilustración 43: Gráfico de la muestra con Media Móvil 4 y 5 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*

En la tabla 10, el periodo candidato 52 se repite en todas las iteraciones de medias móviles con un valor de frecuencia 4, y el periodo 105, el cual es un casi múltiplo de 52, solo se repite en la iteración de medias móviles 1 y 2 con frecuencia 4, luego en las siguientes iteraciones tiene frecuencia 2 y aparece otro nuevo periodo con valor 104 también con valor 2. Esto quiere decir que dicho periodo no es importante dentro de la serie.

*Tabla 12: Periodos candidatos por media móvil consulta "mistletoe". Fuente: Elaboración Propia.*

|  |  |  |
| --- | --- | --- |
| **Medias Móviles** | **Candidatos** | **Frecuencias diferencias** |
| 1 | 52,105 | 4 |
| 2 | 52,105 | 4 |
| 3 | 52 | 4 |
| 4 | 52 | 4 |
| 5 | 52 | 4 |

**Etapa 2 – Encontrando periodos finales**

El último paso es verificar si la frecuencia de todos los periodos candidatos iguala o sobrepasa el número de iteraciones de medias móviles realizado. En este caso se realizaron un total de 5 iteraciones de suavizado.

Como la frecuencia de aparición del candidato 52 es igual a la cantidad de iteraciones de suavizado realizadas, se considera un periodo final. El candidato con valor 105 al no igualar o superar el número de iteraciones realizados se descarta.

*Tabla 13: Periodos candidatos finales y su frecuencia de aparición en las iteraciones de suavizado. Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Candidatos Finales** | **Frecuencia** |
| 52 | 5 |
| 105 | 2 |

Como en la muestra existe un periodo final, 52, la consulta “mistletoe” se declara de carácter **Periódica con periodo 52**.

*Tabla 14: Tabla de periodos finales del método AutoMM. Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Periodos Finales AutoMM:** | 52 |

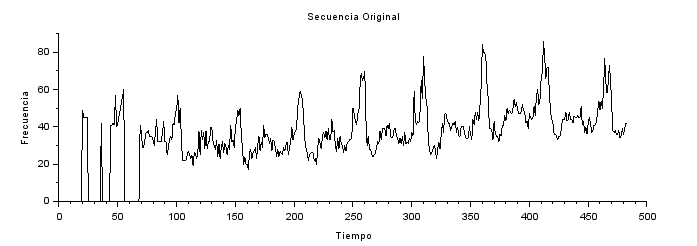
Como conclusión se establece que ambos métodos identifican adecuadamente el tipo de señal.

## Consulta 2: “farm and fleet”

*Serie de tiempo en Semanas:* 2004-01-04 al 2013-04-06

*Cantidad de observaciones (semanas):* 483

La consulta “farm and fleet” tiene relación con una tienda de retail ubicada en E.E.U.U. Por lo que las alzas en la frecuencia de sus consultas tienen relación con los eventos de rebajas de precio de sus productos, ya que las mayores alzas ocurren en la semana de navidad y otra leve alza sucede a mitad de año entre los meses de abril y junio.



*Ilustración 44: Serie de tiempo original en semanas consulta "farm and fleet" Fuente: Elaboración Propia.*

## Método de Vlachos

**Etapa 1 - Descubriendo periodos candidatos**

En la siguiente ilustración se pueden apreciar 3 puntos rojos que representan 3 periodos candidatos.

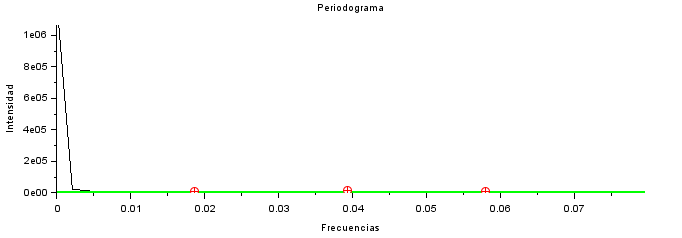


Ilustración : Periodograma consulta "farm and fleet". Fuente: Elaboración Propia.

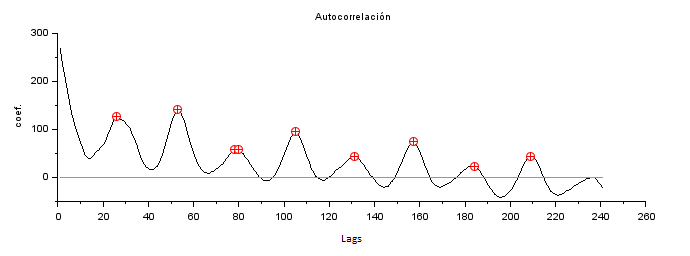
Estos 3 periodos candidatos se almacenan con sus rangos de alcance para luego ser utilizados en la validación de dichos periodos.

*Tabla 15: Periodos candidatos Periodograma y sus rangos de alcance. Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Periodos Candidatos** | **Rangos** |
| 53,7 | [50 – 58] |
| 25,4 | [24 – 27] |
| 17,25 | [16 – 19] |

**Etapa 2 - Validación de Periodos Candidatos (Autocorrelación)**

En el gráfico de autocorrelación se extraen los rezagos importantes que pueden representan periodicidades.



*Ilustración 46: Gráfico de autocorrelación de la serie de tiempo original consulta "farm and fleet". Fuente: Elaboración Propia.*

Como se puede apreciar en la siguiente tabla, los valores de los rezagos son aproximadamente múltiplos del valor 26, lo que nos da un indicio de que en la consulta existe una periodicidad.

*Tabla 16: Rezagos importantes del resultado de la función autocorrelación de la consulta "farm and fleet". Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Rezagos Importantes:** | 26, 53, 78, 80, 105, 131, 157, 184, 209 |

**Etapa 3 - Validación de “picos” cercanos**

Extraídos los periodos candidatos del Periodograma y rezagos importantes de la función de autocorrelación, se procede a realizar la validación de los periodos candidatos con respecto a los rezagos encontrados.

Si los rezagos corresponden a un rango de un periodo candidato, el valor del rezago corresponde a un periodo final de la muestra y se clasifica la consulta como periódica.

