



**UNIVERSIDAD DE CHILE**  
**FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS**  
**DEPARTAMENTO DE INGENIERÍA DE MINAS**

**VALIDACIÓN DE MODELOS NUMÉRICOS EN GEOCIENCIAS A  
PARTIR DE SIMULACIONES GEOESTADÍSTICAS CON ÉNFASIS EN  
SIMULACIONES MULTIPUNTO**

**TESIS PARA OPTAR AL GRADO DE MAGISTER EN MINERÍA**

**ANDRÉS ECHEVERRI LONDOÑO**

**PROFESOR GUÍA:**

**JULIÁN M. ORTÍZ CABRERA**

**MIEMBROS DE LA COMISIÓN:**

**JORGE SILVA SÁNCHEZ**

**ALEJANDRO CÁCERES SAAVEDRA**

**PAULA LARRONDO DURAN**

**SANTIAGO DE CHILE**

**2016**

**RESUMEN DE LA TESIS PARA OPTAR AL  
GRADO DE: MAGÍSTER EN MINERÍA  
POR: ANDRÉS ECHEVERRI LONDOÑO  
FECHA: 16/05/2016  
PROFESOR GUÍA: Julián M. Ortiz C.**

**VALIDACIÓN DE MODELOS NUMÉRICOS EN GEOCIENCIAS A PARTIR DE  
SIMULACIONES GEOESTADÍSTICAS CON ÉNFASIS EN SIMULACIONES  
MULTIPUNTO**

Este documento aborda el tema de la validación de las simulaciones geoestadísticas empleadas para generar modelos numéricos en geociencias. Estos modelos son de gran importancia en la minería, toda vez que muchas de las decisiones que en esta industria se toman, parten del conocimiento generado por estas interpretaciones de la realidad. En ese orden de ideas, el documento examina el estado del arte en cuanto a diversas métricas empleadas para validar las simulaciones geoestadísticas, haciendo énfasis en las simulaciones multi-punto.

En primer lugar se hace un recuento de diversas propuestas efectuadas por algunos autores quienes han presentado sus propuestas o enfoques relacionados con criterios que se deberían tener en cuenta al momento de validar este tipo de modelos. Esto se clasificó en 4 niveles de validación: un primer nivel de condicionamiento de los datos, un segundo nivel de validación de estadística, un tercer nivel de validación geoestadística y finalmente un nivel de validación de predicción.

Se propone acorde a esto una metodología para abordar el tema, consistente en la aplicación a dos casos prácticos de diferentes métricas de validación para determinar su potencial uso en diferentes situaciones.

Un primer caso está relacionado con una variable continua, la cual fue simulada a partir de información extraída de un modelo sintético realístico de plegamientos. Las simulaciones se efectuaron con dos tipos de algoritmos. El primero corresponde a un algoritmo paramétrico que emplea variogramas para su inferencia como es el caso de SGSIM (simulación secuencial gaussiana). El segundo es un algoritmo no paramétrico (simulación multi-punto) que usa una imagen de entrenamiento para inferir la función de distribución y simular las diferentes realizaciones (FILTERSIM – simulación basada en puntuación por filtros).

El segundo caso es una variable categórica de un modelo de formaciones aluviales, del cual fue extraída la información. En este caso se emplearon dos algoritmos no paramétricos. Uno de ellos es SISIM (simulación secuencial de indicadores) que emplea los variogramas de una variable indicador para inferir el modelo. El otro corresponde a SNESIM (simulación ecuación normal simple) que emplea una imagen de entrenamiento para determinar la función de distribución de probabilidad para poder generar el modelo.

A partir de estas simulaciones se ponen en práctica diversas métricas para validar los modelos obtenidos, lo cual permite determinar el desempeño de estas en el proceso de validación. Con dicha información y a partir de revisar los supuestos fundamentales relacionados con la Estacionariedad de las variables y de la inferencia de los modelos, se hace una propuesta de pasos a seguir para validar un modelo a partir de simulaciones geoestadísticas.

# ABSTRACT

This paper addresses the topic of validation of geostatistical simulations used to generate numerical models in geosciences. These models are very important in the mining industry, since many of the decision-making processes are based on the knowledge given by these interpretations of reality. Therefore, this work explores the state of the art of several techniques used to validate the geostatistical simulations, with emphasis on multipoint simulations.

First, proposals or approaches made by some authors that should be taken into account when validating these models are discussed. These proposals are classified into 4 levels of validation: a first level of conditioning data, a second level of statistical validation, a third level of geostatistical validation and finally a fourth level of prediction validation.

Then, a methodology is proposed to address the validation issue. This involves the application of several metrics to two practical cases to determine their potential use in different situations.

The first case is related to a continuous variable, which was simulated using information extracted from a synthetic realistic model of folds. The simulations were performed with two types of algorithms. The first is a parametric algorithm which uses variograms for inference: SGSIM (sequential Gaussian simulation). The second algorithm is nonparametric (multipoint simulation) that uses a training image to infer the distribution function and then simulates several realizations: FILTERSIM (score for filters based simulation).

The second case relies on a categorical variable. The data were extracted from a realistic alluvial formation model. In this case two nonparametric algorithms were used. One of them was SISIM (sequential indicator simulation) which uses variograms of an indicator variable to infer the model. The other algorithm was SNESIM (single normal equation simulation) which uses a training image to calculate the probability distribution function to generate the model.

Once the realizations from the simulation processes were obtained, several validation techniques were implemented to evaluate those models. This procedure allowed to check the performance of those techniques in the validation of simulation models. With these results, and also the review of underlying assumptions of stationarity and inference of models and algorithms; a proposal (step by step) was developed to perform a validation process of geostatistical simulations.

# DEDICATORIA

*A mi esposa, mis padres y mi abuela*

## AGRADECIMIENTOS

En primer lugar quisiera agradecer al profesor Julián Ortíz por su valiosa guía en la realización de este trabajo, al permitirme darle un enfoque amplio al mismo que permitiera su aplicación también en el ámbito ambiental e igualmente por su gran soporte en la realización del Magíster en Minería, desde el momento mismo de la aplicación a la Universidad de Chile.

De igual forma, agradecer a los miembros de la comisión evaluadora, profesores Jorge Silva, Alejandro Cáceres y Paula Larrondo; por sus valiosos aportes para enriquecer la discusión planteada en este trabajo y su invaluable colaboración en los tiempos de revisión de la tesis.

Agradecer al Proyecto de Colaboración entre BHP Billiton y el Laboratorio ALGES que permitieron el financiamiento de los estudios de Magister, así como al AMTC.

Me gustaría agradecer también a los compañeros de Magíster con quienes compartí más tiempo y los cuales me apoyaron incondicionalmente en la obtención de este grado: Yesenia, Iván, Mauricio, Alexander, Mario, Vanesa, Loli, Ccatamayo, Yina, Jose, María Camila. Gracias compañeros por todo su apoyo. Tendré siempre gratos recuerdos de todos ustedes. A Barbara y Nicole quienes siempre tuvieron tiempo para atenderme y aclarar mis inquietudes.

Gracias a mis padres por creer de manera incondicional en mí, y que aunque se encuentren a la distancia, siempre compartirán conmigo la satisfacción de los logros obtenidos. A Clemencia por estar siempre atenta de nuestro bienestar a la distancia y en sus visitas a Chile.

Finalmente, agradecer a mi esposa, quien me impulsó a realizar el Magíster, me brindó su apoyo en todo momento y me da siempre lecciones de como ver la vida de la manera más positiva.

# TABLA DE CONTENIDO

CAPÍTULO 1. INTRODUCCIÓN .....	10
1.1. JUSTIFICACIÓN DEL TRABAJO .....	11
1.2. OBJETIVOS .....	11
1.2.1. Objetivo general .....	11
1.2.2. Objetivos específicos .....	12
1.3. ALCANCE DEL ESTUDIO .....	12
CAPÍTULO 2. ESTADO DEL ARTE .....	14
2.1. ANTECEDENTES GENERALES DE GEOESTADÍSTICA .....	14
2.1.1. Variable aleatoria .....	14
2.1.2. Función aleatoria .....	15
2.1.3. Estacionariedad .....	16
2.2. SIMULACIONES .....	17
2.2.1. SISIM (Sequential Indicator Simulation) .....	20
2.2.2. SNESIM (Single Normal Equation Simulation) .....	20
2.2.3. FILTERSIM (Filter-Based Simulation) .....	21
2.2.4. SGSIM (Sequential Gaussian Simulation) .....	22
2.3. Validación de simulaciones .....	22
2.3.1. Nivel 1. Condicionamiento de los datos .....	23
2.3.1.1. Inspección visual. ....	24
2.3.1.2. Gráfico de dispersión. ....	24
2.3.2. Nivel 2. Validación estadística .....	24
2.3.3. Nivel 3. Validación geoestadística .....	28
2.3.3.1. Validación de estadísticas de dos puntos. ....	28
2.3.3.2. Validación de estadísticas multipunto. ....	30
2.3.4. Nivel 4. Validación de predicción. ....	36
CAPÍTULO 3. METODOLOGÍA .....	39
CAPÍTULO 4. CASO PRÁCTICO VARIABLE CONTINUA .....	41
4.1. MODELO Y ALGORITMOS EMPLEADOS .....	41
4.1.1. Análisis para validación de simulaciones de variable continua .....	44
4.1.1.1. Nivel 1. Condicionamiento de Datos .....	44
4.1.1.2. Nivel 2. Validación estadística .....	47
4.1.1.3. Nivel 3. Validación geoestadística .....	54
CAPÍTULO 5. CASO PRÁCTICO VARIABLE CATEGÓRICA .....	65
5.1. Modelo y algoritmos empleados .....	65
5.1.1. Análisis para validación de simulaciones .....	69
5.1.1.1. Nivel 1. Condicionamiento de Datos .....	69
5.1.1.2. Nivel 2. Validación estadística .....	71
5.1.1.3. Nivel 3. Validación geoestadística .....	74
CAPÍTULO 6. CONCLUSIONES .....	88
6.1. SUPUESTOS FUNDAMENTALES DE LAS VALIDACIONES .....	88
6.2. DISCUSIÓN SOBRE MÉTRICAS DE VALIDACIÓN .....	89
6.3. BUENAS PRÁCTICAS DE VALIDACIÓN .....	92
BIBLIOGRAFÍA .....	95
ANEXOS .....	98
ANEXO A – Parámetros empleados en simulaciones .....	98
ANEXO B – Gráficos de correlación diferida .....	101

ANEXO C – Variograma modelado SGSIM .....	102
ANEXO D – QQ Plots para diferentes realizaciones de FILTERSIM.....	103
ANEXO E – QQ Plots para diferentes realizaciones de SGSIM .....	104
ANEXO F – Reportes de Test estadísticos.....	105
ANEXO G – Parámetros de cumulantes de orden 3 .....	109
ANEXO H – Código MatLab frecuencia de patrones.....	111
ANEXO I – Código MatLab Accuracy plot.....	112
ANEXO J – Variogramas de variable indicador SISIM.....	115

## ÍNDICE DE TABLAS

Tabla 3.1: Software y códigos empleados en el estudio .....	40
Tabla 4.1: Estadísticas descriptivas de información de entrada .....	42
Tabla 4.2: Resultados del test de reproducción de media para FILTERSIM .....	53
Tabla 4.3: Resultados del test de reproducción de media para SGSIM .....	53
Tabla 4.4: Resultados del test de reproducción de varianza para FILTERSIM/SGSIM .....	53
Tabla 4.5: Error medio cuadrático entre variogramas (FILTERSIM/SGSIM) para 85 m .....	56
Tabla 4.6: Test de Hotelling para validación de variogramas FILTERSIM/SGSIM .....	56
Tabla 5.1: Proporciones de categorías variable categórica.....	66
Tabla 5.2: Resultados del test de reproducción de media de muestras para SNESIM .....	73
Tabla 5.3: Resultados del test de reproducción de media de muestras para SISIM .....	73
Tabla 5.4: Resultados del test de reproducción de varianza para SNESIM .....	74
Tabla 5.5: Resultados del test de reproducción de varianza para SISIM .....	74
Tabla 5.6: EMC entre variogramas (SNESIM/SISIM) para categoría 0 (51 m) .....	77
Tabla 5.7: EMC entre variogramas (SNESIM/SISIM) para categoría 3 (51 m) .....	77
Tabla 5.8: Test de Hotelling para validación de variogramas SNESIM/SISIM.....	77



# ÍNDICE DE FIGURAS

Figura 2-1: Inspección visual. ....	24
Figura 2-2. Gráfico de dispersión. ....	25
Figura 2-3: Comparación de histogramas. ....	25
Figura 2-4: Q-Q Plot para comparar distribuciones de datos simulados y datos originales. ....	26
Figura 2-5. Análisis de distribución de medias y varianzas de 40 realizaciones. ....	26
Figura 2-6: Análisis de reproducción del variograma por diferentes realizaciones. ....	29
Figura 2-7: Análisis comparativo de cuerpos obtenidos de simulaciones. ....	31
Figura 2-8: Análisis comparativo de error medio cuadrático de cumulantes de orden dos y orden tres ....	33
Figura 2-9: Función de densidad de múltiples puntos para dos imágenes de entrenamiento. ....	34
Figura 2-10: Validación de función de densidad de múltiples puntos. ....	35
Figura 2-11: Gráfico de frecuencias de patrones para la TI y realización. ....	35
Figura 2-12: Análisis comparativo de conectividad. ....	36
Figura 2-13: Gráfico de dispersión para validar distribuciones de probabilidad. ....	37
Figura 3-1: Esquema metodológico del trabajo. ....	39
Figura 4-1: Modelo continuo de plegamiento ....	41
Figura 4-2: Imágenes extraídas del modelo continuo y empleadas para las simulaciones. ....	41
Figura 4-3: 135 muestras extraídas aleatoriamente de modelo de la realidad. ....	42
Figura 4-4: Box-plot de datos de entrada para simulaciones. ....	42
Figura 4-5: Ejemplo de realizaciones 5 y 74 obtenidas con FILTERSIM (FS) ....	43
Figura 4-6: Ejemplo de realizaciones 41 y 95 obtenidas con FILTERSIM (FS) ....	44
Figura 4-7: Ejemplo de realizaciones obtenidas con SGSIM (SG) ....	44
Figura 4-8: Inspección visual de muestras y realizaciones para FILTERSIM ....	45
Figura 4-9: Inspección visual de muestras y realizaciones para SGSIM ....	46
Figura 4-10: Gráfico de dispersión de valores simulados vs. valores originales ....	47
Figura 4-11: Histogramas para comparar reproducción de estadísticas - FILTERSIM ....	48
Figura 4-12: Histogramas para comparar reproducción de estadísticas - SGSIM ....	49
Figura 4-13: Q-Q plot para análisis de distribuciones FILTERSIM ....	50
Figura 4-14: Q-Q plot para análisis de distribuciones SGSIM. ....	51
Figura 4-15: Curvas de selectividad (Fracción de volumen vs umbral) ....	52
Figura 4-16: Reproducción de variogramas para FILTERSIM. ....	55
Figura 4-17: Reproducción de variogramas para SGSIM. ....	55
Figura 4-18: Esquema de configuración para cálculo de cumulantes de orden 3 ....	57
Figura 4-19: Cumulantes de orden 3 (Realizaciones vs. TI). ....	58
Figura 4-20: Cumulantes de orden 3 (Realizaciones vs. Imagen exhaustiva). ....	58
Figura 4-21: Patrones definidos para construcción de función de densidad de múltiples puntos... ..	59
Figura 4-22: Función de densidad de múltiples puntos – realizaciones vs. TI. ....	60
Figura 4-23: Función de densidad de múltiples puntos – realizaciones vs. realidad. ....	60
Figura 4-24: Gráfico de dispersión de frecuencias de patrones 2x2 TI vs. algoritmos FILTERSIM/SISIM. ....	61
Figura 4-25: Gráfico de dispersión de frecuencias de patrones 2x2 realidad vs. algoritmos FILTERSIM/SISIM. ....	61
Figura 4-26: Gráfico de dispersión de frecuencias de patrones 3x3 TI vs. algoritmos FILTERSIM/SGSIM ....	62
Figura 4-27: Gráfico de dispersión de frecuencias de patrones 3x3 realidad vs. algoritmos FILTERSIM/SGSIM ....	62

