UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

# DETECCIÓN DE ANOMALÍAS EN UN PROCESO DE CARGUÍO AUTÓNOMO

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL ELÉCTRICA

MARÍA FERNANDA FICA TAPIA

PROFESOR GUÍA:
MARCOS ORCHARD CONCHA

MIEMBROS DE LA COMISIÓN:
ANDRÉS CABA RUTTE
CARLOS TAMPIER COTORÁS

SANTIAGO DE CHILE
2018

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL ELÉCTRICA
POR: MARÍA FERNANDA FICA TAPIA
FECHA: 2018
PROF. GUÍA: MARCOS ORCHARD CONCHA

# DETECCIÓN DE ANOMALÍAS EN UN PROCESO DE CARGUÍO AUTÓNOMO

La automatización de la maquinaria dentro de la minería es una tendencia que ha cobrado mayor importancia con el transcurso del tiempo, principalmente por el ambiente peligroso e impredecible en el que deben ser manejadas, lo que genera una serie de riesgos para los operadores manuales de los equipos. Lo anterior sumado a los potenciales beneficios económicos por aumento del factor de utilización, reducción del desgaste de los componentes y operación eficiente de la máquina, hacen del concepto de la minería autónoma una promesa interesante.

El objetivo de este trabajo es crear un detector de anomalías para el proceso de carguío autónomo de maquinaria minera, en donde la anomalía se define como la necesidad de impactar más de una vez contra la pila de recolección para llenar el balde del cargador. Para esto se utilizaron los datos de operación obtenidos de un cargador LHD utilizado en la minera 'San Gerónimo'.

Debido a la complejidad física de estos sistemas, se utilizará un modelo no paramétrico para la modelación del proceso, el cual corresponde a un modelo basado en similitud (SBM de sus siglas en inglés). Se creará una metodología para obtener un modelo que contengan las variables explicativas del proceso y con el que se pueda generar una rutina de detección que avise al operario en caso de que las señales medidas mientras carga sean similares a las caracterizadas en el modelo SBM como anomalía.

La metodología propuesta en este trabajo fue probada con los datos obtenidos en la minera, con los cuales se obtuvo un error de un 7,5 % en la detección de carguíos anómalos. También se probó la metodología con datos de carguío manual, en donde un operador manejaba en la mina el cargador de forma remota. En este caso los resultados no fueron tan positivos como en el caso autónomo, en donde se consiguió que el detector errara en promedio un 23,3 % de las veces que fue probado.

Como principales conclusiones, se verifica el uso de modelos no paramétricos para la caracterización de procesos multivariables, sin embargo la metodología desarrollada sólo es aplicable a la maquinaria estudiada. También se destaca la importancia de datos variados para la creación de los modelos, es decir, que contengan diferentes puntos de operación del proceso. En particular en el caso de los carguíos autónomos, existe una mayor complejidad para caracterizar el carguío, dado que se necesita recolectar los datos de muchos operarios para tener una matriz de entrada rica en información, no así en el caso autónomo donde el hecho de utilizar una secuencia programada reduce la cantidad de datos necesarios para la modelación.

# Agradecimientos

En primer lugar, me gustaría agradecerle a mi familia por el apoyo durante todos los años de estudio. En particular a mi padre y mi madre por darme ánimos y a mi hermana por hacerme pasar rabias, pero también hacerme reír para olvidarlas rápidamente.

A mi profesor guía, Marcos O. por el tiempo y conocimiento entregados, tanto en las salas de clase como durante el transcurso del trabajo de título. A los miembros de mi comisión, Andrés C. por los consejos entregados a lo largo de los últimos años de mi carrera y Carlos T. por la paciencia y disposición para ayudarme durante este trabajo.

También me gustaría agradecer a los miembros del Laboratorio de Robótica de Campo por su gran disposición para responder mis dudas durante todo el trabajo de título.

Les agradezco a mis amigos eléctricos, Alejandro y Matías por todo el tiempo compartido en el departamento, y a mis amigos 'del otro lado' Vicente, Juan Pablo, Ricardo y Hugo [1].

Por último, pero no menos importante, a Arturo por todo el ánimo que me ha entregado estos años, los buenos momentos compartidos y por siempre creer en mí mucho más de lo que yo jamás logré en algunos momentos.

[1] Ahora que hago esto, me arrepiento de no haber hecho más amigas mujeres.

# Tabla de Contenido

# Índice de Tablas

# Índice de Ilustraciones

# Capítulo 1

# Introducción

## 1.1. Motivación

La automatización de la maquinaria dentro de la minería es una tendencia que ha cobrado mayor importancia con el transcurso del tiempo, principalmente por el ambiente peligroso e impredecible en el que deben ser manejadas, lo que genera una serie de riesgos para los operadores manuales de los equipos. Lo anterior sumado a los potenciales beneficios económicos por aumento del factor de utilización, reducción del desgaste de los componentes y operación eficiente de la máquina, hacen del concepto de la minería autónoma una promesa interesante. En este trabajo en particular se trabajará con un cargador LHD (*Load-Haul-Dump*) autónomo, el cual es utilizado para el transporte de material en las minas subterráneas.

Los cargadores LHD son maquinaria diseñada para cargar material o roca fragmentada desde un punto de extracción y transportar éste a una zona de recolección. El transcurso de esta tarea contiene varios imprevistos que afectan al funcionamiento de la máquina, tales como rocas demasiado grandes para ser extraídas por la pala, poca iluminación para ver el lugar de impacto del cargador o sencillamente puntos de extracción con una geometría compleja para un choque efectivo de la pala, entre otros factores. Todo esto genera que la extracción de material no sea tan sencilla como se desearía y que se den situaciones en las cuales el cargador LHD no es capaz de extraer suficiente material luego de un único impacto con la pila. Estas situaciones pueden ser consideradas anómalas y ser evitadas mediante su oportuna detección.

La detección de anomalías juega un rol importante en el cuidado de los componentes y en la operación eficiente, tanto para el caso de los cargadores autónomos como de la maquinaria presente en otras industrias. Poder detectar cuando la operación de la máquina escapa de su condición normal permite detener a tiempo su funcionamiento y así protegerla de situaciones donde se pueda exigir más de lo permitido sus componentes y que, de ser sometidas constantemente a estas situaciones, generen un daño que amerite ser llevadas al servicio técnico y por ende perder tiempo de funcionamiento; o, en casos aún más extremos, deriven a una falla catastrófica y la pérdida total del equipo. Así mismo, detener el proceso ante la presencia de una anomalía y reanudarlo una vez que la máquina se encuentra dentro de

sus condiciones normales puede repercutir en la eficiencia general, para este caso en específico, de extracción de mineral.

Con respecto a los equipos utilizados para el monitoreo en línea y detección de anomalías, el crecimiento tecnológico en los últimos años ha conseguido que la capacidad computacional incremente considerablemente, permitiendo implementar técnicas más avanzadas para estas labores. Adicionalmente, con los recursos computaciones actuales, es posible realizar modelaciones no paramétricas; las cuales no dependen del conocimiento fenomenológico del sistema a analizar, y que mediante los datos medidos pueden obtener una representación matemática de éste, lo que los convierte en una herramienta muy potente para el análisis de sistemas multivariados y no lineales.

Gracias a lo expuesto anteriormente, se abre la posibilidad de poder analizar sistemas de alta complejidad en un menor tiempo gracias al análisis de datos y poder implementar sistemas de monitoreo que alerten cuando su funcionamiento salga de la zona definida como operación normal, aumentando la vida útil de los componentes y repercutiendo en la eficacia de los procesos.

## 1.2. Alcances

El presente trabajo está dentro del marco de un proyecto de *GHH Chile SpA* en conjunto con el *Advanced Mining Technology Center (AMTC)* y el Departamento de Ingeniería Eléctrica de la Universidad de Chile. El proyecto consiste en la automatización de un cargador LHD, en donde parte de ésta es un sistema de detección en línea de anomalías utilizando modelos basados en similitud (SBM). En este contexto, se definirá anomalía como la situación en donde el cargador no es capaz de recolectar suficiente material luego de un único impacto contra la pila.

Para este proyecto se ha utilizado datos recolectados por un cargador perteneciente a la empresa *GHH Chile SpA* y que fue modificado por el equipo del Laboratorio de Robótica de Campo e ingenieros de GHH. Los datos corresponden a la operación del cargador en las instalaciones de la minera 'San Gerónimo', ubicadas en La Serena.

## 1.3. Objetivos

El objetivo principal de este trabajo es la utilización de modelos basados en similitud para crear un detector de anomalías capaz de alertar al usuario en tiempo real mientras se realiza el proceso de carga del LHD provisto por GHH. En este caso, la anomalía se definirá como el momento en que las señales obtenidas del cargador indican que no se podrá recolectar una cantidad de roca adecuada, siendo conveniente detener el proceso y volver a comenzar.

Para poder lograr esta tarea, se proponen objetivos intermedios que en su conjunto permiten la implementación de éste, los cuales corresponden a:

- Estudio del proceso de carga y selección de sus variables relevantes.
- Diseño de una metodología de estudio de eventos para generar modelos que detecten las anomalías registradas en la operación.
- Creación preliminar de un modelo SBM para el proceso de carguío.
- Diseño de una rutina de detección para la implementación en línea.

## 1.4. Estructura del documento

Esta memoria consta de 5 capítulos. En los capítulos 1 y 2 se realiza una pequeña introducción y motivación, además de explicar los conceptos necesarios para contextualizar al lector en el área en que se desarrolla este trabajo, los diferentes métodos que han sido utilizados para resolver estas problemáticas y ayudar a comprender las herramientas empleadas en el desarrollo del detector de anomalías.

El Capítulo 3 describe la implementación de las herramientas utilizadas tanto para crear un modelo SBM para el proceso de carguío autónomo como para la creación de un detector de eventos anómalos. Se detalla la metodología desarrollada para el estudio de anomalías, el algoritmo para la generación de un modelo SBM, cómo validarlo y su aplicación a una rutina de detección en línea.

El Capítulo 4 presenta los resultados obtenidos y el análisis de éstos al probar la metodología descrita en el capítulo anterior con los datos obtenidos en la minera 'San Gerónimo', tanto en la generación del modelo SBM como en el uso de la rutina de detección de anomalías.

Por último en el Capítulo 5 se encuentran las conclusiones del trabajo realizado y propuestas de trabajo futuro nacidas a partir de los resultados obtenidos.

# Capítulo 2

# Revisión bibliográfica

En el presente capítulo se describen y entregan antecedentes previos para contextualizar al lector en la temática abordada por esta memoria. En la Sección 2.1 se encuentran explicados los conceptos básicos relacionados a sistemas de supervisión y detección de fallas o anomalías. En la Sección 2.2 se explica el problema de detección de anomalías, sus dificultades, modos de clasificación y técnicas para abordarlo, profundizando en la Sección 2.3 sobre el método de residuos, donde en las Sección 2.4 se describen métodos de generación de éstos. En la Sección 2.5 se habla brevemente de los casos en los que no se posee información sobre la salida del proceso y cómo detectar anomalías a pesar de ello. En la Sección 2.6 se explica técnicas de validación de modelos y cómo calcular el error de éstos. Finalmente en la Sección 2.7 se encuentran las herramientas utilizadas para el análisis de datos.

## 2.1. Conceptos generales

En este apartado se presentan los conceptos básicos de sistemas de monitoreo, detección y diagnóstico tanto de fallas como de anomalías, los cuales son necesarios para comprender el área en el cual se enmarca este trabajo. También se encuentran las definiciones básicas sobre minería subterránea para contextualizar en qué fue aplicada esta memoria.

### 2.1.1. Conceptos sobre detección y diagnóstico

Los conceptos definidos a continuación pueden ser encontrados en [5], [7], [24], y de forma más resumida en [22]. Corresponden a la terminología utilizada en sistemas de detección y pronóstico, además de terminología la cual es útil para comprender en qué son aplicados los métodos y sistemas mencionados posteriormente.

- **Anomalía:** Patrones encontrados en los datos medidos de un proceso que difieren de su comportamiento normal. Las anomalías pueden caer en distintas categorías como puntuales, contextuales, colectivas, etc.

- **Falla:** Una desviación no permitida, con respecto a la condición usual o normal de funcionamiento, de una de las propiedades características o parámetro del sistema.
- **Evento crítico:** Una interrupción permanente al sistema, deshabilitando a este para seguir su operación normal.
- **Mal funcionamiento:** Una irregularidad intermitente en el desempeño esperado de alguna función del sistema.
- **Perturbación:** Una entrada externa actuando en el sistema que genera una desviación del estado actual de éste.
- **Síntoma:** Un cambio en una variable observada con respecto a su estado normal.
- **Residuo:** Un indicador de fallas o anomalías, basado en la desviación entre mediciones y valores basados en un modelo del proceso.
- **Observación:** Conjunto de mediciones de las variables controladas y manipuladas del proceso.
- **Sistema de monitoreo:** Un sistema a tiempo real determinando la condición física del sistema, guardando datos, reconociendo y alertando sobre las anomalías en un proceso.
- **Tiempo real:** En la industria se considera que un sistema de monitoreo trabaja a tiempo real si éste es capaz de procesar la información y dar una alerta en un tiempo menor o cercano al tiempo de muestreo del proceso.

### 2.1.2. Conceptos sobre minería y proceso de carguío

Un LHD (Figura 2.1), de sus siglas en inglés, *Load-Haul-Dump*, es un vehículo de cuatro ruedas con una articulación central accionada hidráulicamente y un mecanismo de pala con dos grados de libertad ubicado en el frente, como un cargador frontal, pero de bajo perfil, lo que le permite ser utilizado en espacios reducidos. Su función consiste en recoger material (*load*), típicamente de propiedades granulares (como tierra, piedras o roca fragmentada), transportarlo (*haul*) y depositarlo (*dump*) en un lugar de destino. Es utilizado comúnmente en la minería subterránea.

Figura 2.1: Cargador LHD. Imagen obtenida de [20].