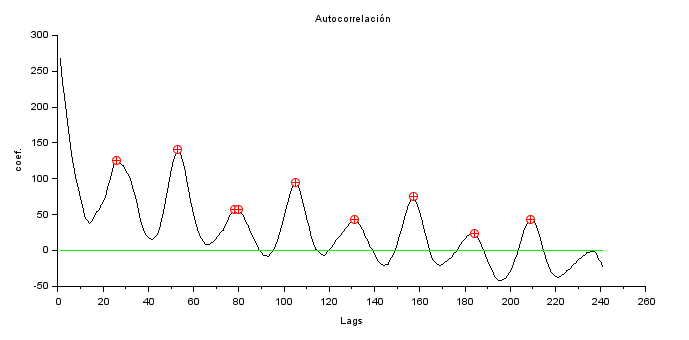
En este caso el valor de los rezagos 26 y 53 corresponden a los rangos de los periodos candidatos 25,4 y 53,7 de la tabla 15. Por lo que se define la consulta “farm and fleet” de carácter Periódico con periodos 26 y 53.

## Método autoMM

**Etapa 1 - Encontrando periodos candidatos.**

En la ilustración 47, el gráfico de autocorrelación extrae los rezagos importantes que pueden representan periodicidades.

Como se aprecia en la tabla 16, los valores de los rezagos son aproximadamente múltiplos del valor 26, lo que nos da un indicio de que en la consulta existe una periodicidad.



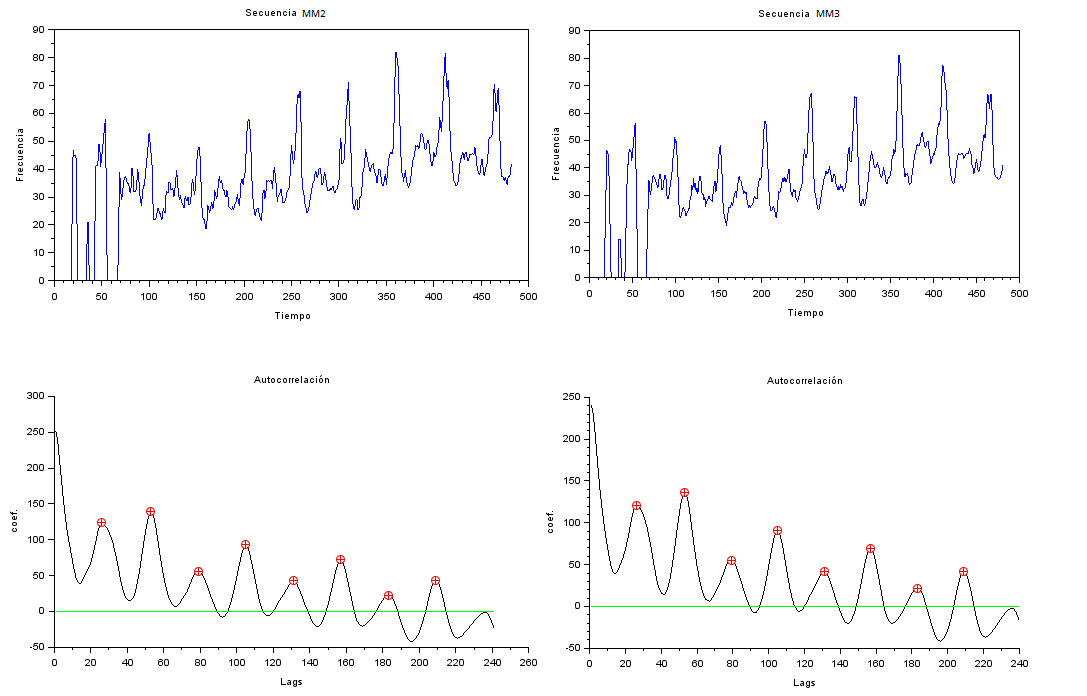
*Ilustración 47: Gráfico de función de autocorrelación de la muestra original consulta "farm and fleet". Fuente: Elaboración Propia.*

Como se descubren varios rezagos, se procede a realizar el proceso de diferenciación de estos con el objetivo almacenar los que tengan el valor de mayor frecuencia de aparición y se declaran como periodos candidatos en cada iteración.

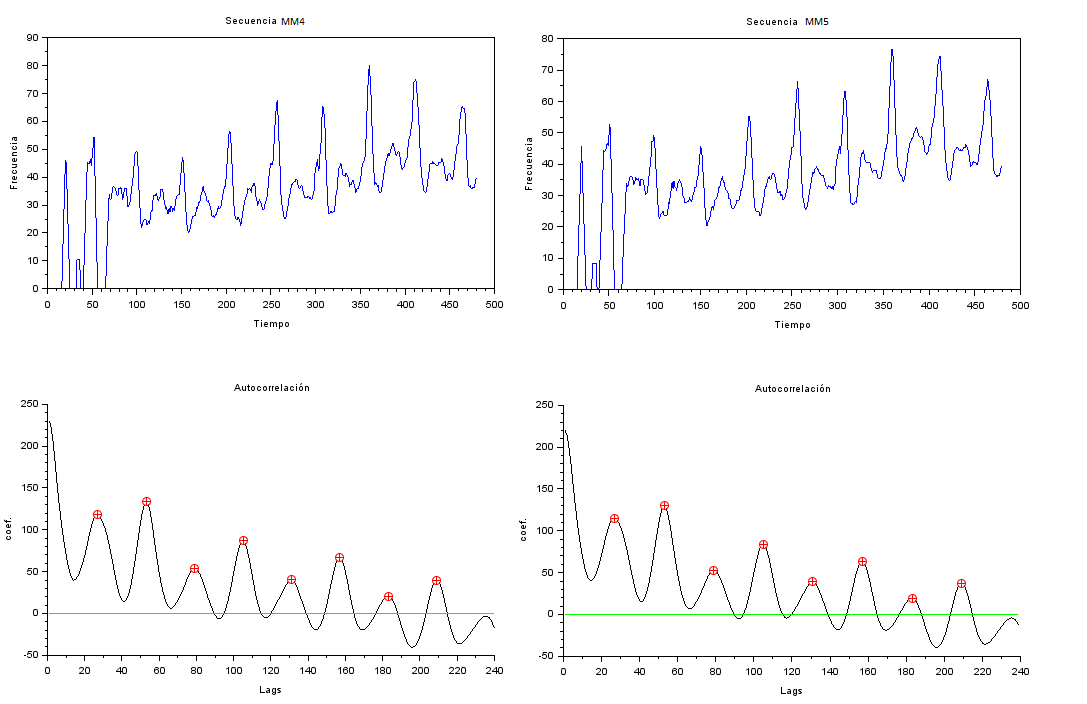
Tabla : Rezagos importantes extraídos de la función de autocorrelación consulta "farm and fleet". Fuente: Elaboración Propia.

|  |  |
| --- | --- |
| **Rezagos Importantes:** | 26, 53, 78, 80, 105, 131, 157, 184, 209 |

En las ilustraciones 48 y 49, el comportamiento de la función de autocorrelación se mantiene a pesar de las iteraciones de suavizado de la muestra mediante medias móviles. Con esto también se mantienen los valores de los rezagos importantes a través de las iteraciones.