Figura 4-28: Accuracy plot realizaciones FILTERSIM/SGSIM para diferentes muestreos .....	63
Figura 4-29: Box-plot datos originales vs muestreos alternativos .....	64
Figura 5-1: Modelo categórico 3D del acuífero. ....	65
Figura 5-2: Imágenes extraídas del modelo categórico y empleadas para las simulaciones.....	65
Figura 5-3: 102 muestras extraídas aleatoriamente del modelo de la realidad .....	66
Figura 5-4: Histogramas con frecuencias de categorías .....	66
Figura 5-5: Ejemplos de realizaciones obtenidas para SNESIM (SN).....	68
Figura 5-6: Ejemplos de realizaciones 17 y 69 obtenidas para SISIM (SS).....	68
Figura 5-7: Ejemplo de realizaciones 52 y 91 obtenidas con SISIM (SS) .....	69
Figura 5-8: Inspección visual de muestras y realización 17 - SNESIM.....	69
Figura 5-9: Inspección visual de muestras y realización 91 - SNESIM.....	70
Figura 5-10: Inspección visual de muestras y realización 17 - SISIM.....	70
Figura 5-11: Inspección visual de muestras y realización 91 - SISIM .....	71
Figura 5-12: Gráfico de dispersión de valores simulados vs. valores originales .....	72
Figura 5-13: Reproducción de proporciones SNESIM/SISIM.....	72
Figura 5-14: Gráfico comparativo de variogramas para SNESIM Categoría 0. ....	75
Figura 5-15: Gráfico comparativo de variogramas para SNESIM Categoría 3. ....	75
Figura 5-16: Gráfico comparativo de variogramas para SISIM Categoría 0. ....	76
Figura 5-17: Gráfico comparativo de variogramas para SISIM Categoría 0. ....	76
Figura 5-18: EMC cumulantes realizaciones vs. TI (Categoría 0) .....	78
Figura 5-19: EMC cumulantes realizaciones vs. realidad (Categoría 0) .....	79
Figura 5-20: Imágenes binarias de categoría 0 .....	79
Figura 5-21: Ejemplo de imágenes binarias de realizaciones obtenidas para categoría 0.....	80
Figura 5-22: EMC cumulantes realizaciones vs. TI (Categoría 3) .....	80
Figura 5-23: EMC cumulantes realizaciones vs. realidad (Categoría 3).....	81
Figura 5-24: Mapa de cumulantes L orden 3 (categoría 3) .....	82
Figura 5-25: Imágenes binarias de categoría 3 .....	82
Figura 5-26: Función de densidad de múltiples puntos cat0 – realizaciones vs. TI.....	83
Figura 5-27: Función de densidad de múltiples puntos cat0 – realizaciones vs. realidad.....	83
Figura 5-28: Función de densidad de múltiples puntos cat3 – realizaciones vs. TI.....	84
Figura 5-29: Función de densidad de múltiples puntos cat3 – realizaciones vs. realidad.....	84
Figura 5-30: Gráfico de dispersión de frecuencias de patrones 2x2 TI cat 0 vs. algoritmos SNESIM/SISIM.....	85
Figura 5-31: Gráfico de dispersión de frecuencias de patrones 2x2 realidad cat 0 vs. algoritmos SNESIM/SISIM.....	85
Figura 5-32: Gráfico de dispersión de frecuencias de patrones 2x2 TI cat 0 vs. algoritmos SNESIM/SISIM.....	86
Figura 5-33: Gráfico de dispersión de frecuencias de patrones 2x2 realidad cat 0 vs. algoritmos SNESIM/SISIM.....	86
Figura 5-34: Gráfico de dispersión de frecuencias de patrones 3x3 TI cat 0 vs. algoritmos SNESIM/SISIM.....	86
Figura 5-35: Gráfico de dispersión de frecuencias de patrones 3x3 realidad cat 0 vs. algoritmos SNESIM/SISIM.....	87
Figura 5-36: Gráfico de dispersión de frecuencias de patrones 3x3 TI cat 3 vs. algoritmos SNESIM/SISIM.....	87
Figura 5-37: Gráfico de dispersión de frecuencias de patrones 3x3 realidad cat 3 vs. algoritmos SNESIM/SISIM.....	87
Figura 6-1: Diagrama de flujo para validar simulaciones .....	94

# CAPÍTULO 1. INTRODUCCIÓN

Los modelos numéricos en geociencias son un insumo primordial en la industria minera para la toma de decisiones. La geoestadística ha sido una rama de la estadística aplicada que ha aportado en gran medida a los avances en esta área y los estimadores lineales como el kriging han sido los abanderados en este proceso. Sin embargo, una fuente importante y creciente en la generación de dichos modelos vienen siendo las simulaciones geoestadísticas dado su potencial para aportar información sobre el modelo de incertidumbre; permitiendo realizar análisis de riesgo, estimaciones a partir de los promedios de las realizaciones y análisis de incertidumbre a través del análisis de la dispersión de las realizaciones.

En igual medida, una rama creciente de estas herramientas vienen siendo las simulaciones multi-punto; las cuales tratan de modelar complejos comportamientos no lineales, que técnicas tradicionales de la geoestadística no permitían hacer dado su formalismo. Estas técnicas no paramétricas emplean una fuente exhaustiva de información complementaria (en la forma de una imagen de entrenamiento), la cual puede corresponder a un modelo conceptual, una interpretación, un escaneo o información obtenida de fuentes adicionales (p. ej. geofísica), y a partir de esta se construye, por así decirlo, la función de distribución acumulada condicional para generar los modelos. Esto implica, por cierto, tener en cuenta ciertas limitaciones dado que esa fuente de información debería ser lo suficientemente grande para poder generar el campo aleatorio representativo de la función aleatoria. Sin embargo, en la práctica, en muchos casos no se cuenta con una fuente tan extensa y también limitaciones asociadas a los recursos computacionales disponibles para procesarlas. Situación similar ocurre con simulaciones paramétricas las cuales también presentan ciertas limitaciones al momento de construir un modelo.

Los anteriores aspectos implican grandes desafíos para los modeladores, que deben construir aproximaciones de la realidad a partir de información escasa, haciendo necesario igualmente realizar supuestos para poder poner en práctica complejos algoritmos, simplificaciones para hacerlos viables computacionalmente, entre muchos otros factores. Todo eso conlleva a la necesidad de efectuar chequeos de los resultados obtenidos, a fin de valorar la validez de los modelos, lo que implica igualmente grandes retos dado el desconocimiento en muchos casos del comportamiento de las variables en estudio.

Grandes esfuerzos se han venido desarrollando por diversos autores, proponiendo mejores y más eficientes algoritmos que tratan de simular de mejor forma los complejos comportamientos encontrados en la naturaleza y que requieren ser modelados para una mejor toma de decisiones en la actualidad. Ello ha conllevado igualmente a proponer diversas métricas que permitan analizar el comportamiento o desempeño de dichos algoritmos conforme a los objetivos planteados.

Este trabajo pretende entonces recopilar toda una serie de herramientas que se encuentran de alguna manera dispersas, pero que pueden ser integradas para construir una propuesta que permita estipular unas prácticas mínimas de validación de modelos a partir de simulaciones geoestadísticas, haciendo énfasis en las simulaciones multi-punto y de esta manera poder tener una mayor confianza en los resultados de los procesos de modelación a través de un protocolo de validación.

## **1.1. JUSTIFICACIÓN DEL TRABAJO**

En la industria minera diversos fenómenos naturales deben ser caracterizados, dada su importancia tanto en el ámbito económico, ambiental y productivo; los cuales determinan en cierta medida la viabilidad de un proyecto minero.

En la actualidad, muchos de estos fenómenos naturales (como la mineralización de un yacimiento, el flujo de aguas subterráneas, el comportamiento de la roca, entre otros) son caracterizados a partir de muestras puntuales, en algunos casos espaciadas a distancias considerables. Para suplir esta falta de información, desde hace varias décadas se han usado modelos numéricos que tratan de reproducir o caracterizar dichos fenómenos de manera continua en el espacio, así como la incertidumbre de estas, y es allí donde la geoestadística ha jugado un papel preponderante logrando avances significativos en este campo. Sin embargo, técnicas convencionales de geoestadística en algunas ocasiones no logran reproducir adecuadamente fenómenos más complejos. Es por ello que han surgido “tendencias” de la geoestadística que trabajan a partir de algoritmos no paramétricos, dentro de las cuales se encuentran las simulaciones multi-punto las cuales emplean información de referencia para la construcción de las diversas realizaciones de un modelo.

No obstante, dichos modelos deben ser validados considerando las limitaciones de la modelación, como suposiciones que en ciertos casos son necesarias asumir, por el tamaño del dominio seleccionado, por las simplificaciones efectuadas de cierto fenómeno para poder resolver un problema en particular, por las decisiones del mismo modelador, por los datos disponibles o por las limitaciones mismas de los algoritmos matemáticos empleados para efectuar dichas modelaciones. Esto puede derivar en ciertas desviaciones respecto a lo que se pretende obtener como resultado.

Sin embargo, para que los modelos numéricos obtenidos respondan al comportamiento de la variable que se pretende estudiar, es fundamental implementar buenas prácticas en la modelación; desde el diseño y la captura de los datos requeridos para modelar, hasta validar los resultados obtenidos.

Este estudio se enfoca principalmente en esta última etapa, revisando y analizando diferentes métricas de validación presentadas por diversos autores, proponiendo un flujo de trabajo con base a la utilidad de dichas métricas según el análisis efectuado; para luego realizar una propuesta que de forma sistémica integre diversos niveles de validación y permita verificar si los resultados de una modelación pueden ser considerados aceptables según los objetivos planteados, lo cual es de suma importancia en la minería y las geociencias dado que gran parte de la toma de decisiones se basan en información aportada por muchos de estos modelos y se requiere una medida de la calidad de los mismos.

## **1.2. OBJETIVOS**

A continuación se presentan el objetivo general y los objetivos específicos del presente trabajo.

### **1.2.1. Objetivo general**

Proponer una metodología mediante un enfoque sistémico de validación de modelos numéricos en geociencias (validación de simulaciones o estimaciones geoestadísticas con énfasis en simulaciones multi-punto) que considere diferentes niveles de verificación; a partir de criterios

mínimos de validación, que tomen en cuenta aspectos como el condicionamiento de los datos, reproducción de estadísticas básicas y reproducción de métricas o herramientas geoestadísticas, así como la predicción.

### **1.2.2. Objetivos específicos**

- Efectuar una revisión bibliográfica de los tipos de validaciones o criterios propuestos por diversos autores para validar modelos numéricos en geociencias (basados en simulaciones geoestadísticas).
- Seleccionar los criterios más robustos para validación conforme a la revisión, para su posterior aplicación en dos casos prácticos.
- Generar y validar un modelo de una variable continua mediante el uso de un algoritmo gaussiano y multi-punto (SGSIM y FILTERSIM) y de igual forma una variable categórica empleando algoritmos no paramétricos (SISIM y SNESIM).
- Discutir los hallazgos encontrados en los casos prácticos, enfatizando los aspectos favorables de la propuesta, sus limitaciones, proponer un flujo de trabajo de validación y posibles aspectos a tener en cuenta en posteriores estudios.

### **1.3. ALCANCE DEL ESTUDIO**

De acuerdo a los objetivos planteados en el numeral anterior, es importante definir los alcances del presente estudio, con los cuales se pretenden definir los límites de actuación o maniobra en búsqueda de dichos objetivos.

Dado que la modelación numérica aplicada a las geociencias implica un panorama muy amplio de técnicas y herramientas, es fundamental aclarar que el presente estudio estará enfocado a la modelación numérica a través de herramientas geoestadísticas, tales como las simulaciones. En ese orden de ideas, el análisis de información bibliográfica o estado del arte estará enfocado a revisar las técnicas o métodos de validación propuestos en ese ámbito de aplicación.

De igual forma, la propuesta que integre varias de estas metodologías de acuerdo a sus bondades estará enmarcada a la validación de las simulaciones geoestadísticas, con énfasis en simulaciones multi-punto; motivo por el cual los algoritmos empleados para simular las variables continua y categórica contendrán algoritmos de simulaciones multi-punto (FILTERSIM para el caso de la variable continua y SNESIM para el caso de la variable categórica). Sin embargo, también se aplicarán las métricas a los resultados de las simulaciones a un algoritmo paramétrico (simulación secuencial gaussiana para el caso de la variable continua) y otro tipo de algoritmo no paramétrico (como simulación secuencial de indicadores) para una variable categórica.

Por otro lado, si bien existen en la literatura gran cantidad de métricas propuestas por diversos autores, este trabajo seleccionará algunas de estas para acotar el estudio y dado a que algunas pueden analizar aspectos similares. Por ejemplo, es el caso de los cumulantes espaciales de orden superior que permiten efectuar análisis de conectividad, tamaños de cuerpos, motivo por el cual no se ponen en práctica otras técnicas relacionadas con estos temas. También, por ejemplo, el análisis de patrones mediante funciones de densidad de múltiples puntos o frecuencias analizan el tema de patrones que bien podrían ser analizados mediante otro tipo de métricas como la propuesta por Pérez (2013).

Finalmente, a partir de la experiencia de aplicación y un mejor entendimiento de las métricas aplicadas a las simulaciones, se propondrá una metodología que integre diversas herramientas de validación; hasta el momento sugeridas de manera independiente por sus autores, mediante un diagrama que permita contar con un flujo de trabajo ordenado y sistemático para abordar el tema de la validación y pueda convertirse en una guía de buenas prácticas para validar simulaciones geoestadísticas.

## CAPÍTULO 2. ESTADO DEL ARTE

### 2.1. ANTECEDENTES GENERALES DE GEOESTADÍSTICA

La geoestadística puede ser definida como la aplicación de modelos probabilísticos a variables regionalizadas (Chilès y Delfiner, 2009), es decir variables (atributos) distribuidas en el espacio y con cierta continuidad espacial.

Por su parte, Pyrcz y Deutsch (2014) definen a la geoestadística como una rama de la estadística aplicada que difiere de la estadística en tres aspectos básicos

- Se centra en el origen geológico de los datos (geológico por su origen asociado a la minería y geología; pero que actualmente es más amplio pudiéndose referir a las geociencias).
- Modelamiento y tratamiento explícito de la correlación espacial entre los datos
- Tratamiento de los datos a diferentes volúmenes y niveles de precisión

Esto conlleva a la necesidad de entender el concepto de variable aleatoria y función aleatoria; dado que la variable regionalizada es entendida como una realización de la función aleatoria y esta a su vez al conjunto de variables aleatorias en el espacio (o dominio D).

#### 2.1.1. Variable aleatoria

Es una variable que puede tomar una serie de posibles valores, cada uno con una cierta probabilidad o frecuencia de ocurrencia (Deutsch y Journel, 1998). Comúnmente denotada por la letra mayúscula Z y sus posibles valores con z minúscula.

En el caso discreto, cada posible valor  $z_i$  tiene asociado una probabilidad de ocurrencia:

$$p_i = \text{Prob}\{Z = z_i\} \in [0,1], \quad \text{con :} \quad \sum_{i=1}^n p_i = 1 \quad (2.1)$$

Para el caso de una variable continua la distribución de probabilidad puede tomar la forma de una función de distribución acumulada  $F(z)$  o una función de densidad de probabilidad  $f(z)$ .

$$F(z) = \text{Prob}\{Z \leq z\} \in [0,1] \quad (2.2)$$

$$f(z) = dF(z)/dz \quad (2.3)$$

Tanto para la función de distribución acumulada, como para la función de densidad de probabilidad, los intervalos pueden ser derivados.

$$\text{Prob}\{Z \in [a, b]\} = F(b) - F(a) = \int_a^b f(z)dz \quad (2.4)$$

De la distribución de la variable aleatoria Z, momentos específicos pueden ser derivados, tales como la media, la mediana y la varianza (Remy et al, 2009).

Media  $m$

$$\begin{aligned} m &= E\{Z\} = \int_{z_{\min}}^{z_{\max}} z dF(z) = \int_{z_{\min}}^{z_{\max}} z f(z) d_z \\ &= \frac{1}{L} \sum_{i=1}^L z^{(i)} \end{aligned} \quad (2.5)$$

Mediana  $M$  (cuantil 0.5)

$$M = q(0.50): \text{valor tal que } Prob\{Z \leq M\} = 0.50 \quad (2.6)$$

Varianza  $\sigma^2$

$$\begin{aligned} \sigma^2 &= Var\{Z\} = E\{(Z - m)^2\} = \int_{z_{\min}}^{z_{\max}} (z - m)^2 f(z) d_z \\ &= \frac{1}{L} \sum_{i=1}^L (z^{(i)} - m)^2 \end{aligned} \quad (2.7)$$

### 2.1.2. Función aleatoria

Una función aleatoria, denotada como  $Z(u)$  es un conjunto de variables aleatorias dependientes  $\{Z(u), u \in D\}$ , donde  $u = \{x, y, z\}$  es un vector de coordenadas en el espacio o dominio  $D$ .