El *proceso de carguío* al que se refiere este trabajo consiste en la recolección del material y su posterior depósito en la zona de destino. Los *puntos de extracción* corresponden a zonas

donde se estima que hay material de interés (como cobre, por ejemplo) que son dinamitadas para fracturar la roca en pedazos de menor tamaño para luego ser recolectados por medio del LHD.

En particular el carguío autónomo corresponde a cuando el cargador sigue una secuencia programada de acciones, entre las cuales se incluye el choque contra la pila de recolección y el movimiento del balde para empujar la roca dentro de éste. Estas acciones producen una serie de cambios en variables de la máquina, tales como las revoluciones por minuto del motor, la presión hidráulica de la pala, el ángulo de la pala a lo largo del tiempo, el derrape de las ruedas, entre muchas otras variables, provocando que este proceso sea difícil de modelar fenomenológicamente dado a la continua interacción de muchas variables.

Adicionalmente, una serie de condiciones, tanto relativas a la maquinaria, como al punto de extracción, afectan el grado de dificultad que conlleva realizar un carguío. Entre los factores que tienen relación con el LHD se encuentran el estado de las cuchillas del balde, la presión del aceite hidráulico y desgaste general del equipo de acuerdo a sus horas de servicio. Respecto del punto de extracción se pueden tener problemas dependiendo del contenido de humedad y tamaño del mineral, entre otros factores; lo cual sólo dificulta la tarea de una extracción exitosa.

## 2.2. Detección de anomalías

El análisis y detección de anomalías se basa principalmente en definir una zona o región de operación normal y considerar anómalo a las observaciones fuera de ésta. La detección de las anomalías sirve para cuidar del estado del equipamiento, prevenir fallas catastróficas y maximizar el beneficio económico del proceso.

Esa labor puede dificultarse debido a que muchas veces no es sencillo definir la barrera entre un comportamiento normal y uno anómalo, las observaciones contienen ruido que puede ser similar a una anomalía, el comportamiento normal de un proceso evoluciona en el tiempo y la distinción entre una condición normal y una anómala puede variar a lo largo del tiempo, entre otros motivos [7].

Para la detección y clasificación de anomalías es necesario tener previamente una base de datos que registre la operación del proceso, en donde los datos poseen una etiqueta refiriendo si lo que representan es un modo normal de operación o una anomalía y el tipo de ésta. Dependiendo de la disponibilidad de las etiquetas, las técnicas de detección de anomalías pueden operar en alguno de los tres siguientes modos:

- *Detección supervisada de anomalías*: En este caso en los datos de entrenamiento se tienen etiquetas claras de cuándo el proceso se encuentra en su condición normal o anómala, y adicionalmente se tiene información del tipo de anomalía que se presenta en cada observación. Por lo general en estos casos se generan modelos predictivos tanto para las clases normales como anómalas, en donde la información medida en línea es contrastada con estos modelos para dar alarma en caso de encontrar una medición

anómala.

- *Detección semisupervisada de anomalías*: En este caso sólo se tienen etiquetas de los comportamientos normales del proceso. Ya que no se requiere etiquetas para las anomalías, suele ser más utilizado que los métodos supervisados.
- *Detección no supervisada de anomalías*: En este caso no se requiere de etiquetas en los datos de entrenamiento. Estos métodos asumen que las operaciones normales son más frecuentes que las anómalas, en donde si no se cumple esto se obtiene un detector propenso a tener una alta tasa de falsas alarmas.

Entre las técnicas comúnmente utilizadas se encuentran herramientas estadísticas para procesos multivariables como análisis de componentes principales o PCA (de sus siglas en inglés, *Principal Component Analysis*) y variaciones de éste [14], [10], el cual en conjunto a estadísticos como el Test de Hotelling han probado ser de gran utilidad para la detección de fallas o anomalías [4]. Otro método es por medio de análisis de discriminante de Fisher (FDA de sus siglas en inglés, *Fisher's Discriminant Analysis*) [8], el cual es útil para el reconocimiento de patrones en los datos. Otros enfoques optan por redes neuronales artificiales o ANN (de sus siglas en inglés *Artificial Neuronal Network*), las cuales pueden discernir sobre el estado anómalo o no en base a características extraídas de los datos [28], [23]. Entre otros métodos se encuentran el uso de sistemas expertos, los cuales intentan emular las decisiones que serían tomadas por los operarios e ingenieros a cargo [2], [9].

Otra técnica de detección utilizada principalmente en sistemas de supervisión en línea es el basado en residuos, el cual consiste en comparar la señal medida en el sistema con una predicha por medio de un modelo matemático, el cual es obtenido por medio de datos pasados del proceso. La diferencia entre la señal medida y la calculada por el modelo se conoce como *residuo* y el análisis de éstos permite determinar si el proceso se encuentra en su zona de operación normal o si se presenta una anomalía. Esta técnica en particular se explicará con mayor detalle a continuación.

## 2.3. Detección de anomalías utilizando residuos

En la Figura 2.2 se encuentra el esquema de un sistema de detección y aislamiento de fallas por medio de residuos. Este proceso se divide en dos grandes partes: obtener los residuos y determinar la anomalía o falla.

Inicialmente deben elegirse las características a modelar del proceso y luego obtener un residuo $r$ definido por:

$$r = y_f - y_{f\_est} \tag{2.1}$$

En donde $y_f$ es la medición obtenida del proceso e $y_{f\_est}$ es el valor estimado mediante un modelo matemático desarrollado en base a datos obtenidos sobre las variables de interés del sistema, las cuales deben ser escogidas antes de entrar a la etapa de modelación.

Luego estos residuos pueden ser analizados para ver si se está en presencia de una anomalía o no, e identificar los diferentes tipos de anomalías que afectan al sistema.

Es importante recalcar lo explicado en la Sección 2.2, en donde el proceso puede evolucionar a lo largo del tiempo y así una situación que anteriormente se consideraría anómala, ahora comprende parte del comportamiento normal, por lo que se debe realizar periódicamente un análisis de los datos obtenidos para comprobar que la operación del sistema no haya evolucionado y sea necesario cambiar el modo de análisis de los residuos.

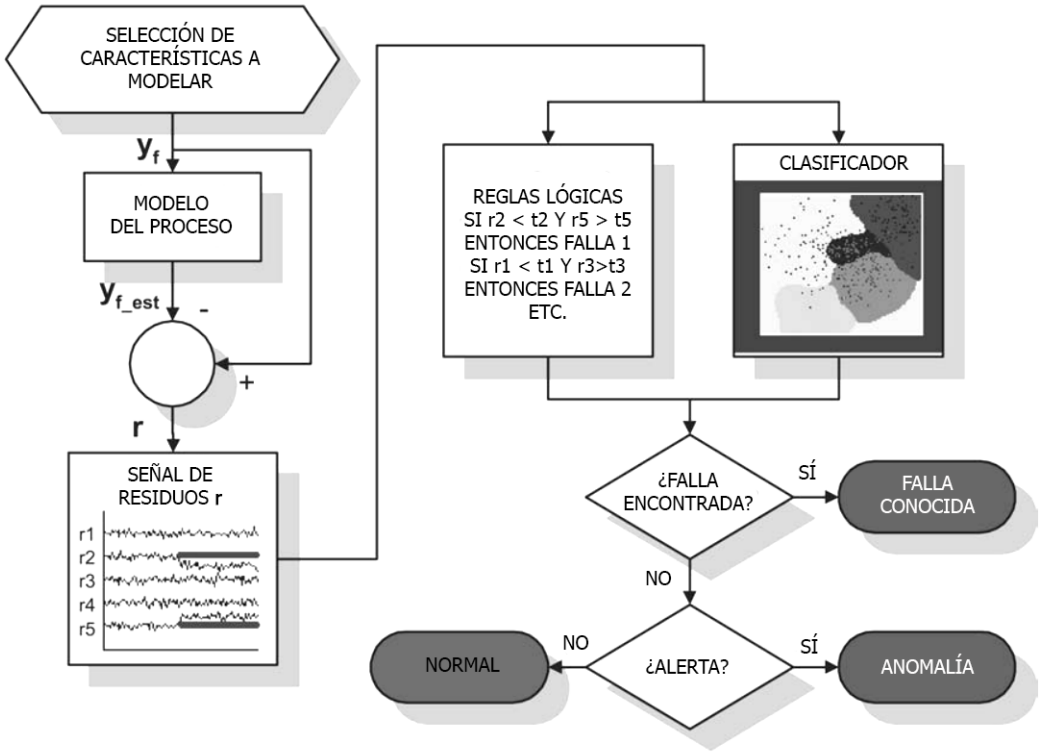



Figura 2.2: Esquema de un sistema de detección e identificación de anomalías, en donde se incluyen las dos grandes etapas del proceso [27].

## 2.4. Métodos de generación de residuos

En los estudios y experimentos, ya sean de índole científica o industrial, de gran o baja escala, suelen generarse datos para ser posteriormente analizados. Éstos pueden ser vectoriales o escalares, mediciones de sensores, caracteres, etc. El objetivo de esta obtención de datos es poder obtener información útil a partir de ellos, en particular poder explicar un proceso en base a éstos [21]; es decir, obtener un modelo.

Como se mencionó en la Sección 2.3, para poder obtener los residuos es necesario tener un

modelo que describa matemáticamente el proceso. Los modelos pueden ser clasificados según la naturaleza de estos [29] en:

- **Modelos de caja blanca:** En este caso el modelo es perfectamente conocido y ha sido construido por completo a partir de información previa y ecuaciones fenomenológicas.
- **Modelos de caja gris:** Para este caso se puede generar el modelo en base de algunas ecuaciones fenomenológicas, pero algunos parámetros deben ser determinados por medio de los datos medidos.
- **Modelos de caja negra:** No es posible una descripción fenomenólogica del sistema, por lo que se utilizan los datos obtenidos y las entradas entregadas para generar un modelo que describa adecuadamente su comportamiento. En particular este tipo de modelación es la que será utilizada en este trabajo.

Como también pueden ser clasificados según las características que tengan. Entre estas clasificaciones se encuentran: modelos continuos, discretos, estáticos, dinámicos, lineales, no lineales, de paramétros concentrados, univariables, multivariables, entre otros [13].

Entre las distinciones de mayor interés para este trabajo se encuentran los modelos paramétricos y no paramétricos, los cuales corresponden a la clasificación de un modelo según su cantidad de parámetros y serán explicados en mayor detalle a continuación.

### 2.4.1. Modelos paramétricos

Cuando un modelo tiene un número fijo de coeficientes, sin importar la cantidad de datos que se utilice para obtenerlo o entrenarlo, entonces se considera paramétrico. Éstos en particular suelen no ser muy costosos computacionalmente, pero tienen la desventaja de tener fuertes supuestos sobre la naturaleza de la distribución de los datos [18]. Entre los más conocidos se encuentran los modelos ARX, ARMAX, OE, BJ, entre otros [16].

### 2.4.2. Modelos no paramétricos

Cuando la cantidad de coeficientes del modelo varía dependiendo de los datos de entrenamiento que se utilicen para obtenerlo, entonces se está en la presencia de un modelo no paramétrico. Estos modelos suelen ser más flexibles y tener menos supuestos sobre la naturaleza de los datos, pero necesitan de una mayor capacidad computacional para ser obtenidos al ser más intrincados cuando se tiene una gran cantidad de datos disponibles [18].

Dado que con los años la tecnología ha avanzado, actualmente se poseen recursos que permiten utilizar estas técnicas en donde no es necesario una gran información previa del proceso a modelar, sino que el enfoque está puesto sobre el análisis de los datos de operación, el cual debe realizarse cada vez que se obtiene nuevas observaciones.

Entre los métodos paramétricos más conocidos y utilizados en la actualidad que han sido aplicados a la detección de fallas y anomalías se encuentran *General Regression Neural*

*Networks* (GRRN) [30], *k-Nearest Neighbors* (kNN) [18], [12] y Similarity-based modeling (SBM) [11], [31].

#### 2.4.2.1. Modelos basados en similitud

SBM (de sus siglas en inglés, *Similarity-based modeling*) es una técnica de modelación no paramétrica que se basa en encontrar similitudes y relaciones entre las variables de un conjunto de observaciones. Este método ha resultado ser efectivo para modelar sistemas multivariables [31] y algunas implementaciones de SBM para la detección de fallas y anomalías se pueden encontrar en [1], [17], [26].

Este método considera un sistema de la forma:

$$y = f(x) \tag{2.2}$$

En donde $x \in \mathbb{R}^m$ corresponde a las variables de entrada e $y \in \mathbb{R}^p$ son las variables de salida del sistema y la función $f(\cdot)$ es desconocida.

Si se posee acceso a mediciones de las entradas y salidas del sistema, se pueden definir las siguiente matrices de entrenamiento:

$$D_{\mathrm{i}} = [x_1, x_2 \cdots x_n] \quad \in \mathbb{R}^{mxn} \tag{2.3}$$
$$D_o = [y_1, y_2 \cdots y_n] \quad \in \mathbb{R}^{pxn} \tag{2.4}$$

Donde $D_{\mathrm{i}}$ corresponde a la matriz de entrenamiento con las observaciones de la entrada y $D_o$ la matriz de entrenamiento con las observaciones de la salida del sistema. Se tiene que $y_{\mathrm{i}} = f(x_{\mathrm{i}})\ \forall \mathrm{i} = 1, ..., n$.

Los pares $[x_{\mathrm{i}}, y_{\mathrm{i}}]_{\mathrm{i}=1,...,n}$ deben ser una base de los puntos de operación que se deseen modelar. Estos puntos pueden ser obtenidos mediantes técnicas como PCA (*Principal Componente Analysis*), PLS (*Partial Least Squares*), Test de Hotelling, entre otros [1].

