UACM
Universidad Autónoma
de la Ciudad de México
Nada humano me es ajeno

COLEGIO DE CIENCIAS Y HUMANIDADES

MAESTRÍA EN CIENCIAS DE LA COMPLEJIDAD

Análisis de series de tiempo caóticas (Estudio de caso: la dinámica de las velocidades de la línea 5 del Metrobús de la Ciudad de México)

TESIS QUE PARA OBTENER EL GRADO DE

MAESTRA EN CIENCIAS DE LA COMPLEJIDAD

PRESENTA

ALEXANDRA GUZMÁN VELÁZQUEZ

Director de tesis
Dr. Juan Antonio Nido Valencia

Codirector
Dr. Carlos Islas Moreno

Ciudad de México, febrero de 2017.

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

La presente obra y cada uno de sus elementos está protegido por la Ley Federal del Derecho de Autor; por la Ley de la Universidad Autónoma de la Ciudad de México, así como lo dispuesto por el Estatuto General Orgánico de la Universidad Autónoma de la Ciudad de México; del mismo modo por lo establecido en el Acuerdo por el cual se aprueba la Norma mediante la que se Modifican, Adicionan y Derogan Diversas Disposiciones del Estatuto Orgánico de la Universidad de la Ciudad de México, aprobado por el Consejo de Gobierno el 29 de enero de 2002, con el objeto de definir las atribuciones de las diferentes unidades que forman la estructura de la Universidad Autónoma de la Ciudad de México como organismo público autónomo y lo establecido en el Reglamento de Titulación de la Universidad Autónoma de la Ciudad de México.

Por lo que el uso de su contenido, así como cada una de las partes que lo integran y que están bajo la tutela de la Ley Federal de Derecho de Autor, obliga a quien haga uso de la presente obra a considerar que solo lo realizará si es para fines educativos, académicos, de investigación o informativos y se compromete a citar esta fuente, así como a su autor ó autores. Por lo tanto, queda prohibida su reproducción total o parcial y cualquier uso diferente a los ya mencionados, los cuales serán reclamados por el titular de los derechos y sancionados conforme a la legislación aplicable.

You don't need to predict the future.
Just choose a future —a good future,
a useful future— and make the kind
of prediction that will alter human
emotions and reactions in such a way
that the future you predicted
will be brought about.
Better to make a good future
than predict a bad one.

ISAAC ASIMOV, Prelude to Foundation

Agradecimientos

Agradezco al área de sistemas del Metrobús de la Ciudad de México por la información proporcionada para la realización de este tesis.

Agradezco al SECITI y a la UACM por la beca de un año (del 16 de junio de 2015 al 15 de mayo de 2016) otorgada al proyecto de investigación con número PI2014-36 que tuvo como resultado de este trabajo.

También agradezco a la UACM por el apoyo en la impresión y empastado de este trabajo, por medio del convenio número UACM-CSE-ITR/09/2017.

Índice general

Resumen

En este trabajo de tesis se analizaron los datos de velocidad promedio —por hora— de la línea 5 del Metrobús de la Ciudad de México, de mayo de 2014. Los datos en bruto consistían en las mediciones de los tiempos de entrada y salida por estación, de ida y vuelta, por día, de todas las unidades. La detección de errores y limpieza de estos datos se realizó utililizando el lenguaje de programación *Python* y la librería *openpyxl*, que ayuda al manejo de lectura y escritura de documentos en *Excel*. Para el estudio de la serie de tiempo obtenida se realizó un breve estudio estadístico y se examinaron las características de la serie. Además se emplearon las técnicas de análisis caótico descritas por el Teorema de Takens. Dentro de este mismo trabajo, se explicó detalladamente en que consisten estos métodos y con que librerías del lenguaje de programación *R* se podían obtener, con el objetivo central de encontrar un intervalo de predicción correcto y de afirmar si este tipo de datos es caótico o estocástico.

Los resultados obtenidos del análisis de las series de tiempo son los siguientes: la media de la serie parece ser estable, con un promedio de 38.94 km/h, y es estacionaria. Para la reconstrucción del espacio fase, el tiempo de retardo τ es de 2 horas y la dimensión de inmersión o encajamiento es de $m = 8$ unidades. La dimensión de la correlación, que cuantifica la dimensión fractal de un objeto geométrico encajado en un espacio fase, es de $D_2 = 2.070023$, un indicador de la conducta caótica de "dimensión menor" de las series de tiempo puesto que éste es un número finito, no entero y mucho menor que la dimensión de encajamiento. El exponente máximo de Lyapunov resulta de $\lambda = 0.03198189$, este valor positivo y finito implica que las trayectorias divergen exponencialmente y que hay un comportamiento caótico. Y aunque sabemos que los sistemas caóticos son infinitamente complejos y extremadamente sensibles a cambios en las condiciones iniciales,

la teoría del caos explota rigurosamente la información contenida en los datos y puede usarse para hacer predicciones, aunque sea en periodos cortos. En este caso, se hicieron predicciones de 21 horas (de un día para el Metrobús), basados en datos de 630 horas. De acuerdo al exponente de Lyapunov máximo y a la gráfica obtenida, el periodo máximo u horizonte de predictibilidad es de 31.2677 horas o día y medio de velocidades promedio.

Casdagli [3], Christophersen [4] y a Shang [15], comentan que con el Teorema Takens se busca el espacio de inmersión o encajamiento en donde se pueda reconstruir el atractor a partir de los datos escalares que preserven las características invariantes del atractor desconocido original. Dadas ciertas condiciones ($m \geq 2d + 1$, donde d la dimensión del atractor desconocido) el Teorema garantiza que el atractor encajado en el espacio de inmersión es "desplegado" sin intersecciones consigo mismo. De hecho, dado un número finito de datos con ruido, la estimaciones de las medidas invariantes y los exponentes de Lyapunov dependen específicamente de m y de τ. Así, dados los valores de ambos elementos, el espacio fase es $Y_t = \{x_t, x_{t-2}, \ldots, x_{t-14}\}$. El atractor encontrado es del tipo caótico, esto debido a la dimensión fractal y al exponente máximo de Lyapunov encontrados y gracias a esto es posible dar un tiempo de predicción u horizonte de predictibilidad, aunque sea limitado. Con la presencia de un atractor estocástico hubiera sido imposible hacer una predicción.

Dado el estudio minucioso de la serie de tiempo con las herramientas matemáticas y computacionales descritas en este trabajo, se sugiere al Metrobús —específicamente a la línea 5— como un sistema de transporte alternativo para la Zona Metropolitana del Valle de México. En las conclusiones de este trabajo se especifican los problemas que se encontraron en el funcionamiento de este transporte y las sugerencias que se proponen para el mejoramiento de éste.

Introducción

La Ciudad de México (CDMX) cuenta con una población de 8.9 millones de habitantes aproximadamente, pero debido a las pocas oportunidades de empleo que tienen más de cinco millones 850 mil personas del Estado de México [33], éstos se trasladan a la ciudad para poder trabajar. La Zona Metropolitana del Valle de México se considera como una de las mayores aglomeraciones urbanas del mundo y la más grande del continente americano y del mundo hispanoparlante [40]. Por lo cual, el tránsito diario es tan difícil que los mexiquenses tardan cuatro horas en el transporte público; mientras que los citadinos pueden tardar poco más de dos horas [32]. Así, la movilidad es un problema fundamental que el gobierno de la CDMX ha ido enfrentando, desde hace varios años, con proyectos sustentados en programas internacionales. Ejemplo de esto es el transporte denominado *BRTs* (sistema integrado de transporte), propuesto en Curitibia, Brasil, en 1972, por el arquitecto y urbanista Jaime Lerner. La idea de Lerner fue crear un sistema nuevo de transporte: un metro sobre suelo, con carriles confinados para autobuses, que tuviera paradas fijas y con un cobro estándar para todo el camino [23]. La misma concepción de *BRTs* fue captada por el Transmilenio en Colombia en el año 2000, pero el confinamiento de éste se hizo en un sólo carril.

El 9 de marzo del 2005, el Jefe de gobierno del antes llamado Distrito Federal, Andrés Manuel López Obrador, creó el Organismo Público Desentralizado Metrobús. Con la asesoría del Centro de Transporte Sustentable de la ciudad, se diseñaron seis rutas de transporte con carriles confinados, con estaciones cada 400 metros aproximadamente, tarjetas de prepago y reglas específicas sobre la circulación de las unidades en los carriles. Decidiendo construir el primer corredor en Av. de los Insurgentes por ser una de las avenidas más largas de la ciudad y una de las que usa un mayor número de personas.

Más adelante, el Metrobús se amplió sobre la misma línea y se han abierto cinco líneas más. Sin embargo, su eficiencia y productividad se ha puesto en duda en todos estos años puesto que debido a las marchas, bloqueos o al tránsito intenso, la velocidad de las unidades es mucho menor en horas y/o días específicos. Para lograr recuperar cierta velocidad promedio durante el día, los conductores de las unidades manejan a velocidades que oscilan entre los 120 y 140 km/h en periodos de tiempo donde la ciudad no está tan congestionada, lo cual pone en riesgo tanto a los usuarios y transeúntes, como a los conductores, generando un problema que también se tiene que regular. Como se menciona en [34], el gerente de Transmilenio Colombia explicó algunas propuestas para que el sistema *BRTs* no se sature: ubicar las zonas con mayor demanda, controlar el número de servicios en hora pico y tener dos carriles, uno de los cuales sea de máxima velocidad y otro convencional, con lo cual también se evitarían accidentes.

El licenciado en Economía Luis Antonio Moreno, en [23], realizó un estudio estadístico sobre la velocidad promedio de todas las líneas en el periodo de enero del 2008 a marzo del 2011 y se percató que la velocidad fue de 15.73 km/h; mientras que la cifra oficial fue de 19.66 km/h. Él expuso que el Metrobús no es necesariamente un transporte rápido y que el resultado de esto es la configuración de las estaciones (que se encuentren en carriles centrales y que las unidades dependan de la coordinación de los semáforos) y recomendó que la distancia entre estaciones debe ser mayor. Incluso, Moreno sugirió que el funcionamiento del Trolebús es mejor puesto que obtuvo una velocidad promedio de 30 km/h, a un menor costo. Así él concluye que el Metrobús no es un servicio cómodo, seguro ni accesible. Sin embargo, menciona que puede ser un transporte alternativo para un volumen de pasajeros menor a los 15,000.

En esta tesis se utilizan los componentes y características de las series de tiempo, la estadística básica y el análisis de las series por medio de la teoría del Teorema inmersión de Takens para explicar el funcionamiento —en términos de la velocidad— del Metrobús, con la finalidad de describir su comportamiento e investigar el mecanismo generador de las series encontradas y buscar posibles modelos temporales que permitan predecir lo que ocurrirá en un periodo de tiempo corto. Para ello se emplean las mediciones de tiempos de entrada y salida por estación de la línea 5, de ida y vuelta, del Metrobús de la Ciudad de México, durante el mes de mayo

del 2014. La importancia de estos datos es debida a que la línea 5, que comenzó sus funciones el 5 de noviembre del 2013, satisface una demanda de 70 mil pasajeros por día, se mueve por Eje 3 Oriente (Ingeniero Eduardo Molina) de San Lázaro a Río de los Remedios, cubriendo una longitud total de 10 km de recorrido con 18 estaciones, según la información detallada en [36] y [37]. Lo cual nos permite conocer la velocidad de las unidades en una avenida conflictiva e identificar las problemáticas en una línea que está muy bien monitoreada las 24 horas del día, los 365 días del año, por el Centro Informativo de Transporte Inteligente (CITI), instancia que controla en tiempo real a los autobuses y a las estaciones de esta línea desde sus inicios. Además, la línea conecta el centro de la ciudad (San Lázaro) con la frontera entre la CDMX y el Estado de México, pues la otra terminal se encuentra entre la delegación Gustavo A. Madero y el municipio de Ecatepec (Río de los Remedios). De acuerdo a [38] la obra de la línea 5 implementó en esta misma vialidad distintas formas de movilidad urbana: banquetas grandes para el peatón, carriles de bicicleta y carriles para tránsito mixto en ambos sentidos, y dos carriles exclusivos para el Metrobús; adecuando espacios para convertirlos en un corredor de parques, juegos infantiles y gimnasios al aire libre. Destacando también la introducción de paneles solares, muros verdes y áreas verdes en sus instalaciones, así como autobuses de baja emisión de gases contaminantes, siendo la primer línea de sistemas *BRTs* en reducir el daño ambiental.

A continuación se explica brevemente lo que se realizó por capítulo.

En el Capítulo 1 se define lo que es una serie de tiempo, se dan ejemplos sobre las series en distintas áreas de investigación, describiendo los componentes y las características de cada una de éstas después de explicar sus definiciones. Una de las propiedades de las series de tiempo es la estacionariedad, de la cual se expone la prueba que la describe, la de Ljung-Box, para posteriormente detallar los modelos lineales estacionarios ($AR(p)$, $MA(q)$ y $ARMA(p,q)$) y no estacionarios ($ARIMA(p,d,q)$), que tienen como objetivo predecir una serie de tiempo estocástica.

En el Capítulo 2, se comienza exponiendo la versión original del Teorema de Takens que explica la reconstrucción del espacio fase, así como su demostración formal. La utilización del teorema permite obtener una versión topológica equivalente del espacio de fases a partir del comportamiento de

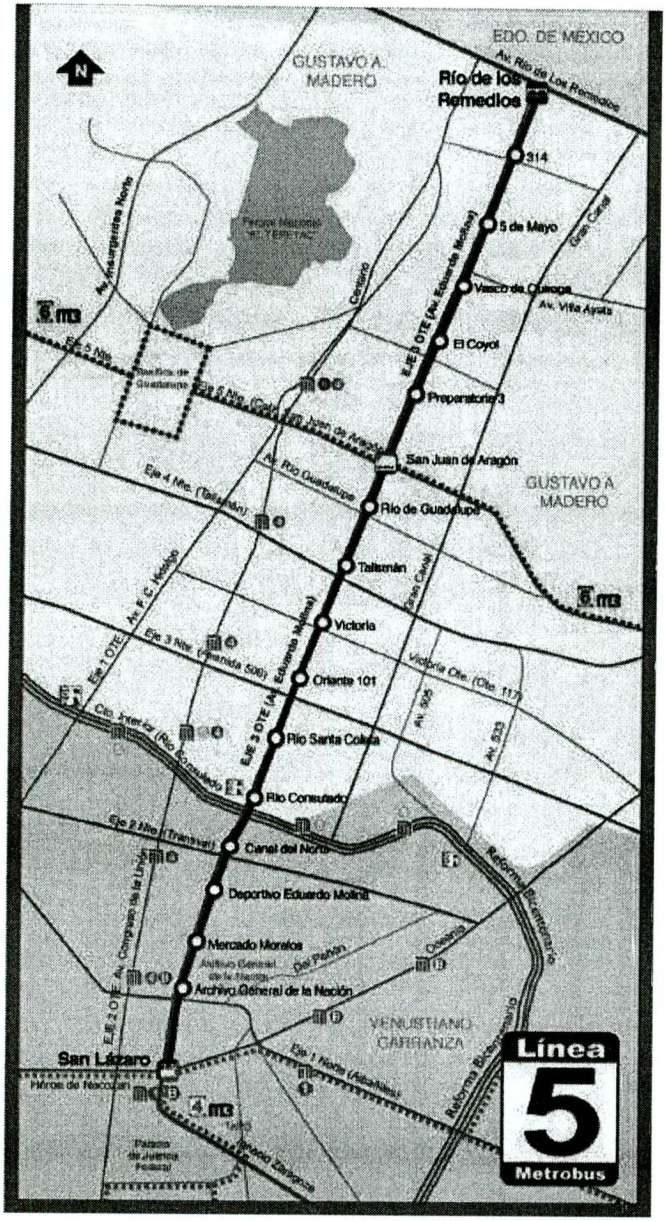


Figura 1: Mapa de la línea 5 del Metrobús de la Ciudad de México. Tomada de [36].

una sola variable del sistema, lo cual resulta de gran utilidad para el estudio de escenario dinámicos caóticos. Se explican además como se calculan las técnicas de análisis de las series temporales caóticas: desde el tiempo de retraso y la dimensión de inmersión necesarias para construir el espacio fase mediante los datos u observaciones de una serie, como la dimensión de correlación, el exponente de Lyapunov máximo y el tiempo de predicción u horizonte de predictibilidad. Finalmente se detalla el modelo innovador de Abbas Golestani y Robin Gras [8] que predice una serie de tiempo caótica calculando la dimensión fractal.

En el Capítulo 3 se explica como se obtiene, utilizando de un programa que se desarrolla en el lenguaje *Python* y con el módulo *openpyxl* (Véase Apéndice A.1), la serie temporal a partir de los datos proporcionados por el área de sistemas del Metrobús de la Ciudad de México.
Las series de las velocidades, correspondientes a las mediciones, se encuentran por día y de acuerdo a éstas se examinan los valores para identificar posibles errores técnicos de detección de las velocidades o del paso de información a las hojas de excel que reunen la información descrita. Los datos se acomodan y calculan en Excel con *Python* y una vez hecho esto, se calculan los valores promedio para encontrar la dinámica subyacente durante todo el mes y aplicar así el Teorema de Takens. La hipótesis inicial del comportamiento de la serie de las velocidades promedio del Metrobús es que ésta es caótica, siguiendo las ideas del artículo de *Chaotic analysis of traffic time series*, [15], el resultado se explica en las conclusiones.
Las técnicas de series caóticas se implementan con paqueterías específicas de *R*, como se pueden leer en el Apéndice A.2. Sin embargo, el Teorema de inmersión y sus algoritmos no puede predecir datos de una serie, sólo puede definir el periodo en el cual pueden existir. Por ello, se explican algunos métodos que pueden encontrar esas cifras en los capítulos anteriores y se presentan los resultados gráficos de éstos.

Capítulo 1

Series de tiempo

El Dr. Alejandro Nava, en [13], explica que la ciencia tiene bases tanto teóricas como experimentales y que las observaciones recabadas de un fenómeno pueden ser útiles en ambas instancias. Las mediciones que se recopilan, observan o registran en intervalos de tiempo regulares (diario, semanal, semestral, anual, etc.) pueden construir una serie de tiempo. Uno de los objetivos del estudio de series de tiempo es el pronóstico de éstas, lo cual significa que se quieren conocer los valores futuros donde no existen mediciones disponibles. Este pronóstico se realiza por ejemplo para optimizar un área específica de una empresa o para encontrar patrones históricos que ayuden en la toma de decisiones, etc. Para realizar un buen pronóstico es importante saber en que periodo de tiempo se está hablando y que tantos periodos (horizonte de predictibilidad) necesitan ser pronosticados. Además se necesita también saber si la serie es determinista, estocástica o caótica.

Definición 1.0.1 *Una serie de tiempo es una secuencia de observaciones de fenómenos físicos, biológicos o químicos, entre otros, registradas en determinados momentos del tiempo, ordenadas cronológicamente y, espaciadas entre sí de manera uniforme.*

Para analizar una serie de tiempo, primero se plantea su gráfica, ya que esto permite realizar una inspección visual identificando la tendencia, así como la estacionariedad y las variaciones irregulares. Después se pueden analizar estadísticamente los datos, para tratar de construir un modelo que explique la estructura y la evolución de una variable a lo largo del tiempo.
Finalmente para intentar hacer una estimación y cálculo de los índices

o propiedades métricas y la aplicación de procedimientos gráficos que permitan determinar si una serie de tiempo aparentemente aleatoria puede corresponder a un comportamiento caótico, es necesario recurrir al Análisis de Sistema Dinámico no Lineal determinista. Esta herramienta permite estudiar el sistema que originó la serie de tiempo y su posible evolución en el tiempo. Por lo que, aunque sabiendo que los sistemas caóticos son infinitamente complejos y extremadamente sensibles a cambios en las condiciones iniciales, la Teoría del Caos explota rigurosamente la información contenida en los datos y puede usarse para hacer predicciones en períodos de tiempo muy cortos. Esto se explicará en el Capítulo 2 y se realizará en el Capítulo 3 de esta tesis. Para la redacción de este capítulo, se siguen las referencias [1], [13], [16] [21] y [39].

1.1 Componente de una serie de tiempo

De acuerdo al análisis clásico de las series temporales, los valores que toma la variable de observación es consecuencia de tres componentes, cuya unión da como resultado los valores medidos, estos componentes son:

(a) Componente tendencia. Se define como un cambio a largo plazo que se produce en relación al nivel medio. La tendencia se identifica con un movimiento suave de la serie a largo plazo.

(b) Componente estacional. Algunas series temporales presentan cierta periodicidad o variación de cierto período (semestral, mensual, etc.). Estos efectos son fáciles de ver y se pueden medir explícitamente o incluso se pueden eliminar de la serie de datos, a este proceso se le llama desestacionalización de la serie.

(c) Componente aleatoria. Esta componente no responde a ningún patrón de comportamiento, sino que es el resultado de factores aleatorios que inciden de forma aislada en una serie de tiempo.

De estos tres componentes, los dos primeros son deterministas, mientras que el último es aleatorio. Así se puede denotar la serie de tiempo como $x_t = T_t + E_t + I_t$, donde T_t es la tendencia, E_t es la componente estacional e I_t es la componente aleatoria.

1.2 Características de las series de tiempo

1.2.1 Medidas de dependencia

Una descripción completa de series de tiempo, observadas como colección de n variables aleatorias y un número arbitrario de puntos $t_1, t_2, \ldots, t_n$, para cualquier número entero positivo n, está dado por la función de distribución conjunta, evaluada como la probabilidad de que los valores de las series estén $F(c_1, c_2, \ldots, c_n)$

Definición 1.2.1 *La función media es definida como*

$$\mu_{x_t} = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx, \tag{1.1}$$

donde E denota el valor esperado del operador.

Definición 1.2.2 *Un modelo de series de tiempo de los datos observados x_t es una especificación de las distribuciones conjuntas, o posiblemente sólo de la media y la covarianza, de una secuencia de variables aleatoria X_t, de las cuales x_t es una postulación de ser una realización*

Definición 1.2.3 *La función de autocovarianza está definida como el producto del segundo momento*

$$\gamma_s(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \tag{1.2}$$

para toda s y t.

La autocovarianza mide la dependencia lineal entre dos puntos de las mismas series observadas a diferentes tiempos. Las series suaves exhiben funciones de autocovarianza que permanecen largas, aún cuando los enteros t y s se encuentren lejos; mientras que las series variables tienden a tener funciones de autocovarianza cercanas a cero para separaciones largas.

Definición 1.2.4 *La función de autocorrelación ACF está definida como:*

$$\rho_j = corr(x_j, x_{j-k}) = \frac{cov(x_j, x_{j-k})}{\sqrt{V(x_j)}\sqrt{V(x_{j-k})}} \tag{1.3}$$

donde $V(x_j) = \sigma^2$ es la varianza.

La *ACF* mide la predictibilidad lineal de las series, al tiempo t. La función ρ_j tiene las siguientes propiedades:

- $\rho_0 = 1$
- $-1 \leq \rho_j \leq 1$
- Simetría $\rho_j = \rho_{-j}$

Definición 1.2.5 *La función de autocorrelación parcial PACF mide la correlación entre dos variables separadas por k periodos cuando no se considera la dependencia creada por los retardos intermedios existentes entre ambas.*

$$\begin{aligned} \pi_j &= corr(x_j, x_{j-k}/x_{j-1}x_{j-2}\ldots x_{j-k+1}) \\ &= \frac{cov(x_j - \tilde{x_j}, x_{j-k} - \tilde{x_{j-k}})}{\sqrt{V(x_j - \tilde{x_j})}\sqrt{(x_{j-k} - \tilde{x_{j-k}})}} \end{aligned} \tag{1.4}$$

1.2.2 Estacionariedad

Definición 1.2.6 *Una serie de tiempo es estrictamente estacionaria si la conducta probabilística de cada colección de valores, $\{x_{t_1}, x_{t_2}, \ldots, x_{t_k}\}$, es idéntica a aquélla con traslación en el tiempo, $\{x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}\}$. Esto es,*

$$P\{x_{t_1} \leq c_1, \ldots, x_{t_k} \leq c_k\} = P\{x_{t_1+h} \leq c_1, \ldots, x_{t_k+h} \leq c_k\},$$

para todo $k = 1, 2, \ldots$, $t_1, t_2, \ldots, t_k$, $c_1, c_2, \ldots, c_k$ y $h = 0, \pm 1, \pm 2, \ldots$.