La función aleatoria permite dar respuesta al requerimiento planteado por las diversas aplicaciones de la geoestadística, el cual es la consideración conjunta de diversas variables aleatorias en el espacio; las cuales en ocasiones se conocen en algunos sitios definidos del espacio a través de un muestreo y desconocidas en el resto del espacio con cierto grado de incertidumbre pero modeladas como funciones aleatorias (Remy et al, 2009).

Al igual que la variable aleatoria  $Z$  está caracterizada por una función de distribución (por ejemplo su función de distribución acumulada  $F(z)$ ); la función aleatoria  $Z(u)$  también puede ser caracterizada por su función de distribución multivariable (es decir, su distribución espacial),

$$F(z_1, \dots, z_n; u_1, \dots, u_n) = Prob\{Z(u_1) \leq z_1, \dots, Z(u_n) \leq z_n \mid u \in D\} \quad (2.8)$$

La función aleatoria también puede ser caracterizada por sus momentos de primer y segundo orden, los cuales resumen la información más relevante (Emery, 2013) tal como se presenta a continuación:

Valor esperado

$$m(u) = E\{Z(u)\} \quad (2.9)$$

$m(u)$  representa la media o promedio en un sitio  $u \in D$ , alrededor de la cual fluctúan los valores que toma la función aleatoria para cada realización.

Varianza

$$\sigma^2(u) = Var\{Z(u)\} = E\{[Z(u) - m(u)]^2\} = E[Z(u)^2] - m(u)^2 \quad (2.10)$$

La varianza es una medida de dispersión entre los valores suministrados por  $Z(u)$  y su media  $m(u)$ .



## Covarianza

$$C(u_1, u_2) = cov[Z(u_1), Z(u_2)] = E[Z(u_1)Z(u_2)] - m(u_1)m(u_2) \quad (2.11)$$

La covarianza mide la relación existente entre los valores de la función aleatoria en los puntos  $u_1$  y  $u_2$ .

## Variograma

$$\gamma(u_1, u_2) = \frac{1}{2} var[Z(u_1) - Z(u_2)] \quad (2.12)$$

Estrictamente semi-variograma, pero por simplificación se denomina variograma. Mide la diferencia entre valores ubicados en dos sitios  $u_1$  y  $u_2$ .

### 2.1.3. Estacionariedad

Como se mencionó anteriormente, la variable regionalizada es una realización de la función aleatoria. Sin embargo, es imposible determinar la función aleatoria  $Z(u)$  a partir de una sola realización  $z(u)$ , es decir, el valor de la variable aleatoria en la posición  $u$ . Esto impide poder inferir la función de distribución.

Surge entonces como respuesta a ello el concepto de estacionariedad, el cual permite postular que la distribución espacial es invariante por traslación en el espacio y que para un conjunto de datos las propiedades solo guardan relación con la distancia relativa entre ellos y no su posición absoluta (reemplazar la repetición sobre las realizaciones por una repetición en el espacio) Esta propiedad es conocida como **estacionariedad estricta**.

$$F_{(u_1, \dots, u_n)}(z_1, \dots, z_n) = F_{(u_1+h, \dots, u_n+h)}(z_1, \dots, z_n), \text{ con } h, n \in \mathbb{N} \text{ y } u_i \in D \quad (2.13)$$

Conforme a dicho postulado, es posible caracterizar la función aleatoria mediante las siguientes simplificaciones (Emery, 2013):

- La distribución univariable no depende del sitio que se está considerando (es decir, no depende de  $u$ ), y en específico la media y varianza no cambian en el espacio

$$F(z_1) = Prob\{Z(u_1) \leq z_1\} \quad (2.14)$$

$$m = E\{Z(u)\} \quad (2.15)$$

$$\sigma^2 = Var\{Z(u)\} \quad (2.16)$$

- La distribución bivariable solo depende de la separación entre los sitios en consideración. Igual situación ocurre para los momentos de segundo orden (covarianza, correlograma y variograma)

$$F_h(z_1, z_2) = Prob\{Z(u+h) < z_1, Z(u) < z_2\} \quad (2.17)$$

$$C(h) = corr\{Z(u+h), Z(u)\} \quad (2.18)$$

$$\rho(h) = \text{cov}\{Z(u+h), Z(u)\} \quad (2.19)$$

$$\gamma(h) = \frac{1}{2} \text{var}\{Z(u+h) - Z(u)\} \quad (2.20)$$

Adicionalmente, como la esperanza o valor esperado es constante el variograma queda definido de la siguiente forma.

$$\gamma(h) = \frac{1}{2} E\{[Z(u+h) - Z(u)]^2\} \quad (2.21)$$

Sin embargo, dado que es muy difícil verificar la Estacionariedad estricta, se suele plantear un relajamiento de la condición inicial y verificar que las ecuaciones 2.15 y 2.18 se cumplen; es decir, se puede afirmar que la función aleatoria es **estacionaria de segundo orden** si su esperanza matemática existe y es constante para cualquier par de puntos y su función de covarianza existe y solo depende del vector de separación de dicho par de puntos.

Sin embargo, en la práctica dicha condición se relaja aún más, y se habla entonces de **casi-estacionariedad de segundo orden**, al restringir el supuesto solamente a la vecindad de búsqueda.

Igualmente, se ha planteado un tipo adicional de Estacionariedad denominado **estacionariedad intrínseca**. Esto es, que para cada vector  $h$ , el incremento definido por  $Y_h(u) = Z(u+h) - Z(u)$  es una función aleatoria estacionaria de segundo-orden en  $u$ , que se caracteriza por las siguientes relaciones (Chilès y Delfiner, 2009):

$$E\{Z(u+h) - Z(u)\} = \langle a, h \rangle \quad (2.22)$$

$$\text{var}[Z(u+h) - Z(u)] = 2\gamma(h) \quad (2.23)$$

Donde  $\langle a, h \rangle$  corresponde a una deriva lineal (deriva del incremento). Se plantea entonces que si la deriva lineal es cero, es decir, que la media es constante, se tendrá:

$$E\{Z(u+h) - Z(u)\} = 0 \quad (2.24)$$

$$E[Z(u+h) - Z(u)]^2 = 2\gamma(h) \quad (2.25)$$

## 2.2. SIMULACIONES

Las simulaciones geoestadísticas permiten caracterizar el comportamiento de la variable regionalizada a partir de los datos conocidos. Este procedimiento lleva a modelar la distribución espacial y poder determinar la función aleatoria, que para el caso específico de las simulaciones representa una variable regionalizada por cada realización obtenida. Sin embargo, esta variable regionalizada es una variable “ficticia” definida como un posible escenario de la realidad.

Si al momento de efectuar las simulaciones se consideran los datos conocidos se habla de simulación condicional. Si por el contrario no son tenidos en cuenta, se hace relación a una simulación no condicional.

La ventaja que plantean las simulaciones respecto a las estimaciones, yace en la posibilidad de medir la incertidumbre originada en el desconocimiento de la variable de interés (medir la variabilidad de los diferentes escenarios generados). Esto permite efectuar una serie de análisis imposibles de desarrollar con un estimador (como por ejemplo kriging); dado su propiedad de suavizamiento. Las simulaciones por el contrario permiten determinar un escenario más negativo o más positivo y un escenario promedio (similar al de la estimación) con lo cual es posible desarrollar análisis de riesgo.

De acuerdo con Chilès y Delfiner (2009) las simulaciones pueden clasificarse en tres tipos:

- Variables continuas: son variables que miden propiedades físicas. Son continuas en cuanto a la continuidad del histograma y espacialmente. Puede aplicarse matemáticas directamente sobre ellas. Las técnicas empleadas para simular este tipo de variables pueden subdividirse en dos grandes grupos:
  - Modelos multigaussianos
  - Modelos no paramétricos
- Variables categóricas: pueden corresponder a variables discretas que representan una clasificación o también el tratamiento de una variable continua mediante truncación con cierto valor o umbral. No se pueden efectuar matemáticas directamente sobre ellas.

Dentro de las técnicas usadas para simular variables categóricas, se pueden destacar (Emery, 2008):

- Simulación de indicadores
  - Modelo gaussiano truncado
  - Modelo plurigaussiano
  - Simulación multi-punto
- Objetos: representan entidades cuya morfología es capturada como un todo y simulado en una sola vez. Algunas técnicas o algoritmos propuestos se basan en:
  - Modelos booleanos
  - Procesos de puntos

Dado que en el presente estudio se empleará un modelo paramétrico asociado a los modelos multi-gaussianos y modelos de simulación multi-punto; se explica a continuación los conceptos básicos asociados a este tipo de simulaciones.

### Modelo multi-gaussiano

Estos modelos reproducen la distribución inferida y la estructura espacial de 2 puntos, impuesta generalmente por un variograma. Requieren que la distribución espacial de la función aleatoria sea multi-gaussiana, es decir, que cualquier combinación lineal de sus componentes sea Gaussiana, o equivalentemente que su densidad de probabilidad sea multi-gaussiana. Dado lo anterior, se tiene una fuerte hipótesis, lo que exige efectuar transformación de la variable mediante una función denominada anamorfosis, que permite llevar una variable con una distribución particular a una

distribución marginal gaussiana estándar, puesto que la mayoría de variables no cumplen dicha hipótesis. Posteriormente, mediante un método aplicado sobre la variable transformada para efectuar la simulación y los resultados son transformados de vuelta mediante la anamorfosis inversa.

La densidad de probabilidad en el conjunto de sitios  $\{u_1, \dots, u_n\}$  tiene la forma:

$$f(y) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(C)}} \exp \left\{ -\frac{1}{2} y^T C^{-1} y \right\} \quad (2.26)$$

Donde:

$y$ : Vector de  $n$  componentes

$C$ : Matriz de varianza-covarianza de las variables ubicadas en los sitios  $\{u_1, \dots, u_n\}$

Dentro de las técnicas que emplean el modelo multi-gaussiano se encuentran el método de descomposición matricial, bandas rotantes, secuencial gaussiano y espectral continuo.

### Simulación Multi-punto

Este tipo de simulación fue inicialmente propuesto por Guardiano y Srivastava en 1993. En esta la función aleatoria es generada a partir de una interpretación previa de la realidad como por ejemplo un modelo conceptual, una interpretación geológica, una imagen de satélite, una imagen geofísica, etc.; que se denominan una imagen de entrenamiento (TI por sus siglas en inglés: Training Image). A partir de esta imagen, Guardiano y Srivastava, proponen generar la función aleatoria como una inferencia probabilística al recorrer la TI a través de un patrón específico y no mediante un modelo teórico matemático como lo hace el modelo multi-gaussiano. Sin embargo, para cada punto en el espacio de cada realización de la variable aleatoria, era necesario recorrer la TI nuevamente.

En 2002, Strebelle, propone utilizar una técnica (SNESIM - single normal equation simulation) que emplea árboles de búsqueda con la finalidad de almacenar la información inferida de la TI disminuyendo el tiempo de procesamiento; pero con dificultades en el almacenamiento cuando la imagen de entrenamiento es muy grande. Surgen por ello nuevas técnicas o variantes que tratan de disminuir los tiempos de cómputo y procesamiento, dentro de las cuales podrían mencionarse (Pérez, 2013), entre muchas otras:

- FILTERSIM (simulation using filter scores) propuesto por Zhang et al (2006) que emplea filtros sobre la imagen de entrenamiento para generar la función aleatoria.
- DS (Direct Sampling) propuesto por Mariethoz et al (2010) el cual muestrea directamente la TI para determinar la ocurrencia de un evento de datos condicionantes dado, evitando la necesidad de almacenamiento de las estadísticas multi-punto inferidas. Para cada nodo a ser simulado, este algoritmo encuentra el número de nodos vecinos más cercanos definidos por el usuario (orden de eventos condicionantes). Las distancias desde el nodo a los nodos más cercanos son calculadas definiendo los eventos de datos condicionantes. Para cada nodo visitado en la TI, el evento más cercano que iguala la fracción de eventos condicionantes definidos por el usuario es buscado y la distancia al nodo analizado es calculada. Si la distancia entre los eventos condicionantes y los eventos muestreados son menores que un umbral dado, el valor del nodo visitado en la TI es asignado al nodo simulado.

Dado que este estudio analiza una variable continua y una variable categórica, sobre las cuales se aplicarán las métricas seleccionadas; se hará una breve descripción de los algoritmos empleados en las simulaciones. En primer lugar se presentarán los algoritmos SISIM y SNESIM (ambos algoritmos no paramétricos empleados para simular una variable categórica. Posteriormente se describirán los algoritmos FILTERSIM y SGSIM (sequential gaussian simulation) los cuales corresponden a algoritmos no paramétrico y paramétrico respectivamente.

### 2.2.1. SISIM (Sequential Indicator Simulation)

La simulación secuencial de indicadores es un tipo de simulación no-paramétrica, que puede ser usada para simular variables categóricas y continuas. Sin embargo, dado que en el presente estudio se hace relación a su uso en una variable categórica, su breve descripción estará enfocada en este aspecto. Para mayor detalle del algoritmo remitirse a Journel y Gómez (1993) y a Remy et al (2009).

Este algoritmo subyace en la propiedad de kriging indicador para inferir la función de distribución acumulada condicional; que para el caso de la variable categórica con  $K$  categorías, se genera una variable indicador en donde 1 corresponde a si está presente la categoría  $k$  y 0 las demás categorías. Para cada una de las categorías se construye dicha variable indicador.

Los pasos del algoritmo se presentan a continuación (Remy, et Al, 2009):

- Definir la ruta o trayectoria para visitar todos los nodos.
- Para cada nodo  $u$  hacer lo siguiente:
  - Obtener los datos categóricos condicionantes  $z(u_\alpha)$ ,  $\alpha = 1, \dots, N$
  - Convertir cada dato  $z(u_\alpha)$  a un vector de valores de datos de indicadores  
 $i(u_\alpha) = [i(u_\alpha, z_1), \dots, i(u_\alpha, z_K)]$
  - Estimar la variable aleatoria indicador  $I(u, k)$  para cada una de las  $K$  categorías mediante un sistema de kriging simple.
  - Se define una función de densidad de probabilidad condicional discreta de la variable categórica  $Z(u)$  mediante los valores estimados  $I^*(u, k) = \text{Prob}(Z(u)=k)$ ; luego de haber efectuado correcciones de relación de orden (es decir que la distribución se encuentre entre 0 y 1) y que  $\sum p_i = 1$ .
  - Llevar el valor simulado (categoría simulada) desde la función de densidad de probabilidad condicional discreta al nodo con posición  $u$ .
- Repetir los pasos para generar las demás realizaciones.

### 2.2.2. SNESIM (Single Normal Equation Simulation)

Desarrollado por Strebelle (2002). Se basa en una estructura dinámica de datos (conocida como árbol de búsqueda) para almacenar eventos de la imagen de entrenamiento (o datos condicionantes) antes de la simulación de la imagen. La construcción requiere escanear la imagen de entrenamiento (TI) una sola vez, lo cual minimiza la demanda de cómputo, permitiendo obtener todas las funciones de densidad de probabilidad condicionantes (proporciones condicionales disponibles desde la TI) para la plantilla de datos (vecindad de búsqueda de datos).

SNESIM es un algoritmo secuencial y básicamente contiene dos partes (Remy et al, 2009): la construcción del árbol de búsqueda y la simulación por sí misma donde las proporciones son leídas y usadas para generar los valores simulados.

El algoritmo en términos generales se puede resumir en los siguientes pasos (Remy et al, 2009):

- Definir una plantilla de búsqueda
- Construir el árbol de búsqueda para la plantilla definida
- Relocalizar los datos condicionantes (datos conocidos) al nodo más cercano de la grilla de simulación (dejarlos congelados durante la simulación)
- Definir una trayectoria aleatoria de visita de todos los puntos a ser simulados
- Para cada punto en el espacio  $u$  a ser simulado, hacer lo siguiente:
  - Encontrar los datos condicionantes  $dcon(u)$  definidos en la plantilla
  - Recuperar la función de distribución de probabilidad condicional desde el árbol de búsqueda  $Prob[Z(u) = k/dcon(u)]$
  - Simular un valor  $z^{(s)}(u)$  desde la probabilidad condicional y adicionarlo a la grilla.