Posteriormente, dado un vector de entrada $x^*$, se estima la salida del sistema $y^* = f(x^*)$ mediante una combinación lineal de las columnas de $D_o$:

$$\hat{y}^* = D_o w \tag{2.5}$$

En donde $\hat{y}^*$ es la estimación dada por el modelo y $w$ es un vector de pesos que se obtiene mediante:

$$w = \frac{\hat{w}}{1^T \cdot \hat{w}} \tag{2.6}$$
$$\hat{w} = (D_{\mathrm{i}}^T \Delta D_{\mathrm{i}})^{-1}(D_{\mathrm{i}} \Delta x^*) \tag{2.7}$$

El símbolo $\Delta$ representa el operador de similitud [31]. Este operador debe cumplir que al tener dos elementos $A$, $B \in \mathbb{R}^n$, $A \Delta B \in \mathbb{R}^+$ debe ser simétrica, alcanzar su máximo cuando

$A = B$ y ser monótonamente decreciente con $\|A - B\|$. Según [31], luego de un estudio preliminar se pudo determinar que el operador que mejor captura la variabilidad de los datos es el operador triangular saturado definido por:

$$A\Delta B = \begin{cases} \mathrm{d} - \|A - B\| & \|A - B\| \leq \mathrm{d} + \varepsilon \\ \varepsilon & \|A - B\| > \varepsilon \end{cases} \tag{2.8}$$

Con $\varepsilon > 0$ un número pequeño para asegurar que $A\Delta B > 0$ y $\mathrm{d} > 0$ un umbral que dependerá de la dispersión de las muestras, por lo general se puede utilizar la distancia promedio entre éstas. A pesar de que el operador triangular saturado será el utilizado para este trabajo, SBM puede ser implementado con cualquier operador de similitud que cumpla las propiedades anteriormente mencionadas.

Si bien SBM es un método no paramétrico que permite estimar la respuesta de un sistema estático en función de una base de datos representativa, este concepto se puede extender sencillamente a sistemas dinámicos discretos en caso de tener una secuencia de mediciones. Ahora la base representativa también contendrá instantes pasados de tanto entradas como salidas del sistema, así este método es capaz de rescatar las dependencias dinámicas de las variables del sistema.

Luego teniendo este método de estimación no paramétrico para sistemas dinámicos, una base de datos representativa del sistema y mediciones de éste, se puede realizar la detección de anomalías mediante la diferencia entre el estimado entregado por SBM y la medición real de las variables. Luego con un criterio o análisis de este residuo se puede determinar si el sistema está comportándose dentro de su rango normal o si se está en presencia de una anomalía.

Algo importante a recalcar es que para disminuir la probabilidad de falsa alarma con este método, es necesario que la base representativa tenga información sobre todos los modos de operación normales del sistema, de lo contrario se pueden dar situaciones en las que el residuo de la señal indique que es una anomalía cuando simplemente es una situación normal del sistema, pero ésta no se encuentra contenida en la base representativa, por lo cual el modelo no es capaz de entregar una estimación correcta.

## 2.5. Detección de anomalías utilizando la estimación

Como se explicó anteriormente, la detección de anomalías se puede realizar mediante la diferencia de un error entre la salida real de la planta y el estimado por el modelo; sin embargo existen ocasiones en las cuales no se tiene acceso a la salida real del proceso.

Para estos casos, las técnicas de modelación descritas anteriormente siguen siendo útiles para la creación de un modelo que intente representar la fenomenología del proceso a través de los datos obtenidos de su operación, pero la detección ya no se puede realizar en función de un error. Ahora, la salida estimada del modelo creado será el único indicio de la existencia o no de una anomalía, por lo cual en base a sólo esta estimación se deberá generar una decisión sobre si alertar o no al operario.

## 2.6. Validación cruzada

Al realizar modelos de cualquier tipo, es importante validar el funcionamiento de éstos, ya que el objetivo de su creación es producir un modelo preciso y creíble [15]. Un tipo particular de validación de modelos corresponde a la validación cruzada, la cual tiene por objetivo evaluar el desempeño de un modelo y garantizar que los resultados son independientes de la separación realizada en la base de datos para obtener los datos de entrenamiento y de prueba.

Este tipo de validación puede ser realizada de diferentes maneras, siendo las más comunes la validación cruzada de $k$ iteraciones y la validación cruzada aleatoria. En ambos casos el error del modelo puede ser calculado por medio de la media aritmética o promedio, la cual se define por la ecuación 2.9.

$$E = \frac{1}{K} \sum_{\mathrm{i}=1}^{N} E_{\mathrm{i}} \tag{2.9}$$

## 2.7. Herramientas útiles para el análisis de datos

En esta sección se explicarán en más detalle las técnicas utilizadas para el análisis de residuos y la obtención de estos mismos.

### 2.7.1. Análisis de componentes principales

El análisis de componentes principales o PCA (de sus siglás en inglés, *Principal Component Analysis*), es un método básico de análisis multivariado que ha sido aplicado exitosamente en diferentes áreas como compresión de datos, extracción de características, procesamiento de imágenes, entre otras [4].

Este método es una técnica lineal de reducción de dimensionalidad, la cual procura capturar la máxima variabilidad de los datos por medio de las componentes principales o **vectores de carga** [25], los cuales son ordenados según la cantidad de variabilidad que explica cada una. Dada una matriz de datos $X \in \mathbb{R}^{nxm}$ siendo $n$ las observaciones y $m$ el número de variables medidas. Esta matriz debe encontrarse previamente normalizada para evitar problemas debido a las unidades de cada variable.

Luego PCA se encarga de resolver el siguiente problema de optimización:

$$\underset{\mathbf{v}\neq 0}{max} = \frac{\mathbf{v}^T X^T X \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \tag{2.10}$$

En donde $\mathbf{v} \in \mathbb{R}^m$. Estos vectores se pueden obtener mediante una descomposición en

valores singuladores o SVD (de sus siglás en inglés, *Singular Value Descomposition*):

$$\frac{1}{\sqrt{n-1}}X = U\Sigma V^T \tag{2.11}$$

Siendo $U \in \mathbb{R}^{nxn}$ y $V \in \mathbb{R}^{mxm}$. La matriz $\Sigma \in \mathbb{R}^{nxm}$ contiene los valores propios no negativos en orden decreciente en su diagonal ($\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{min(m,n)} \geq 0$). Las componentes principales corresponden a los vectores columna ortonormales de la matriz $V$ y la varianza del conjunto de datos proyectado en la i-ésima columna de $V$ es igual a $\sigma_{\mathrm{i}}$. Este cálculo es equivalente a obtener la descomposición en valores propios de la matriz de covarianza S:

$$S = \frac{1}{n-1}X^T X \tag{2.12}$$

Ésta puede ser descompuesta como se muestra en la ecuación 2.13, en donde $\Lambda = \Sigma^T\Sigma \in \mathbb{R}^{mxm}$ es una matriz diagonal que contiene los valores propios no negativos $\lambda_{\mathrm{i}}$ con $\mathrm{i} = 1, ..., m$. Estos valores propios están ordenados de forma decreciente a lo largo de la diagonal; es decir, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Luego la matriz $V$ corresponde a los vectores propios asociados a los valores $\lambda_{\mathrm{i}}$ de la matriz $\Lambda$.

$$S = V\Lambda V^T \tag{2.13}$$

Con el fin de capturar la variabilidad de los datos y reducir la dimensionalidad del problema analizado, se escogen las $a$ componentes principales, las cuales corresponden a los vectores propios asociados a los valores propios de mayor magnitud. Así se logra obtener una matriz $P \in \mathbb{R}^{mxa}$ (con $a < m$), la cual contiene las primeras $a$ columnas de $V$. Posteriormente las proyecciones de $X$ en el nuevo espacio reducido están contenidas en:

$$T = XP \tag{2.14}$$

Y las proyecciones de $T$ en el espacio original se expresan mediante:

$$\hat{X} = TP^T \tag{2.15}$$

La diferencia entre las matrices $X$ y $\hat{X}$ se conoce como la matriz residual $E$ (ecuación 2.16), la cual captura las variaciones de las observaciones asociadas a las componentes no consideradas en el espacio reducido.

$$E = X - \hat{X} \tag{2.16}$$

Se puede definir $t_{\mathrm{i}}$ como la i-ésima columna de la matriz $T$ en el conjunto de entrenamiento, las cuales cumplen con las siguientes propiedades:

- $var(t_{\mathrm{i}}) \geq \cdots \geq var(t_a)$
- $\mathrm{Media}(t_{\mathrm{i}}) = 0;\ \forall\ \mathrm{i} = 1, ..., a$
- $t_{\mathrm{i}}^T t_k = 0;\ \forall\ \mathrm{i} \neq k$

- No existe otra expansión ortogonal de $a$ componentes que capture una mayor variabilidad de los datos.

Luego se puede incorporar un nuevo vector de observaciones $x_n \in \mathbb{R}^m$ a la base de datos mediante su proyección en el espacio reducido:

$$t_n = x_n P \tag{2.17}$$

En caso de que las observaciones sean dependientes de instantes de tiempo pasado, lo cual generalmente sucede en procesos con tiempos de muestreo menor a 2 horas, se pueden agregar los regresores necesarios a la matriz $X$ y encontrar posteriormente las componentes principales de ésta, lo cual se conoce como *PCA dinámico* [25], [4].

Hasta el momento se ha hablado de PCA para reducción de dimensionalidad, pero esta técnica también puede utilizarse para clasificación, lo cual tiene utilidad si se desea identificar distintos tipos de anomalías. Por lo general se asume que los datos proveen de una distribución normal, pero su uso puede expandirse a distintas distribuciones mediante modificaciones del algoritmo, como lo es kernel PCA (kPCA) [19] en donde las variables se representan en un nuevo espacio abstracto $\chi$ para luego ser clasificadas.

### 2.7.2. Test de estadístico de Hotelling

Este test estadístico permite caracterizar la variabilidad de un conjunto de observaciones en función de un umbral escalar, el cual está asociado a cierto grado de confianza. Esto sirve en la detección de anomalías para verificar si el error de estimación se encuentra dentro de una región limitada por un umbral, y así determinar si es aceptable o no.

Al tener una matriz de datos de entrenamiento $X \in \mathbb{R}^{nxm}$, siendo $n$ el número de observaciones y $m$ el número de variables, su matriz de covarianza $S$ se puede escribir mediante la ecuación 2.12 y se puede descomponer como se muestra en la ecuación 2.13. Teniendo un vector de observación $\mathbf{x}$ y asumiendo que la matriz $S$ es invertible, se puede definir el vector $z$ y con éste calcular el estadístico de Hotelling $T^2$:

$$z = \Lambda^{-\frac{1}{2}} V^T \mathbf{x} \tag{2.18}$$

$$T^2 = z^T z \tag{2.19}$$

La matriz $V$ rota los ejes principales de la matriz de covarianza $X$ de tal manera que corresponden directamente a los elementos de $y$, en donde éste corresponde a las proyecciones utilizando el vector de observación $\mathbf{x}$, es decir $y = V^T \mathbf{x}$. Además $\Lambda$ escala los elementos de $y$ produciendo un conjunto de variables con varianza unitaria que corresponden a los elementos del vector $z$.

El estadístico de Hotelling $T^2$ corresponde a una norma-2 de un vector de observación $\mathbf{x}$ de su media, la cual es escalada en dirección de los vectores propios de la matriz de covarianza $S$, y es inversamente proporcional a la desviación estándar a lo largo de los vectores propios.

Esto permite que un umbral escalar pueda caracterizar la variabilidad de los datos en un espacio de observaciones de dimensión $m$.

Este estadístico trabaja en base a niveles de confianza $\alpha$. Un umbral apropiado basado en un nivel de confianza $\alpha$ puede ser determinado asumiendo que las observaciones son muestras aleatorias de una distribución normal multivariable.

Para detectar observaciones anómalas en el conjunto de entrenamiento (u *outliers*) se define el siguiente umbral:

$$T_\alpha^2 = \frac{(n-1)^2 \left(\frac{m}{n-m-1}\right) F_\alpha(m, n-m-1)}{n\left(1 + \left(\frac{m}{n-m-1}\right) F_\alpha(m, n-m-1)\right)} \tag{2.20}$$

En donde $n$ corresponde al número de observaciones y $m$ al número de variables utilizadas para explicar el proceso. $F_\alpha(a, b)$ es el punto crítico superior $(100 \cdot \alpha)\,\%$ de la distribución $F$ de Fisher de $a$ y $b$ grados de libertad.

Los eventos que ocurren en observaciones que no son del conjunto de entrenamiento pueden ser detectados utilizando el umbral:

$$T_\alpha^2 = \frac{m(n-1)(n+1)}{n(n-m)} F_\alpha(m, n-m) \tag{2.21}$$

### 2.7.3. Función *softmax*

Esta función corresponde a una función exponencial normalizada [6] se emplea para comprimir los valores de un vector en un rango de 0 a 1 mediante la ecuación 2.22. Por lo general esta función es utilizada en las capas finales de las redes neuronales.

$$\mathbf{x}_\mathrm{i} = \frac{\mathrm{e}^\mathrm{i}}{\sum_{k=1}^{K} \mathrm{e}^{x_k}} \quad \forall \quad \mathrm{i} = 1, \cdots, K \tag{2.22}$$

# Capítulo 3

# Implementación de herramientas para la detección de anomalías utilizando modelos SBM

Los sistemas de mayor complejidad no son sencillos de modelar fenomenológicamente, debido a que existen dinámicas no modeladas, fenomenologías desconocidas en el proceso, parámetros no estimados, entre otros. Sin un modelo en el cual basarse, la detección de anomalías se reduce a la experiencia del usuario para saber en qué momento el proceso no se comporta como se esperaba.

Afortunadamente, por lo general se disponen de datos de la operación de estos sistemas, en donde entran en juego los modelos no paramétricos o *data-driven modelling*, con lo cual es posible encontrar relaciones o patrones que posteriormente se pueden utilizar para modelar el sistema.

En este capítulo se detalla la implementación del detector de anomalías. En la sección 3.1 se presentan las herramientas y etapas para la construcción de un modelo basado en similitud, en donde se detallan los bloques clave para poder modelar un proceso de forma no paramétrica. En la sección 3.2 se explica el proceso de validación del modelo y las características que se desean cumplir para poder ser utilizado en línea. En la sección 3.3 se detallan los pasos a seguir para la creación de un detector de anomalías en línea, en donde éste utiliza las salidas otorgadas por un modelo SBM.