Esto es, una serie de tiempo X_t, con $t = 1, \ldots$ se dice estacionaria si ésta tiene las propiedades estadísticas similares a aquélla serie en la cual cambia el tiempo, de la forma X_{t+h}, $t = 1, \ldots$, para cada entero h. Se restringe la atención a aquéllas propiedades que dependen sólo de los primer y segundo momento de X_t.

Definición 1.2.7 *Una serie de tiempo es débilmente estacionaria, x_t, es un proceso de varianza finito tal que:*

1. *La función de valor medio, μ_t, definida en (1.1) es constante y no depende del tiempo t y,*

2. *La función de autocovarianza, $\gamma(s,t)$, definida en (1.2), depende de s y t sólo por medio de su diferencia $|s-t|$.*

Por tanto, el término estacionario significa débilmente estacionario.

1.2.3 Prueba de Ljung-Box

Este test prueba en forma conjunta que todos los coeficientes de autocorrelación son simultáneamente iguales a cero, es decir, que son independientes. Está definida como:

$$LB = n(n+2)\sum_{k=1}^{m}\left(\frac{\check{\rho}_k^2}{n-k}\right) \approx \chi^2(m), \tag{1.5}$$

donde n es el tamaño de la muestra y m la longitud del rezago.

1.3 Ejemplos de series de tiempo

Las series de tiempo aquí descritas pueden ser estocásticas o caóticas, el objetivo de presentarlas es ver como a partir de su gráfica se pueden observar detalles importantes sobre el problema que están narrando, destacando los componentes y sus características esenciales que presentamos anteriormente. La figura 1.1 muestra las ganancias trimestrales por acción de la compañía estadounidense Johnson y Johnson, datos calculados por el profesor Paul Griffin de la escuela de Administración de la Universidad de California. En la gráfica se pueden observar las mediciones de 21 años y 84 trimestes, de 1960 a 1980. Para hacer el análisis de la serie de tiempo, se empieza por observar los patrones que ocurren cuando transcurre el tiempo. En este caso, se nota un incremento gradual que fija una tendencia y una variación periódica superpuesta sobre la tendencia que parece repetirse por trimestre.

En el siguiente ejemplo, figura 1.2, se considera un registro de la temperatura media global. Los datos de la serie de tiempo son desviaciones promedio, medidas en grados centígrados, de la temperatura de la Tierra tomada de 1880 al 2009. En esta serie de tiempo se nota una tendencia al alza durante la última parte del siglo XX que ha sido utilizada como argumento para la hipótesis del calentamiento global. Se observa también una nivelación en el año 1935 y luego otra vez la tendencia se incrementa muy levemente

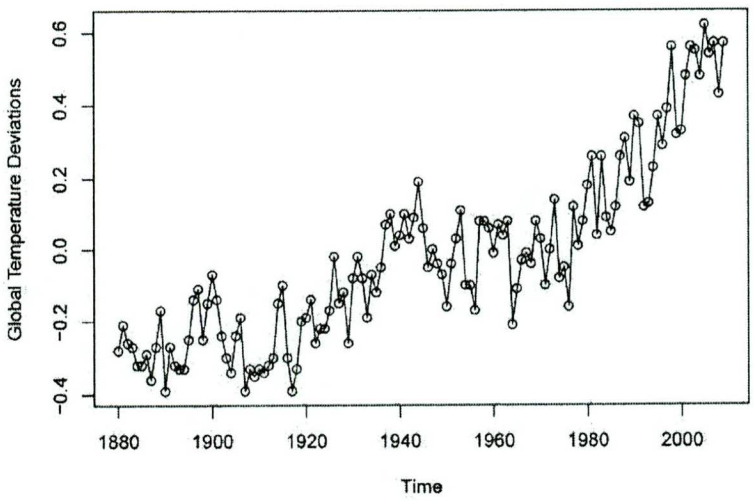


Figura 1.2: Gráfica de las desviaciones promedio anuales de la temperatura global, de 1880 a 2009, en grados centígrados.

La serie de tiempo de la figura 1.3 presenta una muestra de 0.1 s (1000 puntos) de la voz grabada de la frase $aaa \ldots hhh$, se puede notar la naturaleza repetitiva de la señal y las periodicidades regulares. El problema de reconocimiento de voz por medio de la computadora es por demás reelevante, puesto que se requiere convertir una señal particular en la frase grabada $aaa \ldots hhh$. El análisis espectral puede ser usado en este contexto para producir una firma de esta frase, que puede ser comparada con otras firmas de varias bibliotecas de sílabas para buscar una coincidencia. Se puede notar de forma inmediata que ocurre una repetición regular de ondas pequeñas. La separación entre los intervalos se conoce como el período de tono y representa la respuestra del filtro del tracto vocal a una secuencia periódica de impulsos estimulados por la apertura y el cierre de la glotis.

En ejemplo 4 es una serie de tiempo financiera, la gráfica de la figura 1.4 muestra el cambio porcentual de la Bolsa de Valores de Nueva York, de febrero de 1984 al 31 de diciembre de 1991. En esta gráfica se puede detectar fácilmente la caída de la bolsa que ocurrió el día 19 de octubre de 1987. La media de la serie parece ser estable, con un promedio de aproximadamente cero; sin embargo, la volatilidad de datos cambia a lo largo del tiempo. De

hecho, los datos muestran un agrupamiento de la volatilidad, es decir, tiene períodos de variabilidad altos que tienden a agruparse juntos. Un problema en el análisis de este tipo de datos financieros es pronosticar la inestabilidad de los rendimientos futuros.

Las series de tiempo de la figura 1.5 presentan los valores mensuales de dos medidas ambientales: el índice de oscilación del sur (SOI) y el reclutamiento de peces nuevos, datos obtenidos por el Dr. Roy Mendelssohn del grupo de pesca y medio ambiente del Pacífico. Ambas series ocurren en un período de 453 meses, que varía entre 1950 y 1987. Las medidas de los cambios en la presión del aire SOI están relacionadas con las temperaturas de la superficie del mar en el oceano Pacífico central. El centro del Pacífico se calienta de tres a siete años debido al efecto del niño, responsable de las inundaciones en vairas partes del medio oeste de los Estados Unidos, en 1997. Las dos series de la figura 1.5 tienden a exhibir un comportamiento repetitivo, con la presencia de ciclos regulares que son fácilmente visibles. También se puede observar que los ciclos de la primer serie se repiten a un ritmo más rápido que los de la serie de reclutamiento. La segunda serie muestra también varios tipos de oscilaciones a una frecuencia más rápida que parece repetirse aproximadamente cada 12 meses y una frecuencia más lenta que parece se repite cada 50 meses. Los dos series estan relacionadas entre sí, pensando que la población de peces dependa netamente de la SOI. Tal vez incluso tengan una relación temporal. con las modificaciones de la señalización SOI en la población de peces.

Un problema fundamental en la estadística clásica se produce cuando se da una colección de series o vectores de series independientes, generada bajo diferentes condiciones experimentales o diversas configuraciones tratamiento. Esta serie de series se muestra en la Figura 1.6, donde se observan los datos recogidos de varios lugares en el cerebro a través de imágenes de resonancia magnética funcional (fMRI). En este ejemplo, a cinco sujetos se les dio un cepillado periódico en la mano. El estímulo se aplicó durante 32 segundos y luego se detuvo durante 32 segundos. Por lo tanto, el período de la señal es de 64 segundos. La velocidad de muestreo fue una observación cada 2 segundos para 256 segundos ($n = 128$). La serie muestra medidas consecutivas del nivel de sangre oxigenada dependiente (trazo negro) de la intensidad de la señal, que mide zonas de activación en el cerebro. Las periodicidades aparecen en la serie de la corteza motora y con menor intensidad en el tálamo y el cerebelo.

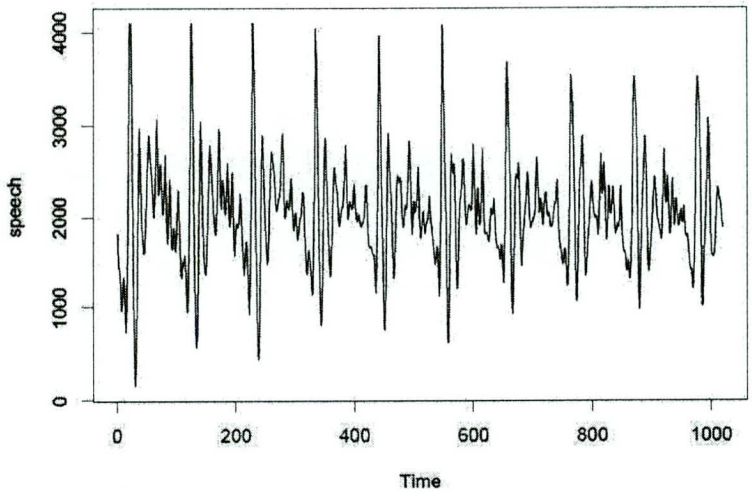


Figura 1.3: Gráfica de la grabación de las sílabas *aaa* . . . *hhh* muestreada con 10,000 puntos por segundo, con n = 1020 puntos.

El hecho de que una tiene las series de diferentes áreas del cerebro sugiere comprobar que áreas están respondiendo diferente al estímulo del cepillo.

Como último ejemplo, las series en la Figura 1.7 representan dos fases o llegadas a lo largo de la superficie, que se denotan por P $(t = 1, \ldots, 1024)$ y S $(t = 1,025, \ldots, 2048)$, a una estación de registro sísmico. El objetivo del equipo de registro en Escandinavia es la observación de terremotos y explosiones mineras. El problema general se centra en distinguir o discriminar entre las formas de onda generadas por los terremotos y las generadas por las explosiones. Las características importantes son las relaciones de amplitud en bruto de la primera fase P a la segunda fase S, que tienden a ser más pequeños para los terremotos que para explosiones. En el caso de los dos eventos en la figura 1.7, la relación de amplitudes máximas parece ser algo menor de 0.5 para el terremoto y aproximadamente 1 por la explosión. De lo contrario, se tiene una sutil diferencia en la naturaleza periódica de la fase S para el terremoto. Se puede usar el análisis espectral de la varianza para probar la igualdad de la componentes periódicos de los terremotos y las explosions. También se pueden clasificar los componentes futuros de P y S de los eventos de origen desconocido.

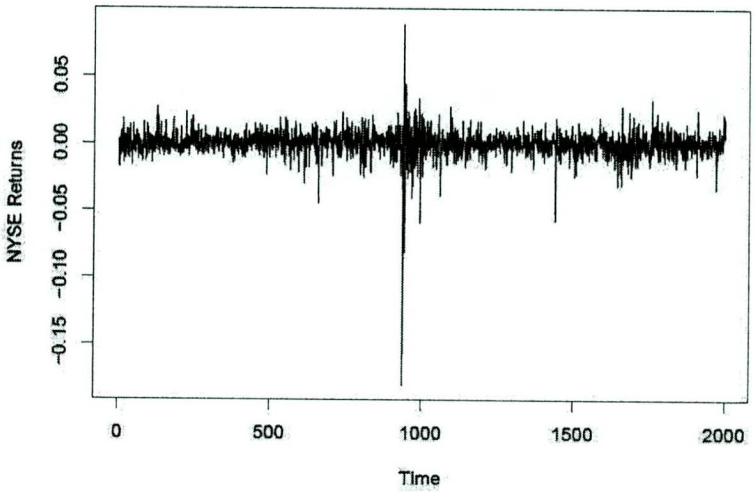


Figura 1.4: Gráfica que muestra los rendimientos del mercado de valores ponderados diarios de la bolsa de Nueva York, del 2 de febrero de 1984 al 31 de diciembre de 1991 (durante 2000 días de negociación). La crisis económica del 19 de octubre de 1987 se produce justo en $t = 938$.

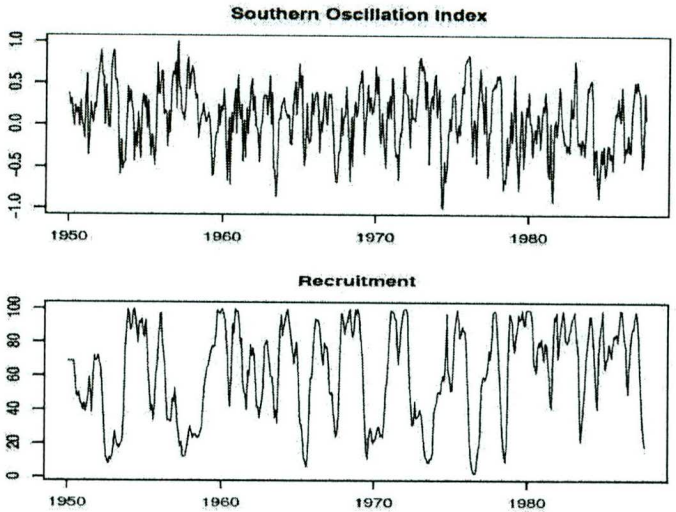


Figura 1.5: Gráfica que muestra el índice de oscilación del sur (SOI) mensual y el reclutamiento estimado de peces nuevos, de 1950 a 1987.

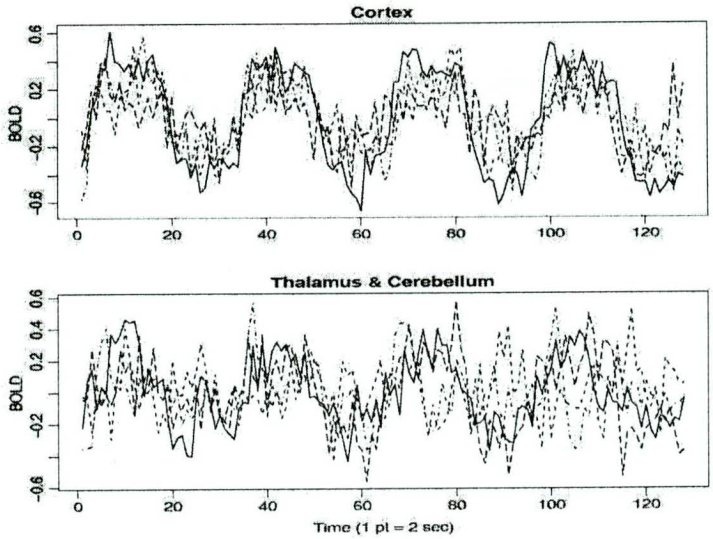


Figura 1.6: Gráfica de las imágenes de resonancia magnética funcional (fMRI) desde varios lugares de la corteza, el tálamo y el cerebelo; n = 128 puntos, una observación tomada cada 2 segundos.

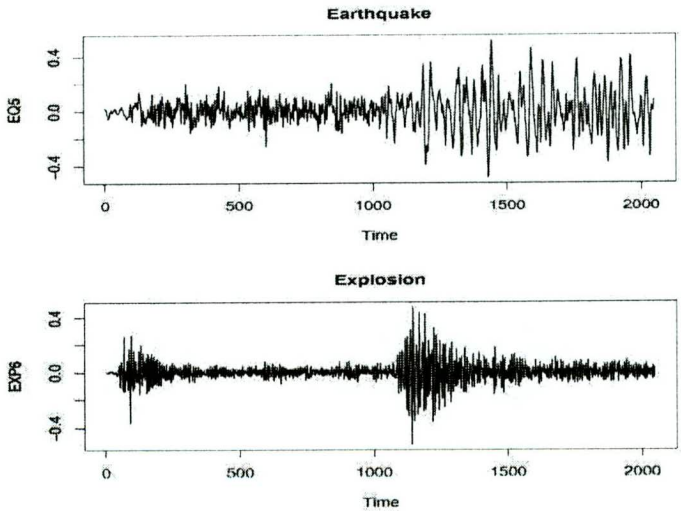


Figura 1.7: Gráfica de las fases de llegada de un terremoto (arriba) y de una explosión (parte inferior) a 40 puntos por segundo.

1.4 Predicción series de tiempo estacionaria

Definición 1.4.1 *Un proceso estocástico es una sucesión de variables aleatorias* $\{x_t\}$, $t = -\infty, \ldots, -2, -1, 0, 1, 2, \ldots, \infty$.

Definición 1.4.2 *Se llama ruido blanco a una sucesión de variables aleatorias con esperanza cero, igual varianza e idependientes en el tiempo. En lo sucesivo, se denota un ruido blanco por* ϵ_t.

1.4.1 Procesos lineales estacionarios.

Procesos autoregresivos $AR(p)$

En estos modelos el valor actual de la serie x_t se explica en función de p valores pasados $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, donde p determina el número de retrasos necesarios para pronosticar un valor actual.

Definición 1.4.3 *El modelo autoregresivo de orden p está dado por:*

$$x_t = \Phi_0 + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \ldots + \Phi_p x_{t-p} + \epsilon_t \tag{1.6}$$

En términos del operador de retardos,

$$(1 - \Phi_1 L - \Phi_2 L^2 - \ldots - \Phi_P L^P) x_t = \epsilon_t$$

donde ϵ_t es un proceso de ruido blanco y las $\Phi' s$ son los parámetros del modelo.

Las condiciones de estacionariedad para los modelos AR de orden 1 y 2 son los siguientes:

- En el proceso AR de órden 1, la variable x_t está determinada por el valor pasado x_{t-1}: $x_t = \Phi x_{t-1} + \epsilon_t$, donde ϵ_t es un proceso de ruido blanco con media cero y varianza constante, Φ es el parámetro.

  Entonces $(1 - \Phi L) x_t = \epsilon_t$, donde el polinomio autoregresivo es $\Phi_1(L) = 1 - \Phi L$, $L = 1/\Phi$.

  La condición de estacionariedad del modelo es: $|L| = |1/\Phi| > 0$, entonces $|\Phi| < 1$

- En el proceso AR de órden 2, la variable x_t está determinada por el valor pasado y el anterior: $x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \epsilon_t$ y ϵ_t es un ruido blanco.

  Entonces $(1 - \Phi_1 L - \Phi_2 L^2)x_t = \epsilon_t$, donde el polinomio autoregresivo es $\Phi_2(L) = 1 - \Phi_1 L - \Phi_2 L^2$, con raíces $L_1, L_2 = \dfrac{\Phi_1 \pm \sqrt{\Phi_1^2 + 4\Phi_2}}{-2\Phi_2}$.

  La condición de estacionariedad del modelo es:

  $|L_1| = \left|\dfrac{\Phi_1 + \sqrt{\Phi_1^2 + 4\Phi_2}}{-2\Phi_2}\right| > 1$ y $|L_2| = \left|\dfrac{\Phi_1 - \sqrt{\Phi_1^2 + 4\Phi_2}}{-2\Phi_2}\right| > 1$, si $(\Phi - 1^2 + 4\Phi_2) > 0$, la raíces son reales y si $(\Phi - 1^2 + 4\Phi_2) < 0$, las raíces son complejas

Proceso de Medias móviles $MA(q)$

Este modelo supone linealidad, el valor actual de la serie, x_t está influenciado por valores de una fuente externa.

Definición 1.4.4 *El modelo de orden q está dado por:*

$$x_t = \Theta_0 - \Theta_1 \epsilon_{t-1} - \Theta_2 \epsilon_{t-2} - \ldots - \Theta_q \epsilon_{t-q} - \epsilon_t \tag{1.7}$$

Expresado en términos del polinomio operador de retardos se tiene:

$$x_t = (1 - \Theta_1 L - \Theta_2 L^2 - \ldots - \Theta_q L^q)\epsilon_t$$
$$x_t = \Theta_q(L)\epsilon_t,$$

donde ϵ_t es un proceso de ruido blanco y μ, $\Theta_1, \Theta_2, \ldots \Theta_q$ son los parámetros del modelo.

Las condiciones de invertibilidad para los modelos MA de orden 1 y 2 son los siguientes:

- El proceso de media móvil de orden 1, $MA(1)$, determina el valor de x_t en función de la innovación actual y su primer retardo: $x_t = \epsilon_t - \Theta\epsilon_{t-1}$, entonces $x_t = (1 - \Theta L)\epsilon_t$

  El polinomio está dado por $\Theta_1(L) = 1 - \Theta L$, con raíz $L = 1/\Theta$

La condición de invertibilidad está dado por $|L| = |1/\Theta| > 1$, esto es, $|\Theta| < 1$

- El modelo de medias móviles de orden 2, $MA(2)$, es: $x_t = \epsilon_t - \Theta_1\epsilon_{t-1} - \Theta_2\epsilon_{t-2}$, donde Θ_1 y Θ_2 son parámetros y ϵ_t es un proceso de ruido blanco.

  Si $x_t = (1 - \Theta_1 L - \Theta_2 L^2)\epsilon_t$, el polinomio de medias móviles esta dado por $\Theta_2(L) = 1 - \Theta_1 L - \Theta_2 L^2$, con raíces $L_1, L_2 = \dfrac{\Theta_1 \pm \sqrt{\Theta_1^2 + 4\Theta_2}}{-2\Theta_2}$

  Las condiciones de inertivilidad están dados por:

$$|L_1| = \left|\frac{\Theta_1 + \sqrt{\Theta_1^2 + 4\Theta_2}}{-2\Theta_2}\right| > 1 \text{ y } |L_2| = \left|\frac{\Theta_1 - \sqrt{\Theta_1^2 + 4\Theta_2}}{-2\Theta_2}\right| > 1$$

Los procesos de medias móviles se denominan procesos de memoria corta, mientras que los autoregresivos son procesos de memoria larga.
Para modelos que tienen representaciones con componentes autoregresivas y de medias móviles, veremos los modelos $ARMA(p, q)$, donde p y q denotan, respectivamente, los órdenes de los componentes autoregresivo y de medias móviles.

Proceso Autoregresivo de medias móviles $ARMA(p, q)$

Definición 1.4.5 *Si x_t sigue un proceso $ARMA(p, q)$, en este habrá p términos autoregresivos y q términos de media móvil.*

$$x_t = c + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \ldots + \Phi_p x_{t-p} + \Theta_1\epsilon_{t-1} + \Theta_2\epsilon_{t-2} + \ldots + \Theta_q\epsilon_{t-q} + \epsilon_t, \tag{1.8}$$

donde ϵ_t es un proceso de ruido blanco y $c, \Phi_1, \ldots, \Phi_p, \Theta_1, \ldots, \Theta_q$ son los parámetros del modelo.

Para un proceso $ARMA(p, q)$ una condición de estacionariedad es la misma que para un proceso $AR(p)$ ($|\Phi| < 1$), del mismo modo que una condición de invertibilidad es la misma que para el proceso $MA(1)$ ($|\Theta| < 1$).

El modelo $ARMA(p, q)$ se puede escribir en términos del operador de retardos como sigue:

$$(1 - \Phi_1 L - \Phi_2 L^2 - \ldots - \Phi_p L^p)x_t = (1 - \Theta_1 L - \Theta - 2L^2 - \ldots - \Theta_q L^q)\epsilon_t$$
$$\Phi_p(L)x_t = \Theta_q(L)\epsilon_t,$$

donde $\Phi_p(L)$ es el polinomio autoregresivo y $\Theta_q(L)$ es el polinomio de medias móviles.

Si el proceso es estacionario, su representación $MA(\infty)$ es $x_t = \dfrac{\Theta_q(L)}{\Phi_p(L)}\epsilon_t$, entonces $x_t = \epsilon_t + \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \phi_3\epsilon_{t-3} + \ldots$

Si el proceso es invertible, una representación $AR(\infty)$ es $\dfrac{\Phi_p(L)}{\Theta_q(L)}x_t = \epsilon_t$, entonces $x_t = \epsilon_t + \phi_1 y_{t-1} + \pi_2 y_{t-2} + \pi_3 y_{t-3} + \ldots$.

Los pesos de la representación $MA(\infty)$, como de $AR(\infty)$, están restringidos a depender del vector finito de parámetros del modelo $ARMA(p,q)$: $\Phi_1, \ldots, \Phi_p, \Theta_1, \ldots, \Theta_q$.
Las condiciones de estacionariedad del modelo $ARMA(p,q)$ vienen impuestas por la parte autoregresiva, dado que la parte de medias móviles finita siempre es estacionaria. Las condiciones de invertivilidad del modelo vienen impuestas por la parte de medias móviles, dado que la parte autoregresiva es siempre invertible, porque siempre está directamente escrita en forma autoregresiva.