SNESIM presenta igualmente una variante que permite simular en múltiples grillas, de manera que se puedan capturar estructuras a gran escala usando plantillas más grandes pero con un pequeño número de nodos e ir reduciendo dicha plantilla a una grilla más pequeña.

Para mayor detalle del algoritmo remitirse a Strebelle (2002), y Remy et al (2009).

### **2.2.3. FILTERSIM (Filter-Based Simulation)**

Desarrollado por Zhang et al (2006), presenta menos requerimientos de memoria que los que presenta SNESIM y puede simular tanto variables categóricas como variables continuas.

FILTERSIM utiliza filtros lineales para clasificar patrones de entrenamiento en un espacio de puntuación de reducidas dimensiones (Remy. et al, 2009). Patrones similares de entrenamiento son almacenados en una clase caracterizada por un patrón promedio llamado prototipo, el cual es determinado durante la simulación de manera que sea el más cercano a un evento condicionante. Posteriormente, un patrón de entrenamiento desde la clase de prototipo es simulado y transferido a la grilla de simulación.

Pero para llevar a cabo este proceso se aplican unos filtros que corresponden a un conjunto de pesos con una configuración definida. Adicionalmente una plantilla de búsqueda se define para determinar los patrones de entrenamiento desde la imagen de entrenamiento y cada uno de estos patrones es representado por un puntaje para cada filtro. Así que es necesario aplicar varios filtros ( $F$  filtros) para poder obtener la información de los patrones y conseguir un vector de  $F$  dimensiones.

Los pasos que sigue el algoritmo FILTERSIM se resumen a continuación (Remy et al, 2009).

- Crear los mapas de puntuación aplicando filtros con un patrón de búsqueda en la imagen de entrenamiento.
- Separar los patrones de entrenamiento dentro de clases y prototipos en el espacio de puntuación.
- Reubicar los datos conocidos a la malla de simulación

- Definir una trayectoria aleatoria de la malla de simulación
- Para cada nodo  $u$  en la trayectoria aleatoria efectuar:
  - Extraer los datos de eventos condicionantes  $dcon$  centrados en  $u$
  - Encontrar el prototipo  $prot_p$  más cercano a  $dcon$
  - Si el prototipo tiene prototipos hijos entonces
    - Encontrar el prototipo hijo  $prot_h$  más cercano a  $dcon$
    - Aleatoriamente crear un patrón  $pat$  desde  $prot_h$
  - Si no tiene prototipos hijos, entonces
    - Aleatoriamente crear un patrón  $pat$  desde  $prot_p$
  - Pegar el patrón dentro de la grilla de simulación y congelar los nodos dentro del nodo central.

Para mayor detalle del algoritmo remitirse a Zhang et al (2006) y Remy et al (2009).

#### 2.2.4. SGSIM (Sequential Gaussian Simulation)

Este algoritmo emplea el formalismo para simular una función aleatoria gaussiana. Para ello se requiere una variable multi-gaussiana de media cero (0) y varianza uno (1.0) y un modelo dado de variograma (de la variable gaussiana). Si la variable no es una variable gaussiana de media cero y varianza uno, esta debe ser transformada de forma previa mediante una función conocida como anamorfosis y se debe efectuar una validación de la hipótesis de multi-gaussianidad (generalmente se verifica la hipótesis de bi-gaussianidad y si dicha hipótesis es aceptada se acepta la multi-gaussianidad).

El algoritmo general de la simulación secuencial gaussiana se presenta a continuación.

- Transformar la variable a una variable gaussiana estándar (media cero y varianza unitaria), en caso de que no corresponde a una variable gaussiana.
- Definir una trayectoria aleatoria para visitar cada nodo de la grilla
- Construir la variografía de la variable gaussiana
- Para cada nodo de la grilla  $u$  a lo largo de la trayectoria aleatoria, efectuar:
  - Determinar los datos condicionantes consistentes en los datos originales disponibles transformados y los datos previamente simulados
  - Estimar la función de distribución acumulada condicional local como una distribución gaussiana de media igual al estimador del kriging y varianza igual a la varianza de estimación del kriging.
  - Obtener un valor desde la función de distribución acumulada condicional y adicionarlo dicho valor simulado a la grilla de simulación.
- Repetir el proceso para las demás realizaciones.

### 2.3. VALIDACIÓN DE SIMULACIONES

En este numeral se efectúa una revisión bibliográfica de las diferentes prácticas empleadas en la validación de modelos numéricos de uso común en geociencias, desde el punto de vista geoestadístico, en especial modelos generados a partir de simulaciones geoestadísticas con énfasis en simulaciones multi-punto. Como es conocido, la modelación en geociencias exige definir en el proceso diversos parámetros, efectuar asunciones diversas con el fin de poder lograr un modelo.

Esto puede conducir a diversas fuentes de error en el proceso, motivo por el cual se hace necesario efectuar una validación del modelo obtenido con el objetivo de identificar esos posibles errores (Pyrz y Deutsch, 2014).

Estas buenas prácticas implican que las validaciones efectuadas deben cubrir diferentes ámbitos o niveles de comprobación de los resultados obtenidos del modelo numérico, respecto a los datos, información o comportamiento de la variable que se pretende modelar. Para ello se propone efectuar una revisión desde los siguientes niveles:

- Nivel 1: Condicionamiento de los Datos. Consiste en verificar si el modelo reproduce los datos empleados (muestras o datos disponibles) como input de la modelación.
- Nivel 2: Validación estadística. Consistente en la verificación de la reproducción de estadísticas básicas relacionadas a la distribución de los datos originales (por ejemplo verificación o comparativo de histograma del modelo generado vs. histograma de datos originales).
- Nivel 3: Validación geoestadística. Consistente en la comprobación de que las simulaciones reproduzcan herramientas que miden la correlación espacial entre dos puntos (como variogramas, variogramas de indicadores), o herramientas que permitan analizar estructuras no lineales (por ejemplo evidenciar sensibilidad a parámetros tales como conectividad) tales como estadísticas multipunto.
- Nivel 4: Validación de la predicción. Por último en este nivel, se efectúa una revisión de metodologías para testear la distribución de incertidumbre mediante una validación cruzada, a fin de determinar la consistencia de la distribución de probabilidad generada con las realizaciones respecto al valor conocido.

### 2.3.1. Nivel 1. Condicionamiento de los datos

Leuangthong et al (2004) plantea que la reproducción de los datos originales por parte de un modelo, debe considerarse como uno de los criterios mínimos de aceptación o validación de éste. En tal sentido define (según lo planteado por Goovaerts, 1997; Journel y Huijbregts, 1978 en anteriores publicaciones), que el valor estimado por Kriging en un punto es exactamente el valor del dato original y la varianza del error es cero. En la ecuación 2.25 se presenta la expresión que condensa esta afirmación.

$$z_k^*(u) = z(u_\alpha) \quad \forall u = u_\alpha, \quad \alpha = 1, 2, \dots, n \quad (2.27)$$

Donde:

$z_k^*(u)$ : Estimador de kriging de la variable aleatoria  $Z$  en la ubicación  $u$  del dominio.

$z(u_\alpha)$ : Es el valor del dato en la ubicación  $u_\alpha$ ,  $\alpha = 1, 2, \dots, n$

Esto aplica igualmente para las simulaciones, dado que estas se basan en el estimador del kriging y su varianza para definir la función de distribución acumulada condicional (Leuangthong et al, 2004). De igual manera, para modelos basados en celdas, Pyrcz y Deutsch (2014) afirman que no deberían encontrarse problemas para reproducir los datos, dado que el kriging tiene como una de sus propiedades la precisión. Lo cual también debería ser aplicable a los métodos de simulación

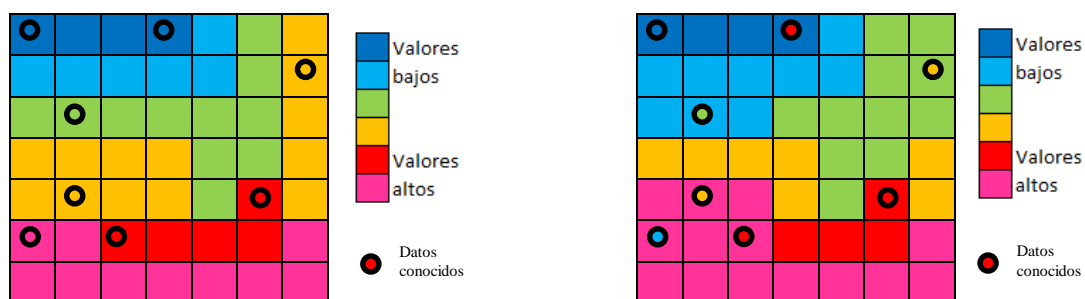


secuencial como el Gaussiano, indicador y métodos multipunto, los cuales asignan el valor del dato original a una celda en particular por la cual inicia la simulación y no se vuelve a calcular el valor en esa celda.

A continuación se describen dos métodos que permiten efectuar el análisis asociado al condicionamiento de los datos, los cuales son mencionados por Leuangthong et al (2004), y Pyrcz y Deutsch (2014).

### 2.3.1.1. Inspección visual.

Como lo que se pretende es determinar la igualdad entre el valor estimado vs. el valor original, una forma de determinar esto es mediante una comprobación visual del modelo (mediante la comparación de algunas de las realizaciones) con los datos originales superpuestos. Esto permite determinar si efectivamente en los sitios donde se tienen datos conocidos, los resultados de la simulación concuerdan con dicho valor conocido. En la Figura 2-1 se presenta un ejemplo de la inspección visual.



**Figura 2-1: Inspección visual.**

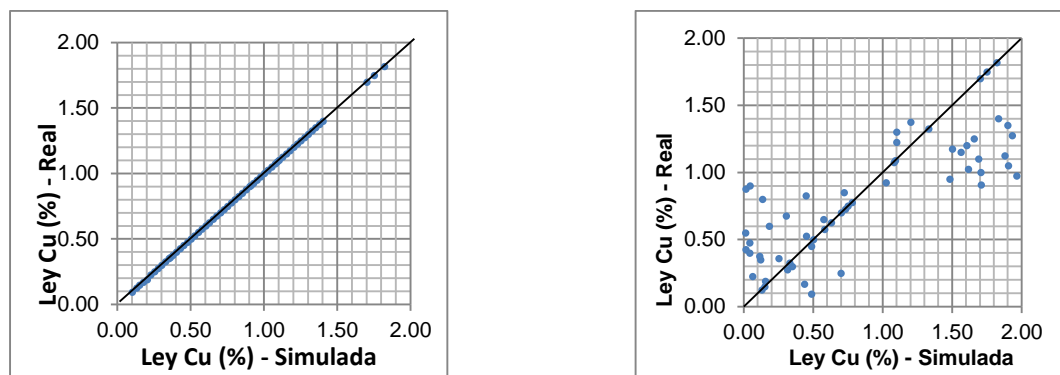
Izquierda: los resultados de la simulación reproducen los datos originales.  
Derecha: los resultados de la simulación no reproducen los datos originales.

### 2.3.1.2. Gráfico de dispersión.

Conocidas las diferentes realizaciones obtenidas mediante la simulación, así como los datos condicionantes, los cuales fueron considerados como input del proceso de modelación; es posible construir un gráfico de dispersión entre los datos simulados (para determinada realización) vs. los datos originales. Los puntos deberán entonces encontrarse sobre la diagonal o bisectriz del gráfico. En la Figura 2-2 se presenta un ejemplo de la verificación que se pretende efectuar. La correlación entre los grupos de datos debería ser muy cercana a 1.0. Leuangthong et al (2004) define que las pequeñas desviaciones respecto a la diagonal podrían deberse a la transformación inversa de los datos gaussianos a datos originales. Grandes desviaciones deberían analizarse para determinar la fuente del error, teniendo en cuenta que la proporción de la desviación podría deberse al efecto del efecto pepa. También puede presentarse problemas al migrar datos densos a una grilla más amplia.

### 2.3.2. Nivel 2. Validación estadística

Otro aspecto de gran relevancia al momento de efectuar una validación de un modelo numérico, es que este reproduzca la distribución de los datos o información inicial. Esto cobra relevancia, dado que al momento de efectuar el análisis exploratorio de datos debe garantizarse que los datos a emplear en la modelación representen adecuadamente la población y por ende sus estadísticas arrojen información fidedigna del dominio de interés.



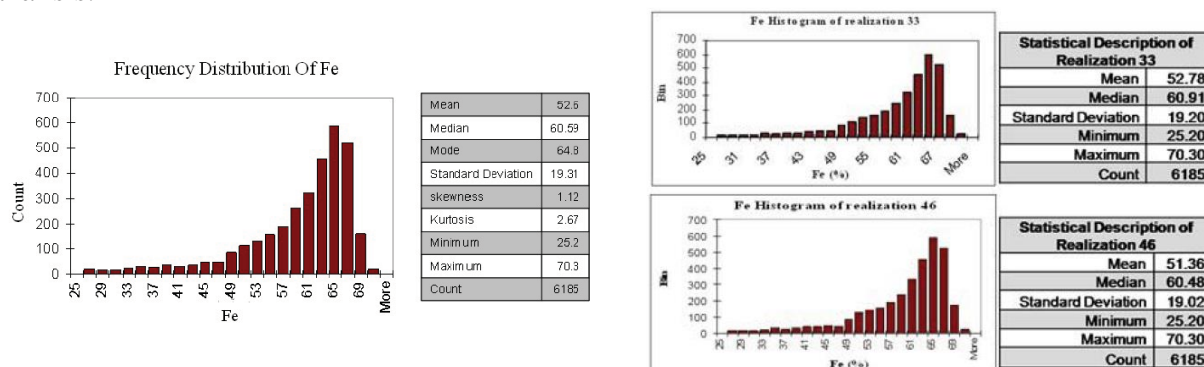
**Figura 2-2.** Gráfico de dispersión.

Izquierda: los resultados de la simulación reproducen los datos originales.  
Derecha: los resultados de la simulación presentan desviaciones respecto a datos originales.

Dada la necesidad de que las estadísticas sean representativas, es necesario tomar en cuenta que la comparación de la distribución de los datos simulados debe efectuarse contra el histograma de los datos originales desagrupados.

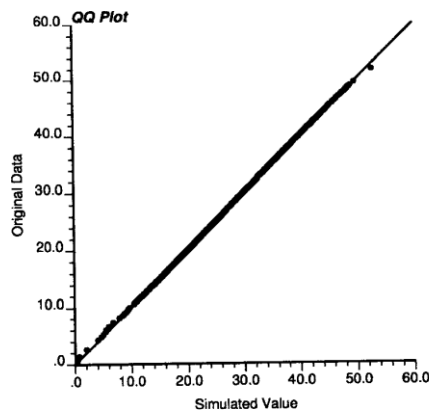
Leuangthong et al (2004) propone efectuar tres comparaciones básicas, retomadas por Pycrz y Deutsch (2014) y que están relacionadas a (1) la forma del histograma, (2) el rango de valores simulados y (3) las estadísticas básicas como media, mediana y varianza (Figura 2-3). Adicionalmente, proponen efectuar comparaciones a partir de la construcción de gráficos cuantil-cuantil (Q-Q plot) dado que los rangos pueden ocultar algunos elementos en el histograma y además pueden incluirse diversas realizaciones en el mismo gráfico para efectuar las diferentes comparaciones (Figura 2-4). Para el caso de las variables binarias (como el caso de una variable indicador) podría verificarse la correspondencia entre las proporciones entre los datos originales y los datos simulados; así como las varianzas (Rezaee et al, 2014).

Este tipo de análisis también permite efectuar comparaciones entre diferentes modelos o algoritmos de simulación para poder comparar los resultados obtenidos y determinar si existen diferencias entre los resultados respecto a los datos originales y seleccionar aquel que represente de mejor manera la variable a simular. Algunos autores como Goovaerts (2001), Olea (2012), Sargents (2005), Asghari (2009), Zhao et al (2014) efectúan algunas validaciones o comparaciones bajo este análisis.