*Ilustración 48: Gráfico de la muestra con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*



*Ilustración 49: Gráfico de la muestra con Media Móvil 2 y 3 con su respectivo gráfico de autocorrelación. Fuente: Elaboración Propia.*

En la tabla 17, la primera iteración tuvo dos valores máximos con la misma frecuencia de aparición a través de la diferenciación, pero en las siguientes iteraciones, el que se mantuvo con mayor valor de frecuencia fue el valor 26.

*Tabla 18: Periodos candidatos por iteración de media móvil. Fuente: Elaboración Propia.*

|  |  |  |
| --- | --- | --- |
| **Medias Móviles** | **Periodos Candidatos** | **Frecuencia diferencia** |
| 1 | 27, 52 | 8 |
| 2 | 26 | 12 |
| 3 | 26 | 12 |
| 4 | 26 | 14 |
| 5 | 26 | 14 |

**Etapa 2 – Encontrando periodos finales**

Teniendo el total de periodos candidatos, se realiza el proceso de unión de valores contiguos de estos, como se aprecia en la tabla 17, el valor 26 y 27 al ser valores contiguos, se realiza un proceso de unión de estos, generando un nuevo periodo candidato con valor 27.

Al unir los valores de los periodos candidatos contiguos, manteniendo sus frecuencias se obtienen los siguientes periodos finales.

*Tabla 19: Periodos finales y su frecuencia de aparición en las iteraciones. Fuente: Elaboración Propia.*

|  |  |
| --- | --- |
| **Periodos Finales** | **Frecuencias** |
| 26 | 5 |
| 52 | 1 |

Como la frecuencia de aparición del periodo 26 es igual al número de iteraciones de medias móviles realizado, se declara la consulta “farm and fleet” como Periódica con periodo de 26 semanas.

## Conclusión métodos consultas tipo Periódica

Como conclusión se establece que ambos métodos identifican adecuadamente el tipo de señal.

El método de Vlachos extrajo periodos candidatos para luego con la función de autocorrelación identificaba rezagos importantes, donde uno de los valores de dichos rezagos correspondía al rango definido de un periodo candidato del Periodograma cumpliendo con la condición de periodicidad del método, declarando correctamente las consultas como Periódicas.

En el método autoMM a pesar de suavizar la muestra, la estructura de la función de autocorrelación se mantuvo y también el número y valor de los rezagos importantes, por lo que los candidatos extraídos en cada iteración de suavizado se mantenían. Luego, en el paso de identificación de periodos finales, al encontrar periodos con frecuencias iguales al valor de la cantidad de iteraciones de suavizado de medias móviles, se cumple con la condición de Periodicidad de la clasificación del método declarando en ambos casos adecuadamente las consultas como Periódicas.

# Análisis Comparativo

Para realizar una evaluación de la calidad en la identificación automatizada de señales periódicas y no periódicas de ambos métodos, se utilizó una herramienta de ***análisis de concordancia*** cuyo resultado es una variable cuantitativa que permite medir el ***grado de acuerdo*** entre dos evaluadores (Uebersax, 1987).

También, se utilizó una herramienta llamada ***Matriz de Confusión*** con el objetivo de determinar el ***grado de acierto y error*** en la clasificación de ambos métodos.

En ambas herramientas se utilizó como referencia las consultas clasificadas en la asesoría realizada por los profesores de la Universidad de Shandong con el objetivo de determinar cuál de los dos métodos clasifica de mejor manera las señales periódicas y no periódicas.

## Proporción de Acuerdo Observado y Coeficiente kappa de Cohen.

Considere la evaluación de 2 valuadores (expertos, jurados, procedimientos de diagnósticos, etc.) en la tabla 19:

Tabla : Tabla Distribución de Acuerdos. Fuente: Diversity of Decision-Making Models and the Measurement of Interrater Agreement, John S. Uebersax, 1987.

|  |  |  |
| --- | --- | --- |
|  | Evaluador 2 | |
| Evaluador 1 | Si | No |
| Si | a | b |
| No | c | d |

Donde, los valores de , , y denotan las frecuencias de cada combinación posible de la evaluación por Evaluador 1 y 2. Y es el número de casos evaluados.

## Proporción de Acuerdo Observado

El *índice de proporción de acuerdos observado*  representa la proporción de casos donde dos evaluadores llegan a un acuerdo. Simplemente expresa la proporción o porcentaje de acuerdo entre ellos, es decir, en qué medida hubo coincidencia en la clasificación entre los dos observadores en relación al total de elementos examinados.

Sin embargo, como indicador tiene el inconveniente de que aun en el caso de que los dos evaluadores clasifiquen con criterios independientes, se produce un cierto grado de acuerdo por azar. Por ejemplo, si dos sujetos, sin ninguna especialización, se dedican a clasificar una serie de sujetos en sanos y sospechosos de enfermedad. En ciertos casos los “diagnósticos” serán coincidentes para cierto número de sujetos, pero no debido a la coincidencia de criterio de los observadores, sino que por simple azar.

Es por este motivo que para evaluar estos métodos, se utiliza otro índice el cual toma en cuenta la parte de acuerdos por azar.

## Índice Kappa de Cohen

Es un índice propuesto por (Cohen, 1960), que permite medir el nivel de concordancia entre dos evaluadores y representa el grado de acuerdo entre ellos incluyendo el factor de decisión por azar entre ellos.

Siendo la proporción general de acuerdo y la proporción de acuerdos por azar.

Si los evaluadores están completamente de acuerdo, entonces k=1. Si k=0 significa existencia de discordancia entre los calificadores.

Para interpretar la evaluación de este índice, (Landis & Koch, 1977) propusieron una escala de medición para la valorización del índice kappa.