El modelo $ARMA(p,q)$ tiene media cero, varianza constante y finita y una función de autocorrelación infinita. La función de autocorrelación decrece rápidamente a cero.

1.4.2 Procesos lineales no estacionarios

Proceso autoregresivo integrado y de media móvil ARIMA(p,d,q)

Los modelos anteriores se basaban en la estacionariedad de una serie de tiempo, es decir, donde se cumple que la media y la varianza son constantes en el tiempo y la covarianza es invariante en el tiempo. Sin embargo, muchas series no son estacionarias, por consiguiente, se debe restar una serie de tiempo d veces para que sea estacionaria y luego aplicar a esta serie diferenciada un modelo $ARMA(p,q)$.

Definición 1.4.6 *El proceso autoregresivo integrado y de media móvil $ARIMA(p,d,q)$, donde p denota el número de términos autoregresivos, d es el número de veces que debe ser diferenciada la serie para que sea estacionaria*

y q es el número de términos de la media móvil invertible, tiene como expresión algebraica:

$$x_t^d = c + \Phi_1 x_{t-1}^d + \ldots + \Phi_p x_{t-p}^d + \Theta_1 \epsilon_{t-1}^d + \ldots + \Theta_q \epsilon_{t-q}^d + \epsilon_t^d \qquad (1.9)$$

En forma del polinomio operador de retardos, el modelo $ARIMA(p, d, q)$ es:

$$\Phi(L)(1 - L)^d x_t = c + \Theta(L)\epsilon_t,$$

donde x_t^d es la serie de las diferencias de orden d, ϵ_t^d es un proceso de ruido blanco y $c, \Phi_1, \ldots, \Phi_p, \Theta_1, \ldots, \Theta_q$ son los parámetros del modelo.

La construcción de los modelos $ARIMA(p, q, d)$ se lleva de forma iterativa mediante un proceso en el que se distinguen cuatro etapas:

1. Identificación. Utilizando los datos ordenados cronológicamente se sugiere un modelo $ARIMA$ que sea investigado. El objetivo es determinar los valores p, d y q que sean apropiados para reproducir la serie de tiempo. Es posible identificar más de un modelo candidato para describir la serie.

2. Estimación. Dado el modelo apropiado para la serie de tiempo, se realiza la inferencia sobre los parámetros.

3. Validación. Se realizan diagnósticos para validar si el modelo se ajusta a los datos, sino se elige otro candidato.

4. Predicción. Una vez seleccionado el mejor modelo candidato, se pueden hacer pronósticos en términos probabilísticos de los valores futuros.

Proceso estacional autoregresivo integrado y de media móvil $ARIMA(p, d, q)(P, D, Q)_s$

Cuando una serie de tiempo tiene intervalos de observación menores a un año, entonces es frecuente que se tengan variaciones o patrones sistemáticos cada cierto periodo. Deben ser captadas en los llamados factores estacionales, dentro de la estructura del modelo a construirse.
Las series de tiempo estacionales pueden ser aditivas o multiplicativas, al mismo tiempo cada una puede ser estacionaria o no estacionaria. Usualmente se presentan con mayor frecuencia los modelos multiplicativos, de esta

manera se combinan términos ordinarios del proceso $ARMA$ y términos estacionales, así como diferencias regulares y diferencias estacionales para trasformar las series en estacionarias. Este tipo de procesos tienen las siguientes características:

- Contiene una componente $ARIMA(p, d, q)$ que modela la dependencia regular, que es la dependencia asociada a observaciones consecutivas.
- Contiene una componente $ARIMA(P, D, Q)$ que modela la dependencia estacional, que esta asociada a observaciones separadas por s periodos.

Definición 1.4.7 *La estructura general de un modelo* $ARIMA(p, d, q)(P, D, Q)_s$ *es:*

$$\begin{aligned} x_t &= c + \Phi_1 x_{t-1} + \ldots + \Phi_p x_{t-p} + \Theta_1 x_{t-s} + \ldots + \Theta_P p x_{t-Ps} \\ &+ \epsilon_t - \phi_1 \epsilon_{t-1} - \ldots - \phi_q \epsilon_{t-q} - \nu_1 \epsilon_{t-s} - \ldots - \nu_Q \epsilon_{t-Qs}, \end{aligned} \tag{1.10}$$

donde $\Phi_1, \ldots, \Phi_p, \Theta_1, \ldots, \Theta_P, \phi_1, \ldots, \phi_q, \nu_1, \ldots, \nu_Q$ *son los parámetros y* $\epsilon_t \approx N(0, \sigma^2)$

Capítulo 2

Técnicas de análisis de series de tiempo caóticas

De acuerdo con Casdagli [3], la modelación no lineal y la predicción de series de tiempo es muy reciente, esto es, la comunidad estadística ha construido modelos no lineales estocásticos desde 1980. Mientras que la comunidad que estudia a los sistemas dinámicos, motivada por el fenómeno del caos, ha construido modelos determinísticos no lineales desde 1987.

El caos se define como la conducta irregular de las soluciones de una ecuación determinista no lineal (de una variable como la logística o de tres como el sistema de Lorenz). Las soluciones caóticas exhiben un espectro de banda amplio, disfrazado como series de tiempo aleatorias cuando son analizadas con técnicas lineales, éstas son solo exactas para un periodo de tiempo gobernado por los errores en las condiciones iniciales y el exponente de Lyapunov del sistema, el cual cuantifica la divergencia exponencial de las trayectorias en el sistema caótico. Sin embargo, cuando se considera el espacio de estado subyacente, en muchos casos la solución caótica se encuentra en un atractor extraño, el cual tiene estructura fractal y típicamente una dimensión no entera.

Existen varios ejemplos de ecuaciones que son no lineales, que tienen pocas variables y que exhiben conducta caótica. En estos sistemas con un grado de libertad menor, las ecuaciones diferenciales ordinarias (EDO) deterministas derivadas de las leyes de la física pueden usarse para establecer numéricamente la existencia del caos y desarrollar una predicción a corto plazo, junto con un análisis de series de tiempo.

Recientemente se ha mostrado que las EDO deterministas, con un número

pequeño de variables, pueden modelar adecuadamente fenómenos físicos de varios grados de libertad cerca de la transición al caos. Motivado por los resultados en la teoría de los sistemas dinámicos, Ruelle y Takens (1971) conjeturaron que la transición a la turbulencia observada en los experimentos de la dinámica de fluidos, podían ser explicados por una bifurcación a un atractor extraño de dimensión menor, confirmada por los resultados de universalidad de Feigenbaum, resultados matemáticos en PDEs, así como con los experimentos físicos y los numéricos. El problema de estimar la dimensión del atractor extraño en el cual subyacen las series de tiempo irregulares, generadas por un sistema de varios grados de libertad como un fluído, fue propuesto por Takens en 1981, sus técnicas de reconstrucción de espacio fase mostraron en principio que es posible dirigir la estimación de dimensión por una observación directa de una serie de tiempo.

2.1 Teorema de Takens

Para la redacción de esta primer sección, se siguen las referencias [9], [10], [12], [19] [20] y [22]. El teorema de Takens muestra como las variables de retraso de una única serie de tiempo pueden ser usadas como variables represantes para construir el atractor de un proceso dinámico subyacente. La reconstrucción del espacio fase, a partir de las series de tiempo, es un enfoque muy importante para el análisis de los sistemas no lineales complejos. La dificultad que conlleva la prueba del teorema es debida a que sus conceptos principales se definen por medio de la topología diferencial y específicamente con la noción de genericidad. Es por ello que antes de presentar y demostrar el Teorema de Takens (como él mismo lo publicó en [19]), se necesitan mencionar algunas definiciones y teoremas importantes.

Teorema 2.1.1 *Teorema de Baire.*
Sea M un espacio métrico completo. Toda intersección numerable de abiertos densos es un subconjunto denso de M.

Definición 2.1.1 *Genericidad*
La noción topológica más común de genericidad es la de que una propiedad que se cumpla en una intersección numerable de abiertos densos (avalados

por el Teorema de Baire).

En la teoría de los sistemas dinámicos el comportamiento asintótico de un sistema disipativo puede ocurrir en sólo cuatro formas típicas: punto fijo, ciclo límite, toro o atractor extraño. El primer atractor extraño reportado en un artículo científico fue el de Edward N. Lorenz en 1963; mientras analizaba la explicación de la imposibilidad de predecir los fenómenos meteorológicos a largo plazo. A continuación su definición.

Definición 2.1.2 *Atractor extraño*
Un atractor extraño cumple las propiedades siguientes:

1. *Para todo tiempo t, las órbitas no salen del espacio de fases, éstas se mueven dentro de un volumen confinado de él. Esto ocurre debido al doblamiento y estiramiento del espacio de fases.*

2. *Algunos de los atractores extraños son fractales puesto que son autosimilares.*

3. *Todo atractor extraño es fuertemente dependiente de cambios pequeños en las condiciones iniciales. Esto es, dos puntos iniciales divergen rápidamente, de modo exponencial, rumbo a un caos aparente.*

Los atractores extraños pueden ser separados en dos categorías: estocástico y caótico, dependiendo de si están asociados con el comportamiento estocástico o caótico del sistema. Los atractores que implican sólo un número finito o infinito de ciclos tipo silla y sus variedades integrales inestables se denominan estocásticos. Los atractores que implican ambos ciclos tipo silla y estables con pequeñas regiones de atracción se denominan caóticos. Todas las trayectorias de fase que forman un atractor estocástico tienen que ser exponencialmente inestables. Un atractor caótico tiene que mantener al menos una trayectoria estable. En particular, los atractores caóticos pueden consistir en un ciclo tipo silla de múltiples revoluciones estables con bobinas suficientemente próximas o en un conjunto denumerable de ciclos límites estables con regiones de atracción suficientemente pequeñas (el número de ciclos puede ser infinito).

Definición 2.1.3 *Variedad*
Una variedad M es un espacio topológico el cual es local en $\mathbb{R}^m$, esto es, cada punto tiene una vecindad abierta la cual es homeomórfica a un subconjunto abierto de $\mathbb{R}^m$

Definición 2.1.4 *Inmersión o encajamiento*
Una inmersión o encajamiento es una transformación multivariada de una variedad que mapea todas las trayectorias sobre la variedad original y sin cruzamientos. Esto es, la inmersión es uno a uno globalmente, que lleva todas las singularidades en trayectorias que define la variedad (las singularidades son puntos sobre la variedad donde las trayectorias se cruzan de manera que las rutas futuras no están determinadas de forma única). Sin embargo, no puede preservar la topología global de una variedad, conserva la topología de cada vencindad local de la variedad original, por lo que cada punto del espacio tangente de la variedad encajada tiene la misma dimensionalidad que tiene la variedad original.

Según [9], los investigadores Sauer, Yorke y Casdagli sugirieron que la inmersión es un concepto que tiene que ver con la extracción de información de características del espacio fase, de las series de tiempo obtenidas por mediciones generales de la evolución de un sistema.

Definición 2.1.5 *Difeomorfismo o campo vectorial*
Sea M una variedad compacta, un sistema dinámico sobre M es un difeomorfismo $\phi : M \to M$ (en tiempo discreto) o un campo vectorial X sobre M (tiempo continuo). En ambos casos, la evolución en el tiempo corresponde con una posición inicial $x_0 \in M$, denotada por $\varphi_t(x_0)$. En el caso discreto $t \in N$ y $\varphi_i = \varphi^i$; mientras que en el caso continuo $t \in R$ y $t \to \varphi_t(x_0)$ es la curva integral por x_0.

Definición 2.1.6 *Observable*
Una función es observable si es una función $y : M \to \mathbb{R}$ de clase C^2 (suave).

El problema a tratar en el teorema es el siguiente: si, para algún sistema dinámico con evolución en el tiempo φ_t, se conocen las funciones $t \to y(\varphi_t(x))$, $x \in M$, entonces se puede obtener información acerca del sistema dinámico original y de la variedad. Los siguientes tres teoremas tratan de esta cuestión.

Teorema 2.1.2 *Sea M una variedad compacta de dimensión m. Para las parejas (φ, y), $\varphi : M \to M$ un difeomorfismo de clase C^2, $y : M \to \mathbb{R}$ una*

función suave, hay una propiedad genérica que dice el mapeo $\Phi_{(\phi,y)} : M \to \mathbb{R}^{2m+1}$, *definido por*

$$\Phi_{(\varphi,y)}(x) = (y(x), y(\varphi(x)), \ldots, y(\varphi^{2m}(x))),$$

es una inmersión o encajamiento.

Demostración 2.1.1 *Si* x *es un punto con periodo* k *de* φ, $k \leq 2m+1$, *todos los valores propios de* $\dfrac{d\varphi^k}{dx}$ *son diferentes entre ellos y diferentes de 1. Además también se considera que no se pueden tener dos puntos fijos diferentes de* φ *al mismo nivel de* y. *Para que* $\Phi_{(\varphi,y)}$ *sea encajamiento cerca del punto fijo* x, *los co-vectores* $\dfrac{dy}{dx}, \dfrac{dy(\varphi)}{dx}, \ldots, \dfrac{dy(\varphi^{2m})}{dx}$ *deben abarcar* $T_x^*(M)$. *Este es el caso para* y *genérico si* $d\varphi$ *satisface la condición anterior en cada punto fijo.*

De la misma forma probamos que $\Phi_{(\varphi,y)}$ *es genéricamente una inmersión cuando se restringen a los puntos periódicos, con período menor o igual a* $2m+1$.
Así, vemos que para un $(\bar{\varphi}, \bar{y})$ *genérico, tenemos que* $\Phi_{(\bar{\varphi},\bar{y})}$, *restringe a una vecindad compacta* V *del conjunto de puntos con período* $\leq 2m+1$ *es un encajamiento, para una vecindad* U *de* $(\bar{\varphi}, \bar{y})$, $\Phi_{(\varphi,y)}|V$ *es un encajamiento para cualquier* $(\varphi, y) \in U$.
Se quiere mostrar que para alguna $(\varphi, y) \in U$, *arbitrariamente cercana a* $(\bar{\varphi}, \bar{y})$, $\Phi_{(\varphi,y)}$ *es un encajamiento.*

Para cualquier punto $x \in M$, *el cual no es un punto de período* $\leq 2m+1$ *para* $\bar{\varphi}$, *los co-vectores* $\dfrac{d\bar{y}}{dx}, \dfrac{d\bar{y}(\varphi)}{dx}, \ldots, \dfrac{d\bar{y}(\varphi^{2m})}{dx} \in T^*(M)$ *pueden ser perturbados independientemente de la perturbación* $\bar{y}$. *Por tanto, arbitrariamente cerca de* $\bar{y}$ *hay un tal* $\bar{\bar{y}}$, *tal que* $(\bar{\varphi}, \bar{\bar{y}}) \in U$ *y tal que* $\Phi_{(\bar{\varphi},\bar{\bar{y}})}$ *es una inmersión.*
Hay entonces una ϵ *positiva tal que para cualquier* $0 \leq \rho(x, x') \leq \epsilon$, $\Phi_{(\bar{\varphi},\bar{\bar{y}})}(x) \neq \Phi_{\bar{\varphi},\bar{\bar{y}}}(x')$, ρ *es una métrica fija sobre* M. *Hay incluso una vecindad* $U' \subset U$ *de* $(\bar{\varphi}, \bar{\bar{y}})$ *tal que para cualquier* $(\varphi, y) \in U'$, $\Phi_{(\varphi,y)}$ *es una inmersión y* $\Phi_{(\varphi,y)}(x) \neq \Phi_{(\varphi,y)}(x')$, *para cualquier* $x \neq x'$ *y* $\rho(x, x') \leq \epsilon$. *Por tanto, cada componente de* V *tiene diámetro menor que* ϵ.

Finalmente, se tiene que mostrar que en U *se tiene una pareja* (φ, y) *con* $\Phi_{(\varphi,y)}$ *inyectiva. Para esto se necesita una colección finita* $\{U_i\}_{i=1}^n$ *de*

conjuntos abiertos de M, cubriendo la cerradura de $M - \{\cap_{j=0}^{2m}\varphi^j(V)\}$ y tal que,

- *para cada $i = 1, \ldots, N$ y $k = 0, 1, \ldots, 2m$, diámetro$(\bar{\varphi}^{-k}(U_i))< \epsilon$,*
- *para cada $i, j = 1, \ldots, N$ y $k, l = 0, 1, \ldots, 2m$, $\bar{\varphi}^{-k}(U_i) \cap U_j \neq \emptyset$ y $\bar{\varphi}^l(U_i) \cap U_j \neq \emptyset$, implica que $k = l$,*
- *para $\bar{\varphi}^j(x) \in M - (U_i U_j)$, $j = 0, \ldots, 2m$, $x' \notin V$ y $\rho(x, x') > \epsilon$, dos puntos de la sucesión $x, \bar{\varphi}(x), \ldots, \bar{\varphi}^{2m}(x), x', \bar{\varphi}(x'), \ldots, \bar{\varphi}^{2m}(x')$ no provienen de la misma U_i*

Sea $\{\lambda_i\}$ una partición de la unidad, es decir, λ_i es una función no negativa con soporte $\bar{U}_i$ y $\Sigma_{i=1}^N \lambda_i(x) = 1$, para toda $x \in \overline{M \backslash V}$.

Sea $\Psi : M \times M \times \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{2m+1} \times \mathbb{R}^{2m+1}$ el mapeo que es definido como $\Psi(x, x', \epsilon_1, \ldots, \epsilon_N) = (\Phi_{(\bar{\varphi}, \bar{\bar{y}}_\epsilon)}(x), \Phi_{(\bar{\varphi}, \bar{\bar{y}}_\epsilon)}(x'))$, donde $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ y $\bar{\bar{y}}_\epsilon = \bar{\bar{y}} + \Sigma_{i=1}^N \epsilon_i \lambda_i$. Definimos $W \subset M \times M$ como $W = \{(x, x') \in M \times M | \rho(x, x') \geq \epsilon\}$ y donde x y x' no están ambos en int(V). Ψ está restringido en una vecindad pequeña de $W \times \{0\}$ en $(M \times M) \times \mathbb{R}^{\mathbb{N}}$, es transversa con respecto a la diagonal de $\mathbb{R}^{2m+1} \times \mathbb{R}^{2m+1}$. Esta transversalidad se sigue inmediatamente de las condiciones impuestas a la cubierta $\{U_i\}_{i=1}^N$. De esta transversalidad concluimos que hay una vecindad muy pequeña $\bar{\epsilon} \in \mathbb{R}^{\mathbb{N}}$ tal que $\Psi(W \times \{\bar{\epsilon}\}) \cap \Delta = \emptyset$. Si esto también para $\bar{\epsilon}$, $(\bar{\varphi}, \bar{\bar{y}}_{\bar{\epsilon}}) \in U'$, entonces $\Phi(\bar{\varphi}, \bar{\bar{y}}_\epsilon)$ es inyectiva y por tanto es un encajamiento.
Esto prueba que para un conjunto denso de parejas (φ, y), $\Phi(\varphi, y)$ es un encajamiento. Ya que el conjunto de todos los encajamientos es abierto en el conjunto de todos los mapeos, entonces hay un conjunto abierto y denso de parejas (φ, y), para el cual $\Phi(\varphi, y)$ es el encajamiento. Lo cual prueba el teorema.

Floris Takens dice que este teorema también funciona si M es no compacto si se restringen los observables a funciones propias.

Teorema 2.1.3 *Sea M una variedad compacta de dimensión m. Para las parejas (X, y), X es un campo vectorial de clase C^2 y y es una función suave sobre M. Se define una propiedad genérica que dice el mapeo $\Phi_{(X,y)} : M \to \mathbb{R}^{2m+1}$, definido por*

$$\Phi_{(X,y)}(x) = (y(x), y(\varphi(x)), \ldots, y(\varphi^{2m}(x))),$$

es una inmersión o encajamiento, donde φ_t es el flujo de X.

BIBLIOTECA UACM

Demostración 2.1.2 *La prueba de este teorema es practicamente la misma que el Teorema 1. En este caso se imponen las siguientes propiedades genéricas sobre X:*

(a) Si $X(x) = 0$, entonces todos los valores propios de $\frac{d\varphi_1}{dx} : T_x(M) \to T_x(M)$ son diferentes y distintas de 1

(b) Ninguna curva integral periódica de X tiene periodo entero menor o igual que $2m + 1$

En este caso φ_1 satisface las mismas condiciones que $\bar{\varphi}$ de la prueba anterior. El resto de la demostración se obtiene de forma inmediata.

Teorema 2.1.4 *Sea M una variedad compacta de dimensión m. Para parejas (X, y), X es un campo vectorial suave y y es una función de al menos clase C^{2m+1} sobre M. Hay entonces una propiedad genérica en la q ue el mapeo $\Phi_{(X,y)} : M \to \mathbb{R}^{2m+1}$, que está definido por $\Phi_{(X,y)}(x) = (y(x), \frac{d(y(\varphi_t(x)))}{dt}\Big|_{t=0}, \ldots, \frac{d^{2m}(y(\varphi_t(x)))}{dt^{2m}})$, es una inmersión. Aquí, φ_t denota el flujo de X.*

Demostración 2.1.3 *Aunque esta prueba es análoga a la del Teorema 1, lo primero que se hace aquí es asumir que el campo vectorial genérico X tiene la propiedad de que $X(x) = 0$, para todo valor propio de $\frac{dX}{dx}$ diferente y distinto de cero. El $Sign(X)$ denota el conjunto de puntos para el cual X es cero, este conjunto es finito.*
Como en la primera demostración, para el campo vectorial X, el conjunto de funciones $y : M \to \mathbb{R}$ es tal que $\Phi_{(X,y)}$ es una inmersión y cuando está restringida a una vecindad de $Sign(X)$, la inmersión es residual.
Finalmente, para obtener la inmersión de $(X, \bar{y})$, $\bar{y}$ cercana a y, no necesitamos una cubierta abierta en el caso actual. Se puede construir un mapeo y_v, con v en algún espacio vectorial de dimensión finita V, el cual es el análogo de y_ϵ, con las siguientes propiedades:

(a) $y_0 = y$

(b) para $x \in Sign(X)$, el 1-jet de y_v es independiente de v

(c) *para* x, $x' \notin Sign(X)$, $x \neq x'$, *el mapeo* $j_x^{2m} \times j_{x'}^{2m} : V \to J_x^{2m}(M) \times J_{x'}^{2m}(M)$ *tiene una derivada suprayectiva para toda* (x, x') *en* $v = 0$. $J_{x'}^{2m}(M)$ *es el espacio vectorial de 2m-jets de funciones sobre* M *en* x, $j_x^{2m}(M)$ *es el 2m-jet de* y_v *en* x.

Usando y_v *definimos el mapeo* $\Phi : M \times M \times V \to \mathbb{R}^{2m+1} \times \mathbb{R}^{2m+1}$ *como antes. El resto de la prueba del Teorema 1 es idéntico.*

Con estos teoremas Takens explicó como un sistema dinámico, con evolución en el tiempo (continuo) φ_t y observable único y (compuesto por un conjunto discreto de valores) es determinado genéricamente por el conjunto de todas las funciones $t \to y(\varphi_t(x))$.

Incorporar una serie temporal caótica de una única variante $x_1, x_2, \ldots, xn$ en el espacio m-dimensional para obtener la ubicación del espacio de fase para N puntos de fase:

$$Y_i = (x_i, x_{i+\tau}, \ldots, x_{i+(m-1)\tau})^T$$

En la fórmula, $i = 1, 2, \ldots, N$, $N = n - (m-1)\tau$, Y_i representa el vector de espacio de fase después de la reconstrucción, τ es el tiempo de retardo, m la dimensión de inmersión, n son el número de puntos de la serie de tiempo original, N es el número de vectores del espacio fase después de la reconstrucción. Así, la matriz de espacio fase es:

$$\begin{aligned} Y_1 &= (x_1, x_{1+\tau}, \ldots, x_{1+(m-1)\tau})^T \\ Y_2 &= (x_2, x_{2+\tau}, \ldots, x_{2+(m-1)\tau})^T \\ &\ldots \\ Y_N &= (x_N, x_{N+\tau}, \ldots, x_{N+(m-1)\tau})^T \end{aligned} \tag{2.1}$$

De acuerdo con el Teorema de Takens, dada una τ apropiada y una m seleccionada, la forma dinámica del sistema original puede ser restaurada con equivalencia topológica para identificar la propiedad básica del sistema dinámico original. Además, durante el proceso de evolución de un sistema dinámico, todas las variantes están inherentemente correlacionadas.
A continuación se describe como se calculan las técnicas de análisis de las series de tiempo caóticas para la reconstrucción del espacio fase.