## 3.1. Herramientas para la construcción de un modelo basado en similitud

En la Figura 3.1 se presentan las etapas a realizar para crear un modelo basado en similitud, al cual se referirá como modelo SBM de aquí en adelante. Este proceso se puede dividir en dos grandes etapas de pre-procesamiento de datos y de modelación no paramétrica como

tal, las cuales se dividen en subetapas que serán explicadas a continuación. Cabe destacar que la programación de los bloques y las funciones detalladas a continuación fueron realizadas en el software MATLAB.

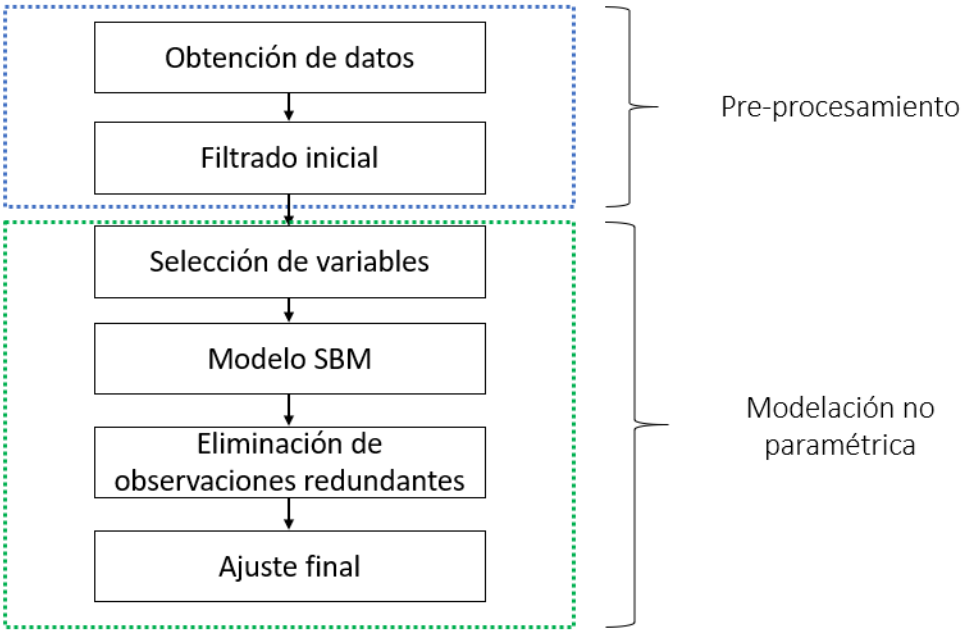



Figura 3.1: Diagrama del algoritmo implementado para la creación del modelo SBM.

### 3.1.1. Obtención de datos

Lo primero a realizar es obtener una base de datos con información relevante del proceso, siendo deseable que ésta contenga todos los tipos de operación posible, para así poder estimar correctamente la operación con el modelo que será realizado.

### 3.1.2. Filtrado inicial

En primera instancia, se eliminan de los datos toda información que no sea numérica; es decir, se eliminan los datos tipo NaN (*Not a Number*) mediante el comando *isnan* de MATLAB.

Luego se normalizan las variables para eliminar lar magnitudes, esta normalización se realiza mediante la siguiente ecuación:

$$x_{\mathrm{i}} = \frac{x_{\mathrm{i}} - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}} \tag{3.1}$$

En donde para cada valor de un vector $x$ de datos, se le restará la media de este vector y se dividirá por la desviación estándar del mismo.

Por último, es importante notar que los datos recopilados no necesariamente corresponden a la parte de interés del proceso, por lo cual en estos casos es importante reconocer qué señales o variables se gatillan cuando sucede ésta parte para así poder reducir los datos y eliminar aquellos que contienen partes que no desean ser modeladas.

### 3.1.3. Selección de variables

En esta etapa se escogen las variables significativas para el modelo, eliminando la información que no aporta a la modelación. Esto se aborda mediante dos métodos, el primero es PCA (*Principal Componente Analysis*) y el segundo es un análisis heurístico.

#### 3.1.3.1. Análisis mediante PCA

El análisis mediante componentes principales corresponde a transformar linealmente las variables del proceso para verlas en un nuevo sistema de coordenadas, donde se presenta la posibilidad de reconocer clústeres que correspondan a las zonas de operación al graficar el *score plot* de los datos, como se ve en la Figura 3.2.

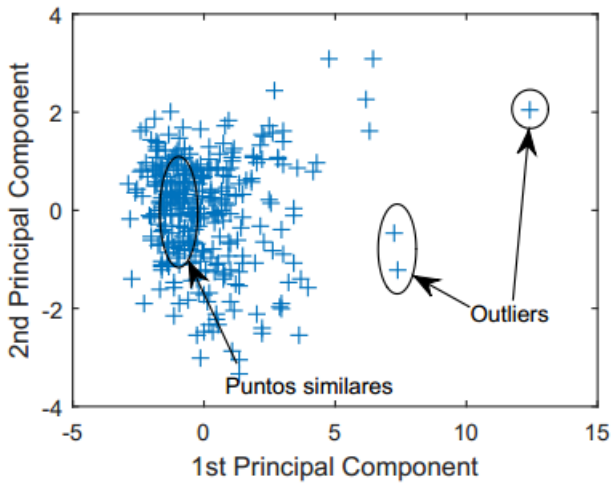



Figura 3.2: Ejemplo de un *score plot* utilizando las primeras y segundas componentes principales. Imagen obtenida de [3].

Una vez encontradas las zonas que definan cuando el proceso se está llevando a cabo con normalidad y cuándo no, se pueden utilizar otras herramientas tales como el Test de Hotelling para la detección de anomalías.

#### 3.1.3.2. Análisis heurístico

Existe la posiblidad de que al realizar un análisis mediante PCA no se logre distinguir claramente las zonas de operación del proceso. En casos como estos, se puede realizar un análisis heurístico de las variables, el cual corresponde a buscar relaciones entre los cambios sucedidos en ellas al momento que se realiza el proceso de interés; es decir, cuáles variables varían en esos momentos, cuál es la correlación entre ellas y cuáles efectivamente corresponden a variables que deberían variar según lo que sucede físicamente.

Si bien este método no llega a ser tan sofisticado como PCA, es un buen comienzo para la selección de variables relevantes y las relaciones entre ellas que definen el proceso.

### 3.1.4. Modelo SBM

Una vez obtenidas las variables descriptivas del proceso, es posible iniciar la construcción del modelo SBM. En la Figura 3.3 se observan las subpartes del modelo, las cuales serán explicadas en detalle a continuación.

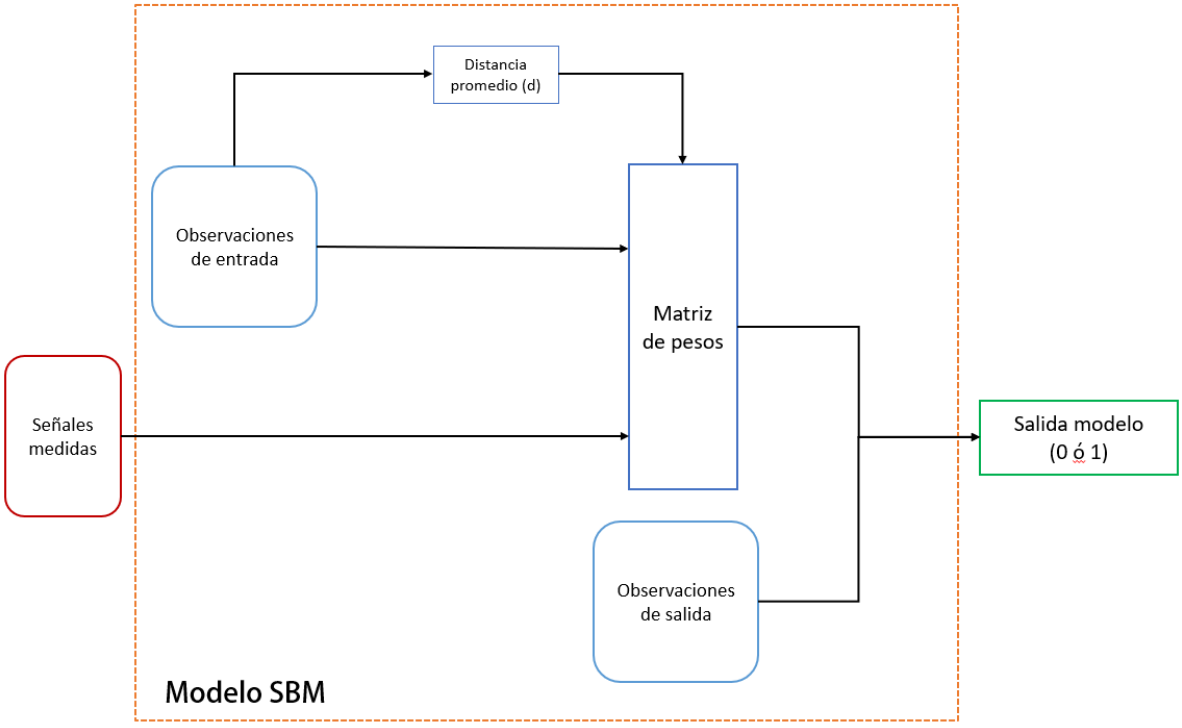



Figura 3.3: Esquema de partes conformantes del modelo SBM y la interacción entre ellas.

#### 3.1.4.1. Matriz de observaciones de entrada

Esta matriz corresponde a las variables seleccionadas de la base de datos inicial. En esta matriz se incorporan tanto valores de las variables como regresores de éstas o relaciones adicionales que se hayan verificado como relevantes para la descripción del proceso. Esta matriz será denotada como $D_{\mathrm{i}}$ de ahora en adelante.

#### 3.1.4.2. Matriz de observaciones de salida

Esta matriz corresponderá a las salidas registradas en la base de datos y será denotada como $D_o$ de ahora en adelante. Cabe destacar que para el modelo desarrollado en este trabajo la salida se considerará binaria, en donde '0' representará el funcionamiento normal y '1' la presencia de una anomalía; sin embargo, en caso de tener etiquetas específicas de los tipos de anomalías existentes, estas categorías pueden expandirse tanto como anomalías haya etiquetadas.

#### 3.1.4.3. Matriz de pesos

Para poder generar la estimación de la salida del proceso mediante el modelo SBM y la entrada de mediciones, es necesario generar una matriz de pesos que será ponderada linealmente con $D_o$. Estos pesos serán calculados a partir de la ecuación 2.7, en donde el operador triangular saturado es definido por la ecuación 2.8.

En particular, es importante el cálculo de la distancia $d$, la cual será calculada como la distancia promedio entre los elementos de la matriz $D_{\mathrm{i}}$; es decir, la sumatoria de la norma euclideana entre cada elemento, dividido por la cantidad de elementos en la matriz. Teniendo el valor $d$, es posible calcular la matriz de pesos y utilizarla para estimar la salida mediante la ecuación 2.5.

### 3.1.5. Eliminación de observaciones redundantes

A pesar de la previa selección de variables explicativas del proceso, existe la posibilidad de que las mediciones realizadas e incluidas en la matriz $D_{\mathrm{i}}$ resulten ser redundantes, lo cual se traduce en que esta matriz tenga una dimensión mayor a la necesaria y que algunas filas puedan ser linealmente independientes, lo cual genera problemas al momento de invertir la matriz de similitudes. Para poder evitar estos problemas, se puede reducir la matriz $D_{\mathrm{i}}$ mediante el análisis de las filas de ésta.

Si la similitud entre las filas $r_{\mathrm{i}}$ y $r_j$ es alta (cercana a $d$), significa que una de éstas no aporta una cantidad considerable de información con respecto a la otra, por lo tanto, con una sola de las filas bastaría para modelar la operación en esas observaciones. Así es posible ordenar las filas en orden decreciente de similitud y eliminar un porcentaje a seleccionar de la matriz $D_{\mathrm{i}}$ y sus salidas correspondientes en la matriz $D_o$.

### 3.1.6. Ajuste final

Es importante notar que la salida del modelo SBM corresponderá a una combinación lineal entre la matriz de pesos $w$ y la matriz de salidas $D_o$. Esta salida la gran mayoría de las veces no resulta un número exacto como 0 ó 1, los cuales corresponden a las salidas esperadas (estado normal o anómalo). Para poder tener una salida perteneciente a los conjuntos designados, es necesario establecer un umbral de decisión, el cual variará dependiendo de la cantidad de datos correspondiente a cada categoría que se posean.

Luego la salida final del modelo SBM estará definida por:

$$y^* = \begin{cases} 1 & D_o w \geq threshold \\ 0 & D_o w < threshold \end{cases} \tag{3.2}$$

Cabe destacar que podrían existir múltiples umbrales de detección en caso de tener más tipos de anomalías clasificadas en el modelo SBM.

## 3.2. Validación del modelo SBM

Para poder comprobar la eficacia del modelo es necesario evaluar su funcionamiento con observaciones que no hayan sido añadidas a la matriz $D_{\text{i}}$, por lo cual generalmente la base de datos se separa en dos conjuntos o tres conjuntos dependiendo de la extensión de ésta. Estos conjuntos corresponden a entrenamiento, validación y prueba, los cuales pueden ser reducidos únicamente a entrenamiento y validación en caso de disponer de pocos datos.

Así es como en esta etapa se utilizan los datos de validación para ver el error que presenta el modelo generado con los datos de entrenamiento, en donde el aceptar o rechazar el modelo dependerá del valor porcentual de errores aceptados. Con estas pruebas también se podrá sacar la tasa de falsos positivos y negativos, con los cuales se podrá analizar cuáles son los tipos de errores que presenta el modelo.

Para poder normalizar los datos de validación, se hará uso de la media y desviación promedio de las observaciones utilizadas para la creación de la matriz $D_{\text{i}}$. Así, los valores que entrarán al modelo SBM se encontrarán normalizados con respecto a las observaciones utilizadas para su construcción.