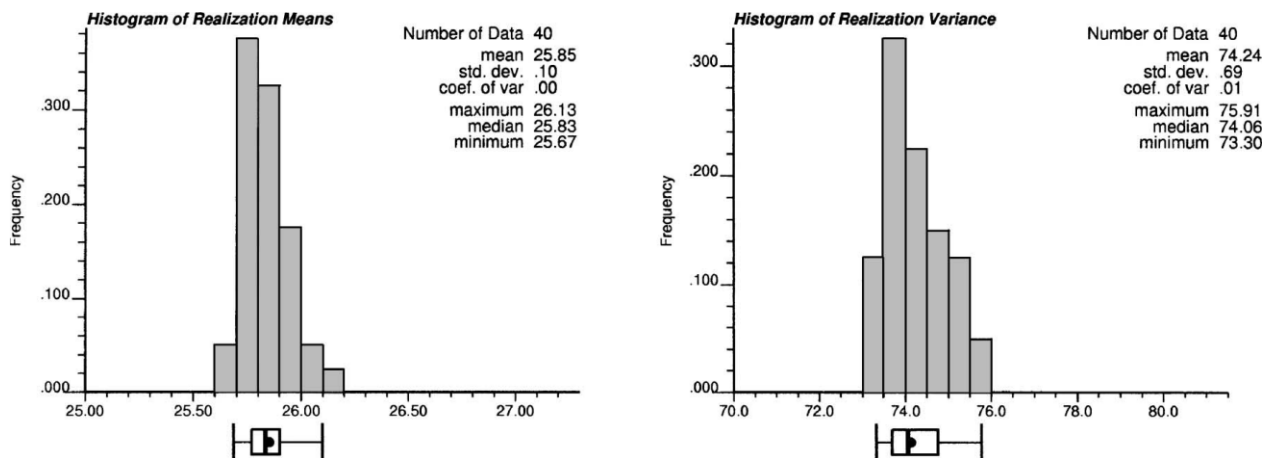
**Figura 2-3:** Comparación de histogramas.

Izquierda: histograma de datos originales de Fe. Derecha: histogramas de dos realizaciones de simulaciones de Fe.  
Fuente: Asghari (2009)



**Figura 2-4:** Q-Q Plot para comparar distribuciones de datos simulados y datos originales.  
Fuente: Leuangthong. et al (2004)

Adicional a verificar la reproducción de las estadísticas por cada una de las realizaciones, Leuangthong et al (2004) propone igualmente la validación de lo que denomina el ensamble de las realizaciones; es decir, verificar que la media de las medias de las realizaciones y la media de las varianzas de las realizaciones reproduzcan respectivamente la media y la varianza de los datos originales. En la Figura 2-5 se presenta un ejemplo de este tipo de validación.



**Figura 2-5.** Análisis de distribución de medias y varianzas de 40 realizaciones  
(punto en el box-plot corresponde a la media de los datos originales).  
Fuente: Leuangthong et al (2004)

Es importante mencionar que las comparaciones deben desarrollarse con respecto a los datos desagrupados; dado que el agrupamiento o muestreo preferencial de los datos pueden dificultar la inferencia de los parámetros de la población.

Sin embargo, es importante tener en cuenta que algunas de las realizaciones en las simulaciones pueden presentar desviaciones respecto a los datos originales o sus estadísticas, lo cual es comúnmente conocido como fluctuaciones ergódicas. Pyrcz y Deutsch (2014) definen que estas fluctuaciones son debidas a la incertidumbre y la extensión espacial limitada del dominio considerado. Cuando esta extensión espacial es lo suficientemente grande respecto al rango de continuidad espacial, las estadísticas de las realizaciones deberían converger a las estadísticas de los datos originales. Debido a esto, los análisis de validación deberán tener en cuenta estos aspectos, considerar las desviaciones esperadas, para determinar la concordancia o no entre los datos de

ingreso y los datos obtenidos. En el mismo sentido, debe tenerse en cuenta, que métodos basados en objetos, o de optimización no garantizan la reproducción precisa de los datos originales.

Wackernagel (2013) resalta la importancia de considerar los problemas asociados a la definición de una proporción que se encuentra sobre un umbral fijo, por ejemplo lo que se denomina funciones de transferencia; que para el caso minero podría corresponder a un tonelaje sobre una ley de corte, o la cantidad de metal sobre una ley de corte. En un caso ambiental podría estar relacionado con la cantidad de suelo contaminado sobre cierto umbral o límite permisible. Esto puede entonces ser analizado al comparar mediante curvas de selectividad los resultados de las realizaciones contra la información disponible como una imagen de entrenamiento y determinar si reproduce dicho comportamiento.

Por su parte, Emery (2008) sugiere que las comparaciones de los datos experimentales (datos originales) y los del modelo pueden ser engañosas. Esto es debido principalmente a las fluctuaciones ergódicas, que son el resultado de simular el campo aleatorio sobre un dominio no infinito. Para ello propuso validar los algoritmos de simulación analizando si las fluctuaciones son consistentes con el campo aleatorio del modelo y el tamaño del dominio donde se efectúa la simulación a través de test estadísticos.

Para el caso particular del nivel 2 de validación (según la subdivisión definida inicialmente en esta revisión bibliográfica), Emery (2008) propone hacer un test para verificar si la reproducción del valor medio y la varianza de dispersión de las realizaciones se encuentran dentro del intervalo de confianza definido.

Para el caso de la media de las realizaciones se analiza según la siguiente expresión.

$$\frac{\bar{Y}(D) - m}{S_{Y(D)}} \sqrt{N-1} \sim T_{N-1} \quad (2.28)$$

Donde:

$\bar{Y}(D)$ :	Media muestral de las N realizaciones
$m$ :	Esperanza de la variable aleatoria (~media de los datos originales).
$S_{Y(D)}^2$ :	Varianza muestral de las N realizaciones
$T_{N-1}$ :	Distribución t student con N-1 grados de libertad

En donde para un grado de confianza definido, por ejemplo del 95%, puede encontrarse un intervalo donde estaría  $\bar{Y}(D)$ . De no encontrarse dicho valor en ese intervalo los resultados del algoritmo deberían cuestionarse. Generalmente, el nivel de confianza debe definirse alto (el mejor error posible) dado los posibles problemas que pueden presentarse cuando los test no son independientes.

En el caso de la varianza, también se propone un test, que se evalúa según la siguiente expresión.

$$\frac{\bar{S}^2(^{\circ}D) - \sigma^2(^{\circ}D)}{S_{S^2(^{\circ}D)}} \sqrt{N-1} \sim T_{N-1} \quad (2.29)$$

Donde:

$S^2(^{\circ} D):$	Varianza muestral de campo aleatorio Y sobre un dominio D
$\bar{S}^2(^{\circ} D):$	Media de N realizaciones de la varianza muestral del campo aleatorio
$\sigma^2(^{\circ} D):$	Varianza de N realizaciones de la varianza muestral del campo aleatorio
$T_{N-1}:$	Distribución t student con N-1 grados de libertad

Respecto a los test estadísticos, es importante considerar que estos se efectúan bajo ciertos supuestos. El primero de ellos es que las muestras corresponden a distribuciones gaussianas, el segundo es la independencia de los datos y finalmente que las características de las poblaciones son constantes (en este caso relacionado a la Estacionariedad). Estos aspectos deberán ser tenidos en cuenta al momento de decidir si los test estadísticos son válidos para ser empleados como métricas de validación de ciertos fenómenos que se quieren modelar, que no cumplen dichos supuestos; como por ejemplo es el caso de simulaciones condicionales donde las realizaciones tendrán cierta correlación asociada a la existencia de los datos condicionantes, lo cual estaría violando el supuesto de independencia de los test estadísticos. De igual forma el comportamiento gaussiano de las variables asociadas a las geociencias, difícilmente podrá ser observado en este tipo de variables y por ende representaría otro aspecto en contra para la aplicación de estos test.

### **2.3.3. Nivel 3. Validación geoestadística**

Respecto a este nivel de validación, diversos autores han efectuado diferentes tipos de validación. Por ejemplo, Pyrcz y Deutsch (2014) hablan de una categoría diferente denominada chequeo de más alto orden, en la cual se incluyen criterios aplicables a métodos de estadísticas multipunto. Rezaee et al (2014) sugieren una validación basada en estadísticas de orden ascendente. En un primer nivel se validan estadísticas de un solo punto, luego estadísticas de dos puntos y finalmente la reproducción de estadísticas multi-punto. Las estadísticas de un solo punto corresponderían a las validaciones efectuadas en el nivel 2 de validación estadística del presente documento, donde se evalúa comparativamente la distribución obtenida con las simulaciones vs. la distribución de los datos originales.

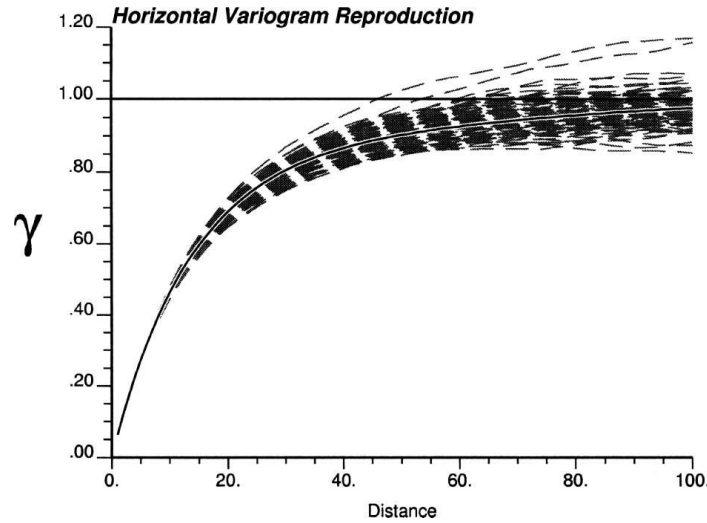
Tomando en cuenta lo propuesto por Rezaee et al (2014) se efectuará una subdivisión de este nivel que permita realizar una validación de las simulaciones en estadísticas de dos puntos (asociada básicamente al análisis de reproducción de variogramas) y posteriormente el análisis de reproducción de estadísticas multi-punto. Finalmente se presentarán otros test o chequeos que permiten igualmente validar las simulaciones geoestadísticas y que no se encuentran dentro de las dos categorías previamente mencionadas.

#### **2.3.3.1. Validación de estadísticas de dos puntos.**

Una de las formas de validar un modelo de simulación geoestadística es mediante el análisis del variograma. Esta herramienta permite analizar el comportamiento espacial de una variable (en un dominio dado). Es por esto que diversos algoritmos emplean esta herramienta para efectuar la construcción de las realizaciones y hace parte de los parámetros requeridos para poder modelar; lo que conlleva a verificar si las realizaciones o el conjunto de estas reproducen adecuadamente el variograma. Al respecto diversos autores han efectuado este tipo de validaciones, tales como: Rezaee et al (2014), Asghari et al (2009), Ortiz y Deutsch (2004), Chiles y Lantuéjoul (2005), Ersoy et al (2008), Caers y Sanjay (2002), Klise et al (2009). En la figura 2.6 se puede observar un ejemplo de los análisis efectuados para validar la reproducción del variograma, en el cual se verifica si en conjunto las realizaciones se encuentran en torno al variograma modelado. Adicionalmente, es importante verificar en las simulaciones multi-punto la reproducción del

variograma de la TI por parte de las realizaciones, puesto que permite analizar igualmente el comportamiento espacial entre la información de entrada y los resultados (consistencia interna).

Contrastando con este tipo de validación, Emery (2008) y Ortiz y Leuangthong (2012) proponen test estadísticos para determinar si las realizaciones se encuentran dentro de intervalos en los cuales pueden considerarse válida la media de los variogramas respecto a su desviación al variograma modelado para diversos pasos.



**Figura 2-6:** Análisis de reproducción del variograma por diferentes realizaciones.  
Líneas puntuadas: realizaciones. Línea continua gris: variograma modelado.  
Fuente: Leuangthong et al (2004)

Ortiz y Leuangthong (2012) y Emery (2008) proponen calcular el Estadístico  $T^2$  de Hotelling y compararlo con el valor de una distribución de Fisher con  $p$  y  $n-p$  grados de libertad. Donde  $p$  hace referencia a un número seleccionado de pasos a los cuales se calcula el variograma experimental. El estadístico  $T^2$  se calcula según la expresión de la ecuación 2.29.

$$T^2 = n(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0) \quad (2.30)$$

Donde:

- $n$ : Número de valores disponibles del variograma experimental en cada paso (número de realizaciones efectuadas)
- $\bar{X}$ : Media muestral del vector de valores del variograma para las realizaciones
- $\mu_0$ : Vector medio basado en el variograma modelado para las mismas  $p$  distancias de paso.
- $S$ : Matriz de covarianza experimental

Finalmente, se plantea como hipótesis nula que la media del vector de variograma para los  $p$  pasos seleccionados es igual a la media del vector basado en el variograma modelado para las mismas  $p$  distancias seleccionadas como paso. Esta hipótesis se rechaza si se cumple la siguiente desigualdad. Ver 2.30.

$$T^2 > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \quad (2.31)$$

Donde:

$F_{p,n-p}(\alpha)$ : Distribución Fisher con  $p$  y  $n-p$  grados de libertad.

Emery (2008) propone adicionalmente un test para verificar si la varianza del variograma experimental es significativamente diferente de la varianza del variograma modelado para diversos pasos, de acuerdo a la siguiente expresión.

$$\sum_{n=1}^N \left\{ \left( \Gamma_D^{(n)} - \bar{\Gamma}_D \right)^T (\Sigma_D^2)^{-1} \left( \Gamma_D^{(n)} - \bar{\Gamma}_D \right)^T \right\} \sim \chi_{K(N-1)}^2 \quad (2.32)$$

Donde:

$\bar{\Gamma}_D$ : Vector muestral medio de la matriz varianza-covarianza sobre N realizaciones  
 $\Gamma_D^{(n)}$ : Vector regional del semi-variograma calculado sobre la n realización  
 $\chi_{K(N-1)}^2$ : Variable aleatoria Chi cuadrado con N-1 grados de libertad  
 $\Sigma_D^2$ : Matriz varianza-covarianza teórica del vector  $\Gamma_D$

### 2.3.3.2. Validación de estadísticas multipunto.

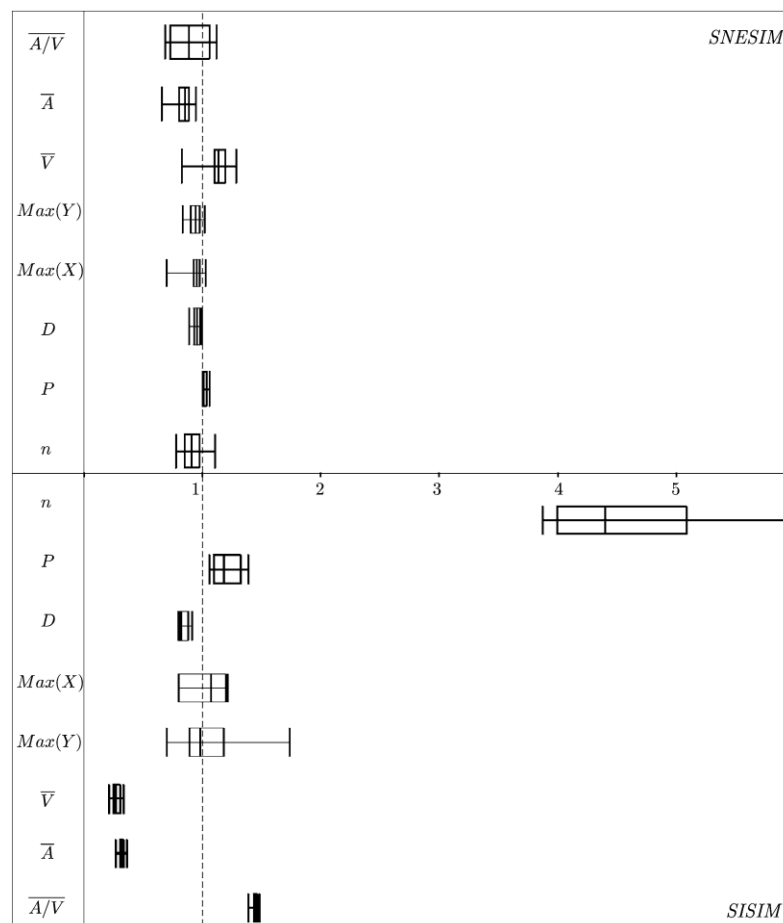
Dado que las estadísticas de dos puntos no permiten capturar o caracterizar funciones curvilíneas (Pyrz y Deutsch, 2014), las cuales están presentes comúnmente en fenómenos de la naturaleza y más especialmente en las geociencias, se han introducido métodos de simulación que emplean estadísticas multipunto, que permiten considerar este tipo de comportamientos no lineales y modelarlos a través de estadísticas de orden superior tomadas a partir de una imagen de entrenamiento (TI por sus siglas del inglés - *Training Image*) (Boisvert et al, 2010).