2.2 Técnicas de análisis

Para el estudio de las series de tiempo, se emplean algunas técnicas de análisis caótico, como son: la reconstrucción del espacio fase (explicada en la sección anterior), la determinación de tiempo de retraso y de la dimensión encajada, el cálculo de la dimensión de la correlación y del exponente más grande de Lyapunov.
En esta sección se explican detalladamente en que consisten estos métodos, con el objetivo central de encontrar un intervalo de predicción correcto y de afirmar si los datos que se exponen en el Capítulo 2 son caóticos o estocásticos. Para la redacción de este capítulo, se utiliza en el artículo *Chaotic analysis of traffic time series* [15] y las referencias: [5], [11], [17] y [18].

2.2.1 Espacio Fase equivalente

De acuerdo al Teorema de Takens antes expuesto y explicado, la evolución de cualquier variable del sistema está determinada por las otras variables con las cuales interactúa, entonces, la información de las variables relevantes está contenida implícitamente en la historia de cualquier otra variable.

De esta forma, un espacio fase "equivalente" puede ser reconstruído mediante la asignación de un elemento de la serie de tiempo x_t y sus sucesivos retrasos como coordenadas de una nueva serie de tiempo vectorial

$$Y_t = \{x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(m-1)\tau}\},$$

donde τ es el tiempo de retraso y m es la dimensión encajada (ambos términos dependen de lo datos dados).

2.2.2 Tiempo de retraso (τ)

El tiempo de retraso τ es el valor del tiempo para el cual se puede reconstruir el espacio fase y debe ser calculado apropiadamente para que se capture la estructura del atractor. El criterio básico para estimar el tiempo de retraso (denominado *time lag* en el lenguaje *R*) se basa en el siguiente razonamiento:

1. Si τ tiene un valor muy pequeño, los vectores de retardo construídos no son independientes, por lo cual la inmersión tendrá una acumulación de puntos alrededor de la diagonal en el espacio fase. Esto tiene como consecuencia la pérdida de información y de las características del atractor.

2. Si τ es muy grande, las coordenadas resultantes son no correlacionadas y la inmersión será muy complicada de realizar.

Así, el valor de τ debe ser el adecuado para conocer el atractor. Existen varias formas de obtener el tiempo de retraso, se detallan dos de ellas, la función de autocorrelación y la información mutua promedio.

Función de autocorrelación *acf*

Calcular la Función de autocorrelación (***acf***) de los datos, es decir, la función que nos indica que tan dependientes son las variables. Seleccionar a τ como el tiempo para el cual la función es muy cercana a cero (denominada el método de *"first.zero"* en *R*). La razón de ello es que el momento en que la función de autocorrelación toma el valor de cero, la muestra $x_{t+\tau}$ está completamente decorrelacionada (ya no depende) de x_t.
La función de autocorrelación de orden k, para un período estacionario Y_t, denotada por ρ_k, es la correlación de la serie separada k períodos:

$$\rho_k = Corr(Y_t, Y_{t-k}) = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{V(Y_t)}\sqrt{V(Y_{t-k})}},$$

donde la covarianza, $Cov(Y_t, Y_{t-k}) = E[(Y_t - \mu)(Y_{t-k} - \mu)]$, la media, $\mu = E(Y_t) = E(Y_{t-k})$ y la varianza, $V(Y_t) = \sigma^2$.

Información mutua promedio *AMI*

Calcular la Información mutua promedio (***AMI***) definida por:

$$I(\tau) = \sum_{ij} p_{ij}(\tau) ln(p_{ij}(\tau)) - 2\sum_{i} p_i(\tau),$$

donde p_i es la probabilidad de que x_t toma un valor dentro de los i-ésimos intervalos de un histograma, p_{ij} es la probabilidad de que x_t está en el intervalo i y $x_{t+\tau}$ está en el intervalo j.

El punto mínimo de la gráfica de la función $I(\tau)$ se considera como el valor más adecuado para τ, ya que este es el tiempo para el cual $x_{t+\tau}$ agrega la información máxima para el conocimiento que tenemos de x_t.
Una vez obtenido el tiempo de retraso, se calcula la dimensión de inmersión.

2.2.3 Dimensión de inmersión (m)

Un sistema dinámico es fractal si contiene estructuras similares a cualquier escala, a esto se le conoce como *autosimilaridad*. El atractor resultante de los sistemas deterministas caóticos puede exhibir un tipo inusual de *autosimilaridad*. Una dimensión de inmersión o encajamiento apropiada debe ser buscada de tal forma que el atractor sea invariante, esto es, que no sea sensible a perturbaciones pequeñas de las condiciones iniciales. De acuerdo al Teorema de Takens, un atractor d dimensional puede ser inmerso en un espacio fase m dimensional que estima y describe las características del sistema dinámico. Según varios investigadores se generaliza el teorema de inmersión, enfatizando la importancia de la dimensión fractal del atractor para estimar la dimensión mínima del espacio de encajamiento, ésta sería de $m > 2d$, muchos otros científicos sugieren que $m > d$ es suficiente. Para calcular la dimensión de inmersión se cuenta con dos algoritmos, el *FNN* y el *ANN*, se describen a continuación.

Algoritmo *FNN*

Calcular la dimensión de inmersión, es decir, la dimensión necesaria para reconstruir el estado fase, implica obtener el porcentaje de vecinos falsos, dada una cierta distancia x. Este método, denominado "*False Nearest Neighbors*" o *FNN*, se basa en suponer que para sistemas deterministas, los puntos suficientemente cercanos que se encuentran en el espacio de fase reconstruido permanecen igualmente cercanos en espacios de mayor dimensión. Esto es cierto si la dimensión de inmersión es bastante grande como para desplegar la estructura del atractor. El algoritmo *FNN* que calcula esta dimensión es el siguiente:

Dados los puntos x^a y x^b "suficientemente cercanos" en el espacio fase, se hace una comparación de su distancia euclideana $|x^a - x^b|$ en dos dimensiones de inmersión consecutivas d y $d+1$. Para la dimensión de inmersión $\boldsymbol{m}$ y el tiempo de retraso $\boldsymbol{\tau}$, las distancias están dadas por:

$$R_d^2 = \sum_{m=0}^{d-1} [x^a(t + m\tau) - x^b(t + m\tau)]^2$$

Pasar de la dimensión d a la $d + 1$ significa que una nueva coordenada igual a $x(t + d\tau)$ está siendo añadida en cada vector de retardo, por lo que la distancia euclideana de los dos puntos en la dimensión $d + 1$ es:

$$R_{d+1}^2 = R_d^2(t) + |x^a(t + d\tau) - x^b(t + d\tau)|^2$$

La distancia relativa entre los dos puntos en dimensión d y $d+1$ es la relación:

$$\sqrt{\frac{R_{d+1}^2 - R_d^2}{R_d^2}} = \frac{|x^a(t + d\tau) - x^b(t + d\tau)|}{R_d}$$

Mediante las relaciones siguientes se detectan los vecinos falsos:

1. Si,

$$\sqrt{\frac{R_{d+1}^2 - R_d^2}{R_d^2}} > x,$$

2. Si,

$$R_{d+1} > \sigma/x$$

donde x es una relación de la distancia (un valor umbral heurístico) y σ es la desviación estándar de los datos, considerada como la medida representativa del tamaño del atractor.

Al repetir este procedimiento para todos los puntos del atractor se obtiene la proporción de vecinos falsos para la dimensión $\boldsymbol{m}$ que se está analizando. Se grafica esta proporción para todas las dimensiones necesarias, obteniendo la curva de falsos vecinos. Se escoge el valor de $\boldsymbol{m}$ para el cual el porcentaje de falsos vecinos alcanza el valor nulo.

Con el fin de aplicar el método a la serie de tiempo de tráfico se selecciona un valor adecuado para la relación de la distancia x. Según algunos investigadores, este valor es cercano a 15 para varios sistemas no lineales. Sin embargo, el algoritmo *FNN* tiene una desventaja puesto que el valor del umbral x es subjetivo, en orden de asegurar una distinción entre datos caóticos de baja dimensión y ruido. Las series de tiempo

pueden tener distintos valores umbrales, lo cual implica que es muy difícil e incluso imposible dar un valor umbral razonable, independientemente de la dimensión $\boldsymbol{m}$ y de cada punto de la trayectoria. En 1997, Cao[2] modificó el algoritmo en uno conocido como "Averaged False Neighbors" o *AFN*.

Algoritmo *ANN*

En el algoritmo *ANN*, se calculan dos parámetros E_1 y E_2, los cuales provienen de las cantidades definidas por el algoritmo *FNN*. Basándose en la construcción del vector $\boldsymbol{m}$ dimensional definido por $y_i(m) = (x_i, x_{i+\tau}, x_{i+2\tau} + \ldots + x_{i+(m-1)\tau})$, donde $i = 1, 2, \ldots, N - (m-1)\tau$ y $\boldsymbol{\tau}$ el tiempo de retraso. Al igual que en el método *FNN*, la aproximación *ANN* define la cantidad:

$$a(i,m) = \frac{||y_i(m+1) - y_{n(i,m)}(m+1)||}{||y_i(m) - y_{n(i,m)}(m)||}$$

donde $||\cdot||$ es la norma máxima, $y_i(m+1)$ es el vector i-ésimo reconsruído para la dimensión de inmersión $\boldsymbol{m}$ y $n(i,m)$ es un entero tal que el vector de retardo $\boldsymbol{m}$-dimensional $y_{n(i,m)}(m)$ es el vecino cercano de $y_i(m)$. E_i está formulado como el valor medio de todas las distancias *FNN*, $a(i,m)$:

$$E(m) = \frac{1}{N - m\tau} \sum_{i=1}^{N-m\tau} a(i,m)$$

$E(m)$ depende sólo de $\boldsymbol{m}$ y de $\boldsymbol{\tau}$. La variación de $\boldsymbol{m}$ a $\boldsymbol{m+1}$ se puede investigar con la ecuación:

$$E_1(m) = \frac{E(m+1)}{E(m)}$$

La función $E_1(m)$ deja de cambiar cuando $\boldsymbol{m}$ es mayor que algún valor $\boldsymbol{m_0}$ ($m \geq m_0$ y m es cercano al valor 1) si la serie de tiempo proviene de un atractor. Entonces $\boldsymbol{m_0 + 1}$ es la dimensión de inmersión mínima.

E_2 es la cantidad que se usa para distinguir entre una serie de tiempo estocástica y una determinista, está formulada como sigue:

$$E_2^*(m) = \frac{1}{N - m\tau} \sum_{i=1}^{N-m\tau} |x_{i+m\tau} - x_{n(i,m)+m\tau}|$$

Si la serie de tiempo es determinista, entonces existe alguna m tal que $E_2(m)! = 1$; mientras que es estocástica si $E_2(m)$ es aproximadamente 1 para todos los valores calculados.

Tanto E_1 como E_2 se calculan para todos los valores diferentes que se van incrementando, estos correspondientes al a dimensión de inmersión $\boldsymbol{m}$. Las conductas globales de E_1 y de E_2 como funciones de $\boldsymbol{m}$ son usadas para estimar tanto la dimensión de inmersión mínima como la naturaleza del proceso dinámico subyacente que generan las series de tiempo.

Ahora, vemos como se calcula la dimensión de correlación.

2.2.4 Dimensión de correlación, D_2

La dimensión de correlación cuantifica la dimensión fractal de un objeto geométrico encajado en un espacio fase, o también se define como la medida de la extensión de la región en la cual la presencia de los datos afectan la posición de los puntos que se sitúan en el atractor. El método para calcular esta dimensión tiene mayor importancia, ya que con este se puede distinguir entre un sistema estocástico o caótico.

Según Takens y varios científicos, para caracterizar un sistema dinámico con una dimensión de atractor d y un espacio de fase dimensional $\boldsymbol{m}$, la función de correlación $C(r)$ está dada por:

$$C(r) = \lim_{n\to\infty} \frac{2}{N(N-1)} \sum_{i,j=1}^{N} H(u)$$

donde H es la función de Heaviside, con $H(u) = 1$ para $u > 0$ y $H(u) = 0$ para $u \leq 0$, $u = r - |Y_i - Y_j|$, N es el número de puntos sobre el atractor reconstruído, r es el radio de la esfera centrada en Y_i o Y_j.

Si las series de tiempo están caracterizadas por un atractor, entonces para valores positivos de r, la función de correlación $C(r)$ está relacionada con el radio r por la ley de potencia:

$$C(r) \approx \alpha r^{D_2}$$

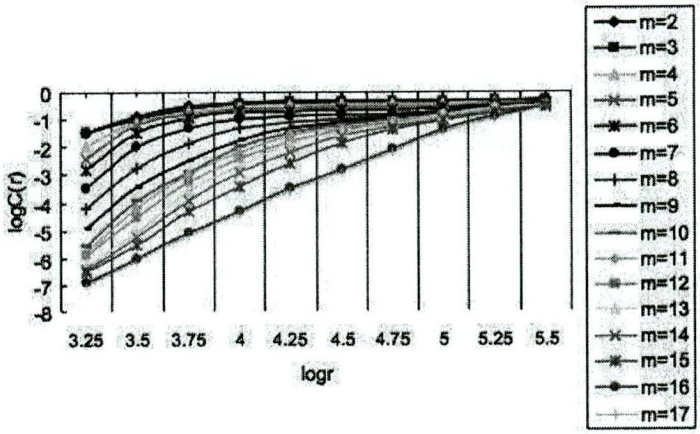


Figura 2.1: Gráfica de $logC(r)$ *versus* $logr$. Tomada de [15].

donde α es una constante, $\boldsymbol{D_2}$ es el exponente de correlación o la pendiente de la gráfica $\log C(r)$ versus $\log r$ (véase la figura 2.1). La pendiente puede ser estimada por medio del método de mínimos cuadrados sobre un rango de r (escalas de longitud), conocido como la región de escalamiento. Para un proceso aleatorio, $\boldsymbol{D_2}$ varía linealmente cuando $\boldsymbol{m}$ se incrementa, sin alcanzar un valor de saturación. Para un proceso determinista, el valor de $\boldsymbol{D_2}$ se satura y se vuelve independiente de $\boldsymbol{m}$, para una dimensión de inmersión creciente.

$$D_2 = \lim_{r \to 0} \frac{logC(r)}{logr}$$

Para ver si el caos existe, los valores del exponente de correlación se grafican contra los valores de dimensión de inmersión. Véase la figura 2.2. Si el valor del exponente de correlación $\boldsymbol{D_2}$ es finito, pequeño y no entero, el sistema exhibe una dinámica caótica de "dimensión menor". Si $\boldsymbol{D_2}$ es un valor entero finito no pequeño, la dinámica subyacente del sistema es dominada por un determinismo periódico fuerte. Por el contrario, si el exponente de correlación aumenta sin límite, con respecto al aumento de la dimensión de inmersión, el sistema sujeto a investigación se considera como estocástico.

En el caso de este estudio[15], la región de escalamiento existe y las series de tiempo tienen una característica caótica. La pendiente de la recta en la región de escalamiento es la dimensión de la correlación, como puede observarse en

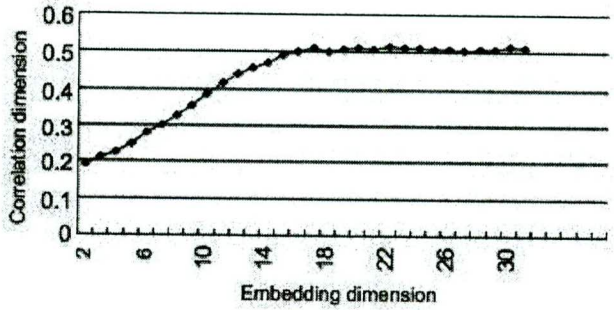


Figura 2.2: Gráfica que muestra la relación entre dimensión de inmersión y dimensión de correlación. Tomada de [15].

la figura 2.2.
La dimensión de correlación se incrementa cuando la dimensión de inmersión está por arriba de cierto valor y después de ese valor se satura, esto es un indicador de la existencia de la dinámica determinista. La dimensión saturada es aproximadamente de $D_2 = 0.51$ y la dimensión de inmersión $m = 16$. La dimensión de correlación es finita y pequeña, lo cual es un indicador de que la serie de tráfico exhibe una conducta caótica.
A continuación, calculamos el exponente de Lyapunov máximo.

2.2.5 Exponente de Lyapunov máximo, λ

Otra de las propiedades esenciales de los sistemas deterministas caóticos es la impredictibilidad o predictibilidad limitada de la evolución futura del sistema. Los exponentes de Lyapunov se caracterizan por la inestabilidad exponencial o por la tasa promedio de divergencia o de convergencia de trayectorias cercanas en el espacio fase, así miden la predictibilidad del sistema dinámico. Del análisis de estabilidad se puede ver que el exponente de Lyapunov máximo indica la expansión y la divergencia exponencial de las trayectorias cercanas. Además, la existencia de un exponente de Lyapunov máximo es relevante en la existencia de un atractor extraño a uno no caótico. Cuando no existe este exponente máximo, la predictibilidad a lo largo del tiempo, del sistema dinámico, está garantizada.
Para un proceso determinista de dimensión chica, el exponente deberá ser un número positivo finito, para un proceso lineal deberá ser cero y para un

estocástico, infinito. Véase el Cuadro 2.1.

Dinámica	Exponente máximo de Lyapunov
Punto fijo estable	$\lambda < 0$
Ciclo límite estable	$\lambda = 0$
Caos determinista	$0 < \lambda < \infty$
Ruido (estocástico)	$\lambda = \infty$

Cuadro 2.1: Dinámicas posibles de los sistemas estudiados y los correspondientes exponentes máximos de Lyapunov.

En general, para un espacio fase m-dimensional, la tasa de expansión o contracción de las órbitas es descrita, para cada dirección, por un exponente de Lyapunov, resultando m exponentes de Lyapunov diferentes, de los cuales algunos son cero o negativos.

Uno de los métodos para calcular el exponente máximo de Lyapunov es el de Rosenstein. Dados x_{n1} y x_{n2}, dos puntos en el espacio fase, con distancia euclideana $|x_{n1} - x_{n2}| = \delta_0$, entonces después de un tiempo t se espera que la nueva distancia δ sea igual a $\delta = \delta_0 e^{\lambda t}$, donde $\lambda > 0$, este último llamado el exponente de Lyapunov.

Después de reconstruir el espacio fase usando los valores deseados de τ y m, sea x_{n0} un punto en este espacio y x_n sus puntos vecinos (lo suficientemente cerca) con una distancia entre ellos menor que r. Se calcula la distancia promedio del punto en cuestión y se repite el proceso para N número de puntos a lo largo de la órbita, con la finalidad de calcular la cantidad promedio S conocida como el factor de extensión.

$$S = \frac{1}{N}\Sigma_{n_0=1}^{N}\left(ln\frac{1}{|u_{x_{n0}}|}\Sigma|x_{n0} - x_n|\right)$$

Donde $|u_{x0}|$ es el número de vecinos encontrados alrededor del punto x_{n0}.
Una gráfica de S *versus* N muestra una curva con un incremento lineal, seguido de una región sin grandes cambios. Esto se puede observar en la figura 2.3. La región donde la curva es lineal representa el incremento exponencial de S; mientras que la segunda representa el efecto de saturación de la divergencia

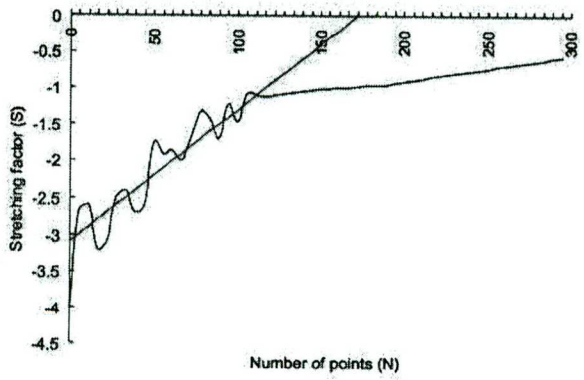


Figura 2.3: Estimación del exponente mayor de Lyapunov usando el método de Rosenstein. Tomada de [15].

exponencial debida al tamaño finito del atractor. El exponente de Lyapunov máximo, λ_{max} es la pendiente de la recta que se ajusta a los datos por medio del método de mínimos cuadrados.

Utilizando el exponente máximo de Lyapunov se puede encontrar un rango, aunque mínimo, para el cual se predice lo que pasará en el sistema que se estudie.

2.2.6 Tiempo de predicción

Ya que el caos es fundamentalmente determinista no se puede predecir, pero si se puede conocer el comportamiento de la serie de tiempo en un intervalo pequeño. El período límite de una predicción u horizonte de predictibilidad de un sistema caótico, Δ_{max}, depende del exponente máximo de Lyapunov.

$$\Delta_{max} = \frac{1}{\lambda_{max}}$$

Para ser caótico, el exponente de Lyapunov deberá estar entre 0 y 1. Para hacer predicciones, se comienza con el atractor desdoblado en un espacio m-dimensional y tiempo τ. Dado un vector inicial $y(t_1)$, se seleccionan las k trayectorias más cerca del atractor y los k puntos más cercanos a $y(t_1)$, uno sobre cada trayectoria. Un promedio de estas trayectorias es usado para

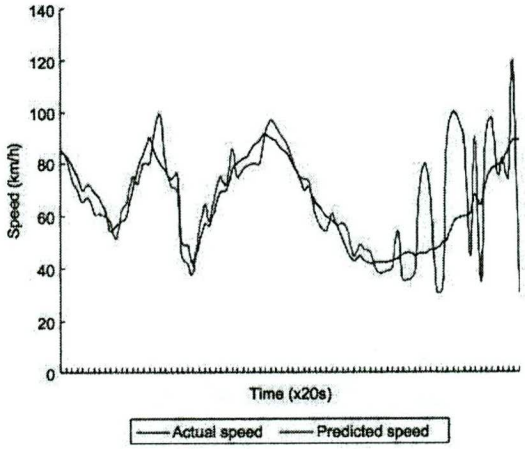


Figura 2.4: Comparación de la serie de tiempo de velocidad promedio de la autopista de Beijing y de la que se predice. Tomada de [15].

encontrar el siguiente punto de la trayectoria predicha, $y(t_1 + m\tau)$. El punto predicho es entonces el nuevo vector desde donde se empieza y el proceso se repite.

Aunque el fin de este trabajo no es hacer una predicción precisa de la serie de tiempo estudiada, es reelevante revisar de forma breve como se calculan algunos datos a partir de una serie caótica, así como se explicó en el primer capítulo como se hacen predicciones de series estocásticas.

2.3 Predicción de la serie de tiempo caótica

El método *GenericPred* de predicción a largo plaza de series de tiempo no lineales complejas está basado en los conceptos de la teoría del caos y en procesos de optimización, desarrollado por los investigadores Abbas Golestani y Robin Gras en su artículo del 2014, *Can we predict the unpredictable*[8].

La idea general de este modelo es extraer una característica única de la serie de tiempo estudiada que represente el comportamiento de la serie y de la cual se generarán valores sucesivos nuevos que continuarán los datos de la serie, cada valor miminiza la diferencia entre la característica de la nueva serie y

la inicial.

Definición 2.3.1 *Sea* $S_N = \{x_1, x_2, \ldots, x_N\}$ *una serie de tiempo, en la cual se calculan dos medidas no lineales* $V()$ *de* S_N*: la dimensión fractal y el exponente de Lyapunov. Se construye un mapeo, formando una serie nueva* $S_N^m = \{y_L, y_{L+1}, \ldots, y_N\}$*, para diferentes aplicaciones, como sigue:* $y_i = V(S_{i-L+1,i})$*,* $L \leq i \leq N$*, donde* $S_{i-L+1,i} = \{y_{i-L+1}, y_{i-L+2}, \ldots, y_i\}$.

Por otro lado, $S_N^m = S_N$, donde $0 < L < N$ es el tamaño de la ventana que se desplaza, usada para calcular el nivel local de caos medido por $V()$. Así, cuando el mapeo es aplicado, la nueva serie de tiempo S_N^m corresponde a la variación en el tiempo de la medida local no lineal, en la serie inicial S_N. Se considera $V(S_N^m)$ como un valor de referencia que se usa para predecir el siguiente valor k de la serie de tiempo: y_{N+i}, $1 \leq i \leq k$.