Adicionalmente, otra restricción que se utilizará para validar el modelo es el tiempo de procesamiento. Si este es menor o igual al tiempo de muestreo de los datos recolectados, entonces el modelo se considerará útil. De lo contrario, se deberá revisar los pasos anteriores para reducir el tiempo de procesamiento total. Esta restricción se añade para poder utilizar posteriormente el modelo SBM en línea con las mediciones recibidas del proceso en tiempo real.

Luego en la Figura 3.4 se puede ver un diagrama de flujo de las etapas explicadas en la

primera y segunda sección. La etapa crítica del modelo preliminar corresponde a la búsqueda de filas linealmente independientes, en donde en caso de repetirse varias veces la dinámica de eliminar observaciones iniciales, entonces se recomienda volver a la selección de variables. En cuanto a la validación del modelo, su etapa crítica corresponde a la verificación del modelo al ser probado con los datos de validación. Nuevamente, si se repite en exceso la dinámica de volver a crear un modelo SBM ya que no se cumplen las expectativas deseadas, es recomendable replantear las variables explicativas encontradas anteriormente para la modelación del proceso.

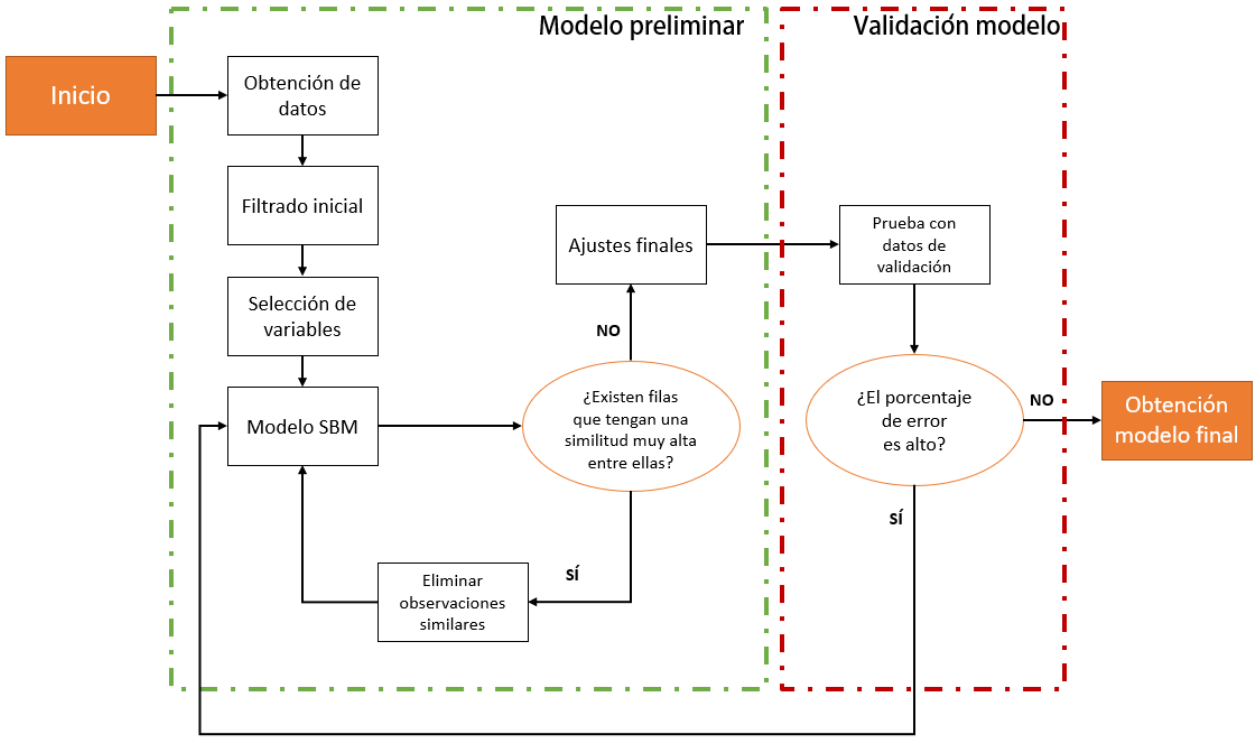



Figura 3.4: Diagrama construcción de modelo SBM y sus diferentes etapas: la creación preliminar del modelo y su validación.

## 3.3. Generación de rutina de detección

En el caso más básico, se puede crear una rutina de detección únicamente utilizando la salida entregada por el modelo SBM, en donde al momento que éste detecte una entrada que correspondería a una salida anómala, se alerte inmediatamente al usuario, como se muestra en la Figura 3.5. Esta configuración de detector presenta el problema de que es muy sencillo otorgar falsas alarmas si es que el modelo SBM no es extremadamente exacto, lo cual va contra el propósito del detector de alertar únicamente cuando se encuentra una situación anómala y finalmente estas falsas alarmas serían más perjudiciales que beneficiosas para el desarrollo del proceso.

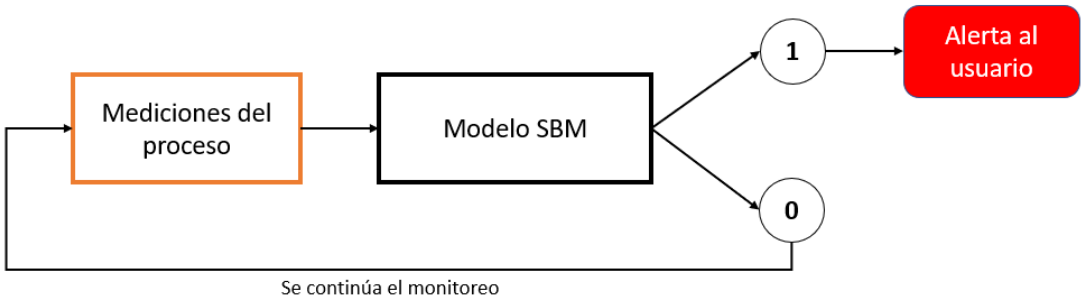



Figura 3.5: Sistema de detección utilizando únicamente la salida actual del modelo SBM.

Otro método para el monitoreo del proceso y la detección de anomalías es no utilizar directamente la salida del modelo SBM en el momento $t$, sino usar tanto la salida actual, como las pasadas. Esta forma tiene la desventaja de tener un *delay* mayor a realizar una alerta directa, pero tiene el beneficio de contrarrestar los errores puntuales cometidos por el modelo SBM. De esta manera, se puede reducir el número de falsas alarmas a cambio de tener que esperar un mayor tiempo para dar una respuesta definitiva sobre el estado del proceso.

Si bien en primera instancia esta metodología puede arreglar los problemas de falsas alarmas, es importante recalcar que es de utilidad sólo si es posible reconocer si el proceso presenta o no una anomalía mientras este se está llevando a cabo; es decir, en el caso de tener un proceso muy veloz en donde las mediciones realizadas de éste no sean obtenidas con una frecuencia de muestreo mucho mayor al periodo del proceso, entonces es complicado utilizar este método de monitoreo y es conveniente revisar el modelo para obtener una representación más precisa del proceso.

En el caso que se den las condiciones de tener un apropiado equipamiento para la medición del proceso, es posible utilizar una función *softmax* para restar importancia a los resultados incorrectos otorgados por el modelo. Esta función se encarga de otorgar un valor entre 0 y 1, el cual equivale al valor relativo de cada elemento del vector entregado con respecto al valor de los demás elementos componentes del vector. Así, teniendo en consideración los resultados obtenidos de esta función y los retornados por el modelo SBM se puede obtener una decisión a partir de la siguiente ecuación:

$$y_t^* = \frac{e^{\overrightarrow{x}}}{\sum_{i=1}^{t} e^{\overrightarrow{x}_i}} \times \overrightarrow{x} \tag{3.3}$$

En donde $y_t^*$ corresponde a la salida de la rutina de detección en el instante $t > 1$, la cual corresponderá a la ponderación lineal entre el vector de salidas del modelo SBM ($\overrightarrow{x}$) y la función *softmax* aplicada sobre éste.

Luego bastará utilizar nuevamente un umbral sobre la salida de la rutina para distinguir las situaciones anómalas de las normales. El proceso anteriormente explicado se encuentra en la Figura 3.6, en donde se muestra el flujo de etapas a realizar para la detección de anomalías con una rutina de detección que considere tanto la salida actual del modelo como las pasadas para la toma de decisión.

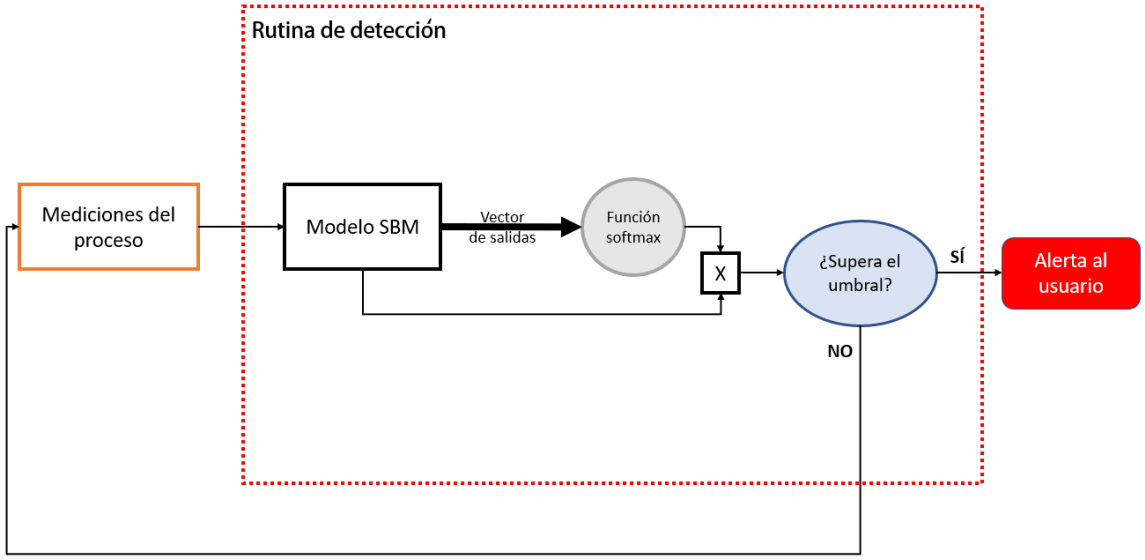



Figura 3.6: Diagrama de flujo de las operaciones a realizar para detectar anomalías en función a las salidas de un modelo SBM y el uso de una función *softmax* sobre éstas.

Por último, el ajuste final de la rutina se realizará para escoger con cuántas salidas del detector SBM se tomará la decisión de alertar o no.

# Capítulo 4

# Pruebas y resultados obtenidos

En este capítulo se procede a entregar los resultados obtenidos aplicando la metodología descrita en el Capítulo 3. En la sección 4.1 se describe brevemente el proceso de carguío y la base de datos a utilizar para el desarrollo del detector. En la sección 4.2 se encuentra el procedimiento realizado para la creación del modelo SBM, en donde se explican en detalles las subpartes realizadas que corresponden principalmente al filtrado y análisis de datos para elegir las variables relevantes del proceso y crear un modelo preliminar, el cual será posteriormente refinado hasta cumplir con las especificaciones deseadas. En la sección 4.3 se encuentran los resultados al probar el modelo propuesto con los datos de validación y el análisis de éstos. En la sección 4.4 se explica la rutina de detección aplicada, los parámetros de ésta y sus resultados. Por último, en la sección 4.5 se explica cómo utilizar la metodología diseñada para la detección de anomalías en los carguíos realizados manualmente y los resultados obtenidos utilizar la rutina de detección sobre estos datos.

## 4.1. Descripción del proceso de extracción y base de datos utilizada

El carguío de material con un cargador LHD se realiza mediante el impacto directo de éste contra la pila de extracción. Esta labor muchas veces se realiza de forma manual, pero dada la creciente automatización en la industria minera, existen cargadores capaces de ser manejados a distancia. En particular para este trabajo se utilizaron los datos obtenidos de un cargador LHD perteneciente a *GHH Chile SpA* y modificado en conjunto por ingenieros del Laboratorio de Robótica de Campo del *AMTC* e ingenieros pertenecientes a *GHH Chile SpA*. Es por esto que la máquina posee sensores adicionales para que tanto la navegación dentro de la mina como el carguío de material se realice de manera autónoma siguiendo una secuencia programada.

Dada esta situación, los datos recolectados en la minera 'San Gerónimo' incluyen las mediciones de los sensores tanto cuando se está navegando como cuando se está cargando material. El primer paso a realizar es dividir estas situaciones en conjuntos diferentes, en

donde para este trabajo los datos relevantes serán los de carguío autónomo.

En la Tabla 4.1 se encuentra el detalle de los datos recolectados durante las pruebas en la minera. Es importante destacar que los carguíos exitosos se definen como aquellos en los que se logra cargar suficiente material luego de un impacto con la pila, mientras que los fallidos corresponden a cuando se necesitan dos o más impactos contra ésta, lo cual corresponde a la anomalía que se detectará. El etiquetado de los carguíos fue realizado con la información recolectada por cámaras montadas en el cargador LHD y consideró únicamente el primer intento de recolección al chocar contra la pila; es decir, si en una prueba se chocó dos veces contra la pila de extracción para llenar el balde, sólo se contará el primer impacto y se clasificará como una anomalía.

| | **Número de carguíos** | **Carguíos exitosos** | **Carguíos fallidos** | **Tasa de muestreo** |
|---|---|---|---|---|
| Carguíos autónomos | 15 | 7 | 8 | 5 Hz |

Tabla 4.1: Tabla con información sobre la base de datos utilizada para la creación del detector de anomalías para carguío autónomo.

En la Figura 4.1 se grafica una de las variables sensadas durante las pruebas, en donde la línea naranja corresponde a la variable durante un comportamiento anómalo y la azul durante un comportamiento normal. A pesar de ser situaciones diferentes, no se notan mayores diferencias a primera vista, más que un desfase temporal, por lo cual será necesario estudiar con mayor detalle el comportamiento del proceso.