De acuerdo entonces a este tipo de estadísticas de orden superior empleadas, es necesario efectuar validaciones de acuerdo a estas, que en general no se aplican según Boisvert et al (2010) debido al desconocimiento general. Este tipo de validaciones comprenden diversas técnicas propuestas por diferentes autores. Por ejemplo, De Iaco y Maggio (2011) proponen el análisis general de formas o cuerpos y validación de estadísticas de orden superior conocidas como cumulantes que permiten analizar sistemas complejos, no Gaussianos y con comportamientos o patrones no lineales de conectividad espacial (Dimitrakopoulos et al, 2010).

Respecto a la validación a través del análisis general de formas De Iaco y Maggio (2011) analizan las realizaciones basados en cantidades medibles de facies, tales como:

- Número de cuerpos,  $n$
- Fracción del volumen global,  $p$  (medida de abundancia)
- Dirección de máxima variabilidad,  $D$
- Máxima extensión en la dirección x,  $Máx(X)$  – medida de conectividad
- Máxima extensión en la dirección y,  $Máx(Y)$  – medida de conectividad
- Volumen medio,  $\bar{V}$
- Área media,  $\bar{A}$
- Razón  $\bar{A}/\bar{V}$

Cada uno de estos parámetros son analizados a través del programa *faprop* propuesto por Soleng et al (2006). En la Figura 2-7 se presentan diagramas de caja comparativos de cada uno de los parámetros descritos para diferentes simulaciones, los cuales han sido normalizados por los valores de los parámetros obtenidos para la TI de manera que mientras más cercano sean los valores a la unidad, más será su similitud en cuanto a formas con la imagen de entrenamiento.



**Figura 2-7:** Análisis comparativo de cuerpos obtenidos de simulaciones.  
Fuente: De Iaco y Maggio (2011)

Respecto a las validaciones de estadísticas de orden superior, De Iaco y Maggio (2011) proponen efectuar análisis del error medio cuadrático entre los cumulantes espaciales de orden dos, de orden tres y de orden cuatro de las simulaciones vs. la imagen de entrenamiento. Mientras más cercano sean los errores medios cuadráticos a cero, más similares serán las características espaciales entre las realizaciones y la imagen de entrenamiento. Sin embargo, para tener una mayor claridad sobre esta propuesta es importante comprender a que hace relación este término.

Los cumulantes espaciales de orden superior corresponden a una propuesta presentada por Dimitrakopoulos et al (2010) como alternativa para capturar comportamientos más complejos en geología, tales como comportamientos curvilíneos, relaciones geométricas, heterogeneidad espacial compleja y conectividad de valores extremos. Dichos cumulantes son definidos como combinaciones de momentos estadísticos que permiten caracterizar variables aleatorias no gaussianas que al ser considerados en el espacio permiten grandes combinaciones posibles de variables aleatorias de puntos, sugiriendo un vínculo con patrones de conectividad espacial compleja en un sentido geológico (Dimitrakopoulos et al, 2010).



El formulismo matemático de los cumulantes y su relación con los momentos se presenta a continuación.

Dada una variable aleatoria  $Z$ , su función generadora de momentos  $M$  está definida por:

$$M(w) = E[e^{wZ}] = \int_{-\infty}^{+\infty} e^{wz} f_Z(z) dz \quad (2.33)$$

La función generadora de cumulantes de  $Z$  es el logaritmo natural de la función generadora de momentos  $M$

$$K(w) = \ln(E[e^{wZ}]) \quad (2.34)$$

Y dado que una función aleatoria está completamente determinada por su función de densidad de probabilidad, su función de distribución acumulada, su primera o segunda función característica (momentos). El  $r$  momento de  $Z$  (con  $r \geq 0$ ) es  $M[Z, \dots, Z] = E[Z^r]$ . Con lo cual, siempre que la función generadora de momentos  $M$  tiene una expansión de Taylor alrededor del origen, entonces el  $r$  momento de  $Z$  es la  $r$  derivada de  $M$  en el origen. Por lo tanto, los cumulantes de  $Z$  son los coeficientes en la serie de expansión de Taylor de la función generadora de cumulantes  $K$  en el origen, tal como se presenta a continuación:

$$M(w) = E[e^{wZ}] = \sum_{r=0}^{\infty} \frac{w^r \text{Mom} \overbrace{[Z, \dots, Z]}^r}{r!} \quad (2.35)$$

$$K(w) = \ln(E[e^{wZ}]) = \sum_{r=0}^{\infty} \frac{w^r \text{Cum} \overbrace{[Z, \dots, Z]}^r}{r!} \quad (2.36)$$

La relación entonces, entre los momentos y los cumulantes, obtenidos de las ecuaciones 2.35 y 2.36 son las siguientes para los primeros 3 cumulantes:

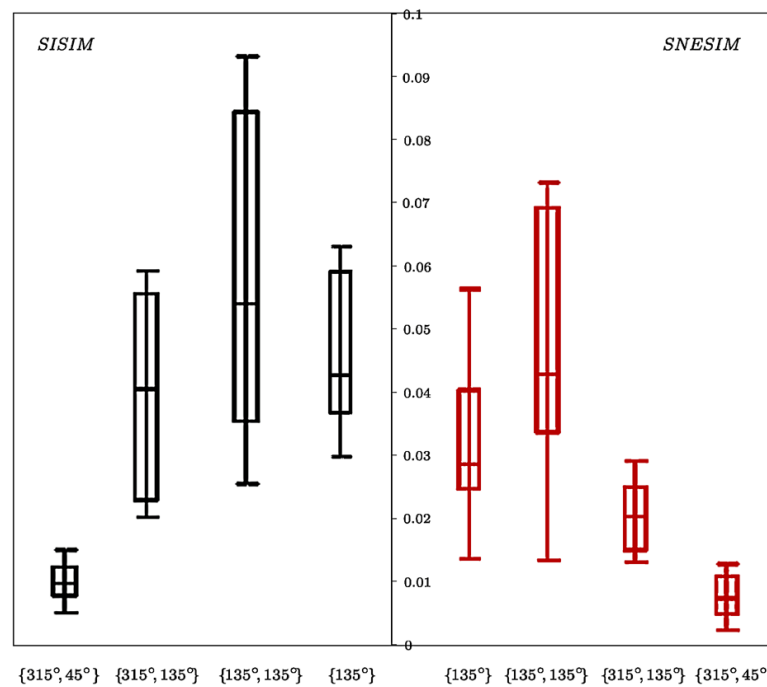
$$\text{Cum}[Z] = \text{Mom}[Z] \quad (2.37)$$

$$\text{Cum}[Z, Z] = \text{Mom}[Z, Z] - \text{Mom}[Z]^2 \quad (2.38)$$

$$\text{Cum}[Z, Z, Z] = \text{Mom}[Z, Z, Z] - 3\text{Mom}[Z, Z]\text{Mom}[Z] + 2\text{Mom}[Z]^3 \quad (2.39)$$

Para un mayor detalle del formulismo de los cumulantes espaciales de orden superior, referirse a Dimitrakopoulos et al (2010) y Mustapha y Dimitrakopoulos (2009).

Explicado el concepto de cumulantes de orden superior, se presenta en la Figura 2-8 los box-plot de los errores medios cuadráticos para cumulantes de segundo y tercer orden efectuados por De Iaco y Maggio (2011) para sus simulaciones mediante dos algoritmos teniendo en cuenta la anisotropía de su modelo.



**Figura 2-8:** Análisis comparativo de error medio cuadrático de cumulantes de orden dos y orden tres  
Fuente: De Iaco y Maggio (2011)

Por su parte, Boisvert et al (2010) proponen cinco diferentes métricas para comparar la reproducción de las estadísticas multipunto de los datos de entrada (de la imagen de entrenamiento) respecto a las simulaciones geoestadísticas, las cuales son:

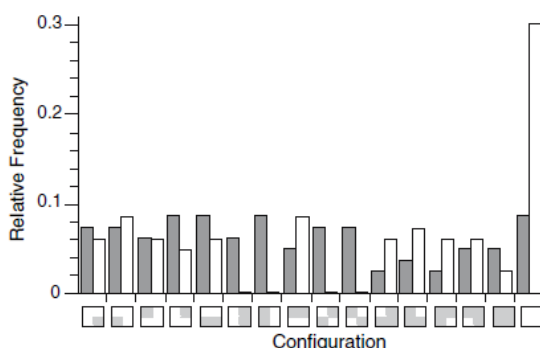
- Estadísticas de primer y segundo orden
- Tendencias
- Histogramas multiescala
- Función de densidad de múltiples puntos
- Configuraciones perdidas en la función de densidad de múltiples puntos

Las estadísticas de primer y segundo orden, hacen referencia a los niveles 1 y 2 de validación presentados en este documento y al de validación de estadísticas de dos puntos (verificación de reproducción de variogramas).

Según Boisvert et al (2010), cuando una tendencia es usada para forzar una proporción no estacionaria de facies que no pueden ser capturadas por la imagen de entrenamiento, el uso de dicha tendencia debe quedar representada en las realizaciones y por ende debe verificarse este aspecto. Esto se efectúa en dos caminos: (1) globalmente para una simple realización o (2) localmente sobre un set de realizaciones.

En cuanto a los histogramas multiescala, Boisvert et al (2010) manifiestan que dado que las realizaciones de una simulación geoestadística generalmente son usadas para obtener un promedio a una escala más grande que el tamaño de bloque simulado (rebloqueo a un tamaño de interés como por ejemplo unidades selectivas de explotación), el comportamiento del histograma dependerá de la estructura espacial subyacente de la imagen de entrenamiento y las realizaciones. Por esto, dado que en teoría el histograma de la imagen de entrenamiento y las realizaciones de la simulación a diferentes escalas, deberían ser idénticos, se efectúan chequeos para determinar diferencias.

Al igual que la verificación de la reproducción del variograma, debe chequearse la reproducción de la función de densidad de múltiples puntos, entendida esta como la frecuencia de ocurrencia de una cierta configuración multipunto (Boisvert et al., 2007). Sin embargo, esto no es fácil de lograr y fue resuelto por Boisvert et al (2007), mediante la suma de los valores absolutos de las diferencias de las frecuencias de cada configuración presente en la realización y en la TI. Mientras más pequeño sea esta suma o más cercano esté a cero, más grado de similaridad existirá entre la TI y la realización. En la Figura 2-9 se presenta una función de densidad de múltiples puntos para dos imágenes de entrenamiento, de la cual es posible obtener la suma de las diferencias para cada configuración.

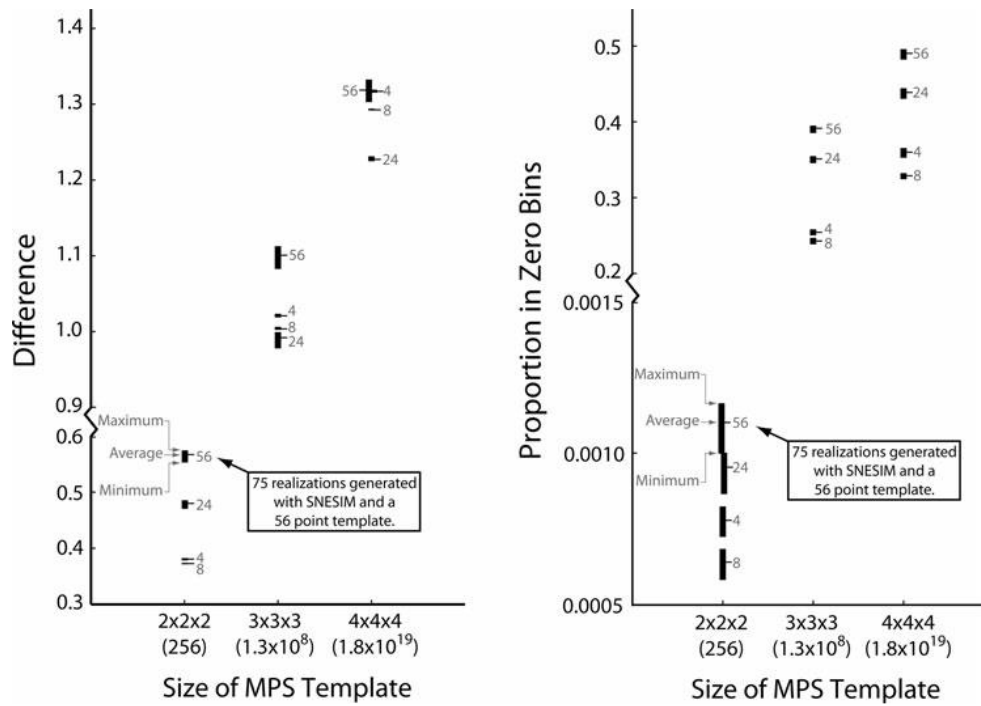


**Figura 2-9:** Función de densidad de múltiples puntos para dos imágenes de entrenamiento.  
Fuente: Boisvert et al (2007)

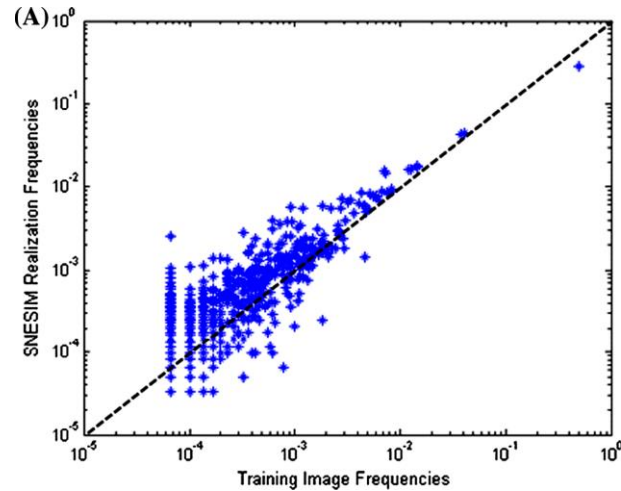
Finalmente, Boisvert et al (2010) analizan configuraciones perdidas en la función de densidad de múltiples puntos. Dado que se asume que la imagen de entrenamiento es representativa del dominio a modelar, si cierta configuración no aparece en la imagen de entrenamiento entonces no debería aparecer en las realizaciones.

En la Figura 2-10 se presentan las estadísticas de los análisis efectuados por Boisvert et al (2010) para las métricas de función de densidad de múltiples puntos y configuraciones perdidas de la función de densidad de múltiples puntos. En la imagen de la izquierda se presenta la suma de las diferencias absolutas de las funciones de densidad de múltiples puntos. En la derecha la proporción de configuraciones o patrones no presentes. Mientras más cercanos estén dichos valores a cero mejor sería la reproducción de las estadísticas de orden superior de las realizaciones respecto a la imagen de entrenamiento.

Rezaee et al (2014) proponen otra forma de validación de las realizaciones obtenidas mediante las simulaciones. En un primer análisis validan las realizaciones en función de las frecuencias de patrones similares entre la imagen de entrenamiento y las realizaciones. Para ello definen una ventana de búsqueda que permite construir una base de datos de patrones, cada uno de los cuales es almacenado, así como su frecuencia. Dado que el análisis efectuado es sobre una variable binaria, se asigna 1 y 0 para determinado valor y X cuando no hay información disponible. Luego se identifican los patrones similares y para cada par se grafican las frecuencias obtenidas en la imagen de entrenamiento vs. las frecuencias obtenidas para las realizaciones; permitiendo comparar si las realizaciones reproducen adecuadamente la imagen de entrenamiento. En la Figura 2-11 se presenta este tipo de gráfico, donde se puede observar, según el análisis efectuado por Rezaee et al (2014), que hay menos similitud entre las realizaciones y la imagen de entrenamiento para bajas y medias frecuencias.