Tabla : Métricas de medición del índice de kappa de cohen. Fuente: The Measurement of Observer Agreement for Categorical Data, Landis & Koch, 1977.

|  |  |
| --- | --- |
| **Kappa** | **Grado de acuerdo** |
| <0,00 | Sin acuerdo |
| 0,01 – 0,20 | Débil |
| 0,21 – 0,40 | Aceptable |
| 0,41 – 0,60 | Moderado |
| 0,61 – 0,80 | Sustancial |
| 0,81 – 0,99 | Casi perfecto |

También, (Fleiss, 1981) y (Cicchetti & Sparrow, 1981) caracterizan los valores índices de kappa de la siguiente manera:

Tabla : Clasificación índice kappa por Fleiss, Cicchetti & Sparrow. Fuente: Statistical Methods for Rates and Proportions, Fleiss J, Levin B & Paik M, 2003.

|  |  |
| --- | --- |
| **Kappa** | **Clasificación** |
| <0,40 | Pobre |
| 0,41 – 0,49 | Razonable |
| 0,60 – 0,74 | Bueno |
| 0,75< | Excelente |

Se toma en cuenta este último método de clasificación y nos enfocamos en realizar una **evaluación de acuerdo** entre ambos métodos, con respecto a la clasificación realizada por el grupo de profesores de la Universidad de Shangdong.

## Matriz de confusión.

Cuando uno se refiere al rendimiento de un algoritmo de clasificación, uno está interesado en la capacidad del algoritmo de clasificar o separar sus clases de manera correcta o incorrecta (Powers, 2001).

Cuando buscamos errores que hace el algoritmo de clasificación, la Matriz de Confusión proporciona dicha información.

Cada columna de la matriz representa los casos que el algoritmo predijo, mientras que cada fila representa los casos en una clase real.

Considere la evaluación de un algoritmo que predice la clase a la que pertenece un conjunto de individuos, comparando la predicción versus la clase a la que realmente pertenecen.

La siguiente tabla muestra la matriz de confusión para un clasificador de dos clases Si y No.

Tabla : Tabla matriz de confusión para un clasificador. Fuente: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, Powers, David M. W. ,2001.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicción Clasificador | |
|  |  | Si | No |
| Verdad | Si | TP | FN |
| No | FP | TN |

Las notaciones para la matriz de confusión son:

* **TP:** Verdaderos Positivos, cuando lo predicho por el algoritmo es Si, cuando la verdad, es Si.
* **FN:** Falso Negativo, cuando lo predicho por el algoritmo declara No, cuando la verdad es Si.
* **FP:** Falsos Positivos, cuando lo predicho por el algoritmo declara Si, cuando la verdad es No.
* **TN:** Verdaderos Negativos, cuando lo predicho por el algoritmo declara No, cuando la verdad es No.

## Métricas de la matriz de confusión.

A continuación se describirán algunas métricas que proporcionan la información necesaria para determinar el acierto y error de la clasificación realizada por el algoritmo o evaluador.

**Sensitivity, Recall o True Positive Rate (TPR):** Corresponden a la proporción de instancias que corresponden a la clasificación verdadera.

**Precision o Positive Predictive Values (PPV):** Corresponde a la proporción de valores predichos verdaderos del conjunto de verdaderos totales.

**Specificity o True Negative Rate (TNR):** Corresponde a la proporción de instancias negativas que corresponden a la clasificación negativa.

Algunas de las interpretaciones que es posible dar a estos indicadores son las siguientes:

La proporción **Sensitivity/Recall** indica que tan bueno es el algoritmo detectando positivos y cuantifica la evasión de falsos negativos. **Specificity** cuantifica que tan bueno es el algoritmo en la evasión de falsas alarmas o falsos positivos. **Precision** indica cuantos casos positivos que fueron relevantes.

La matriz de confusión proporciona una visión de los aciertos y errores de ambos métodos tomando como referencia de verdad lo clasificado por la agrupación de la Universidad de Shangdong.

## Preparando los casos de prueba para la Evaluación de Métodos

Para realizar la evaluación de los métodos se consideraron 200 consultas al azar de cada tipo de consultas clasificadas por los asesores de la Universidad de Shandong (200 para AQ, 200 para OBQ, 200 para SQ y 200 para PQ), haciendo un total de 800 consultas.

Estas consultas, se segmentaron en conjuntos, con el objetivo de evaluar las metodologías en diferentes escenarios, dichos conjuntos fueron ordenados de la siguiente manera:

Tabla : Conjuntos de tipos de consultas para evaluar la calidad de resultados los métodos VL y AutoMM. Fuente: Elaboración Propia.

|  |  |  |
| --- | --- | --- |
| **Conjuntos** | **Descripción** | **Número de Consultas** |
| PQ-AQ-SQ-OBQ | Representa el conjunto de resultados de consultas Periódicas, Aperiódicas, Estables y Ráfaga. | 800 |
| AQ-PQ | Representa el conjunto de resultados con consultas Aperiódicas y Periódicas. | 400 |
| SQ-PQ | Representa el conjunto de resultados con consultas Estables y Periódicas. | 400 |
| OBQ-PQ | Representa el conjunto de resultas con consultas en Ráfaga y Periódicas | 400 |

Con estos conjuntos se determinará la efectividad de discriminación de la clasificación de consultas Periódicas y no periódicas en cada escenario.

Utilizando los índices *proporción de acuerdo observado* y *Kappa de Cohen* se realizó una medición de la fiabilidad en la identificación de consultas de tipo Periódicas de las metodologías AutoMM y VL.

Con las métricas de la Matriz de Confusión se realizó un análisis de cuantificación de errores de los mejores resultados en la clasificación de señales periódicas obtenidos por cada método en los índices kappa de Cohen, tomando como referencia de verdad, lo clasificado por la agrupación de la Universidad de Shangdong.

## Procedimiento para la Evaluación

Primero para el Método AutoMM, se realizó un *benchmark,* variando la cantidad de iteraciones de suavizado de medias móviles de 1 a 40, con el objetivo de identificar la cantidad de iteraciones que entrega mejores resultados a la hora de identificar series periódicas en cada uno de los conjuntos detallados en la tabla anteriormente.

Para el Método de Vlachos se realizaron 40 iteraciones en cada uno de los segmentos para extraer los valores máximos y mínimos de los índices de concordancia nombrados anteriormente.

Una vez extraídos los resultados de máximos de los índices de concordancia, se utilizó la matriz de concordancia para conocer los índices de error y aciertos con el fin de detectar las deficiencias de ambos métodos.

Extraídos los índices de concordancia y mediciones de la matriz de confusión de cada una de estas metodologías, se realizó un análisis y comparación de dichos resultados, con el objetivo de evaluar ambos métodos en el reconocimiento de señales de tipo periódica.