El parámetro σ, de una distribución normal $N(y_i, \sigma^2)$, es estimado calculando la variación entre cada dos valores consecutivos y_i, y_{i+1} de las series S_N^m. Esta distribución representa la distribución de probabilidad $P(y_i|y_{i-1})$. Se consideran varios conjuntos de datos para determinar si la distribución normal es una buena aproximación de la real. Sin embargo, el mismo método puede ser aplicado usando otras distribuciones sin afectar drásticamente la predicción.

Para predecir y_{N+i}, se genera un conjunto de valores aleatorios $Pos(y_{N+i})$ de N_{rand}, siguiendo la distribución $N(Y_{N+i-1,\sigma^2})$: $Pos(y_{N+i}) = \{y_{N+i}^j, 1 \leq j \leq N_{rand}\}$. N_{rand} es un parámetro que puede impactar la calidad de la predicción pues cuando se tienen más valores se incrementa la oportunidad de encontrar un valor óptimo. Sin embargo, no hay una mejora significante en los datos cuando N_{rand} es mayor que 10. Por esta razón, se escoge a 10 como el valor de N_{rand} para cada experimento. y_{N+i} es calculado seleccionando y_{N+i}^j, que hace la medida no lineal nueva más cercana a $V(S_N^m)$.

$$j_{min} = arg\ min_j(|V(S_{N+i-1}^m + y_{N+i}^j) - V(S_N^m)|),$$
$$y_{N+i} = y_{N+i}^{j_{min}},$$

donde $(S_{N+i-1}^m + y_{N+i}^j = \{y_1, y_2, \ldots, y_{N+i-1}, y_{N+i}^j\})$.

El valor y_{N+i}^j se escoge para hacer que $V(S_{N+i-1}^m + y_{N+i}^j)$ esté lo más cercano a $V(S_N^m)$. El punto importante es que el valor de referencia es siempre $V(S_N^m)$,

la medida no lineal calculada de las series originales.

Así, el método *GenericPred* usa dos reglas básicas:

1. Mantener siempre el valor de la medida no lineal constante durante la predicción.

2. El valor nuevo debe ser elegido de un conjunto de valores generados a partir de una distribución de probabilidad.

La predicción debe seguirse paso a paso debido a que el valor predicho en el paso actual se necesita para determinar un rango válido en el siguiente paso. El modelo de predicción clásico aproxima considerando un valor único para el siguiente paso, mientras que en el método *GenericPred*, varios puntos son considerados simultáneamente. Este método también es capaz de ajustar constantemente la información sobre la series de tiempo actual, mientras que los métodos clásicos de predicción se aplican al modelo sin tomar en cuenta la concordancia entre la serie original y la que se predice. Técnicamente, cualquier medida no lineal puede ser usada para caracterizar la serie de tiempo. Sin embargo, se usa el método $P\&H$[1] ya que se ha mostrado que puede detectar eficientemente entre distintos tipos de conducta no lineal.

[1]El método $P\&H$ tien como objetivo distinguir señales aleatorias de señales determinísticas. Véase la referencia [7]

Capítulo 3

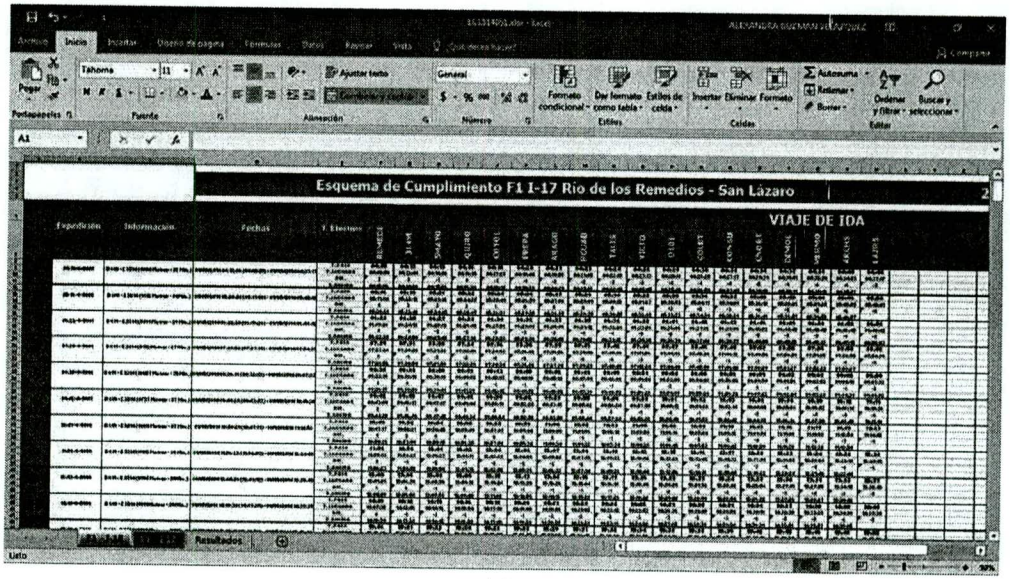
Datos de velocidad promedio

3.1 Construcción de la serie de tiempo

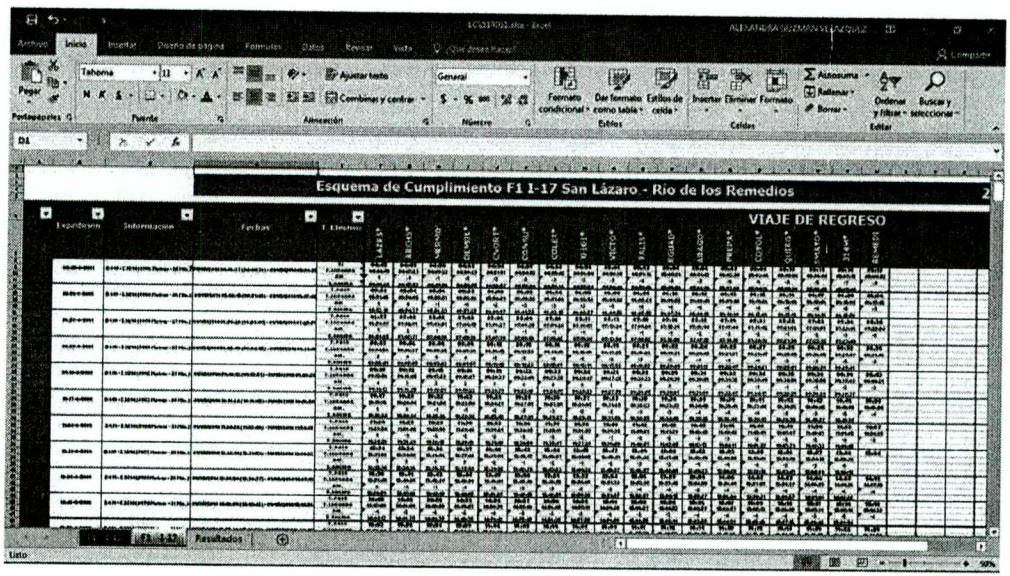
El área de sistemas del Metrobús de la Ciudad de México proporcionó la información de los tiempos de llegada y salida entre cada estación, de la Línea 5, de la ida y el regreso de las unidades del Metrobús, del mes de mayo de 2014. Estos datos se encuentran organizados en 31 libros de Excel: para cada día se tiene un sólo libro con dos hojas, la primera corresponde al viaje de Río de los Remedios a San Lázaro, la segunda al viaje de vuelta. Véase la figura 3.1. En un día circulan aproximadamente 17 unidades, una sóla puede hacer de 6 a 21 corridas por día y el tiempo de espera entre unidades, en las estaciones terminales, es de 6 a 10 minutos, dependiendo de la disponibilidad y del horario.

Para conocer los datos de velocidad promedio entre estaciones, en el mes de mayo, se usa el lenguaje de programación *Python*[1] y la libreria denominada *openpyxl* para ordenar los datos de todas las hojas de tal forma que se pueda ver la dinámica de todos los días de mayo, de 4 *am* a las 0 horas, en un sóla gráfica. La idea de organizar la lista de esta forma es para ver si existen datos anómalos, limpiar la lista y después construir la serie de tiempo buscada con la ayuda de los datos calculados.

[1]El programa que realizamos en Python se puede ver en el Apéndice A.1 (programatesis.py).



(a)



(b)

Figura 3.1: (a) Hoja de Excel donde se muestran los tiempos de llegada y salida por estación, en el viaje de Ida. (b) Hoja de Excel donde se muestran los tiempos de llegada y salida por estación, en el viaje de Regreso.

Se empiezan quitando las hojas ocultas (sin información) de cada libro de excel, se calcula el número de filas en cada hoja y de acuerdo al valor máximo de filas por libro, se toman los datos de las horas de llegada y de salida por estación y se depositan en una tercera hoja (*Resultados*), por cada libro de Excel. Ahí mismo, se cuenta el tiempo total de recorrido y la velocidad promedio tanto en m/s como en km/h, usando las distancias entre estación que se encuentran con Google maps y que se comparan con las medidas proporcionadas por el Metrobús. Véase ahora la figura 3.2.

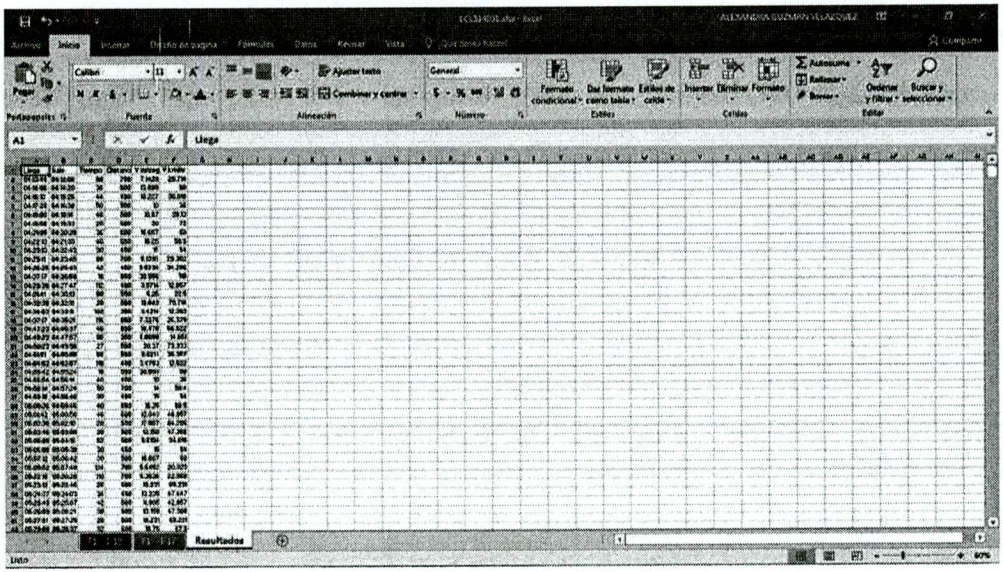


Figura 3.2: Hoja de Excel donde se muestran las distancias y las velocidades promedio entre estación, en m/s y km/h.

Una vez obtenida la hoja *Resultados* en cada libro, se crea un libro nuevo denominado *serie.xlsx*, donde se recolectan los datos, de todos los libros, de acuerdo al número de renglones o filas de cada uno de ellos (pues varían mucho, hay mayor número de filas por la noche y en lunes y viernes). Ya teniendo toda la lista, se quitan los datos nulos, es decir, los renglones donde la columna A y/o la columna B estén en blanco, se obtiene el tiempo promedio de llegada y salida por estación y se ordenan las cifras tomando como base este tiempo promedio. El resultado es una serie de $270,807$ datos de velocidad en km/h, acomodado de 4 am a las 0 horas del día siguiente,

esto se puede observar en la figura 3.3.

Como se puede observar en la figura 3.3 (a), existen valores muy grandes de la velocidad promedio en km/h que físicamente no son probables. Esto también lo podemos constatar en la gráfica boxplot, figura 3.4. La razón de que ocurran estas cifras "erróneas" es debida únicamente a la captura de la información en los libros de Excel. Esto se constata viendo las diferencias entre las horas de llegada y salida de los libros 9 y 27 de Excel. Por ejemplo: en el libro 27, figura 3.5 (a) y fila 6986, se pueden ver la hora de llegada $17:57:16$ y la de salida $17:57:15$, su diferencia es de tan sólo 1 segundo y esto da como resultado una velocidad de 2160 km/h para una distancia de 600 m. Ambos datos, pertenecientes a la hoja Resultados del libro 27 corresponden a los valores que podemos observar en la hoja de Viaje de Ida del mismo libro, figura 3.5 (b).

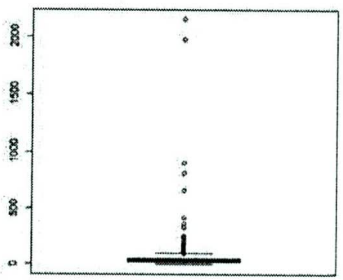
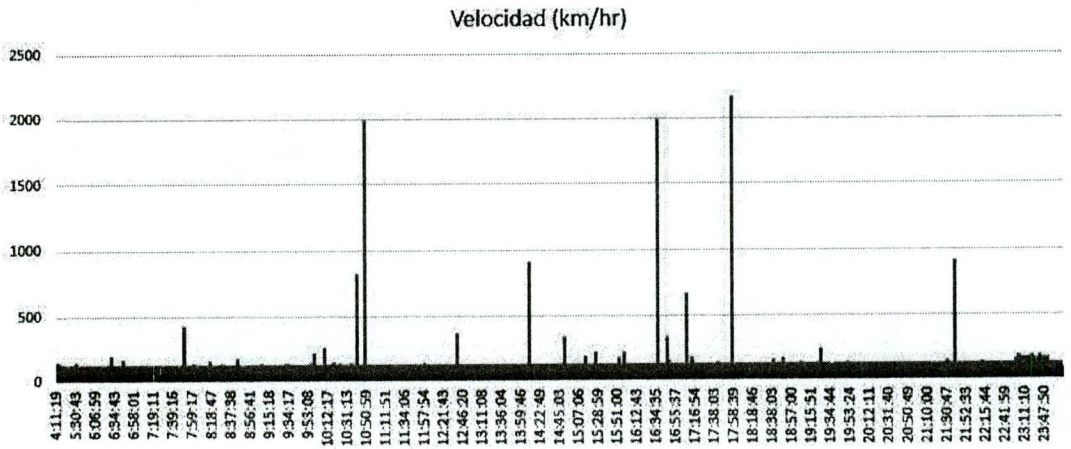
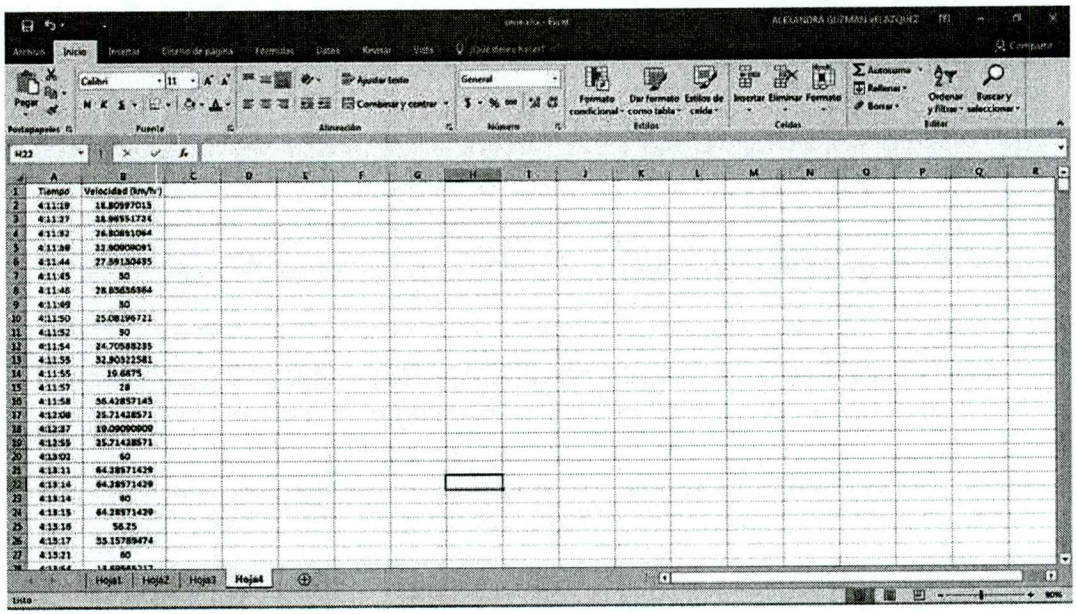


Figura 3.4: Diagrama de caja de la serie de tiempo de 270,807 datos.

En el libro 9, figura 3.6 (a) y fila 5834, se pueden apreciar que los datos obtenidos de llegada y de salida son $16:36:49$ y $16:36:48$, respectivamente, lo cual genera una diferencia de tan sólo 1 segundo nuevamente, pero como la distancia es en este caso de 550 m, la velocidad resultante es de 1980 km/h. Pero éstos no son los únicos casos anómalos, utilizando nuevamente *Python*, se encuentran 48 datos de velocidad promedio que son mayores a 140 km/h. Todos estos datos se quitan porque fijamos como velocidad máxima a $140km/h$ y así la serie queda con $270,763$ datos. Véase la figura 3.7.

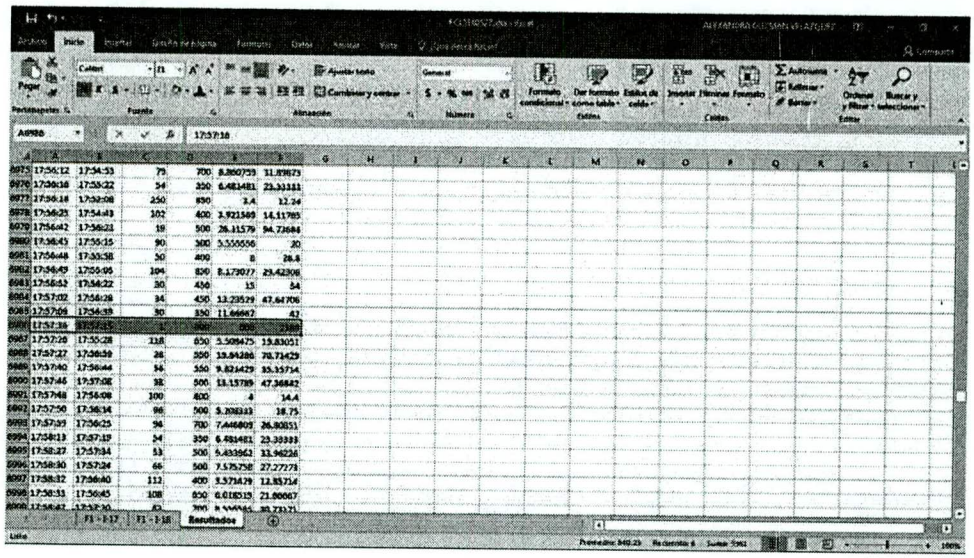


(a)

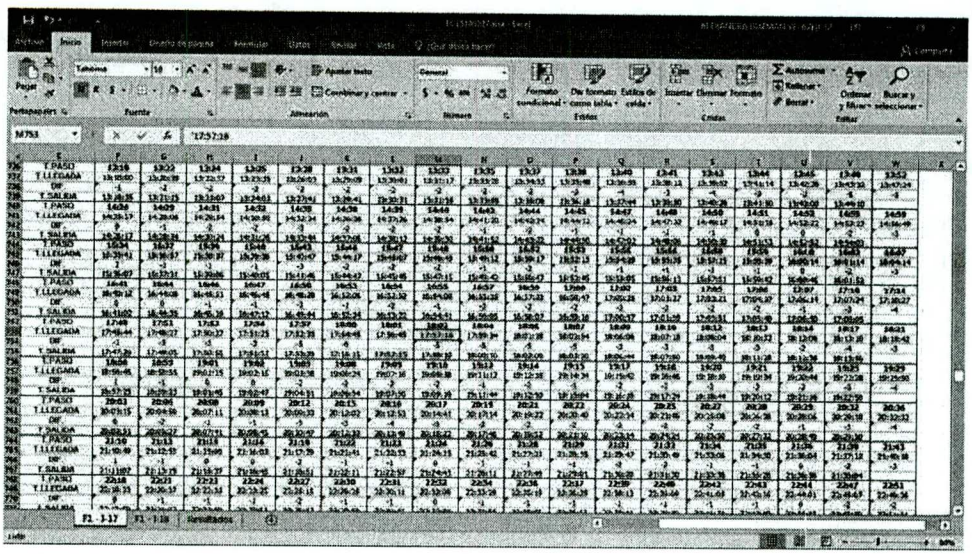


(b)

Figura 3.3: (a) Gráfica de la serie de tiempo obtenida. (b) Hoja de Excel donde se muestra la serie de tiempo construída con base a las hojas *Resultados* de todos los libros.

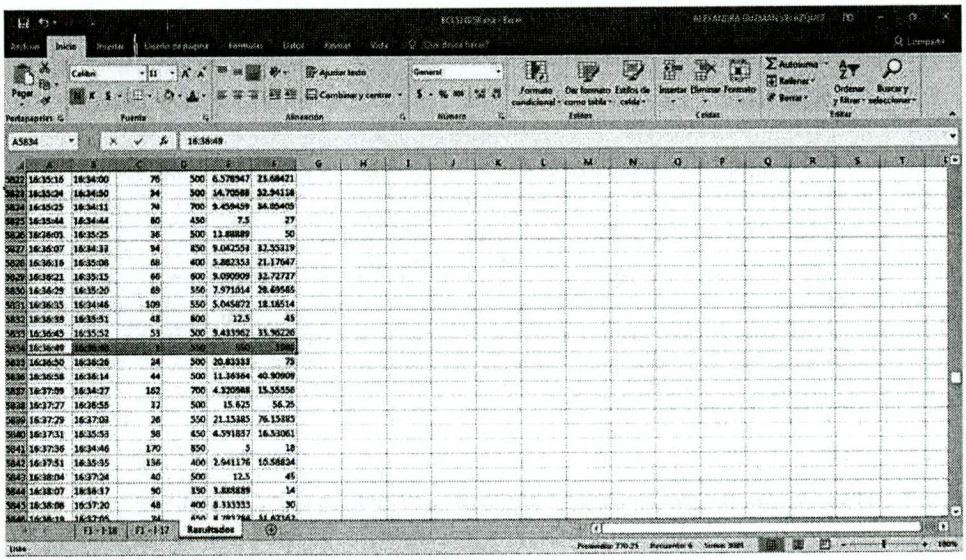


(a)

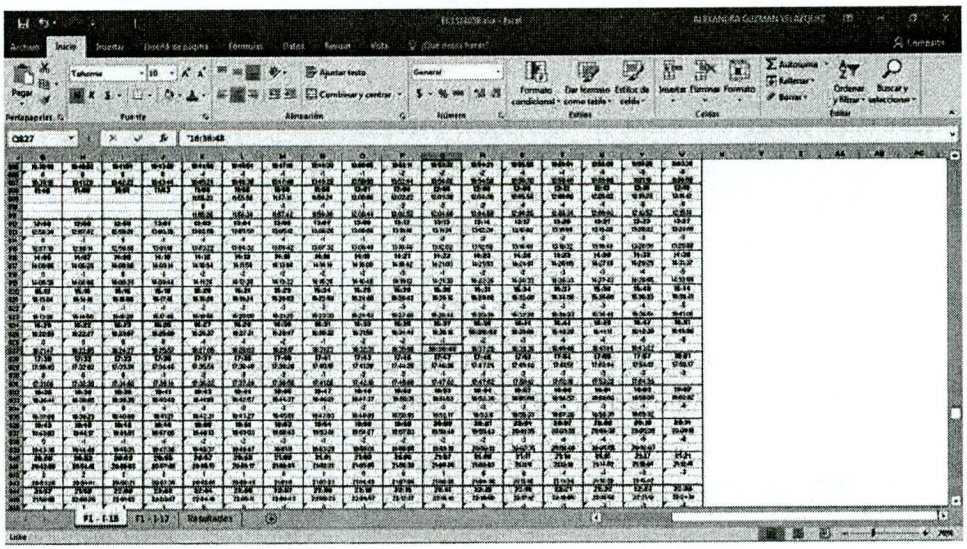


(b)

Figura 3.5: Libro 27



(a)



(b)

Figura 3.6: Libro 9

Se destaca que se escogió $140km/h$ como valor máximo de la velocidad promedio porque existen muchos valores entre 120 y 140 km/h (65 datos), sobre todo entre las 23 y 0 horas, lo cual tiene sentido si tomamos en cuenta

Velocidad
Min. : 0.00731
1st Qu.: 22.10526
Median : 34.61538
Mean : 39.02729
3rd Qu.: 52.10526
Max. : 140.00000

Cuadro 3.1: La estadística de los datos.

que a esas horas ya no hay tránsito vehicular ni peatonal. La lista queda entonces como se puede observar en la figura 3.7. La estadística de estos datos se puede observar en el cuadro 3.1 y el diagrama de caja de la figura 3.8.