**Figura 2-10:** Validación de función de densidad de múltiples puntos.  
Fuente: Boisvert et al (2010)



**Figura 2-11:** Gráfico de frecuencias de patrones para la TI y realización.  
Fuente: Rezaee et al (2014)

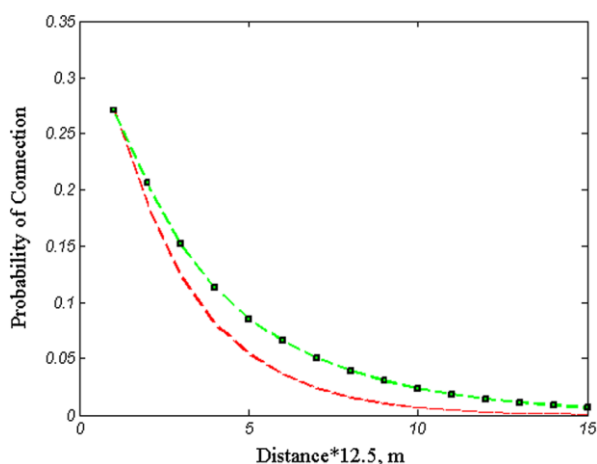
Por su parte Pérez (2013) propone una herramienta de validación de simulaciones a partir de la reproducción de patrones definidos desde las realizaciones (evento de datos) y que son encontrados en la imagen de entrenamiento o viceversa. Para ello, los patrones o evento de datos se construyen a partir de una espiral de búsqueda que define el usuario (que puede ser obtenido desde las realizaciones o desde la TI) y una vez se tenga este, se realiza una búsqueda de estos patrones en la imagen de entrenamiento para determinar la proporción de patrones encontrados o si los patrones son obtenidos desde la TI la proporción de patrones que se encuentran en la realización. Sin embargo, dado la demanda computacional, esto se desarrolla para unos pocos patrones. Esta

herramienta está desarrollada a partir del método Direct Sampling propuesto por Mariethoz et al (2010).

Adicionalmente Rezaee et al (2014) efectúan una validación en términos de conectividad espacial. Esta se efectúa evaluando la probabilidad de que dos nodos estén conectados a una distancia pre-determinada en una dirección en particular. A partir de éste análisis se construye una gráfica permitiendo determinar si existe menor o mayor conectividad de la realización respecto a la imagen de entrenamiento.

Western et al (2001), Krishnan y Journel (2003) y Schlüter y Vogel (2011) desarrollan este mismo análisis de conectividad en su trabajo; el cual consiste en determinar la probabilidad de que dos pixeles separados por una distancia  $h$  estén conectados por pixeles con la misma propiedad (por ejemplo un valor de 1 para una variable binaria). Dicho análisis se efectúa para diferentes distancias con lo cual es posible obtener una curva de la distancia vs. la función de conectividad.

En la Figura 2-12 se presenta un ejemplo de los posibles resultados a obtener con los cuales se puede comparar la imagen de entrenamiento y las realizaciones.



**Figura 2-12:** Análisis comparativo de conectividad.  
Rojo: realizaciones. Verde: Imagen de entrenamiento  
Fuente: Rezaee et al (2014)

#### 2.3.4. Nivel 4. Validación de predicción.

Al momento de efectuar la validación de un modelo de estimación (en donde la estimación entrega un único valor), las técnicas de validación cruzada y Jack Knife han sido empleadas para determinar la calidad al comparar el valor estimado con el valor conocido. De manera equivalente, en las simulaciones (las cuales entregan una distribución de incertidumbre) la técnica empleada es similar a una validación cruzada pero teniendo en cuenta que la comparación será entre pares de valores conocidos y distribuciones de incertidumbre generadas por los modelos sobre múltiples realizaciones (Pyrzcz y Deutsch, 2014). Ello exige entonces evaluar la bondad de la distribución de incertidumbre.

Para ello se analiza la precisión y exactitud del modelo de incertidumbre, basado en la fracción real de valores conocidos que cae dentro de intervalos de probabilidad simétricos de anchura variable  $p$  entre 0 y 1 ( $p \in [0,1]$ ). Pyrcz y Deutsch (2014) citan el trabajo previo de Deutsch (1996) y afirman

que la distribución de probabilidad es exacta si la fracción de valores verdaderos que caen en el intervalo  $p$  exceden  $p$  para todo  $p \in [0,1]$ . La precisión de una distribución de probabilidad exacta se mide por la cercanía de la fracción de valores verdaderos de  $p$  para todos  $p \in [0,1]$ . Gráficamente se puede efectuar el análisis, mediante una nube de dispersión

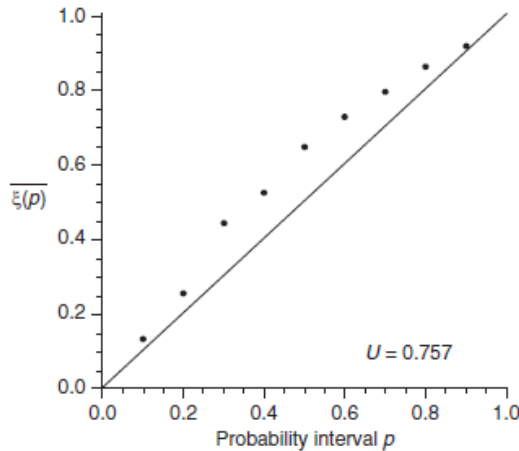
El procedimiento propuesto es el siguiente:

- Calculan las probabilidades asociada a los valores verdaderos  $z(u_i)$ ,  $i=1, \dots, n$ , usando el modelo de incertidumbre  $F^*(u_i, z(u_i) | n(u_i))$ ,  $i=1, \dots, n$ . Donde  $z(u_i)$  es el valor conocido en un punto  $u_i$  y  $n(u_i)$  son los restantes puntos conocidos con los cuales se calcula el punto  $z(u_i)$ .
- Luego se definen intervalos de probabilidad simétricos  $p$ , con valores de probabilidad superior e inferior dados por  $(1+p)/2$  y  $(1-p)/2$  respectivamente.
- Posteriormente, se define una función indicador en cada posición  $u_i$  según la ecuación 2.40.

$$\xi(u_i, p) = \begin{cases} 1, & \text{Si } F(u_i, z(u_i) | n(u_i)) \in \left[ \frac{1-p}{2}, \frac{1+p}{2} \right] \\ 0, & \text{En caso contrario} \end{cases} \quad (2.40)$$

- El promedio de la función indicador sobre las  $n$  ubicaciones  $u_i$ , será la proporción de ubicaciones donde el valor verdadero (probabilidad real) cae dentro del intervalo de probabilidad simétrico  $p$  (probabilidad teórica).

De manera práctica emplean un gráfico de dispersión para efectuar los análisis. Si los puntos caen sobre la línea diagonal (línea de 45°) la distribución de probabilidad será exacta y precisa. Si los puntos caen por encima de la línea diagonal son exactos pero no precisos, y si caen por debajo no serán ni exactos ni precisos. En la figura 2.13 se presenta un gráfico de este análisis presentado por Pyrcz y Deutsch (2014).



**Figura 2-13:** Gráfico de dispersión para validar distribuciones de probabilidad.

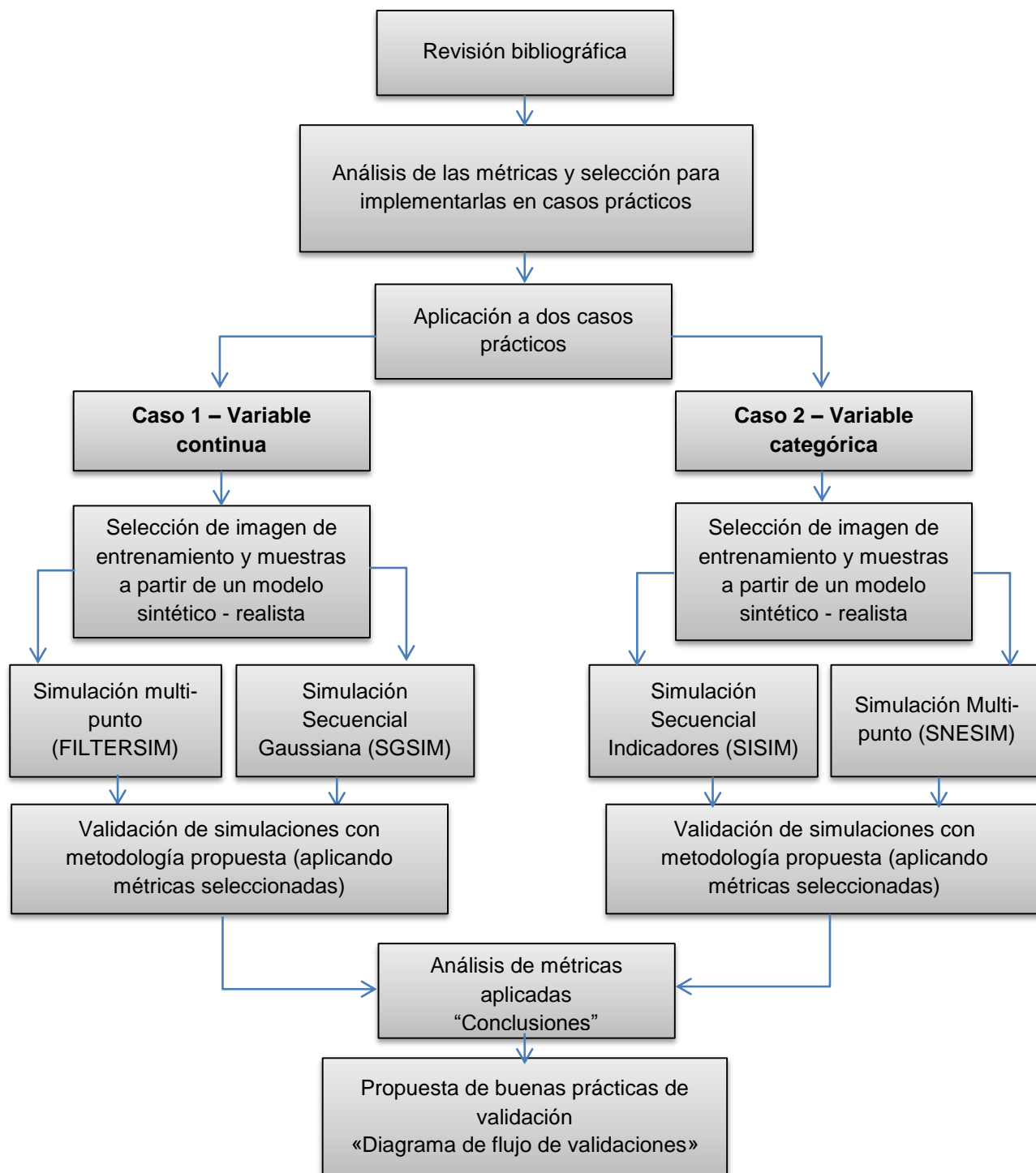
Fuente: Pyrcz y Deutsch (2014).

Olea (2012) desarrolla un trabajo similar para determinar la consistencia de las distribuciones de incertidumbre y comparar diversos parámetros de input de los modelos.

Finalmente, Mariethoz y Caers (2014) citan el trabajo desarrollado por Tan et al (2014). En este se propone analizar la incertidumbre espacial asociada a las simulaciones multi-punto, las cuales generan dos tipos de variabilidad. La primera relacionada con una variabilidad dentro de una realización que correspondería a la variabilidad de patrones dentro de la realización y como se compara con la variabilidad de patrones de la TI; y la segunda a una variabilidad entre realizaciones o lo más comúnmente conocido como incertidumbre espacial. Es deseable entonces encontrar una baja variabilidad dentro de la realización (es decir, una buena comparación con la TI), pero a la vez una gran variabilidad entre las realizaciones (es decir, no crear múltiples copias de la TI). Para abordar este tema se emplea la metodología ANODI (Analysis of Distances) basada en distancias entre realizaciones y distancias entre estadísticas de una realización y estadísticas de la TI objetivo.

## CAPÍTULO 3. METODOLOGÍA

La metodología propuesta para desarrollar este estudio se presenta en el diagrama de flujo siguiente con las diferentes actividades y procedimientos para lograr los objetivos planteados.



**Figura 3-1:** Esquema metodológico del trabajo.



Es importante aclarar que si bien las métricas de validación que se pondrán en práctica en los capítulos subsiguientes compararán en algunos casos el desempeño de dos tipos de algoritmos, el ejercicio no pretende presentar un algoritmo como mejor que el otro; sino más bien mostrar como dichas métricas pueden validar o verificar aspectos importantes que deberían reproducir las realizaciones y que tal vez cierta característica fue reproducida de mejor manera por parte de uno de los algoritmos (acorde a los parámetros de modelación inicial empleados en las simulaciones). Bien podría hacerse el ejercicio como método para mejorar el desempeño de un mismo algoritmo respecto a un análisis de sensibilidad al ajustar o calibrar los parámetros de modelación.

- **Software**

Para desarrollar los casos prácticos diferentes programas o códigos construidos por otros autores o instituciones fueron empleados, bien sea en el caso de las simulaciones o para las validaciones. A continuación se presenta un listado de estos con los respectivos créditos a sus creadores.

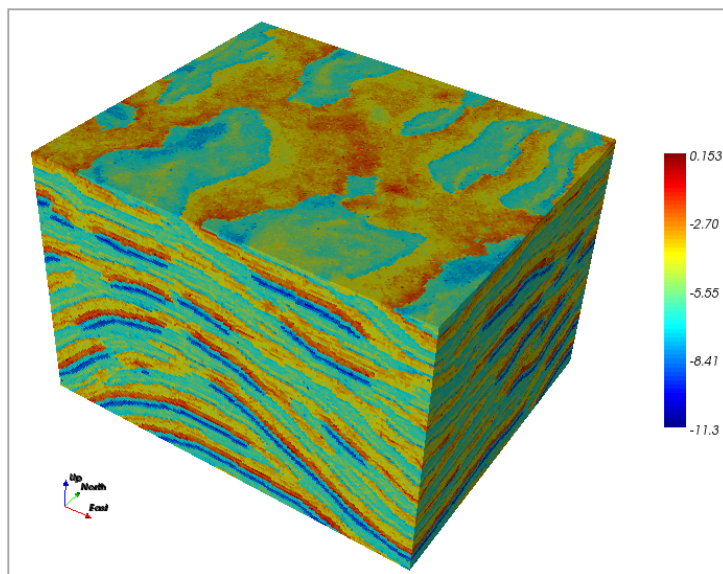
**Tabla 3.1:** Software y códigos empleados en el estudio

<b>Software/Código</b>	<b>Uso</b>	<b>Autores</b>
SGeMS	Simulaciones	Universidad de Stanford
Gslib	Modelación de variogramas y sus validaciones, estadísticas	Universidad de Stanford
HOSC	Cálculo de cumulantes de orden 3	Dimitrakopoulos y Musthapa. Universidad McGill
Programas MatLab Student_mean.m Student_var.m Hotelling_variog.m	Test estadísticos	Emery. Universidad de Chile
Programas MatLab distribution.m pixelp1t.m scatp1t.m	Gráficos de dispersión, despliegue de realizaciones.	Emery. Universidad de Chile
Programas MatLab Grid2Pts.m LoadGrid.m LoadPts.m NewGrid.m WriteGrid.m WritePts.m YXZ2XYZ.m ExtractSlice.m	Selección de imagen exhaustiva, imagen de entrenamiento.	Mariethoz. University of Lausanne (UNIL)
Programas MatLab Accuracy.m SamplesAP.m SearchPatternsGeneral.m	Obtención de probabilidades reales para generación de accuracy plot Obtención de frecuencias de patrones observados en realizaciones, imagen exhaustiva y TI.	Programados para el presente estudio.

## CAPÍTULO 4. CASO PRÁCTICO VARIABLE CONTINUA

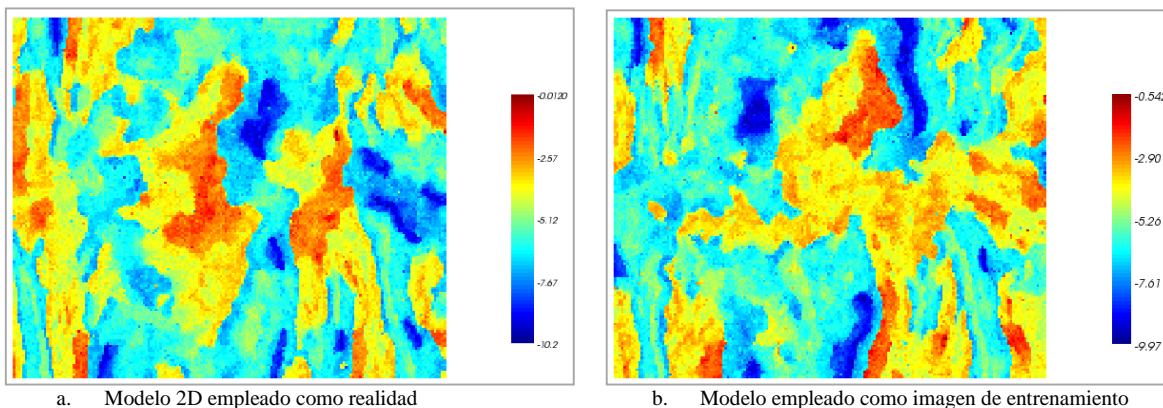
### 4.1. MODELO Y ALGORITMOS EMPLEADOS

Para efectuar validaciones de una variable continua, se empleó un modelo 3D de plegamientos<sup>1</sup>. Este modelo está compuesto por una grilla de 180x150x120 unidades, el cual puede apreciarse en la figura 4.1.