## Resultados Proporción de Acuerdo Observado para Método AutoMM

En las siguientes tablas, se muestran los resultados obtenidos de la proporción de acuerdo observado por el Método AutoMM por cada número de iteraciones de medias móviles y conjuntos de tipo de consultas:

Tabla : Tabla de Resultados Índice de Acuerdo Observado por media móvil del 1 al 20. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AutoMM | PQ-AQ-SQ-OBQ | AQ-PQ | SQ-PQ | OBQ-PQ |
| MM1 | 0,336 | 0,540 | 0,505 | 0,598 |
| MM2 | 0,615 | 0,670 | 0,680 | 0,730 |
| MM3 | 0,685 | 0,703 | 0,695 | 0,788 |
| MM4 | 0,698 | 0,720 | 0,693 | 0,798 |
| MM5 | 0,709 | 0,733 | 0,700 | 0,805 |
| MM6 | 0,699 | 0,718 | 0,695 | 0,795 |
| MM7 | 0,690 | 0,703 | 0,703 | 0,775 |
| MM8 | 0,690 | 0,705 | 0,715 | 0,775 |
| MM9 | 0,711 | 0,725 | 0,733 | 0,790 |
| MM10 | 0,723 | 0,755 | 0,725 | 0,800 |
| MM11 | 0,733 | 0,760 | 0,748 | 0,798 |
| MM12 | 0,746 | 0,763 | 0,773 | 0,793 |
| MM13 | 0,749 | 0,763 | 0,780 | 0,795 |
| MM14 | 0,761 | 0,775 | 0,785 | 0,808 |
| MM15 | 0,780 | 0,780 | 0,805 | 0,815 |
| MM16 | 0,779 | 0,780 | 0,803 | 0,815 |
| MM17 | 0,790 | 0,785 | 0,805 | 0,820 |
| MM18 | 0,800 | 0,795 | 0,815 | 0,835 |
| MM19 | 0,801 | 0,798 | 0,818 | 0,833 |
| MM20 | 0,813 | 0,808 | 0,825 | 0,838 |

Tabla : Tabla de Resultados Índice de Acuerdo Observado por media móvil del 21 al 40. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AutoMM | PQ-AQ-SQ-OBQ | AQ-PQ | SQ-PQ | OBQ-PQ |
| MM21 | 0,825 | 0,815 | 0,838 | 0,838 |
| MM22 | 0,830 | 0,820 | 0,838 | 0,838 |
| MM23 | 0,840 | 0,825 | 0,843 | 0,843 |
| MM24 | 0,841 | 0,823 | 0,845 | 0,840 |
| MM25 | 0,848 | 0,830 | 0,853 | 0,848 |
| MM26 | 0,856 | 0,838 | 0,858 | 0,858 |
| MM27 | 0,859 | 0,845 | 0,855 | 0,858 |
| MM28 | 0,861 | 0,840 | 0,848 | 0,860 |
| MM29 | 0,863 | 0,838 | 0,843 | 0,855 |
| MM30 | 0,863 | 0,835 | 0,840 | 0,850 |
| MM31 | 0,864 | 0,838 | 0,843 | 0,848 |
| MM32 | 0,869 | 0,840 | 0,843 | 0,855 |
| MM33 | 0,874 | 0,840 | 0,840 | 0,858 |
| MM34 | 0,873 | 0,838 | 0,838 | 0,855 |
| MM35 | 0,875 | 0,835 | 0,838 | 0,853 |
| MM36 | 0,875 | 0,835 | 0,833 | 0,853 |
| MM37 | 0,878 | 0,833 | 0,838 | 0,850 |
| MM38 | 0,874 | 0,825 | 0,828 | 0,840 |
| MM39 | 0,876 | 0,828 | 0,830 | 0,840 |
| MM40 | 0,875 | 0,823 | 0,825 | 0,838 |

Ilustración : Gráfico de Resultados Índice de Acuerdo Observado por conjuntos de clasificación, Medias móviles del 1 al 40. Fuente: Elaboración Propia.

Como se aprecia en la ilustración 50 relacionada a las tablas 25 y 26, a medida que se aplican iteraciones de suavizado de medias móviles a la serie de tiempo mejora la identificación de señales tipo periódicas.

Si analizamos por conjunto de consultas se observar con detalle el grado de efectividad en la discriminación de consultas periódicas con respecto a las de tipo ráfaga, estable y aperiódicas.

El conjunto de OBQ-PQ inicio con un porcentaje de aciertos de 60%, luego entre las iteraciones 4 y 5 de medias móviles el porcentaje de aciertos subió considerablemente a un 80%, manteniendo este valor hasta llegar a la media móvil 14 donde luego empezó a subir de apoco hasta llegar a un valor máximo de 86% en la media móvil 24.

El conjunto de consultas AQ-PQ comenzó con un porcentaje de aciertos de 54%, luego en la media móvil 3 alcanzó un 70% manteniéndose hasta la media móvil 9 donde comenzó a tener una subida hasta llegar al porcentaje máximo de 84,5% en la media móvil 27.

El conjunto de consultas SQ-PQ comenzó con un porcentaje de aciertos del 50,5%, luego entre las medias móviles 3 y 8 se mantuvo cerca del 70% comenzando a subir hasta llegar a un porcentaje máximo de aciertos del 85,8% en la media móvil 26.

Dentro del conjunto general, llegando a la media móvil 5 ya se obtuvo un 70% en la proporción de acuerdo en los resultados de reconocimiento de señales periódicas y no periódicas. Luego este índice se mantiene estable y tendiendo a aumentar, llegando a un valor máximo en promedio de 87,8% de acuerdo cuando se aplica un número de 37 iteraciones.

También se puede decir que el Método AutoMM comienza teniendo mejores resultados reconociendo señales tipo periódicas dentro del conjunto del tipo Ráfaga, ya que en un número menor de iteraciones, en media móvil 4 y 5, bordea el 80% de acuerdo con lo establecido por los asesores.

## Resultado Índice de Proporción de Acuerdo para Método de Vlachos

Se realizaron 40 iteraciones y se extrajo los valores máximos y mínimos de los resultados en los índices de acuerdo entre el método y la agrupación de la Universidad de Shangdong, dando un resultado del 68,1% como porcentaje máximo y un 66,3% mínimo.

Tabla : Valor máximo y mínimo índice de acuerdo observado por conjunto de método Vlachos. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PQ-AQ-SQ-OBQ | AQ-PQ | SQ-PQ | OBQ-PQ |
| max | 0,681 | 0,760 | 0,738 | 0,768 |
| min | 0,663 | 0,733 | 0,705 | 0,740 |
| promedio | 0,672 | 0,747 | 0,722 | 0,757 |

Para el método de Vlachos el resultado del grado de discriminación entre conjuntos se puede apreciar que el grado de aciertos promedios es mayor en el conjunto de Ráfagas con un 75,7%, luego viene el conjunto de aperiódicas con un 74,7% y finalmente las estables con un 72%, con esto podemos decir que para el método de Vlachos es más fácil discriminar las consultas periódicas de las ráfaga, luego vienen las aperiódicas y las estables.