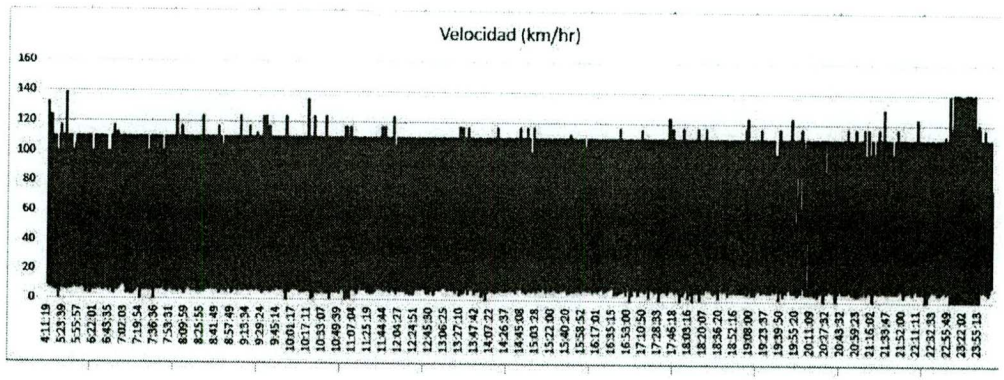


Figura 3.7: Lista de 270,763 datos de velocidad promedio.

Una vez visto el comportamiento de los datos, se construye la serie de tiempo, de tal forma que sea evidente la dinámica del Metrobús en cada uno de los días del mes de mayo. Para ello se utilizan las cifras ya calculadas de las hojas denominadas *Resultados*, donde se obtienen los promedios de las velocidades por hora, de 4 *am* a las 0 horas, por día. Véase la figura 3.9 (a), Así, teniendo 21 horas por día y 31 días, se obtiene una serie de tiempo de 651 datos, como se puede observar en la gráfica de la figura 3.9 (b).

La serie de tiempo encontrada tiene como valor mínimo, 33.64 (km/h, valor máximo, 43.71 km/h), punto medio 38.94 km/h, lo cual satisface los datos

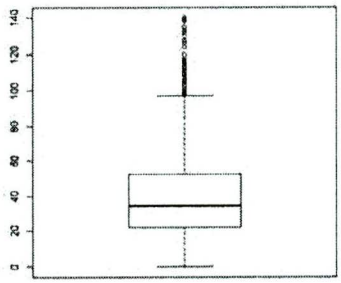


Figura 3.8: Diagrama de caja de la lista de 270,763 datos.

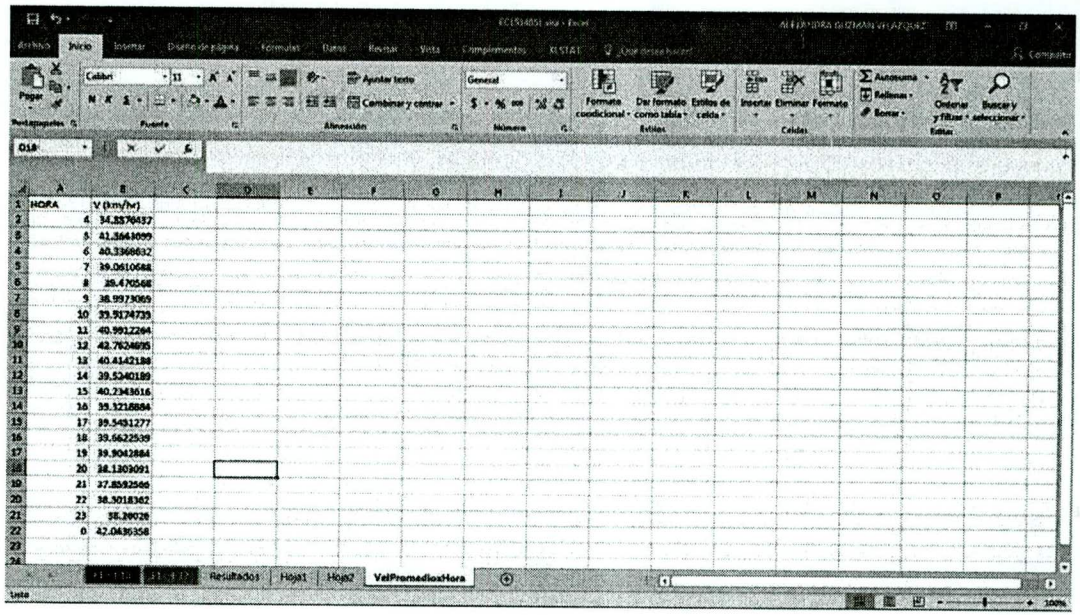
Velocidad
Min. : 33.64
1st Qu.: 38.28
Median : 38.81
Mean : 38.94
3rd Qu.: 39.55
Max. : 43.71

Cuadro 3.2: Datos estadísticos más relevantes de la serie de tiempo.

estándares que maneja el Metrobús de la Ciudad de México[2]. Véase el cuadro 3.2.

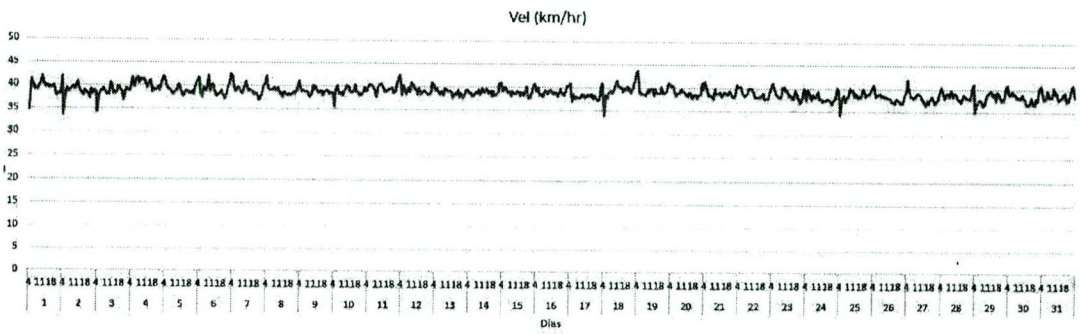
Para checar si la serie de tiempo es estacionaria, se pueden graficar las funciones de autocorrelación y autocorrelación parcial, si las funciones van decreciendo cuando el tiempo de retardo aumenta, entonces lo es. Se observan las gráficas de la figura 3.10 y se ve que efectivamente la serie de tiempo obtenida es estacionaria. Incluso, existen varias pruebas para probar esto, una de ellas se llama Ljung-Box (definida en la sección 1.2.3) y lo que hace es examinar si existe evidencia de correlaciones distintas de cero en un período de retardo de 1 a 20 por ejemplo, si los valores de p son pequeños ($p \leq 0.05$), entonces la serie es estacionaria. En este caso se

[2]Para mayor información, regresar a la introducción de esta tesis o checar las referencias [36] y [37].



	A	B
1	HORA	V (km/hr)
2	4	34.8376437
3	5	41.3643099
4	6	40.3368632
5	7	39.0610588
6	8	39.470568
7	9	38.9973069
8	10	39.9174739
9	11	40.9912264
10	12	42.7624695
11	13	40.4142186
12	14	39.5340189
13	15	40.2343616
14	16	39.3218884
15	17	39.5491277
16	18	39.6622539
17	19	39.9042884
18	20	38.1303091
19	21	37.8592566
20	22	38.3018362
21	23	38.29028
22	0	42.0436358

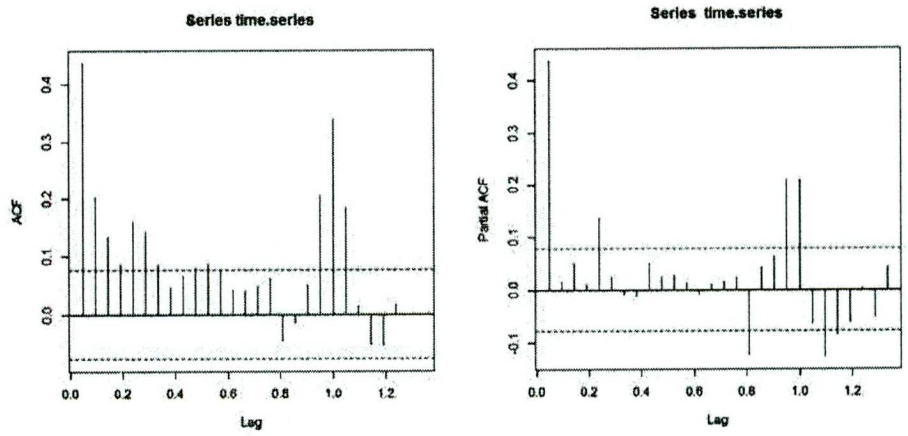
(a)



(b)

Figura 3.9: (a) Libro 1 en donde se muestran los datos de velocidad promedio del Metrobús por hora, en un día de mayo de 2014. (b) Serie de tiempo de 651 datos, en el eje del tiempo podemos ver que.

calcula en *R*, *Box.test(time.series,lag=20,type="Ljung-Box")*, cuyo resultado es $p - value < 2.2e - 16$, verificando la estacionariedad. Sin embargo, si se quiere demostrar esto mediante la definición formal, vista en el Capítulo 1, se pueden utilizar los vectores de Takens y calcular tanto la media, la varianza y la autocovarianza de las series de tiempo $x_{t-\tau}$ y checar que los valores no cambian. Al final de esta sección, cuando se encuentre la predicción de la serie de tiempo, se checará la definición formal.



(a) (b)

Figura 3.10: (a) Gráfica de la función de autocovarianza de la serie de tiempo. (b) Gráfica de la función de autocorrelación de la serie de tiempo. (c) Gráfica de la función de autocorrelación parcial de la serie de tiempo.

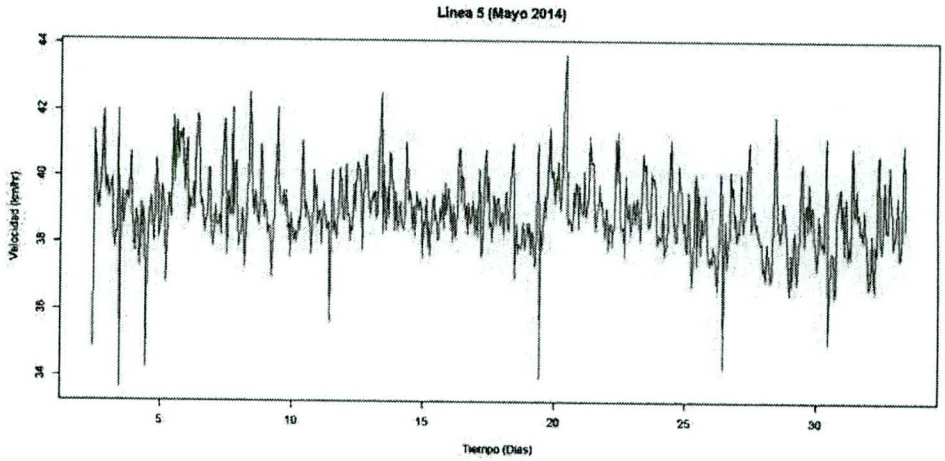


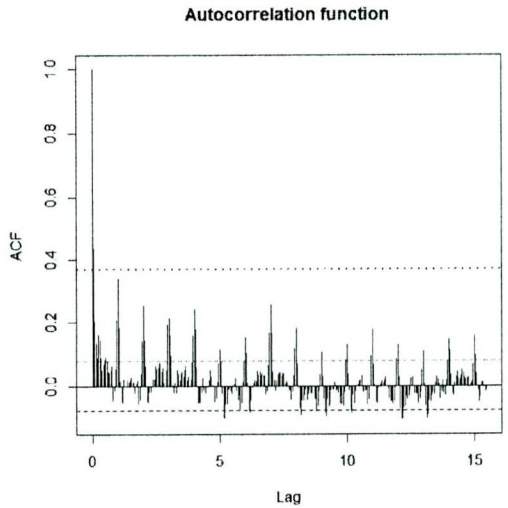
Figura 3.11: Serie de tiempo de velocidad promedio del Metrobús, línea 5, del mes de mayo de 2014.

3.2 Técnicas de análisis de series de tiempo caóticas

Una vez obtenida la serie de tiempo formada por los datos de velocidad promedio del mes de mayo, de la Línea 5 del Metrobús de la Ciudad de México, figura 3.11, se reconstruye el espacio fase utilizando el método dado por el Teorema de Takens como se revisó en el segundo capítulo. Para ello, se utilizan los paquetes *tseriesChaos*[27], *nonlinearTseries*[29], *Tseries* y [30] del lenguaje *R*, se obtienen así: el tiempo de retraso, la dimensión de inmersión o de encajamiento, la dimensión de correlación, el exponente de Lyapunov máximo y el tiempo de predicción para esta serie. El programa en *R* se puede ver en el Apéndice *A*.2 y lleva por nombre *proyectoTesis.R*.

3.2.1 Tiempo de retraso (τ)

Para el tiempo de retraso se utiliza la función ***acf*** y el método *"first.zero"* para calcular τ, como lo se definió en la sección 2.2.2. De acuerdo a este método,



(a)

Figura 3.12: Gráfica del tiempo de retraso.

τ=*time.lag=timeLag(time.series, "acf",method="first.zero", lag.max=NULL, do.plot = TRUE)*=2. Véase figura 3.12.

3.2.2 Dimensión de inmersión (m)

Se calcula la dimensión con el algoritmo *ANN* de Cao, puesto que si obtenemos los vecinos falsos con el algoritmo *FNN* y la función *false.nearest* de la libreria *tseriesChaos*[27], éste no proporciona información suficiente, véase la figura 3.13 (a). En esta gráfica se puede observar que el porcentaje de vecinos falsos no desciende nunca y así no llegará a ser cero, como el mismo método lo define, pero esto no implica que el *software* este mal programado o que no se estén metiendo bien los datos, lo que ocurre es que el método ya no puede hacer más con los valores que se le dan si se usa un valor umbral x subjetivo, esto último se explicó en la sección 2.2.3.

Como se puede ver en la figura 3.13 (b), la función $E_1(d)$ toca a la recta punteada cuando $d = 8$. Mientras que la función $E_2(d)$ exhibe que la serie de tiempo es determinista, puesto que existen varios valores de d para los cuales

$E_2(d)! = 1$. Así,

m=embedding.dim=estimateEmbeddingDim(time.series,number.points=651,time.lag, max.embedding.dim=15,threshold=0.95,max.relative.change=0.1,do.plot= TRUE)=8

Se destaca que algunos valores de los parámetros que pide la función *estimateEmbeddingDim* son indispensables para lograr que se pueda calcular la dimensión y que se minimice el tiempo en el que se realicen estas cuentas, esto se puede encontrar en la referencia [29]. Por ejemplo, si *number.points=length(time.series)* es muy grande, los cálculos computacionales son excesivos. Sin embargo, si el valor de este parámetro es chico, la dimensión no cambia y el tiempo en el que corre la función no es tan grande. Con respecto al valor de *threshold=0.95* y al de *max.relative.change=0.1*, ambos son básicos para calcular la dimensión con el método *ANN*, puesto que el primero es el valor umbral en el cual la función $E_1(d)$ para de cambiar y además su valor es muy cercano a 1. El segundo, es el cambio entre ambas funciones y éste necesariamente es un poco más grande que la diferencia entre la unidad y 0.95 (el valor por default es de 0.01). La dimensión máxima o *max.embedding.dim* se utiliza de 15 unidades puesto que en las referencias leídas ([15],[17] y [29]), éste o 16 son los números encontrados en los experimentos.

3.2.3 Espacio fase

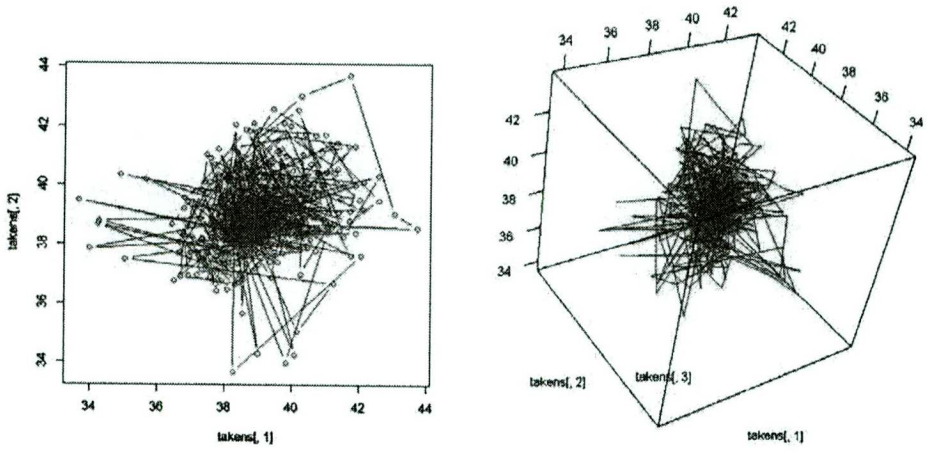
Dados los valores de τ y de m, entonces la reconstrucción del espacio fase estaría dada por $Y_t = \{x_t, x_{t-2}, \ldots, x_{t-14}\}$, donde x_t es la serie de tiempo de 651 datos de velocidad promedio de la figura 3.11.

Y_t=takens=buildTakens(time.series,embedding.dim,time.lag)

La gráfica de la proyección del atractor en el plano $Y_t[x_t]Y_t[x_{t-2}]$ se puede observar en la figura 3.14 (a) y en el espacio tridimensional $Y_t[x_t]Y_t[x_{t-2}]Y_t[x_{t-4}]$, se ve en la figura 3.14 (b). Las gráficas de las series de tiempo $x_t, x_{t-2}, \ldots, x_{t-14}$, se pueden observar en la figura 3.15.

3.2.4 Dimensión de correlación (D_2)

Para calcular la dimensión de inmersión m con R, la función es la siguiente:



(a) (b)

Figura 3.14: (a) Proyección del atractor en el plano $Y_t[x_t]$ *versus* $Y_t[x_{t-1}]$ (b) Proyección del atractor en el espacio trimensional $Y_t[x_t]Y_t[x_{t-1}]Y_t[x_{t-2}]$.

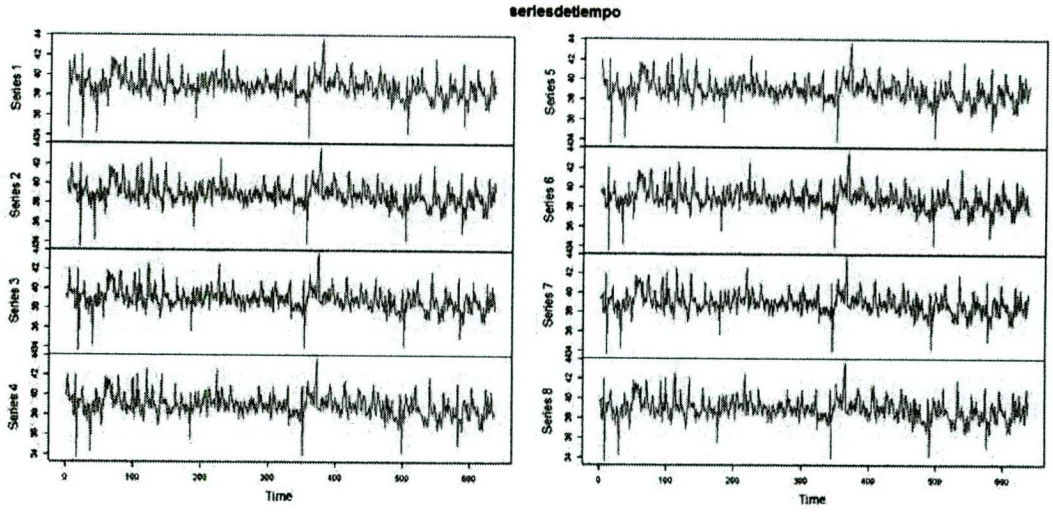
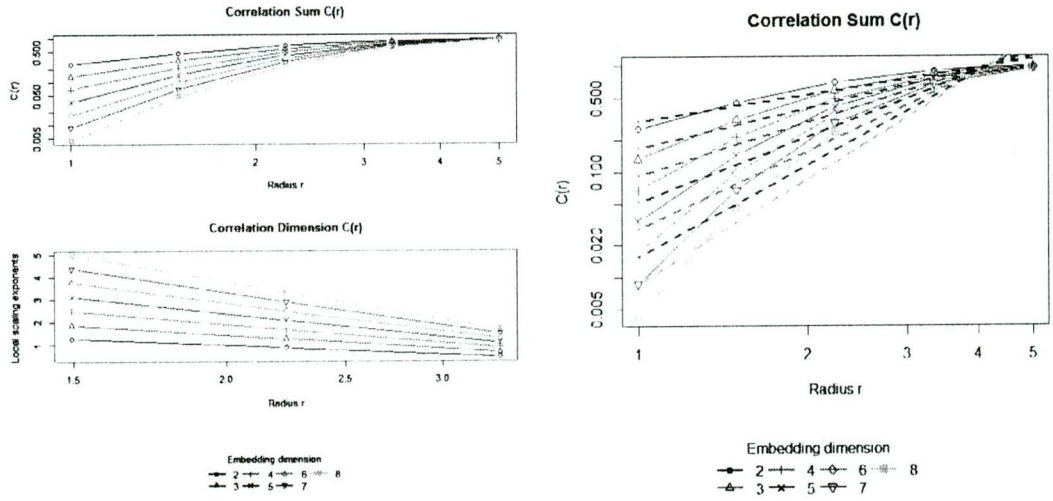


Figura 3.15: Gráficas de cada una de las series de tiempo del espacio fase $Y_t = \{x_t, x_{t-2}, \ldots, x_{t-14}\}$.



(a) (b)

Figura 3.16: (a) Gráfica de $C(r)$ y de su reescalamiento logarítmico. (b) Estimación de la dimensión de correlación con un rango de r de $[0, 5]$.

$\boldsymbol{D_2}$=*DimCorr=corrDim(time.series, min.embedding.dim = 2,embedding.dim, time.lag, min.radius=1,max.radius=5, corr.order = 2,n.points.radius =5,theiler.window = 100, do.plot = TRUE,number.boxes = NULL)*= 2.070023.

Véase la figura 3.16. Es importante hacer notar que si las rectas graficadas no convergen en un rango específico de r, entonces hay que ampliar el rango y el número de puntos en el que se va calculando la suma de correlación $C(r)$. En este caso el rango es de tan sólo 5 unidades, puesto *max.radius=5*.

El valor de la dimensión de correlación dice que efectivamente, el sistema dinámico que se esta estudiando es determinista caótico, puesto que éste es finito, no entero y mucho menor que la dimensión de inmersión. Véase la sección 2.2.4. El valor del parámetro *theiler.window* es reelevante pues indica la separación mínima en el tiempo entre vecinos cercanos que está permitido tener, en orden de tomar en cuenta puntos correlacionados temporales para buscar los vecinos cercanos. Se cambió el valor de 40 a 100 y se obtuvieron mejores resultados.