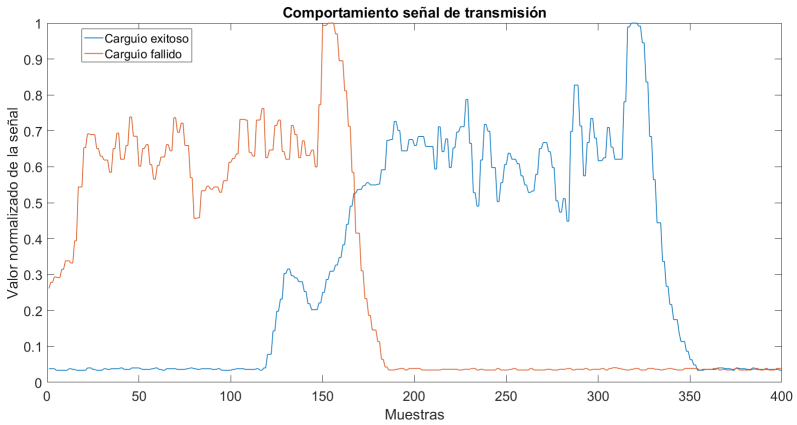



Figura 4.1: Ejemplo del comportamiento de una variable del proceso durante un carguío normal y uno anómalo.

## 4.2. Desarrollo del modelo SBM

### 4.2.1. Filtrado preliminar

El primer filtro de la base de datos corresponde a guardar únicamente las variables que presenten variaciones al momento de cargar material. Esto se puede realizar fácilmente al

excluir en un principio todas las señales que no presenten cambios durante toda la operación, las cuales con gran probabilidad corresponden a señales de alarmas.

Luego, en conjunto a conocimiento previo del proceso, es posible reconocer señales que presentan cambios importantes al momento de chocar con la pila tales como las revoluciones por minuto del motor, presión de transmisión, uso del pedal, entre otras. Conociendo de antemano las variaciones de estas señales, es posible verificar las variaciones de los demás sensores en los periodos en que se carga material y así eliminar mediciones adicionales. Así, de un promedio de 75 sensores [1], es posible disminuir a 20 sensores que presentan variaciones al momento de cargar material.

Posteriormente, como los datos guardados pueden corresponder a más instantes que únicamente cuando se carga, los datos son cortados en la duración del proceso de carguío. Para esto se utilizó la información de las señales del pedal y la transmisión, las cuales muestran un patrón al momento de impactar contra la pila.

En la Figura 4.2 se observa que la duración de la prueba es 1200 muestras o 240 segundos, pero el momento de la recolección se encuentra únicamente al final de la prueba, en donde se puede ver que la señal del pedal y la señal de transmisión alcanzan sus valores máximos.


Figura 4.2: Comportamiento señales del pedal y la transmisión hidráulica al momento de realizarse las pruebas de carguío.

Luego, utilizando condicionales sobre estas señales se logró cortar las pruebas realizadas a los momentos de interés, los cuales corresponden en promedio a unos 40 segundos, los cuales corresponden a cuando se llega frente a la pila, se impacta contra esta y posteriormente se retrocede. Por último, los datos son normalizados y guardados.

### 4.2.2. Análisis inicial de los datos

Luego de filtrar los datos de las pruebas realizadas, se comienza el análisis de éstos. Para ello se utilizó tanto PCA como un análisis heurístico de las señales, los cuales serán explicados

[1] En los datos obtenidos, no todas las pruebas tienen datos guardados de todos los sensores, por esto se habla de un promedio de número de sensores y no un valor exacto.

a continuación.

#### 4.2.2.1. Análisis mediante componentes principales

Una buen punto de partida para analizar sistemas multivariables es utilizar PCA o *Principal Component Analysis*, en donde se puede reducir la dimensionalidad de los datos y ser representados en dos o tres componentes.

En la Figura 4.3 se muestran las dos primeras componentes al utilizar PCA sobre las pruebas de carguío. Estas componentes explican aproximadamente el 62 % de la variabilidad de los datos, en donde se alcanza a apreciar una zona en donde se presenta una agrupación o clúster, sin embargo los puntos fuera de esta zona no corresponden a ningún transiente de la operación, sino a partes del proceso de carguío. Además, si bien es posible ajustar una zona donde sólo se encuentren los carguíos exitosos, el caso ideal sería tener un clúster que contuviese la gran mayoría de las proyecciones de los carguíos exitosos y que las proyecciones de los carguíos fallidos se encuentren lejanas a éste, pero en este caso proyecciones de los carguíos exitosos también se encuentran muy lejanas del clúster que se puede formar.

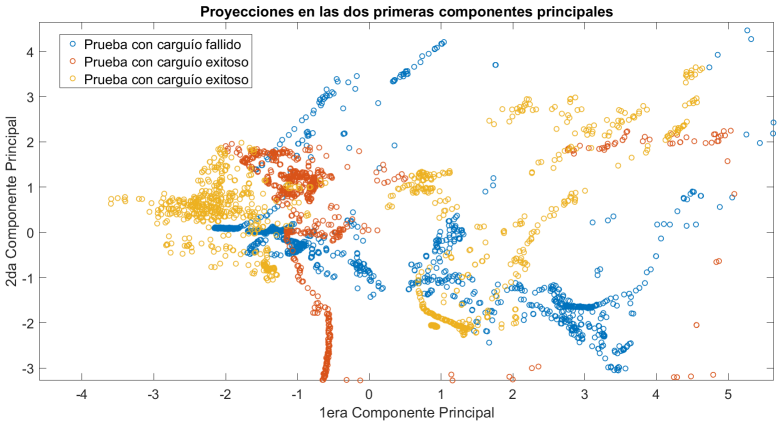



Figura 4.3: Gráfico con las primera y segunda componente del análisis PCA de carguíos tanto exitosos como fallidos.

También se probaron las proyecciones en 3 componentes principales, sin embargo se repitió la misma situación, es por esto que se decidió analizar de forma heurística las señales para encontrar las variables relevantes del proceso de carguío.

#### 4.2.2.2. Análisis heurístico

Para este análisis de observó el comportamiento de los sensores al momento de cargar. Dado que los datos fueron cortados incluyendo sólo el momento del carguío y posteriormente fueron normalizados, no se puede obtener inmediatamente cuáles señales contienen las mayores variaciones; sin embargo, al volver a analizar las señales en crudo sí es posible obtener estadísticos de éstas como lo es su desviación estándar. De esta manera se puede reducir el

número de variables, en donde es posible reducir de 20 a 16 variables, eliminando las que presentan una menor desviación estándar y reduciendo en un 20 % el número de variables.

### 4.2.3. Elección de variables

Tomando en consideración el análisis tanto por medio de PCA como de forma heurística, se puede realizar un análisis entre las correlaciones de las variables pre-seleccionadas para eliminar aquellas que sean redundantes entre sí. En la Figura 4.4 se observa que existen variables que tienen alta correlación entre ellas; es decir, añadir ambas al modelo resultaría redundante, por lo cual se puede eliminar una.

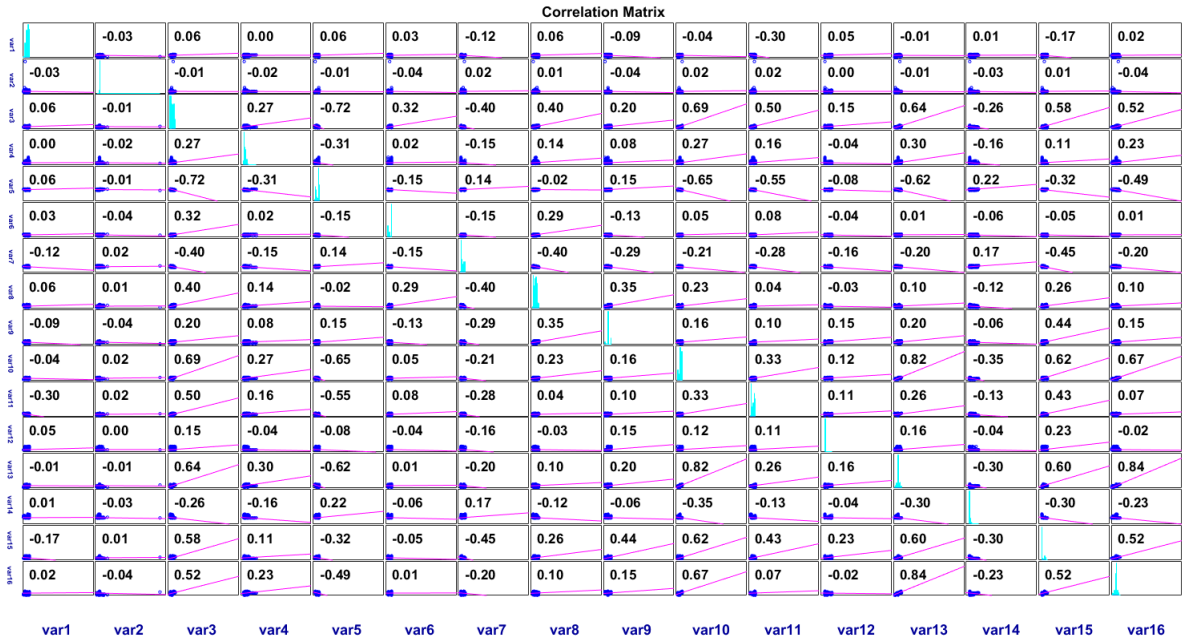



Figura 4.4: Matriz de correlación entre las variables escogidas previamente para la modelación del proceso de carguío.

Bajo este criterio, se seleccionó un total de 13 variables, las cuales corresponden a: *angle*, *angle speed*, *dump down pressure*, *dump up pressure*, *laser distance*, *lift down pressure*, *lift up pressure*, *manual pedal*, *manual bucket*, *transmission pressure forward*, *transmission pressure backward*, *rpm*, *skid* [2].

### 4.2.4. Modelo SBM

Para la creación del modelo SBM, se necesita construir la matriz de entradas $D_{\text{i}}$ y la matriz $D_o$ con la respectiva etiqueta de la observación. En particular, para la matriz $D_{\text{i}}$ se desea

[2]Se mantuvieron los nombres en inglés de las variables utilizadas dado que la traducción de estas podría no ser tan clara en español

que cada fila de ésta contenga las observaciones de las variables explicativas en el instante $t$, pero muchas veces esta información no es suficiente para poder explicar en su totalidad un proceso, sobretodo si este proceso es dinámico. Es por esto que las filas de la matriz $D_{\mathrm{i}}$ deben contener regresores de las variables explicativas y operaciones sobre éstas que puedan ser relevantes.

#### 4.2.4.1. Formación matriz $D_{\mathrm{i}}$

Tras prueba y error, se escogió que se utilizarían las mediciones actuales de las variables relevantes, el promedio de estas en una ventana de 11 muestras, la desviación estándar de las variables en una ventana de 11 muestras y el centro de masa de las mediciones en una ventana de 11 muestras. El centro de masa corresponde al punto donde se concentra el promedio tanto en el eje x como y (considerando cada variable graficada en 2D).

Un ejemplo de esto se puede ver en la Figura 4.5, en donde cada fila está compuesta por las variables anteriormente mencionadas y sus regresores, en donde al avanzar el tiempo la ventana temporal avanza con éste. Es importante destacar que para el entrenamiento, y por ende la formación de la matriz $D_{\mathrm{i}}$, se utilizaron 10 de las 15 pruebas realizadas.

| | | | | |
|---|---|---|---|---|
| Instante t | Mediciones actuales | Promedio medición t a t-10 | Desviación estándar de las mediciones t a t-10 | Centro de masa de las mediciones t a t-10 |
| Instante t+1 | Mediciones actuales | Promedio medición t+1 a t-9 | Desviación estándar de las mediciones t+1 a t-9 | Centro de masa de las mediciones t+1 a t-9 |
| Instante t+2 | Mediciones actuales | Promedio medición t+2 a t-8 | Desviación estándar de las mediciones t+2 a t-8 | Centro de masa de las mediciones t+2 a t-8 |

Figura 4.5: Ejemplo de la creación de filas de la matriz $D_{\mathrm{i}}$.

#### 4.2.4.2. Formación matriz $D_o$

Esta matriz es conformada por la etiqueta asociada a cada prueba de carguío, por lo cual corresponderá a un vector Nx1 de ceros o unos dependiendo de la etiqueta y de largo N dependiendo de las muestras tomadas en la prueba.

#### 4.2.4.3. Eliminación de observaciones redundantes

Dado que la matriz de similitud ($D_{\mathrm{i}}^T \Delta D_{\mathrm{i}}$) puede contener vectores muy parecidos entre ellos, estos dificulta la inversión de ésta, la cual es necesaria para obtener la matriz de pesos $w$. Para solucionar este problema se puede utilizar la técnica explicada en la sección 3.1.5, la cual fue aplicada a la matriz $D_{\mathrm{i}}$ generada. Luego de obtener en orden decreciente la similitud de los vectores, se eliminó el 10 % de los vectores que podrían ser considerados redundantes al poseer una alta similitud.

Una solución alternativa que fue encontrada al momento de realizar las pruebas en el modelo SBM es utilizar el comando *pinv* de MATLAB, el cual dará una aproximación de la inversión de la matriz de similitud, por lo cual no es tan exacto, pero reduce considerablemente el tiempo de cada prueba en la etapa de validación.

#### 4.2.4.4. Ajustes finales

Finalmente resta encontrar un umbral que se aplicará a la salida entregada por el modelo SBM, el cual determinará si la salida obtenida corresponde a una medición normal o una que presenta una anomalía. Para la obtención de este umbral se utiliza el promedio de de la matriz $D_o$, el cual dará un indicio sobre qué tipo de observaciones predominan en la matriz $D_{\mathrm{i}}$, las anómalas (1) o de operación normal (0).

## 4.3. Validación del modelo

Para validar el modelo, se utilizaron las 5 pruebas que no fueron usadas para la creación de las matrices $D_{\mathrm{i}}$ y $D_o$. Ya que la cantidad de pruebas es pequeña, se utilizó el método de validación cruzada para obtener resultados cuantitativos del desempeño del modelo creado. Este método fue utilizado debido a que, al tomar un porcentaje aleatorio de las muestras obtenidas por los sensores para la creación del modelo y la posterior validación, se puede romper la dependencia temporal que se utiliza en las filas de la matriz $D_{\mathrm{i}}$, por lo cual cada prueba realizada debe ser considerada por separado, a pesar de que cada una contenga varias mediciones de los sensores.