En general se puede decir que la identificación de señales periódicas es similar dentro de los conjuntos independientes, AQ-PQ, SQ-PQ, y OBQ-PQ, pero el rendimiento general de identificación no es del todo bueno ya que no supera el 70%.

## Resultados Índices de kappa de Cohen para Método AutoMM

Tabla : Tabla de Resultados Índice Kappa de Cohen de AutoMM por tipo de consulta, media móvil del 1 al 20. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AutoMM | PQ-AQ-SQ-OBQ | AQ-PQ | SQ-PQ | OBQ-PQ |
| MM1 | 0,051 | 0,080 | 0,010 | 0,195 |
| MM2 | 0,274 | 0,340 | 0,360 | 0,460 |
| MM3 | 0,352 | 0,405 | 0,390 | 0,575 |
| MM4 | 0,370 | 0,440 | 0,385 | 0,595 |
| MM5 | 0,388 | 0,465 | 0,400 | 0,610 |
| MM6 | 0,370 | 0,435 | 0,390 | 0,590 |
| MM7 | 0,354 | 0,405 | 0,405 | 0,550 |
| MM8 | 0,359 | 0,410 | 0,430 | 0,550 |
| MM9 | 0,393 | 0,450 | 0,465 | 0,580 |
| MM10 | 0,413 | 0,510 | 0,450 | 0,600 |
| MM11 | 0,429 | 0,520 | 0,495 | 0,595 |
| MM12 | 0,449 | 0,525 | 0,545 | 0,585 |
| MM13 | 0,455 | 0,525 | 0,560 | 0,590 |
| MM14 | 0,476 | 0,550 | 0,570 | 0,615 |
| MM15 | 0,506 | 0,560 | 0,610 | 0,630 |
| MM16 | 0,504 | 0,560 | 0,605 | 0,630 |
| MM17 | 0,520 | 0,570 | 0,610 | 0,640 |
| MM18 | 0,542 | 0,590 | 0,630 | 0,670 |
| MM19 | 0,544 | 0,595 | 0,635 | 0,665 |
| MM20 | 0,564 | 0,615 | 0,650 | 0,675 |

Tabla : Tabla de Resultados Índice Kappa de Cohen de Auto MM por tipo de consultas, medias móviles del 21 al 40. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AutoMM | PQ-AQ-SQ-OBQ | AQ-PQ | SQ-PQ | OBQ-PQ |
| MM21 | 0,586 | 0,630 | 0,675 | 0,675 |
| MM22 | 0,594 | 0,640 | 0,675 | 0,675 |
| MM23 | 0,612 | 0,650 | 0,685 | 0,685 |
| MM24 | 0,613 | 0,645 | 0,690 | 0,680 |
| MM25 | 0,628 | 0,660 | 0,705 | 0,695 |
| MM26 | 0,647 | 0,675 | 0,715 | 0,715 |
| MM27 | 0,652 | 0,690 | 0,710 | 0,715 |
| MM28 | 0,654 | 0,680 | 0,695 | 0,720 |
| MM29 | 0,653 | 0,675 | 0,685 | 0,710 |
| MM30 | 0,651 | 0,670 | 0,680 | 0,700 |
| MM31 | 0,653 | 0,675 | 0,685 | 0,695 |
| MM32 | 0,664 | 0,680 | 0,685 | 0,710 |
| MM33 | 0,673 | 0,680 | 0,680 | 0,715 |
| MM34 | 0,669 | 0,675 | 0,675 | 0,710 |
| MM35 | 0,672 | 0,670 | 0,675 | 0,705 |
| MM36 | 0,671 | 0,670 | 0,665 | 0,705 |
| MM37 | 0,675 | 0,665 | 0,675 | 0,700 |
| MM38 | 0,663 | 0,650 | 0,655 | 0,680 |
| MM39 | 0,668 | 0,655 | 0,660 | 0,680 |
| MM40 | 0,663 | 0,645 | 0,650 | 0,675 |

Ilustración : Gráfico índices kappa de cohen por conjuntos de clasificación de consultas, iteración de medias móviles del 1 al 40. Fuente: Elaboración Propia.

En las tablas 28 y 29 relacionada la ilustración 51, a medida que se aumenta el número de iteraciones de suavizado de medias móviles, el índice de proporción kappa de Cohen va en aumento llegando a un valor máximo de 0,675 en la iteración 37. Según la tabla, de las métricas establecidas por Fleiss, Cicchetti & Sparrow, el valor llega al rango de nivel “Bueno”.

También se aprecia en la ilustración que existe un alza notoria entre las medias móviles 3 y 5 en la identificación de señales periódicas dentro del conjunto de consultas tipo Ráfaga (OBQ-PQ), donde en la media móvil 5 da como resultado 0,61 alcanzando el nivel de evaluación “Bueno”. Esto quiere decir que el método tiene un mejor rendimiento en la identificación de señales periódicas dentro del conjunto de tipo ráfaga.

## Resultados Índice de Kappa de Cohen para Método de Vlachos

Para el Método de Vlachos se iteró 40 veces los 800 casos de prueba de las consultas y se extrajo el rango mínimo y máximo de los índices de kappa de cohen para cada uno de los conjuntos.

Tabla : Valor máximo y mínimo índice de acuerdo observado. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PQ-AQ-SQ-OBQ | AQ-PQ | SQ-PQ | OBQ-PQ |
| max | 0,370 | 0,520 | 0,475 | 0,535 |
| min | 0,338 | 0,465 | 0,410 | 0,480 |

Los resultados de los índices arrojan un resultado máximo de 0,370 y mínimo 0,338, donde según las métricas de la tabla, es “pobre”.

Para la identificación de señales periódicas dentro de los conjuntos independientes AQ-PQ, SQ-PQ y OBQ, están dentro del rango “moderado”.

## Resultados Matriz de Concordancia

Para los resultados de matriz de confusión los casos utilizados fueron los siguientes:

* Para el Método de Vlachos se utilizaron los resultados que representaba los máximos de kappa de Cohen para cada conjunto de tipos de consultas, se denotará como “Max Vlachos”.
* Para el Método AutoMM se utilizó la tabla de resultados que representan los mejores resultados en el índice de Kappa de Cohen en el conjunto total de tipos de consultas, la cual fue de 37 iteraciones, está la denotaremos como AutoMM 37.