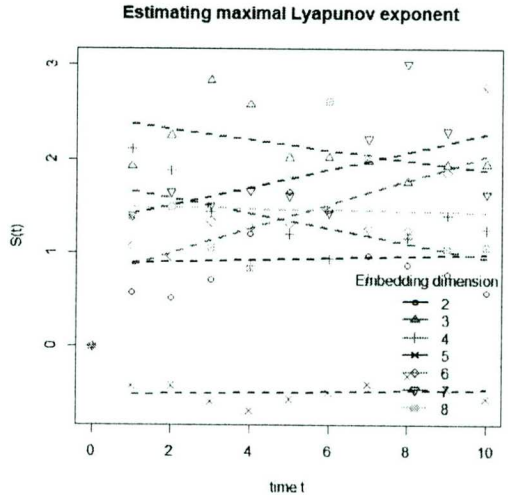


Figura 3.17: Estimación del exponente máximo de Lyapunov.

3.2.5 Exponente máximo de Lyapunov (λ)

Si usamos la función *maxLyapunov*:

λ=*ExpLyap=maxLyapunov(time.series,min.embedding.dim=2,embedding.dim, time.lag,radius=5)*= 0.03198189

Este resultado dice que efectivamente, el sistema dinámico es determinista caótico, pues $0 < \lambda < \infty$. Véase la sección 2.2.5. En este caso, los parámetros fundamentales para obtener el exponente de Lyapunov máximo son solamente la dimensión de inmersión y el radio.

El valor máximo del exponente de Lyapunov es la pendiente de la recta obtenida por medio de regresión lineal, que coincide con la dimensión de inmersión. Véase la gráfica 3.17.

3.2.6 Tiempo u horizonte de predictibilidad

Por último, el valor de la predicción se encuentra de la siguiente forma:

$$\Delta_{max} = \frac{1}{\lambda_{max}} = 31.2677 \; hrs \approx 1 \text{ día y medio.}$$

La serie de tiempo es entonces determinista caótica y tiene un tiempo de predicción de 31.2677 hrs. Para hacer una predicción se pueden utilizar varios métodos, como el descrito en el Capítulo 2 de este trabajo.

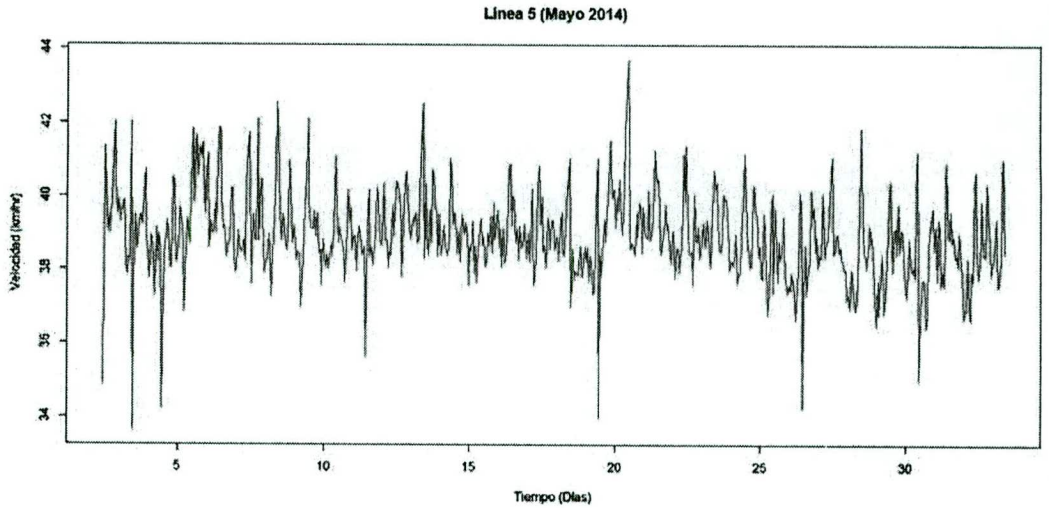
Los resultados de las técnicas de análisis son los siguientes:

Dinámica	Exponente máximo de Lyapunov
Tiempo de retardo	$\tau = 2$
Dimensión de inmersión	$m = 8$
Takens	$Y_t = \{x_t, x_{t-2}, \ldots, x_{t-14}\}$
Dimensión de correlación	$D_2 = 2.070023$
Exponente máximo de Lyapunov	$\lambda = 0.03198189$
Tiempo de predicción	$\Delta_{max} = 31.2677\ hrs$

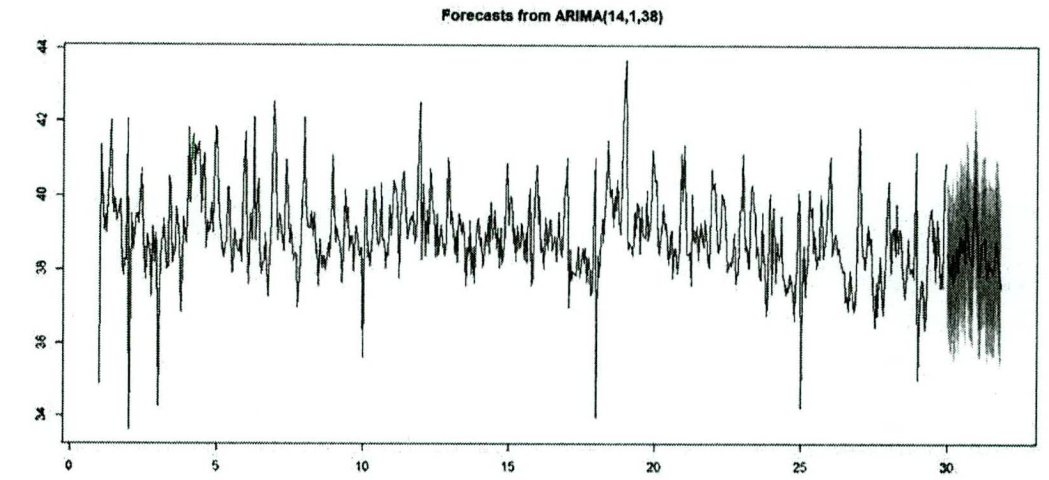
Cuadro 3.3: Dinámicas posibles de los sistemas estudiados y los correspondientes exponentes máximos de Lyapunov.

3.3 Predicción

Los paquetes que se utilizan para calcular la predicción de la serie de tiempo son *forecast*[31] y *tsDyn*[28] del lenguaje *R*.
Para poder checar el valor de la predicción, se necesita saber si la serie de tiempo obtenida es determinista o aleatoria y si es aleatoria si es estacionaria o no. De acuerdo a la sección anterior, la serie de tiempo es determinista caótica, pero si esta hubiera sido aleatoria y habiendo probado su estacionariedad, se recurriría a los métodos: autoregresivo integrado y de media móvil $ARIMA(p, d, q)$ o al estacional autoregresivo integrado y de media móvil $ARIMA(p, d, q)(P, D, Q)_s$, para pronosticar los siguientes 21 valores de la serie de 630 datos (21 horas por 30 días). El programa con el que se corren estos datos es también *proyectoTesis.R* Véanse las figuras 3.18 y 3.19.
La predicción para series de tiempo caóticas que se explicó en el segundo capítulo se muestra en la figura 3.20 y se realizó utilizando el programa *timeseriesforecastingusingchaos.R* (puede verse en el Apéndice A.2), realizado por el Dr. Daniel Fernández en *http://mechanicalforex.com.*

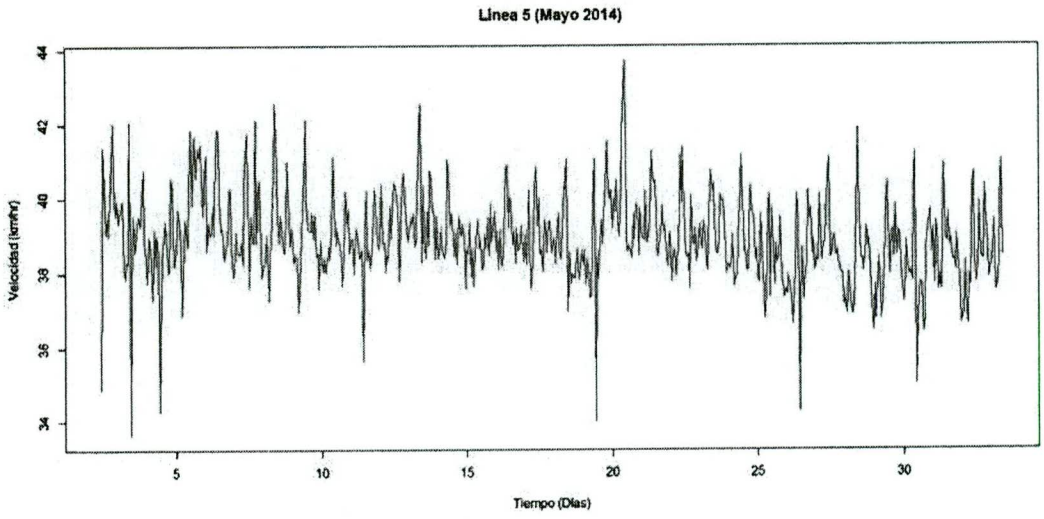


(a)

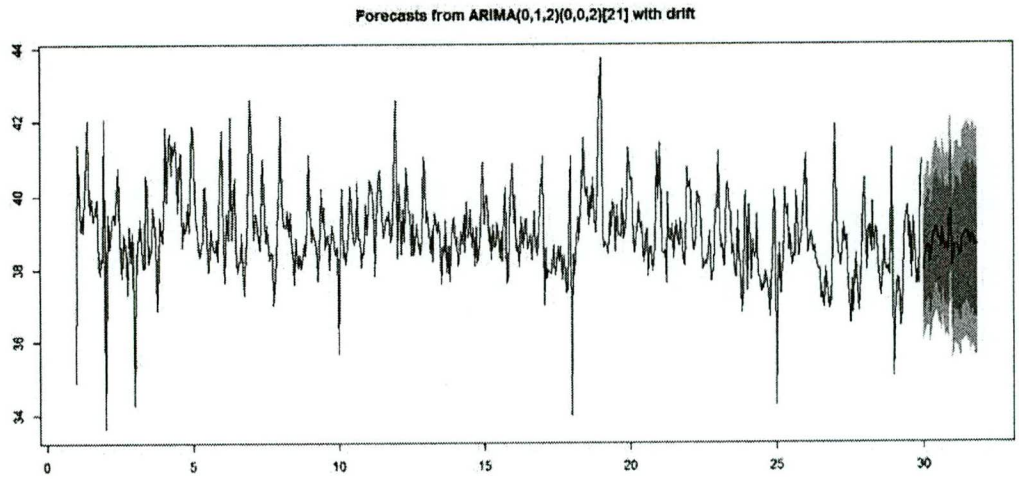


(b)

Figura 3.18: (a) Gráficas de la serie de tiempo y los residuos. (b) Gráficas del ACF de la serie de tiempo y de los residuos.

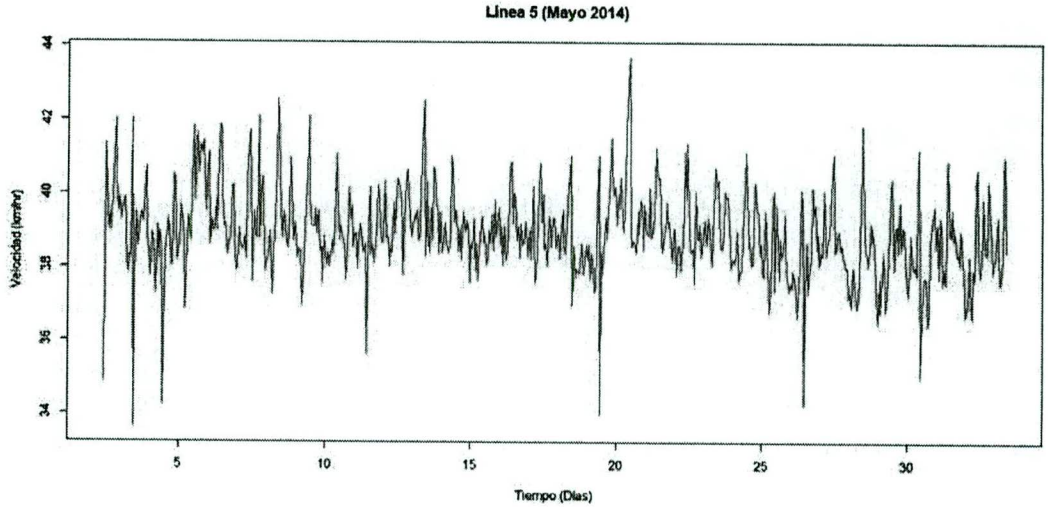


(a)

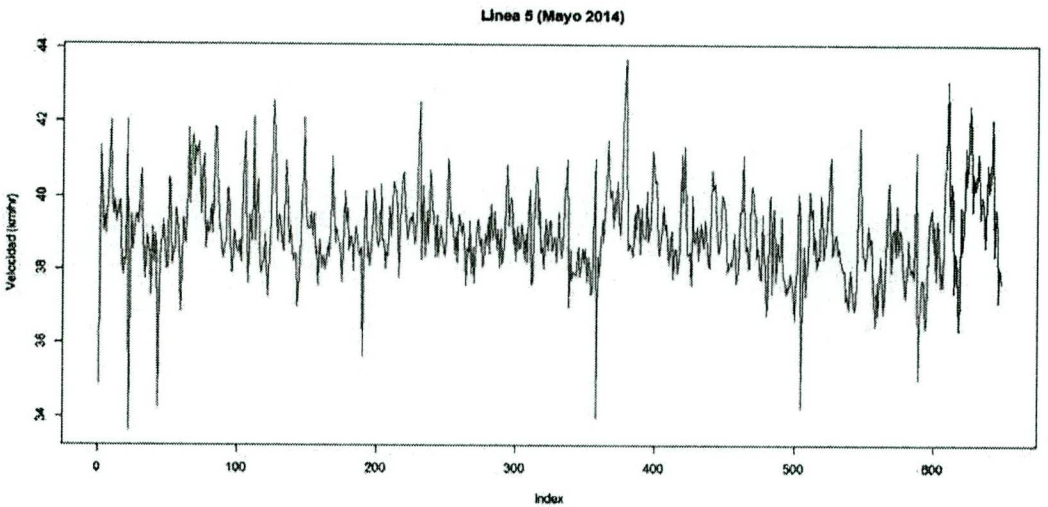


(b)

Figura 3.19: (a) Gráficas de la serie de tiempo y los residuos. (b) Gráficas del ACF de la serie de tiempo y de los residuos.



(a)



(b)

Figura 3.20: (a) Gráficas de la serie de tiempo y los residuos. (b) Gráficas del ACF de la serie de tiempo y de los residuos.

Conclusiones

Como se mencionó en la introducción de este trabajo, para explicar el funcionamiento —en términos de la velocidad— de la línea 5 del Metrobús de la Ciudad de México se recurrieron a las propiedades de las series de tiempo, a la estadística básica y a la teoría de inmersión o encajamiento de Takens. Durante los primeros dos capítulos se definieron y expusieron los métodos de análisis que se utilizarían en el tercer capítulo en donde se construyó la serie de tiempo buscada. En seguida se explicará lo que se realizó por capítulo y se irán exponiendo las conclusiones en cada caso.

En el primer capítulo de este trabajo se revisaron las series de tiempo para poder entender el contexto de lo que significa, los tipos de serie que existen y sus elementos, para que se utilizan e incluso de la predicciones que pueden obtenerse de acuerdo a las características que éstas tengan. Las definiciones y procedimientos de esta parte se retoman en el último capítulo para obtener una predicción de la serie construída.

El segundo capítulo de la tesis es fundamental puesto que de aquí derivan todas las técnicas de análisis caótico que se utilizan en la última parte del trabajo para analizar los datos de velocidad promedio y entender la dinámica de éstos. La inclusión de la versión original del Teorema de Takens y de sus demostraciones es importante, puesto que la forma en como se aborda nos da una idea geométrica clara de la reconstrucción del espacio fase. Los algoritmos descritos en la sección de técnicas de análisis se retoman uno a uno en el último capítulo y explican detalladamente como se obtuvieron los resultados, permitiendo entender claramente lo que las rutinas de *R* están calculando.

En el tercer capítulo, se utilizaron las mediciones de velocidad obtenidas por

los datos de la línea 5, del mes de mayo del 2014. Se construyó una sola serie de tiempo mensual y se hizo uso de las técnicas de análisis de series de tiempo caóticas para obtener el tiempo de predicción y los datos en ese periodo. La hipótesis principal de este trabajo, basada en el artículo [15], es cierta. Es decir, la serie de tiempo si es caótica, gracias a lo cual pudimos encontrar el horizonte de predictibilidad buscado.
Con respecto a la predicción, ésta nunca fue el objetivo principal de este trabajo de tesis, pero siempre es interesante saber que tipos de valores se obtienen con los distintos métodos estudiados. Además fue muy representativo observar las gráficas con los datos obtenidos para una serie estocástica por medio los métodos $ARIMA$ y por el modelo *GenericPred* de predicción de series de tiempo complejas; puesto que comparando las gráficas de $ARIMA(p, d, q)$ y la de la predicción caótica, ambas conservan la dinámica subyacente de la gráfica original, pero en la segunda si aparecen los valores más grandes (cercanos a los 43 km/h) y más pequeños (alrededor de 37 km/h) que se observan en la serie de tiempo real.

Las problemáticas enfrentadas durante todo el trabajo fueron del tipo computacional, de cotejo y de construcción de los datos como se explica a continuación:

- Hubiera sido recomendable tener más datos con los cuales trabajar porque, según Christophersen [4], los métodos existentes para estimar la dimensión de un atractor o de los exponentes de Lyapunov no se aplican adecuadamente a series de tiempo con menos datos, lo cual limita la información sobre las propiedades del sistema subyacente, tales como los grados de libertad y el nivel de complejidad. Es decir, la fiabilidad de los resultados puede ser sensible a la calidad y a la cantidad de los datos.

- Al comenzar a revisar los datos en bruto, existen muchos errores que probablemente se cometieron al llenar las hojas de excel o al momento de tomar las mediciones reales. Para un estudio netamente estadístico, como el del Lic. Moreno, en [23], lo anterior pueden repercutir en los resultados obtenidos, viendo a las series de tiempo por día o por semana. Sin embargo, en el caso de revisar la serie de tiempo mensual total, las repercusiones son mínimas.

- Con respecto a la construcción de datos de velocidad, fue difícil el

trato directo con las hojas de Excel que el Metrobús nos proporcionó, para ello la ayuda del lenguaje *Python* fue indispensable. Pero con el modulo usado resultó complicado el paso de información del lenguaje a la página, sobre todo cuando se trata de fórmulas y su resultado en tiempo real.

- El uso de las paqueterías de *R* fue igualmente muy conveniente, pero como en todo lenguaje y paqueterías, existen muchos problemas sino se conocen a fondo los algoritmos programados. En el caso de esta tesis, se tenían muchos otros métodos para resolver algunas de las técnicas descritas, pero se utilizaron las que venían precargadas en *R*.

Los resultados del análisis del funcionamiento del Metrobús son los siguientes:

- La velocidad promedio de la línea 5 del Metrobús oscila alrededor de los 38 km/h, valor cercano a los 40 km/h que señala la página oficial del Metrobús [37] y mucho mayor a la obtenida por el Lic. Moreno en [23] para todas las líneas del Metrobús. Lo cual corresponde a una avenida en la que los automóviles no pueden ir a una velocidad mayor a los 50 km/h, de acuerdo a las nuevas disposiciones de movilidad reglamentadas desde el 2015 por el Jefe de Gobierno Miguel Ángel Mancera, pues la avenida Eje 3 Oriente es una vía primaria según [35].

- Una vez que se limpian los datos erróneos (velocidades mayores a los 140 km/h, al sobreponer todas las series de tiempo semanales, nos fijamos que en la madrugada o en la noche (de 11 en adelante) la velocidad del Metrobús aumenta demasiado (de 120 a 140 km/h). Pero incluso en horas durante el día que no son muy difíciles, hay unidades que alcanzan dichas velocidades, lo cual es un peligro para la movilidad de usuarios, transeúntes y automóviles que pasen cerca de los carriles confinados (en la Introducción se menciona que en el mismo eje se encuentran otros medios de transporte).

- Se destaca que las hojas de excel de los días lunes y viernes tenían muchas más filas, lo cual indica que las unidades del Metrobús dan muchas más vueltas que los fines de semana por ejemplo.

- En las noches, los recorridos de las unidades son muy tardados debido al tránsito pesado y los tiempos de espera en la estación también aumentan.

- Ya que la serie obtenida es caótica, los resultados arrojados por el Teorema de Takens nos dan un valor de predicción de un día y medio aproximadamente y los datos obtenidos por las distintas técnicas son cualitativamente parecidos a los datos reales.

Lo anterior nos lleva a deducir que la línea 5 del Metrobús sigue siendo una muy buena opción de transporte urbano entre el Estado de México y la ciudad, no sólo porque en sus instalaciones y en sus autobuses se introdujeron sistemas que no dañan el medio ambiente (como se explicó en la Introducción), sino porque su velocidad se encuentra en el rango de parámetros impuestos en el nuevo Reglamento de Tránsito. Sin embargo, existen todavía muchas dificultades que pueden resolverse haciendo modificaciones en las estaciones, como las mencionadas también en la Introducción (construyendo otro carril de mayor velocidad y quitar o ajustar los semáforos para este transporte) y en la regulación de las velocidades en horas de poca circulación para evitar accidentes, de la misma forma en que se imponen castigos a los automóviles particulares. Con respecto a los valores de predicción encontrados, éstas técnicas nos pueden ayudar a conocer la dinámica del Metrobús de un día a otro y con ello ayudar a entender y tal vez proponer medidas en días y horas complicadas, como lo son los lunes y viernes y las horas de entrada y salida de escuelas y/o oficinas.