Por lo tanto, para cada prueba de validación de la metodología propuesta, se realizó una nueva matriz de entrada $D_{\mathrm{i}}$ y de salida $D_o$, se calculó la distancia $d$ correspondiente a los datos de entrenamiento utilizados y los promedios y desviaciones estándar de las variables explicativas para luego poder normalizar con estos valores los datos de validación. En estas pruebas para la eliminación de observaciones redundantes se utilizó el comando *pinv* de MATLAB.

Así, en la Tabla 4.2 se encuentran los resultados obtenidos al realizar estas pruebas, en donde se observa que el modelo no posee un porcentaje de error tan bajo como se desearía, lo cual se atribuye a que la base de datos no es suficiente para caracterizar en un gran porcentaje el proceso de carguío, pero aún así logra caracterizar la esencia de éste. Aún así, el tiempo de ejecución de una iteración del código es considerablemente menor al tiempo de muestreo de las señales, por lo cual el modelo puede ser utilizado sin problemas para estimar la salida del proceso en línea.

En la Figura 4.6 se muestra el histograma de las salidas del modelo SBM previas a la aplicación del umbral de detección. Se puede observar que las salidas varían en un rango mayor a [0,1], lo cual se debe a que los pesos dados por el modelo no necesariamente deben sumar 1, sino que pueden tener un valor continuo aleatorio; por lo tanto, los valores negativos corresponden a observaciones contrarias a algunas de la matriz $D_{\mathrm{i}}$, las cuales podrían ser

| | |
|---|---|
| **Número de pruebas realizadas** | 40 |
| **Porcentaje de error promedio** | 16,2 % |
| **Máximo porcentaje de error obtenido** | 41,9 % |
| **Mínimo porcentaje de error obtenido** | 4,3 % |
| **Tasa promedio de falsos negativos** | 9,9 % |
| **Tasa promedio de falsos positivos** | 6,3 % |
| **Tiempo promedio de ejecución** | 0,02 [s] |

Tabla 4.2: Resultados obtenidos de la validación cruzada del modelo SBM para carguíos autónomos.

agregadas a ésta para refinar el modelo.

Lo descrito anteriormente no fue realizado en las pruebas realizadas debido a que se necesita de una secuencia de mediciones para poder utilizar la estructura propuesta del modelo SBM, por lo tanto, si las observaciones que entregaron las salidas menores a cero corresponden a diferentes pruebas de carguío, al añadir éstas a la matriz $D_{\text{i}}$ se corre el riesgo de quedar con una distribución desbalanceada de datos de entrenamiento y validación.

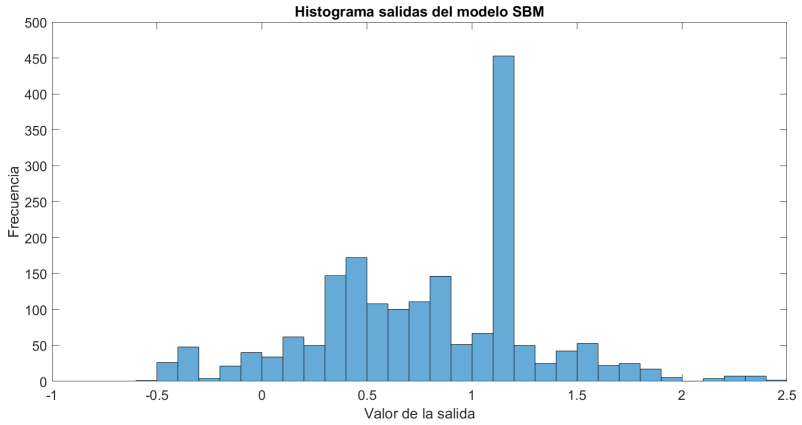



Figura 4.6: Histograma de los valores obtenidos como salida del modelo SBM antes de aplicar el umbral en una prueba de validación.

Adicionalmente, se debe tener en consideración que los errores obtenidos en el modelo pueden ser puntuales, tal como se muestra en la Figura 4.7, en la cual se nota una clara tendencia de que la salida corresponde a cero; es decir, un comportamiento normal, pero debido a que el porcentaje de error es obtenido mediante el número total de errores cometidos en cada estimación de la salida del proceso, el porcentaje de error crece. Esto puede ser arreglado al implementar la rutina de detección.

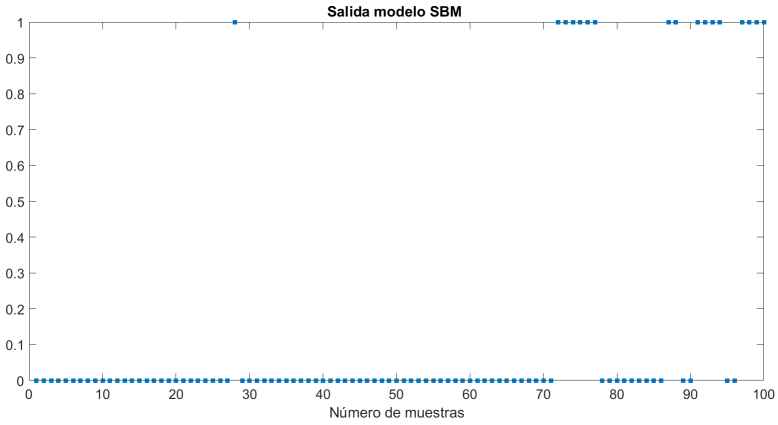
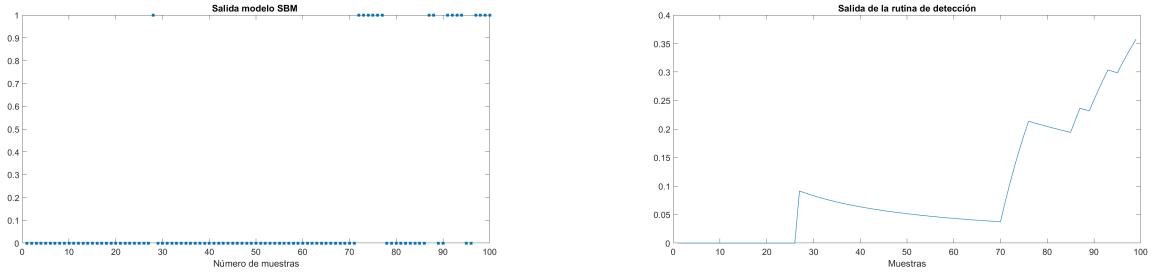



Figura 4.7: Salida obtenida por el modelo SBM en una prueba de validación.

## 4.4. Implementación rutina de detección

Para la implementación de la rutina de detección, se utilizó de la función *softmax* como fue explicado en la sección 3.3. Con esto es posible disminuir la influencia de los errores puntuales, tomando en consideración la tendencia de las salidas pasadas. En la Figura 4.8 se observa como al aplicar la función *softmax* y tener en consideración las salidas pasadas del detector, la salida final obtenida cambia al momento de tener un error de detección, pero su valor sigue estando por debajo del umbral de detección básico que sería 0.5.




(a) Salida modelo SBM

(b) Salida función *softmax*

Figura 4.8: En (a) se encuentra la salida del modelo SBM para una prueba de validación y en (b) la salida de la función *softmax* aplicada sobre la salida obtenida por el modelo.

Luego de solucionar los errores puntuales, es necesario elegir un umbral de detección y un momento en el cual la rutina dará su decisión final. Cabe destacar que este momento debe ser antes del término del proceso, en donde para el caso del carguío autónomo, debe ser antes de los 20 [s], que es la duración promedio de éstos.

Para escoger la cantidad de muestras a esperar antes de tomar la decisión final, se probó la salida otorgada por la función *softmax* en las primeras 100 muestras (equivalente a 20 segundos), utilizando un umbral de detección de 0,5, resultados que se encuentran en la Figura 4.9.

En base a esta se concluyó que sobre las 44 muestras de espera para la decisión (equivalente a 8,8 segundos), el porcentaje de error se mantiene estable hasta aumentar a un total de 78 muestras de espera (equivalente a 15,6 segundos); por lo cual se escogió éste valor como el momento de la toma de decisión.

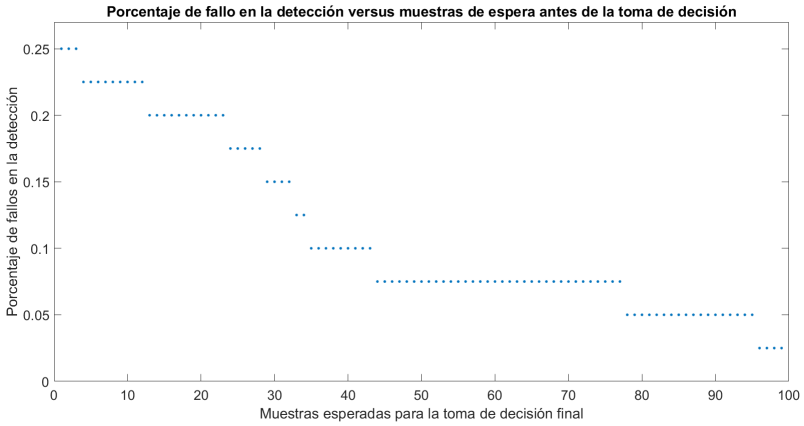



Figura 4.9: Porcentaje de error en la detección versus las muestras de espera antes de tomar la decisión final sobre la existencia de una anomalía o no.

Con lo anteriormente descrito se obtuvieron los resultados de la Tabla 4.3, en los cuales se presenta una mejora con respecto a los obtenidos en la Tabla 4.2, debido a que porcentaje de los errores contabilizados en las primeras pruebas son contrarrestados con el uso de la rutina de detección. En cuanto a la tasa de falsos negativos y positivos, se tiene una mayor tasa del primero, por lo tanto el detector entrega una mayor cantidad de alertas falsas, lo cual perjudica el tiempo en que el cargador se encontraría en operación.

| | |
|---|---|
| **Número de pruebas realizadas** | 40 |
| **Porcentaje de error promedio** | 7,5 % |
| **Tasa promedio de falsos negativos** | 5 % |
| **Tasa promedio de falsos positivos** | 2,5 % |

Tabla 4.3: Resultados obtenidos al correr la rutina de detección en las pruebas de validación realizadas.

Sin embargo, siguen existiendo casos como los mostrados en la Figura 4.10, en donde en ningún momento la salida del modelo es capaz de estimar la salida real del proceso, por lo cual la rutina de detección no puede corregir el error otorgado inicialmente por el modelo SBM ya que este no es puntual como en la Figura 4.8, sino mantenido en el tiempo, lo que significa que el modelo no es capaz de estimar la salida correcta con los datos incluidos en la matriz $D_{\text{i}}$.

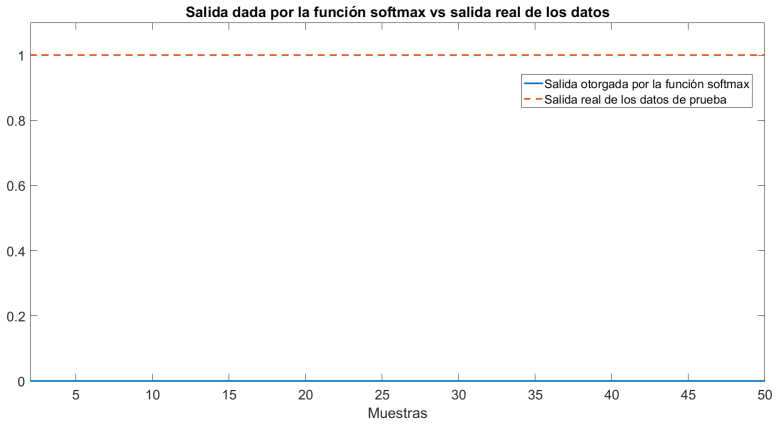



Figura 4.10: Error de detección otorgado tanto por el modelo SBM como por la salida de éste utilizando la función *softmax*.

## 4.5. Extensión a carguíos manuales

Si bien el trabajo realizado se enfoca en la detección de anomalías para el carguío autónomo de un cargador LHD, se probó la metodología planteada a una base de datos de carguíos manuales, los cuales fueron obtenidos a partir del mismo cargador LHD mientras éste era operado de forma remota por el personal de la minera 'San Gerónimo'. A continuación se detallan brevemente los detalles de los datos y las pruebas realizadas con éstos.

### 4.5.1. Base de datos utilizada

En la Tabla 4.4 se encuentra la información de las pruebas que componen la base de datos a utilizar. Cabe destacar que el etiquetado de los datos corresponde a si se llenó o no el balde del cargador luego del primer impacto contra la pala, pero no se tiene información del operario que utilizó el cargador en cada prueba.

| | **Número de carguíos** | **Carguíos exitosos** | **Carguíos fallidos** | **Tasa de muestreo** |
|---|---|---|---|---|
| Carguíos autónomos | 25 | 17 | 8 | 5 Hz |

Tabla 4.4: Tabla con información sobre la base de datos de carguíos manuales utilizada para probar el detector de anomalías diseñado.

En la Figura 4.11 se observan las diferencias del comportamiento de las variables entre los tipos de carguío. Para el carguío autónomo, dado que la secuencia a seguir es fija, tienen una duración menor, mientras que el carguío manual presenta mayores variaciones al momento de chocar contra la pila de recolección, las cuales cambiarán dependiendo del operador que haya utilizado el cargador.

Al igual que para el caso de los carguíos autónomos, se consideró únicamente el primer intento de recolección al chocar contra la pila para el etiquetado de los datos.