Tabla : Tabla de resultados matriz de concordancia conjuntos PQ-SQ-AQ-OBQ. Fuente: Elaboración Propia.

|  |  |  |  |
| --- | --- | --- | --- |
| PQ-SQ-AQ-OBQ | Precisión (PPV) | Recall o sensibility (TPR) | Specificity (TNR) |
| Max Vlachos | 0,433 | 0,885 | 0,613 |
| AutoMM 37 | 0,750 | 0,765 | 0,915 |

Tabla : Tabla de resultados matriz de concordancia conjuntos PQ-SQ. Fuente: Elaboración Propia.

|  |  |  |  |
| --- | --- | --- | --- |
| PQ-SQ | Precisión (PPV) | Recall o sensibility (TPR) | Specificity (TNR) |
| Max Vlachos | 0,682 | 0,890 | 0,585 |
| AutoMM 37 | 0,895 | 0,765 | 0,910 |

Tabla : Tabla de resultados matriz de confusión conjuntos PQ-AQ. Fuente: Elaboración Propia.

|  |  |  |  |
| --- | --- | --- | --- |
| PQ-AQ | Precisión (PPV) | Recall o sensibility (TPR) | Specificity (TNR) |
| Max Vlachos | 0,708 | 0,885 | 0,635 |
| AutoMM 37 | 0,884 | 0,765 | 0,900 |

Tabla : Tabla de resultados matriz de confusión PQ-OBQ. Fuente: Elaboración Propia.

|  |  |  |  |
| --- | --- | --- | --- |
| PQ-OBQ | Precisión (PPV) | Recall o sensibility (TPR) | Specificity (TNR) |
| Max Vlachos | 0,713 | 0,895 | 0,640 |
| AutoMM 37 | 0,922 | 0,765 | 0,935 |

Dentro del análisis por conjuntos AutoMM 37 identifica y clasifica como no periódicas las consultas en ráfaga de mejor manera ya que tiene un porcentaje de Specificity del 93,5%, luego le sigue las consultas estables con un 91% y 90% las estables. También se puede decir que las autoMM37 alcanzo un porcentaje igual en todos los conjuntos, por lo que identifica las periódicas de similar manera dentro de todos los conjuntos.

Siguiendo con el análisis por conjuntos Vlachos identifica y clasifica como no periódicas las consultas en ráfaga de mejor manera con un porcentaje de Specificity del 64%, luego le siguen las Aperiódicas con un 63,5 y luego las estables con un 58,5%. También, se puede decir que Vlachos reconoce mejor las consultas tipo Periódicas dentro del conjunto de Ráfagas con un 89,5% en Recall, luego sigue las Estables con un 89% y las aperiódicas con un 88,5%.

En general comparando los resultados entre ambos métodos, el método de Vlachos tiene mejores resultados a la hora de identificar las señales que efectivamente son de *tipo periódica*, ya que tiene una mejor proporción de Recall. Sin embargo, su precisión es menor al Método AutoMM 37 por el hecho de clasificar de manera errónea las que no lo son, ya que el valor de Specificity para cada conjunto de prueba, es mucho menor a la conseguida con el Método AutoMM37.

## Análisis de resultados

Al comparar los resultados obtenidos con ambos métodos sobre el mismo universo de datos entregado por la asesoría de expertos realizada por la Universidad de Shangdong, el Método AutoMM arrojó mejores resultados en los aciertos de clasificación de señales periódicas y no periódicas con un máximo de 87% de aciertos, a diferencia del método de Vlachos con un resultado de 68% aciertos.

También, al observar el comportamiento de los índices de acuerdo por conjunto se puede concluir que para el método autoMM es más fácil discriminar las consultas periódicas de las consultas ráfaga, ya que obtuvo buenos resultados con pocas iteraciones de suavizado.

Ahora, la discriminación de periódicas dentro del conjunto de tipos aperiódicas y estables se puede decir que a menor suavizado el método discrimina mejor las consultas tipo periódica de las aperiódicas que de las estables y a mayor suavizado el método discrimina mejor las consultas tipo periódica de las estables que de las aperiódicas, esto quiere decir que para determinar cual tiene mejor grado de discriminación en la identificación de la consultas periódicas entre los conjuntos de aperiódicas y estables dependen del grado de suavizado de la muestra.

Gracias al Análisis de Matriz de Confusión se observó que el Método de Vlachos clasifica mejor las señales que son periódicas con respecto al Método AutoMM, sin embargo, tiende a etiquetar de periódicas las señales que no lo son, es por esto que el Método de Vlachos tiene un porcentaje de aciertos menor al Método AutoMM. También se pudo comprobar que ambos métodos discriminan de mejor manera las señales de tipo Ráfaga de las Periódicas, ya que su índice de Recall y Specificity es el más alto de los conjuntos.

En cuando a los resultados del índice de Kappa de Cohen, el Método AutoMM alcanza una calificación máxima de 0,675 calificándose como “bueno”, mientras que el Método de Vlachos fue calificado como “pobre”, concluyéndose que el Método AutoMM es más fiable que el de Método de Vlachos.

## Métricas Objetivos Específicos

## Métrica Objetivo Especifico 1

Tabla 35: Objetivo Especifico 1. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obj.Esp | Situación Actual | Resultados Esperados | Métricas | Criterios de Éxito |
| [OE1] | No se tienen métodos implementados ni tampoco herramientas de software que permitan automatizar la detección de señales periódicas. | Un conjunto de métodos implementados para realizar una etapa experimental para la detección de señales periódicas. | Numero de métodos implementados | >=2 |

La investigación cumple con el criterio de éxito del objetivo específico [OE1] implementado dos métodos automatizados.

* Método de Vlachos, &u & Castelli.
* Método de AutoMM.

## Métrica Objetivo Especifico 2

Tabla 36: Objetivo Especifico 2. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obj.Esp | Situación Actual | Resultados Esperados | Métricas | Criterios de Éxito |
| [OE2] | No existe en la literatura un dataset consistente para evaluar señales periódicas. | Un dataset de series de tiempo en el contexto web para la evaluación experimental. Este dataset está compuesto por distintos tipos de señales. | Número de series de tiempo. | >=400 |

La investigación cumple con el criterio de éxito del objetivo específico [OE2] utilizando 800 consultas del motor de búsqueda de Google representando muestras de series de tiempo relacionadas al contexto de la web, clasificadas por agrupación de la facultad de ciencias y tecnología de la universidad de Shandong.