Apéndice A

Programas

A.1 Código en Python

```
#librerias Excel con Python
import openpyxl
import win32com.client
import pandas as pd
import numpy as np

#nombre del archivo
lista="ECL51405"

#funcion que publica las hojas de los libros de excel
def hojaslib(lib):
    wb=openpyxl.load_workbook(lib)
    print(wb.get_sheet_names())
    return

#ejecuta la funcion hojaslib en todos los libros de mayo
for m in range(1,32):
    wb=0
    lib=lista+str(m)+'.xlsx'
    print(lib)
    hojaslib(lib)

#funcion que limpia libros de excel
def quitarlib(lib):
    wb=openpyxl.load_workbook(lib)
    print(wb.get_sheet_names())
    if wb.sheetnames[2]=='D2-I18':
```

```
        print("yes")
29      wb.remove_sheet(wb['D2-118'])
        wb.save(lib)
31  else:
                print("no")
33  return

35 #ejecuta la funcion quitarlib en todos los libros de mayo
   for m in range(1,32):
37   wb=0
     lib=lista+str(m)+'.xlsx'
39   print(lib)
     quitarlib(lib)
41
   #calcular el numero maximo de filas de ambas hojas del libro
43 def calculofilas(lib):
         wb=openpyxl.load_workbook(lib)
45       ws=wb.worksheets[0]
         ws1=wb.worksheets[1]
47       M1=ws.max_row
         M2=ws1.max_row
49       print(M1)
         print(M2)
51       return

53 for m in range(1,32):
     wb=0
55   lib=lista+str(m)+'.xlsx'
     print(lib)
57   calculofilas(lib)

59 # funcion que realiza lo siguiente:
   # Crea una hoja de resultados en cada libro con todos los datos
       necesarios para obtener
61 # los tiempos de llegada y de salida entre las estaciones, de
       ida y de vuelta, por dia
   # Calcula la diferencia entre horas de llegada y salida, asi
       como la velocidad entre
63 # estaciones en metros/segundos
   def libros(lib):
65     wb=openpyxl.load_workbook(lib)
       wb2=openpyxl.load_workbook("distancias.xlsx")
67     ws3=wb2.worksheets[0]
       #wb.remove_sheet(wb['Resultados'])
69     ws2=wb.create_sheet(title='Resultados')
```

```
    ws=wb.worksheets[0]
    ws1=wb.worksheets[1]
    ws2=wb.worksheets[2]
    ws2['A1']='Llega'
    ws2['B1']='Sale'
    ws2['C1']='Tiempo'
    ws2['D1']='Distancia'
    ws2['E1']='V m/seg'
    ws2['F1']='V km/hr'
    M1=ws.max_row
    M2=ws1.max_row
    c1=int((M1-7)/4)
    c2=int((M2-7)/4)
    c=(c1+c2)*17
    print(M1,M2,c1,c2,c)
    if c2>=c1:
      for j in range(0,c2):
                                    for i in range(0,17): #Viaje de
Ida
                                        ws2.cell(row=i+2+34*j,
column=1).value=ws.cell(row=9+4*j,column=i+7).value
                                        ws2.cell(row=i+2+34*j,
column=2).value=ws.cell(row=11+4*j,column=i+6).value
                                        ws2.cell(row=i+2+34*j,
column=4).value=ws3.cell(row=i+1,column=1).value
                                    for i in range(0,17): #Viaje de
vuelta
                                        ws2.cell(row=i+19+34*j,
column=1).value=ws1.cell(row=9+4*j,column=i+7).value
                                        ws2.cell(row=i+19+34*j,
column=2).value=ws1.cell(row=11+4*j,column=i+6).value
                                        ws2.cell(row=i+19+34*j,
column=4).value=ws3.cell(row=i+1,column=2).value
                                    c_=i+19+34*j
     else:
                            for j in range(0,c1):
                                    for i in range(0,17): #Viaje de
Ida
                                        ws2.cell(row=i+2+34*j,
column=1).value=ws.cell(row=9+4*j,column=i+7).value
                                        ws2.cell(row=i+2+34*j,
column=2).value=ws.cell(row=11+4*j,column=i+6).value
                                        ws2.cell(row=i+2+34*j,
column=4).value=ws3.cell(row=i+1,column=1).value
```

```
                                        for i in range(0,17): #Viaje de
    vuelta
103                                                     ws2.cell(row=i+19+34*j,
    column=1).value=ws1.cell(row=9+4*j,column=i+7).value
                                                        ws2.cell(row=i+19+34*j,
    column=2).value=ws1.cell(row=11+4*j,column=i+6).value
105                                                     ws2.cell(row=i+19+34*j,
    column=4).value=ws3.cell(row=i+1,column=2).value
                                            c_=i+2+34*j
107     print(c,c_)
        for i in range(2,c_+1):
109       ws2['C'+str(i)]='=abs((A'+str(i)+'-B'+str(i)+')*86400)'
          ws2['E'+str(i)]='=D'+str(i)+'/C'+str(i)
111       ws2['F'+str(i)]='=E'+str(i)+'*(18/5)'
        wb.save(lib)
113     return

115 #ejecuta la funcion lib en todos los libros de mayo
    for m in range(1,32):
117   wb=0
      ws=0
119   ws1=0
      ws2=0
121   lib=lista+str(m)+'.xlsx'
      print(lib)
123   libros(lib)

125 #cuenta las filas o renglones de Resultados de cada libro
    def drow(lib):
127   wb=openpyxl.load_workbook(lib)
      ws = wb['Resultados']
129   M=ws.max_row
      print(M)
131   return

133 for m in range(1,32):
      wb=0
135   ws=0
      ws1=0
137   ws2=0
      lib=lista+str(m)+'.xlsx'
139   print(lib)
      drow(lib)
141
    ### UN DIA DE MAYO
```

```
#funcion que crea un libro nuevo en donde se vacian los datos de
    todos los libros de Excel y se
#obtiene la serie de tiempo
 def completar(lib):
  wb=openpyxl.load_workbook('serie.xlsx')
  ws=wb.worksheets[0]
  ws.cell(row=1, column=1).value='Llega'
  ws.cell(row=1, column=2).value='Sale'
  ws.cell(row=1, column=3).value='Tiempo'
  ws.cell(row=1, column=4).value='Distancia'
  ws.cell(row=1, column=5).value='Velocidad'
  wb2=openpyxl.load_workbook(lib)
  ws2=wb2.worksheets[2]
  for j in range(1,7):
      for i in range(2,10373):
        ws.cell(row=i+10370*(m-1),column=j).value=ws2.cell(row=i
    ,column=j).value
  wb.save('serie.xlsx')
  return

#ejecuta la funcion completar en todos los libros de mayo para
    formar la serie de tiempo
#velocidad promedio en mayo
for m in range(1,32):
  wb2=0
  ws2=0
  lib=lista+str(m)+'.xlsx'
  print(lib)
  completar(lib)

### SERIE DE TIEMPO
#funcion que crea un libro nuevo y se vacian los datos de todos
    los libros de Excel y se obtiene
#la serie de tiempo chica
def completar2(lib):
  wb=openpyxl.load_workbook('seriedtiempos.xlsx')
  ws=wb.worksheets[0]
  ws.cell(row=1, column=1).value='DÃas'
  ws.cell(row=1, column=2).value='Hora'
  ws.cell(row=1, column=3).value='Vel (km/hr)'
  wb2=openpyxl.load_workbook(lib)
  ws2=wb2.worksheets[5]
  wb3=openpyxl.load_workbook('dias.xlsx')
  ws3=wb3.worksheets[0]
  for i in range(1,32):
```

```
                    ws.cell(row=2+21*(i-1),column=1).value=ws3.
    cell(row=i,column=1).value
    for i in range(2,24):
        ws.cell(row=i+21*(m-1),column=2).value=ws2.cell(row=i,
    column=1).value
        ws.cell(row=i+21*(m-1),column=3).value=ws2.cell(row=i,
    column=2).value
    wb.save('seriedtiempos.xlsx')
    return

#ejecuta la funcion completar2
for m in range(1,32):
    wb2=0
    ws2=0
    lib=lista+str(m)+'.xlsx'
    print(lib)
    completar2(lib)

#quitar las celdas en blanco
def quitarcel():
    archivo='C:\\Users\\ALDI\\AppData\\Local\\Programs\\Python\\
    Python35\\seriet.xlsx'
    xl = win32com.client.DispatchEx('Excel.Application')
    wb = xl.Workbooks.Open(Filename=archivo)
    ws2=wb.worksheets[2]
    inirow = 1
    finrow = ws2.UsedRange.Rows.Count
    for row in range(inirow,finrow+1):
        if ws2.Range('A{}'.format(row)).Value is None:
            ws2.Range('A{}'.format(row)).EntireRow.Delete(Shift=-4162)
    wb.Save()
    wb.Close(True)
    xl.Quit()
    return

for m in range(1,32):
    wb2=0
    ws2=0
    lib=lista+str(m)+'.xlsx'
    print(lib)
    quitarcel(lib)

#Ordenar de forma descendente

rang = sheet.Range("A1","A4")
```

```
225 sheet.Sort.SetRange(rang)
    sheet.Sort.Apply()
227
    #Obtener dos columnas en la hoja serie, en la primera se observa
        el tiempo promedio y en la segunda la velocidad
229 def acomodar(serie):
            wb=openpyxl.load_workbook("serietiempo.xlsx")
231         ws=wb.worksheets[0]
            ws2=wb.create_sheet(title='serie')
233         ws2.cell(row=1,column=1).number_format = 'hh:mm:ss'
            for i in range(1,231851):
235                 ws['C'+str(i)]='=IF(((hour(A'+str(i)+')=0)*AND(
    hour(B'+str(i)+')=23)),abs((A'+str(i)+'"1"-B'+str(i)+')*
    86400),abs((A'+str(i)+'-B'+str(i)+')*86400))'
                    #ws['C'+str(i)]='=SI(((HORA(A'+str(i)+')=0)*Y(
    HORA(B'+str(i)+')=23)),abs((A'+str(i)+'"1"-B'+str(i)+')*
    86400),abs((A'+str(i)+'-B'+str(i)+')*86400))'
237                 ws['H'+str(i)]='=(A'+str(i)+'+B'+str(i)+')/2' #
    tiempo promedio
                    ws2.cell(row=i,column=1).value=ws.cell(row=i,
    column=8).value #tiempo promedio en hoja serie
239                 ws2.cell(row=i,column=2).value=ws.cell(row=i,
    column=6).value #velocidad km/hr en hoja serie
        ws.auto_filter.ref = 'A1:A231851' #ordenar columna con
    respecto al tiempo promedio
241     return
```

programatesis.py

A.2 Código en R

```
## LIBRERIAS
library(tseries)
library(deSolve)
library(tseriesChaos)
library(nonlinearTseries)
getwd()
## SERIE DE TIEMPO
datos<-read.table("st.txt",head=TRUE)
time.series<-ts(datos)
datos<-read.table("seriedtiempos.txt",head=TRUE)
View(datos)
time.series<-ts(datos,frequency=21,c(1,31))
plot.ts(time.series,xlab="Tiempo (Días)",ylab="Velocidad (km/hr)
    ",main="Línea 5 (Mayo 2014)",col=c("red"))
#plot.ts(time.series,xlab="Tiempo",ylab="Velocidad (km/hr)",main
    ="Línea 5 (Mayo 2014)",col=c("red"),xlim=c(260000,270763))

bds.test(time.series, m=8)
nLP<-nonLinearPrediction(time.series, embedding.dim=8, time.lag
    =2, prediction.step=40,
radius=5, radius.increment=1)

## DATOS ESTADÍSTICOS
mean(time.series)
sd(time.series)
var(time.series)
summary(time.series)
boxplot(time.series)
acf(time.series,type="covariance",plot=TRUE)
acf(time.series,plot=TRUE)
pacf(time.series,plot=TRUE)

## SERIE ESTACIONARIA, MÉTODOS
library(fpp)
Box.test(time.series,lag=50,type="Ljung-Box")
adf.test(time.series,alternative="stationary")
kpss.test(time.series)

## TIME LAG
time.lag<-timeLag(time.series,"acf",method="first.zero", lag.max
    =NULL,do.plot = TRUE)
time.lag
#1
```

```
2
## EMBEDDING DIM
embedding.dim<-estimateEmbeddingDim(time.series,number.points
    =651,time.lag,15,0.95,0.1,TRUE)
embedding.dim
#12
8
eps<-sd(time.series)/10
eps
fn<-false.nearest(time.series, 12, 1, t=100, rt=20, eps=0.1)
plot(fn)

## TAKENS
takens<-buildTakens(time.series,embedding.dim,time.lag=2)
plot(takens[,1],takens[,2],type='b',col='red')
plot3d(takens[,1],takens[,2],takens[,3],type='l',col='red')
seriesdetiempo<-ts(takens)
plot(seriesdetiempo,type='l',col='red')

## SERIE ESTACIONARIA
m_=list()
v_=list()
sd_=list()
for(i in 1:12)
m_[i]<-mean(takens[,i])
for(i in 1:12)
sd_[i]<-sd(takens[,i])
for(i in 1:12)
v_[i]<-var(takens[,i])
print(m_)
print(v_)
print(sd_)

## DIM CORR
DimCorr<-corrDim(time.series, min.embedding.dim = 2,max.
    embedding.dim=embedding.dim, time.lag, min.radius=1,max.
    radius=5, corr.order = 2, n.points.radius = 5,theiler.window
    = 100, do.plot = TRUE,number.boxes = NULL)
estimDimCorr<-estimate(DimCorr,regression.range=NULL,do.plot=
    TRUE)
estimDimCorr
2.070023

## EXP LYAPUNOV
```

```
ExpLyap<-maxLyapunov(time.series,min.embedding.dim=2,max.
    embedding.dim=embedding.dim,time.lag,radius=5,theiler.window
    =100)
ExpLyap
estimExpLyap<-estimate(ExpLyap,regression.range = NULL,use.
    embeddings=2:8,do.plot=TRUE)
estimExpLyap
0.03198189
## NONLINEAR PREDICTION
nlP<-nonLinearPrediction(time.series=time.series[609:651],
    embedding.dim=11, time.lag=2,prediction.step=1, radius=5,
    radius.increment=1)
plot(nlP)
cat("real value: ",time.series,"Vs Forecast:",nlP)
38.93801 hrs

## TABLA LATEX
library(xtable)
#tabla<-read.table("sttyv.txt")
tabla<-read.table("summary1.txt",head=TRUE,sep=",")
print(xtable(tabla),include.rownames=FALSE)

## PREDICCIÓN ARIMA
require(forecast)

datos<-read.table("seriedtiempos.txt",head=TRUE)
time.series<-ts(datos,frequency=21,c(1,31))
plot(time.series,xlab="Tiempo",ylab="Velocidad (km/hr)",main="
    Línea 5 (Mayo 2014)",col=c("red"))
plot(time.series,col=c("red"),xlim=c(28,32))

View(time.series)

datos2<-read.table("seriedtiempos2.txt",head=TRUE)
time.series2<-ts(datos2,frequency=21)
plot(time.series2)
View(time.series2)
acf(time.series2,plot=FALSE)
pacf(time.series2,plot=FALSE)

fit<-Arima(time.series2,order=c(14,1,38))
forc<-forecast(fit,h=40)
plot(forc,include=40)
plot(forc)
```

```
fit2<-auto.arima(time.series2)
forc2<-forecast(fit2,h=40)
plot(forc2,include=40)

## PREDICCIÓN SETAR
library(tsDyn)
#Test for nonlinearity
nonlinearityTest(time.series, verbose = TRUE)

mod.set <- setar(time.series2, m=8, d=2)
plot(mod.set)
fit3<-predict(mod.set,n.ahead=32)
plot(fit3)
plot(time.series2,xlim=c(28,32))
lines(fit3,type="l",col="blue")
```

proyectoTesis.R

```
library(quantmod)
library(fractaldim)

#Programmed by Dr. Daniel Fernandez 2014-2016
#https://Asirikuy.com
#http://MechanicalForex.com

datos<-read.table("seriedtiempos.txt",head=TRUE)
time.series<-ts(datos,frequency=21,c(1,31))
endingIndex<-609
time.series_TEST<-time.series[1:endingIndex]
View(time.series_TEST)

#These are the fractal dimension calculation parameters
#see the fractaldim library reference for more info

method <- "rodogram"

#number of samples to draw for each guess
random_sample_count <- 50

Sm<-0
Sm_prediction<-0
Sm<-as.data.frame(time.series_TEST,row.names=NULL)
#View(Sm)
#do 40 predictions of next values in Sm
for(i in 1:40){
  delta <- c()

  # calculate delta between consecutive Sm values to use for the
  # building of the normal distribution to draw guesses
  for(j in 2:length(Sm$time.series_TEST)){
    delta <- rbind(delta, Sm$time.series_TEST[j]-Sm$time.
    series_TEST[j-1])
  }

  # calculate standard deviation of delta
  Std_delta <- apply(delta, 2, sd)

  #update fractal dimension used as reference
  V_Reference <- fd.estimate(Sm$time.series_TEST, method=method,
    trim=TRUE)$fd
  V_Reference
  # create N guesses drawing from the normal distribution
  # use the last value of Sm as mean and the standard deviation
```

```
    # of delta as the deviation
45  Sm_guesses <- rnorm(random_sample_count , mean=Sm$time.
      series_TEST[length(Sm$time.series_TEST)], sd=Std_delta )

47  minDifference = 1000000

49  # check the fractal dimension of Sm plus each different guess
      and
    # choose the value with the least difference with the
      reference
51  for(j in 1:length(Sm_guesses)){
      new_Sm <- rbind(Sm, Sm_guesses[j])
53    new_V_Reference <- fd.estimate(new_Sm$time.series_TEST,
      method=method, trim=TRUE)$fd

55    if (abs(new_V_Reference - V_Reference) < minDifference ){
        Sm_prediction <- Sm_guesses[j]
57      minDifference = abs(new_V_Reference - V_Reference)
      }
59  }

61  print(i)
    #add prediction to Sm
63  Sm <- rbind(Sm, Sm_prediction)
}
65

67 plot(Sm$time.series_TEST, type="l")
   lines(as.data.frame(time.series_TEST[1:(endingIndex+40)], row.
     names = NULL), col="red")
```

timeseriesforecastingusingchaos.R

Bibliografía

[1] Brockwell, P., David, R. (2002): *Introduction to time series and forecasting*. Springer. E.U.A.

[2] Cao, L. (1997): Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena*, **110(1-2)**, 43-50

[3] Casdagli, M. (1991): Chaos and deterministic versus stochastic nonlinear modeling. Santa Fe Institute. *Journal of the Royal Statistical Society.* Series B. **54** (2): 303?328.

[4] Christophersen, N., Kugiumtzis, D., Lillekjendlie, B. (1994): Chaotic time series. Part I: Estimation of invariant properies in state space. *Modeling, Identification and Control.* **15** (4): 205-224.

[5] Chua, L., Parker, T. (1989): *Practical Numerical Algorithms for Chaotic Systems*. Springer-Verlag. New York, E.U.A.

[6] Coghlan A. (2015): *A little book of R for time series*. Cambridge, Reino Unido

[7] Golestani, A., Jahed Motlagh, M.R., Ahmadian, K., Omidvarnia, A.H., Mozayani, N. (2009): A new criterion for distinguish stochastic and deterministic time series with the Poincaré section and fractal dimension. *Chaos* **19,** 013137

[8] Golestani, A., Gras, R. (2014): Can we predict the unpredictable? *Nature: Scientific reports* **4:** 6834

[9] Huke, J.P. (2006): Embedding Nonlinear Dynamical Systems: A Guide to Taken's Theorem. Manchester Institute for Mathematical Sciences School of Mathematics.

[10] Jänich, K., Bröcker, T.H. (2007): *Introduction of differential topology.* Cambridge University Press. Reino Unido

[11] Kantz H., Schreiber, T. (2003): *Nonlinear Time Series Analysis.* Cambridge University Press. Reino Unido

[12] Landa, P.S. (1996): Attractors and repellers. Reconstruction of attractors from an experimental time series. Quantitative characteristics of attractors. *Nonlinear Oscillations and Waves in Dynamical Systems.* Series Mathematics and Its Applications **360**: 22-27

[13] Nava, A. (2002): *Procesamiento de series de tiempo.* Fondo de Cultura Económica. Ciudad de México

[14] Novales, A. (1993): *Econometría.* Mc. Graw Hill /INTERAMERICA S.A. España. Segunda edición.

[15] Shang, P., Li, X., Kamae, S. (2004): Chaotic analysis of traffic time series. *Chaos, Solitons and Fractals.* Elsevier. **25**, 121-128

[16] Shumway, R., Stoffer, D. (2011): *Time series analysis and its applications. With R examples.* Springer. E.U.A.

[17] Siek, M. (2011): *Predicting storm surges. Chaos, computational intelligence, data assimilation, ensembles.* CRC Press. Holanda

[18] Strogatz, S. (1994): Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. Perseus Books Publishing, LLC.

[19] Takens, F. (1981): Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Lecture Notes in Mathematics.* **898**, 366:381.

[20] Zhang, J. Zhang, P. (2017): *Time Series Analysis Methods and Applications for Flight Data.* Springer. Berlín, Alemania

Tesis:

[21] Bermejo, M.A. (2011): *Métodos estadísticos en series temporales no lineales, con aplicación a la predicción de energía eólica.* (Tesis doctoral de Estadística). Universidad Carlos III de Madrid. Madrid, España

[22] García, C.J. (1993): *La teoría del caos: algunas implicaciones en el área de la metodología de la ciencia* (Tesis para obtener el grado de Maestría en metodología de la ciencia). Universidad Autónoma de Nuevo León. Monterrey, N.L.

[23] Moreno, L. (2011): *Metrobús: estudio de caso.* (Tesis para obtener el título de Licenciado en Economía). Universidad Autónoma Metropolitana Azcapotzalco. México, D.F.

Libros electrónicos:

[24] Prof. Rafael de Arce. Prof. Ramón Mahía. Modelos Arima. Dpto. de Economía Aplicada. Programa Citius. Técnicas de previsión de variables financieras [en línea] [fecha de consulta: 2016-08-05]. Disponible en: $< https : //www.uam.es/personal\_pdi/economicas/anadelsur/pdf/-Box - Jenkins.PDF >$

[25] PhD Daniel Fernandez. Using R in Trading: Time series forecasting using chaos, Part 1. MECHANICAL FOREX [en línea] 2016 [fecha de consulta: 2016-08-05]. Disponible en: $< http : //mechanicalforex.com/2016/03/using - r - in - trading - time - series - forecasting - using - chaos - part - 1.html >$

[26] John Villavicencio. Introducción a las series de tiempo. [en línea] [fecha de consulta:2016-08-05]. Disponible en: $< http : //www.estadisticas.gobierno.pr/icpr/LinkClick.aspx?-fileticket = 4\_BxccUaZmg\%3D¿$.

Manuales electrónicos de algunas paqueterías de R:

[27] Antonio, Fabio Di Narzo. Analysis of nonlinear time series. [en línea] 2013-04-29 16:08:47. [fecha de consulta: 2016-06-13]. Disponible en: $< https : //cran.r - project.org/web/packages/-tseriesChaos/tseriesChaos.pdf >$

[28] Antonio Fabio Di Narzo. Nonlinear Time Series Models with Regime Switching. [en línea] 2016-05-22 22:56:44. [fecha de consulta: 2016-07-

04]. Disponible en:
< *https* : *//cran.r − project.org/web/packages/tsDyn/tsDyn.pdf* >

[29] Constantino A. Garcia. Nonlinear Time Series Analysis. [en línea] 2015-07-25 17:43:38. [fecha de consulta: 2016-06-13].
Disponible en: < *https* : *//cran.r − project.org/web/packages/-nonlinearTseries/nonlinearTseries.pdf* >

[30] Kurt Hornik. Time Series Analysis and Computational Finance. [en línea] 2016-05-02 13:58:30. [fecha de consulta: 2016-06-13]. Disponible en:
< *https* : *//cran.r − project.org/web/packages/tseries/tseries.pdf* >

[31] Rob Hyndman. Forecasting Functions for Time Series and Linear Models. [en línea] 2016-04-14 14:53:40. [fecha de consulta: 2016-06-13].
Disponible en: < *https* : *//cran.r − project.org/web/packages-/forecast/forecast.pdf* >

Periódicos digitales:

[32] Detter Hanzel Forteza Maya. En la CDMX se pierde hasta hora y media en transporte público. [en línea] 2016-12-13 [fecha de consulta: 2016-12-15]. Disponible en:
< *https* : *//elsemanario.com/metropoli/*169765*/la − cdmx − se − pierde − hora − media − transporte − publico/* >

[33] Claudia Solera y Laura Toribio. Se les va la vida trasladarse a diario al DF. [en línea] 2011-06-27 [fecha de consulta: 2016-12-15]. Disponible en:
< *http* : *//www.excelsior.com.mx/*2011*/*06*/*27*/nacional/*747939 >

[34] Ilich Valdez. El Metrobús de Bogotá, menos camiones por un mejor servicio. [en línea] 2015-11-19 [fecha de consulta: 2016-12-15].
Disponible en:
< *http* : *//www.milenio.com/df/metrobus*$_{B}$*ogotaColombia-saturacion_Metrobus_DFTransmilenio_Bogota_*0*_*631137010*.html/* >

[35] Redacción El Universal. GUÍA Límites de velocidad en avenidas de la Ciudad de México. [en línea] 2015-12-25 [fecha de consulta: 2016-12-15].
Disponible en:

< *http* : *//www.eluniversaldf.mx/home/guia* − *limites* − *de* − *velocidad* − *en* − *avenidas* − *de* − *la* − *cdmx.html* >

Páginas de internet:

[36] Metrobús de la Ciudad de México. [en línea] 2016-12-15 [fecha de consulta: 2016-12-15].
Disponible en: < *http* : *//www.metrobus.cdmx.gob.mx* >

[37] Metrobús (Ciudad de México). [en línea] 2016-12-15 [fecha de consulta: 2016-12-15]. Disponible en:
< *https* : *//es.wikipedia.org/wiki/Metrob%C3%BAs-_(Ciudad_de_M%C3%A9xico)* >

[38] Metrobús, décimo aniversario. Líneas 5 y 6. [en línea] 2016-12-15 [fecha de consulta: 2016-12-15]. Disponible en:
< *http* : *//www.metrobus.cdmx.gob.mx/docs/libro/MB10*$_p$*6.pdf* >

[39] Villavicencio, J. Introducción a series de tiempo [en línea] 2016-12-15 [fecha de consulta: 2016-12-15]. Disponible en:
< *http* : *//www.estadisticas.gobierno.pr/iepr/LinkClick.aspx?-fileticket* = *4_BxccUaZmg%3D* >

[40] Ciudad de México. [en línea] 2016-12-15 [fecha de consulta: 2016-12-15]. Disponible en:
< *https* : *//es.wikipedia.org/wiki/Ciudad_de_M%C3%A9xico* >

Diseño en tesis

Arquitectura #56 Local 4
Copilco Universidad
Delegación Coyoacán
México D.F. C.P. 04360
Email: tesisenarquitectura@hotmail.com
Tel. 53 39 60 53
Diseño en Tesis