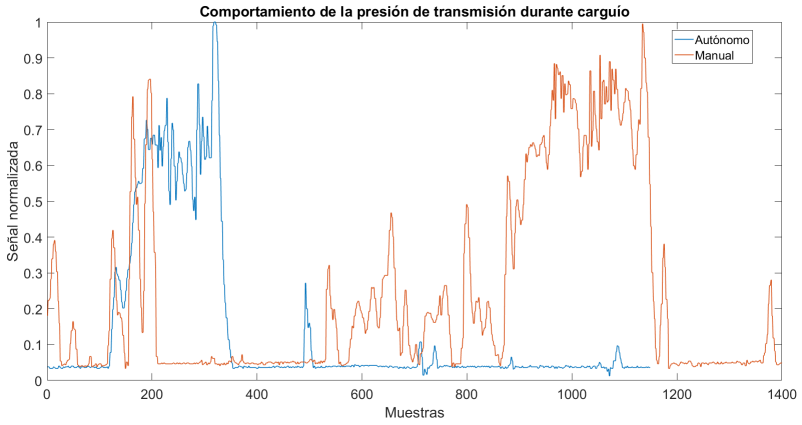



Figura 4.11: Gráfico del comportamiento de la tranmisión hidraúlica durante un proceso de carguío autónomo y otro manual.

### 4.5.2. Estructura modelo SBM

Para el desarrollo del modelo SBM se siguió de las mismas etapas realizadas para el carguío autónomo, utilizando las mismas variables explicativas del proceso e igual distribución en las filas de la matriz $D_{\mathrm{i}}$, lo cual fue explicado en la Sección 4.2. Para la creación del modelo SBM se utilizó un total de 15 pruebas, dejando las 10 restantes para la validación.

### 4.5.3. Validación del modelo

Para la validación del modelo nuevamente se utilizó del método de validación cruzada, en donde se realizaron 30 pruebas por separado con conjuntos de entrenamiento y validación diferentes, los cuales otorgaron los resultados expuestos en la Tabla 4.5, los cuales son considerablemente peores a los resultados obtenidos validando el modelo con los carguíos autónomos. Esto se debe a que la operación manual del cargador contiene características propias del usuario; es decir, cada persona maneja de forma distinta la máquina.

| | |
|---|---|
| **Número de pruebas realizadas** | 30 |
| **Porcentaje de error promedio** | 36,4 % |
| **Máximo porcentaje de error obtenido** | 68,1 % |
| **Mínimo porcentaje de error obtenido** | 10,6 % |
| **Tasa promedio de falsos negativos** | 19,7 % |
| **Tasa promedio de falsos positivos** | 16,7 % |
| **Tiempo promedio de ejecución** | 0,021 [s] |

Tabla 4.5: Resultados obtenidos de la validación cruzada del modelo SBM para carguíos manuales.

También se puede notar en la Tabla 4.5 una ligera diferencia en el tiempo de ejecución para los carguíos manuales. Esto se asocia a que la duración de los carguíos manuales es mayor a la de los autónomos y que además la base de datos presenta más pruebas de este tipo de carguío, por lo cual se puede crear la matriz $D_{\text{i}}$ con más observaciones. Esto genera que la multiplicación matricial esté compuesta de más elementos, pero al ser una operación sencilla el tiempo de ejecución aumenta levemente (un 5 % comparado al caso anteriormente estudiado).

### 4.5.4. Implementación rutina de detección

Para evitar que el error del modelo se vea afectado por los errores puntuales de estimación, se utilizó la función *softmax*, tal como fue utilizada en la Sección 4.4. En la Tabla 4.6 se encuentran los resultados obtenidos luego de aplicar la función a las salidas del modelo SBM y esperar 44 muestras antes de tomar la decisión final de si existe o no una anomalía.

Se decidió mantener el número de muestras de espera debido a que se desconoce si al momento de utilizar la máquina existirá un modo únicamente autónomo y otro manual, o bien el detector tendrá que funcionar para ambos por igual. En este último caso, conviene utilizar la misma configuración de la rutina de detección para ambos tipos de carguíos.

| | |
|---|---|
| **Número de pruebas realizadas** | 30 |
| **Porcentaje de error promedio** | 23,3 % |
| **Tasa promedio de falsos negativos** | 15,7 % |
| **Tasa promedio de falsos positivos** | 7,6 % |

Tabla 4.6: Resultados obtenidos al correr la rutina de detección en las pruebas de validación realizadas con los carguíos manuales.

Es importante destacar de la Tabla 4.6 que el porcentaje de error, a pesar de utilizar la rutina de detección, es considerablemente mayor al obtenido en el caso autónomo (7,5 %). Esto se explica mediante la dificultad de modelar la operación de cada usuario diferente, en donde se desconoce si la misma persona utilizó la máquina para las pruebas o si lo realizó de manera similar en todas éstas, situación que no se da en el caso autónomo debido a que el cargador obedece una secuencia de acciones. En cuanto a la tasa de falsos negativos y positivos, se tiene una mayor tasa del primero, lo que indica que el detector está inclinado a alarmar en casos donde no hay anomalías.

# Capítulo 5

# Conclusiones

De las pruebas realizadas con la metodología propuesta se logran diversas conclusiones, pero es importante destacar que éstas aplican para la maquinaria estudiada; y particularmente, para la utilizada en las pruebas realizadas en la minera 'San Gerónimo', pero la metodología diseñada puede ser extrapolada a otros usos en donde sea complejo obtener un modelo fenomenológico y exista una base de datos sobre el proceso a estudiar.

Como primera conclusión, se comprueba la factibilidad del uso de modelos no paramétricos para la modelación de procesos fenomenológicamente complejos, en donde la alta cantidad de variables de éstos no permite utilizar métodos paramétricos como lo sería un modelo ARX, por ejemplo. Sin embargo, es importante recalcar que la eficacia de los métodos basados en datos reside en las bases de datos disponibles para modelar los procesos. Para el trabajo realizado en esta memoria, los datos recolectados eran limitados debido al costo de las pruebas como a las limitaciones físicas presentadas en el momento de la toma de datos. Esto no significa que no se pueda realizar un modelo a partir de una base de datos pequeña, pero sí reduce el espectro de puntos de operación que se pueden representar con el modelo realizado, dado que es altamente probable que al momento de ser utilizado en condiciones reales, sobretodo en el caso de la minería que posee muchos imprevistos, la detección no sea tan precisa como lo fue en las pruebas de validación.

Segundo, se comprobó la efectividad de la metodología para el desarrollo de una rutina de detección de anomalías, en donde para el caso del carguío autónomo se consiguió bajar el error del modelo a un 7,5 % al utilizar las muestras pasadas obtenidas a partir del modelo SBM para tomar una decisión pasado un tiempo de la obtención de datos. En cuanto al tiempo de espera antes de la decisión, este corresponde aproximadamente a un tercio del tiempo que conlleva realizar el proceso de carguío, por lo cual el determinar si el detener la operación luego de los primeros 9 segundos genera un aumento de la eficacia del equipo (llámese un menor desgaste de los componentes), requiere un análisis más fino e incluir los factores económicos que podría significar detener la carga a la mitad versus dañar partes de la maquinaria como podría ser el daño de los neumáticos de ésta.

También es importante destacar el uso de heurísticas para la selección de variables relevantes cuando las técnicas más avanzadas no dan resultados concluyentes. En el caso de esta

memoria, al utilizar PCA no se logró llegar a una clara distinción de los estados anómalos y los normales, por lo cual se procedió a buscar relaciones entre las variables estudiadas en conjunto a conocimiento previo del proceso de carguío, con el cual fue posible encontrar variables que explicaran correctamente el proceso y generar un modelo SBM a partir de éstas.

Por otra parte, la metodología diseñada también pudo ser comprobada con datos de pruebas de carguío manual, en donde ahora el cargador LHD no seguía una secuencia programada, sino que era utilizado por un operador de forma remota. Para este caso se obtuvo que la metodología propuesta genera un modelo con un error del 23,3 %, considerablemente mayor al caso autónomo, a pesar de tener una base de datos de mayor tamaño para la creación del modelo. Esto comprueba que es más sencillo modelar en base a similitud un proceso que presenta características similares, como es en el caso del carguío autónomo en que la secuencia a seguir siempre es la misma. En cambio, al intentar realizar un modelo SBM sobre la operación del cargador LHD en manos de un humano, resulta considerablemente más complejo debido a que cada operador tendrá una forma particular de usar la máquina, por lo cual se necesita una base de datos considerablemente mayor para poder incluir estas características en las matrices de entrada y salida del modelo.

Como trabajo futuro, se plantea probar esta metodología con datos pertenecientes a otros cargadores para comprobar si únicamente sirve para el cargador LHD que fue diseñada. También se propone un estudio más profundo de los carguíos manuales, desarrollando un modelo que no necesariamente utilice las mismas variables que el carguío autónomo. Además, si se tiene la posibilidad de obtener más datos del uso de la máquina en terreno, se propone revalidar el modelo realizado y añadir a la matriz $D_{\text{i}}$ las observaciones que no sea capaz de estimar correctamente.

# Bibliografía


[1] León A. Detección de anomalías en procesos industriales usando modelos basados en similitud. Master's thesis, Universidad de Chile, 2012.

[2] S. Altug, Mo-Yuen Chen, and H. J. Trussell. Fuzzy inference systems implemented on neural architectures for motor fault detection and diagnosis. *IEEE Transactions on Industrial Electronics*, 46(6):1069–1079, Dec 1999.

[3] Villegas C. Araya L., Casado F. Señales y sistemas ii - apuntes del curso, 2017.

[4] Ding S. (auth.). *Data-driven Design of Fault Diagnosis and Fault-tolerant Control Systems*, pages 73–93. Advances in Industrial Control. Springer-Verlag London, 1 edition, 2014.

[5] Isermann R. (auth.). *Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault-tolerant Systems*. Springer-Verlag Berlin Heidelberg, 1 edition, 2011.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information science and statistics. Springer, 1st ed. 2006. corr. 2nd printing edition, 2006.

[7] Kumar V. Chandola V., Banerjee A. Anomaly detection: A survey. *ACM Computing Surveys*, 41, July 2009.

[8] Leo H. Chiang, Mark E. Kotanchek, and Arthur K. Kordon. Fault diagnosis based on fisher discriminant analysis and support vector machines. *Computers and Chemical Engineering*, 28(8):1389 – 1401, 2004.

[9] S. M. El-Shal and A. S. Morris. A fuzzy expert system for fault detection in statistical process control of industrial processes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(2):281–289, May 2000.

[10] Yong Gao, Xin Wang, Zhenlei Wang, and Liang Zhao. Fault detection in time-varying chemical process through incremental principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 158(Supplement C):102 – 116, 2016.

[11] L. Gong and D. Schonfeld. Space kernel analysis. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1577–1580, April 2009.

[12] Q. P. He and J. Wang. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, 20(4):345–354, Nov 2007.

[13] David Mautner Himmelblau. *Process Analysis by Statistical Methods*, pages 3–8. John Wiley and Sons, 1970.

[14] Jian Huang and Xuefeng Yan. Dynamic process fault detection and diagnosis based on dynamic principal component analysis, dynamic independent component analysis and bayesian inference. *Chemometrics and Intelligent Laboratory Systems*, 148(Supplement C):115 – 127, 2015.

[15] Barry L. Nelson David Nicol Jerry Banks, John Carson. *Discrete-Event System Simulation*. 4th Edition. Prentice Hall, 4 edition, 2004.

[16] Lennart Ljung. *System Identification: Theory for User*, pages 69–106. Prentice Hall, 2 edition, 1999.

[17] Eugenín M. Uso de modelos de similitud para detección de anomalías y modelos de predicción en procesos de concentración de minerales. Master's thesis, Universidad de Chile, 2015.

[18] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, pages 16–24. Adaptive Computation and Machine Learning. The MIT Press, 1 edition, 2012.

[19] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, pages 479–512. Adaptive Computation and Machine Learning. The MIT Press, 1 edition, 2012.

[20] Página oficial CAT. Underground mining load-haul-dump. `https://www.cat.com/en_US/products/new/equipment/underground-hard-rock/underground-mining-load-haul-dump-lhd-loaders.html`.

[21] Kjell A. Doksum Peter J. Bickel. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I (2nd Edition)*, pages 1–15. Prentice Hall, 2nd edition, 2000.

[22] Isserman R. Trends in the application of model-based fault detection and diagnosis of technical processes. *Control Engineering Practice*, 5:709–719, May 1997.

[23] S. Rajakarunakaran, P. Venkumar, D. Devaraj, and K. Surya Prakasa Rao. Artificial neural network approach for fault detection in rotary system. *Applied Soft Computing*, 8(1):740 – 748, 2008.

[24] Chiang L.H. (auth.) Russell E.L., Braatz R.D. *Fault Detection and Diagnosis in Industrial Systems*, chapter 1, pages 3–10. Advanced Textbooks in Control and Signal Processing. Springer-Verlag London, 1 edition, 2001.

[25] Chiang L.H. (auth.) Russell E.L., Braatz R.D. *Fault Detection and Diagnosis in Industrial Systems*, chapter 1, pages 35–54. Advanced Textbooks in Control and Signal Processing. Springer-Verlag London, 1 edition, 2001.

[26] Fuentealba S. Diseño e implementación de un sistema supervisor con modelos basados en similitud para la detección y aislamiento de fallas en turbina a gas natural. Master's thesis, Universidad de Chile, 2012.

[27] Wegerich S. Similarity-based modeling of vibration features for fault detection and identification. *Sensor Review*, 25:114–122, 2005.

[28] B. SAMANTA and K.R. AL-BALUSHI. Artificial neural network based fault diagnostic of rolling element bearings using time-domain features. *Mechanical Systems and Signal Processing*, 17(2):317 – 328, 2003.

[29] Jonas Sjöberg, Qinghua Zhang, Lennart Ljung, Albert Benveniste, Bernard Delyon, Pierre-Yves Glorennec, Håkan Hjalmarsson, and Anatoli Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691 – 1724, 1995.

[30] D. F. Specht. A general regression neural network. *IEEE Transactions on Neural Networks*, 2(6):568–576, Nov 1991.

[31] F. A. Tobar, L. Yacher, R. Paredes, and M. E. Orchard. Anomaly detection in power generation plants using similarity-based modeling and multivariate analysis. In *Proceedings of the 2011 American Control Conference*, pages 1940–1945, June 2011.