## Métrica Objetivo Especifico 3

Tabla 37: Objetivo Especifico 3. Fuente: Elaboración Propia.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obj.Esp | Situación Actual | Resultados Esperados | Métricas | Criterios de Éxito |
| [OE3] | No existe en la literatura referencia en términos de calidad para evaluar la efectividad de los métodos. | Un método con efectividad alta para la discriminación entre señales periódicas y no periódicas. | RAW | >=80% |
| Kappa de Cohen | >=60% |
| Precision, Recall, Specificity | >60% |

La investigación cumple con los criterios de éxito del objetivo específico [OE3], donde el método propuesto AutoMM cumple con un RAW de 87,8%, Índice de Kappa de 67,5%, (Bueno), Precision de 75%, Recall 88,5% y Specificity de 91,5%.

# Conclusión

En la investigación, utilizando distintas herramientas dedicadas a encontrar periodos dentro de las muestras de series de tiempo, fue posible establecer métodos automatizados que logran identificar señales tipo periódicas.

Además, utilizando un número considerable de muestras clasificadas por asesores de la universidad de Shangdong se obtuvo un conjunto importante de casos de prueba para determinar la fiabilidad de ambos métodos implementados en la investigación.

Con respecto a los resultados obtenidos en la evaluación se pudo comprobar que el esquema propuesto de promedios móviles ayuda en la detección y clasificación de señales periódicas y no periódicas, ya que permite suavizar las variaciones de la serie de tiempo, disminuyendo sus irregularidades y su componente aleatoria, manteniendo su estructura principal.

En la evaluación por conjuntos de tipos de consultas se logró observar que ambos métodos discriminaban de manera más efectiva las señales tipo ráfaga de las periódicas, a diferencia de las aperiódicas y estables que su grado de discriminación era similar y dependía del grado de suavizado de la muestra.

Con esto, la investigación establece un avance en el estudio en la identificación y clasificación de señales periódicas y no periódicas a través de la implementación de métodos automatizados utilizando un conjunto de herramientas propuestas por distintas publicaciones encontradas en el estudio realizado, las cuales permitieron cumplir con los objetivos específicos propuestos en la investigación.

**BIBLIOGRAFIA**

Barlett, M. (1955). *An Introduction to Stochastic Processes with Special Reference to Methods and Applications.* Cambridge: Cabridge University Press.

Chen, Z., Yang, H., Ma, J., Lei, J., & Gao, H. (2011). *Time-based Query Classification and its Application for Page Rank.* Shanghai: Binary Information Press.

Chen, Z., Yang, H., Ma, J., Lei, J., & Gao, H. (2011). *Time-based Query Classification and its Application for Page Rank.* China: Binary Information Press.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing interrater releability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, 127-137.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 37-46.

Cooley, J. W., & Tukey, J. W. (1965). An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation*, 297-301.

Fleiss, J. L. (1981). The Measurement of Interrater Agreement. En B. L. Joseph L. Fleiss, *Statistical methods for rates and proportions* (págs. 212-236.).

Friss-Cristensen, E., & Lassen, K. (1991). *Length of solar cycle- An indicator of solaractivity closely related with climate.* In Science, 254, pages 698-700.

*Google Trends*. (s.f.). Obtenido de https://www.google.com/trends/

Hamilton, J. D. (1954). *Time Series Analysis.* Princeton University Press.

Jones, R., & Diaz, F. (2007). Temporal profiels of queries. *ACM Transactions on Information Systems (TOIS)*.

Landis, R. J., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics 33*, 159-174.

Metzler, D., Jones, R., Peng, F., & Zhang, R. (2009). *Improving Search Relevance for Implicitly Temporal Queries.* Boston, Massachusetts, USA: SIGIR '09.

Molinero, L. M. (Enero de 2004). *Análisis de series temporales.* Obtenido de www.seh−lelha.org: www.seh−lelha.org/stat1.htm

Moral Carcelo, J. (2000). *Introduccion al tratamiento de series temporales mediante filtros.* Madrid.

Morón, J. (2011). *Señales y Sistemas.* Maracaibo,Venezuela: Universidad Rafael Urdaneta.

Parthasarathy, S., Mehta, S., & Srinivasan, S. (2006). Robust Periodicity Detection Algorithms. *CIKM '06 Proceedings of the 15th ACM international conference on Information and knowledge management*, 874-875.

Peralta, I. M. (s.f). *Universidad Carlos III de Madrid.* Obtenido de Material Docente Series Temporales: http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/SeriesTemporales/Tema2SeriesEstud.pdf

POLLOCK, D. (s.f). *STATISTICAL FOURIER ANALYSIS.*

Powers, D. M. (2001). Evaluation: From Precision, Recall and Fmeasure to RoC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 37-63.

Radinsky, K., Svore, K., Dumais, S., Teevan, J., Bocharov, A., & Horvitz, E. (2012). Modeling and Predicting Behavioral Dynamics on the Web. *WWW '12 Proceedings of the 21st international conference on World Wide Web*, 599-608.

Ríos, G. (14 de Noviembre de 2008). Series de Tiempo.

Rizzo, R. C., & Levich, R. M. (1998). *Alternative Tests for Time Series Dependence Based on Autocorrelation Coefficients.*

Shokouhi, M. (2011). Detecting Seasonal Queries by Time-Series Analysis. *SIGIR '11 Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 1171-1172.

Subasic, I., & Castillo, C. (s.f.). The Effects of Query Bursts on Web Search.

*Text Retrieval Context*. (s.f.). Obtenido de http://trec.nist.gov/

Tulena, J., Azzolini, M., de Vriesa, J., Groeneveld, W., Passchier, J., & van de Wetering, B. (1997). *Quantitative study of spontaneous eyes blinks and eye tics in Gilles de la Tourette's.* In Journal of Neurol. Neurosury. Psychiatry 1999,67:800-802.

Uebersax, J. S. (1987). Diversity of Decision-Making Models and the Measurement of Interrater Agreement. *Psychological Bulletin*, 140 - 146.

Vasco, U. d. (s.f). *Universidad del País Vasco - Procesado de Señales.* Obtenido de http://www.ehu.eus/Procesadodesenales/tema7/ty3.html

Villavicencio, J. (s.f). *Introduccion a Series de Tiempo.* Obtenido de http://www.estadisticas.gobierno.pr/iepr/LinkClick.aspx?fileticket=4\_BxecUaZmg%3D

Vlachos, M., Meek, C., Vagena, Z., & Gunopulos, D. (2004). Identifying Similarities, Periodicities and Bursts for Online Search Queries. *SIGMOD '04 Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, 131-142.

Vlachos, M., Yu, P., & Castelli, V. (2005). *On Periodicity and Structural Periodic Similarity.* New York: IBM T.J Watson Research Center.

Weisstein, E. W. (s.f). *MathWorld - A Wolfram Web Resource.* Obtenido de http://mathworld.wolfram.com/Covariance.html