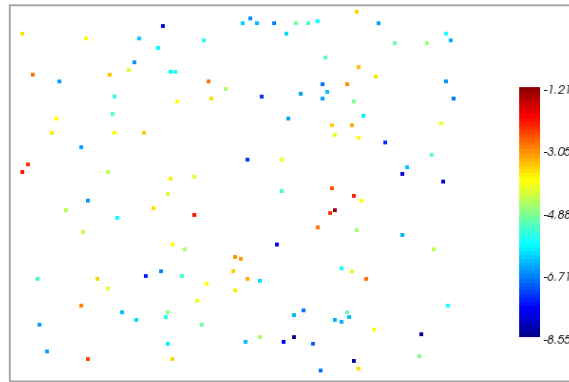
**Figura 4-1:** Modelo continuo de plegamiento

De este modelo se seleccionó un nivel el cual fue considerado como imagen de entrenamiento (nivel 60) y un nivel levemente inferior (nivel 55), el cual fue empleado como la realidad. De este último se seleccionaron 135 muestras; las cuales son empleadas para efectuar las posteriores simulaciones. En las Figura 4-2 y Figura 4-3 se presentan la imagen de la realidad, la imagen de entrenamiento y las muestras seleccionadas aleatoriamente.



**Figura 4-2:** Imágenes extraídas del modelo continuo y empleadas para las simulaciones

<sup>1</sup> Disponible en <http://trainingimages.org/training-images-library.html>

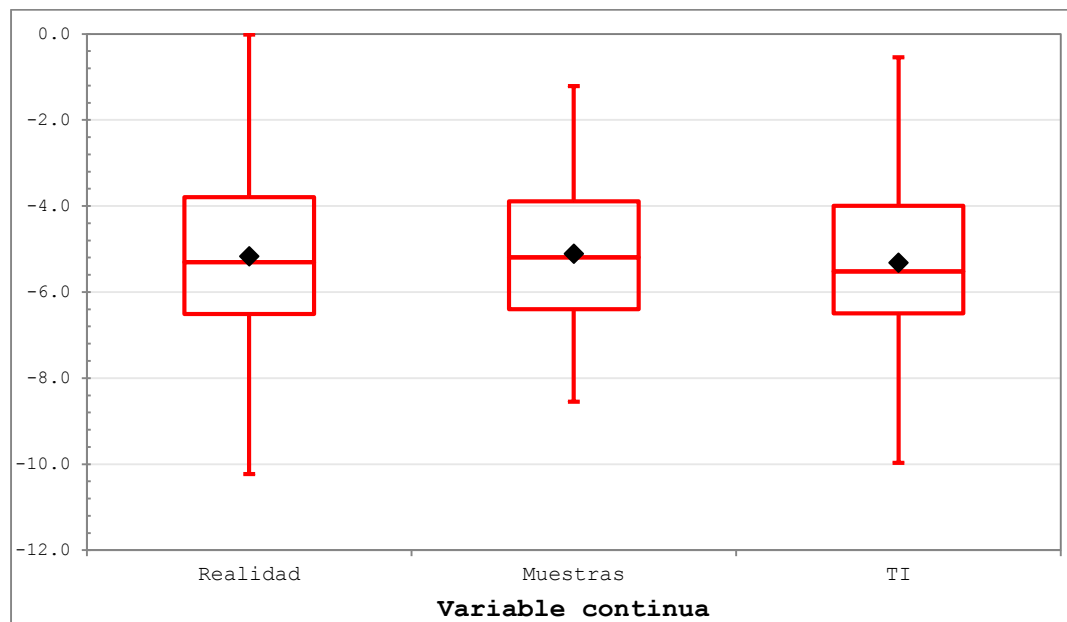


**Figura 4-3:** 135 muestras extradías aleatoriamente de modelo de la realidad

Las estadísticas descriptivas de los datos empleados (imagen de entrenamiento, imagen exhaustiva – realidad- y las muestras se presentan en la Tabla 4.1. En la figura 3 se presentan los box-plot que describen la distribución de estos conjuntos de datos.

**Tabla 4.1:** Estadísticas descriptivas de información de entrada

	<b>Realidad</b>	<b>Muestras</b>	<b>TI</b>
<b>Datos</b>	27000	135	27000
<b>Media</b> ♦	-5.16759	-5.10669	-5.31489
<b>Varianza</b>	2.88545	2.52742	2.51415
<b>Mínimo</b>	-10.2293	-8.54945	-9.96908
<b>1<sup>er</sup> cuartil</b>	-6.51048	-6.39906	-6.49721
<b>Mediana</b>	-5.30432	-5.19469	-5.52066
<b>3<sup>er</sup> Cuartil</b>	-3.79609	-3.88948	-3.9935
<b>Máximo</b>	-0.01204	-1.21058	-0.542249



**Figura 4-4:** Box-plot de datos de entrada para simulaciones

Como se puede observar en la Figura 4-4 y estadísticas de la Tabla 4.1, existe una mayor similitud entre las distribuciones de la imagen de entrenamiento y la imagen exhaustiva (asumida como realidad); sin embargo con algunas diferencias que sería lo lógico tener al momento de emplear una imagen de entrenamiento que nunca será igual a la realidad de nuestra variables de interés a simular; sino una aproximación a esta. Por otro lado, los datos que constituyen las muestras presentan una distribución más estrecha que refleja una situación muy posible de lo que acontecería en un caso real, donde se tienen muestras que de alguna manera pretenden caracterizar la realidad; pero que posiblemente no logrará reproducir exactamente esta e inclusive podría presentar sesgos dado, por ejemplo, por muestreos preferenciales, no uniformes, ni aleatoriamente distribuidos en el espacio.

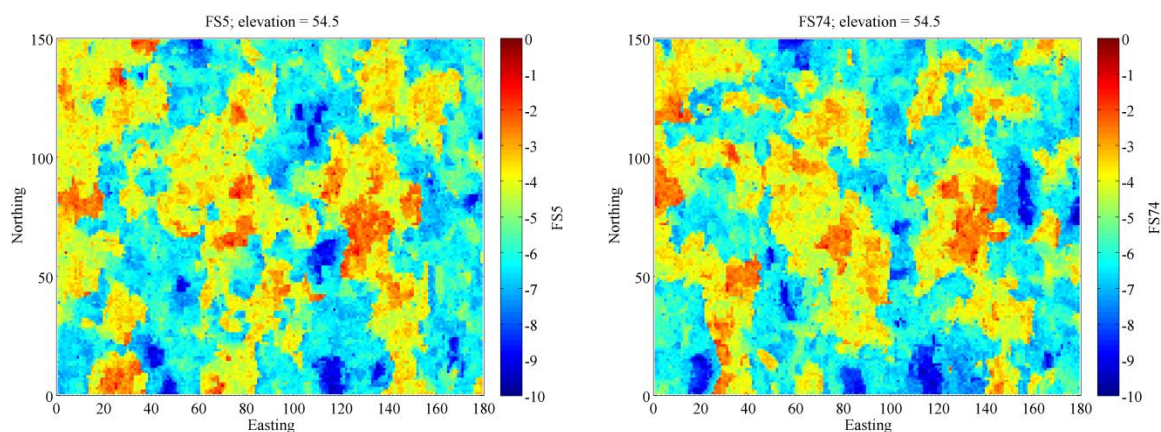
- **Algoritmos a emplear para las simulaciones**

Teniendo claro las características de la información de entrada, se seleccionaron como algoritmos de simulación FILTERSIM y SGSIM, cuya descripción se presentó en el marco teórico de este informe. Ambos algoritmos están disponibles en el software SGeMS de la Universidad de Stanford<sup>2</sup>. Los parámetros empleados para efectuar las simulaciones de cada uno de estos algoritmos puede verse en el anexo A.

En el caso del SGSIM se efectuó adicionalmente la transformación de los datos a una distribución gaussiana y se comprobó la hipótesis de bi-gaussianidad (con lo cual es posible asumir que se tienen distribuciones multi-gaussianas y por ende es posible emplear este algoritmo en las simulaciones). En el anexo B se presentan los gráficos de correlación diferida a diferentes distancias donde se valida la hipótesis de bi-gaussianidad. Igualmente se modeló el variograma para los datos transformados a gaussianos lo cual es un requisito del algoritmo. El variograma modelado se presenta en el anexo C.

Se generaron 100 realizaciones por medio de estos algoritmos, con las que se pretenden poner en práctica las diferentes métricas para validar los resultados obtenidos, en los diferentes niveles de validación propuestos.

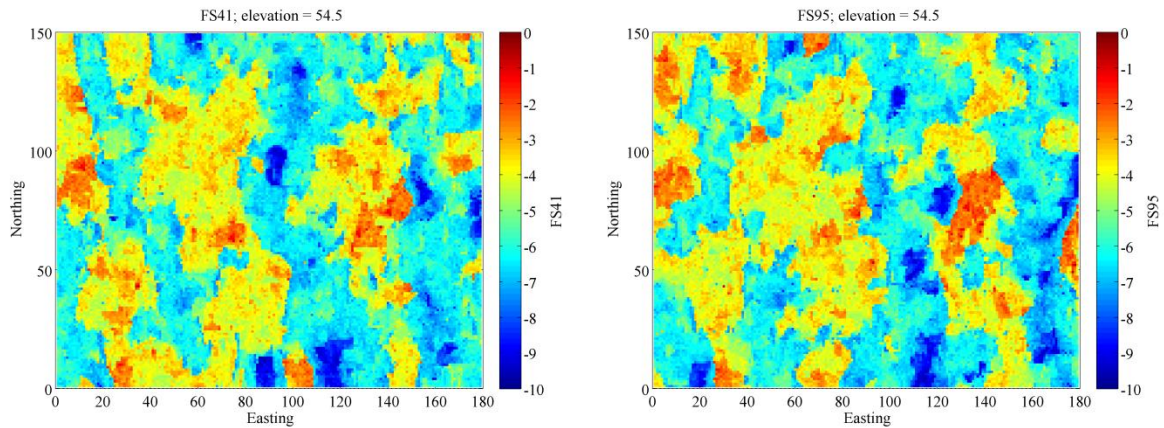
Entre las Figura 4-5 a Figura 4-7 se pueden observar algunos ejemplos de los resultados obtenidos en diferentes realizaciones tanto para FILTERSIM, como para SGSIM.



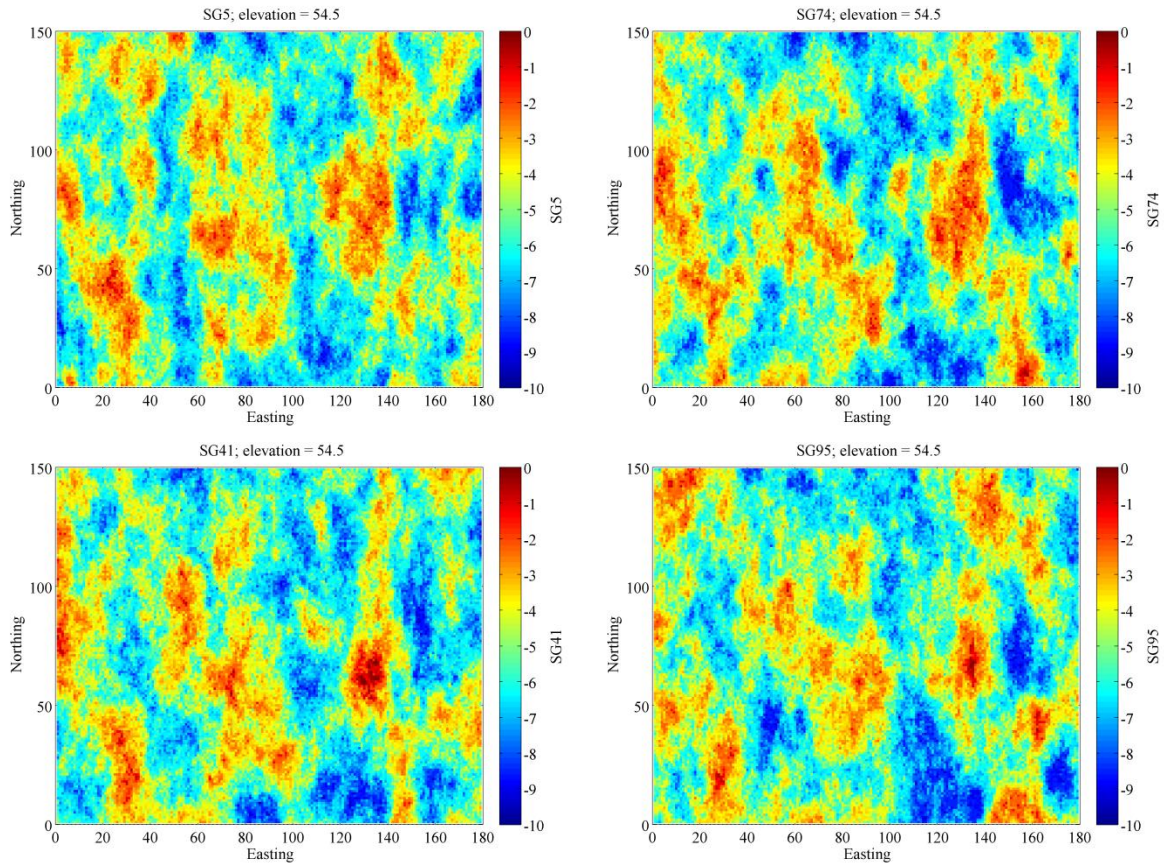
**Figura 4-5:** Ejemplo de realizaciones 5 y 74 obtenidas con FILTERSIM (FS)

<sup>2</sup> Disponible en <http://sgems.sourceforge.net/>





**Figura 4-6:** Ejemplo de realizaciones 41 y 95 obtenidas con FILTERSIM (FS)



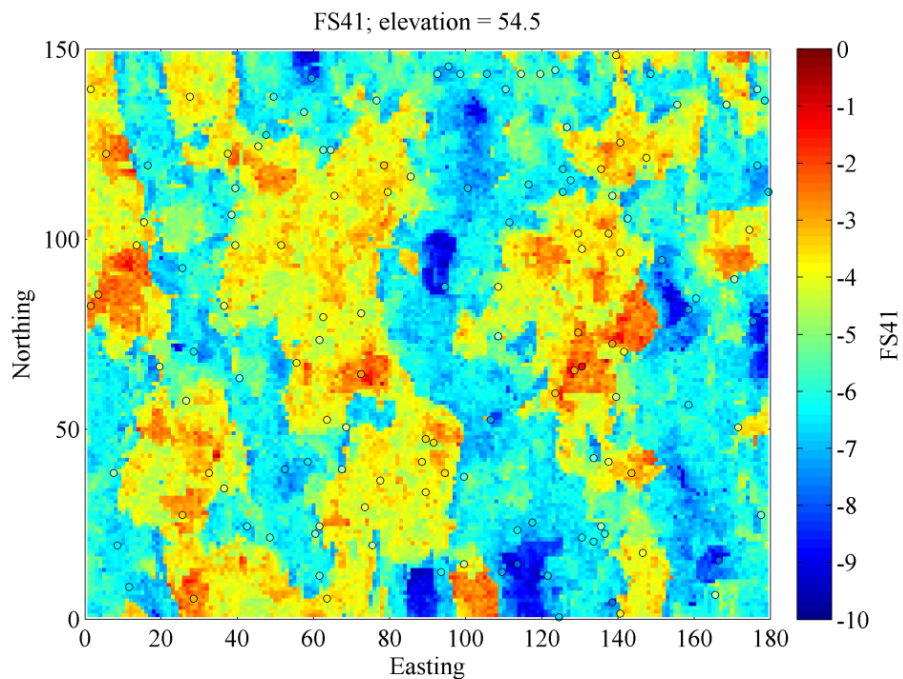
**Figura 4-7:** Ejemplo de realizaciones obtenidas con SGSIM (SG)

#### 4.1.1. Análisis para validación de simulaciones de variable continua

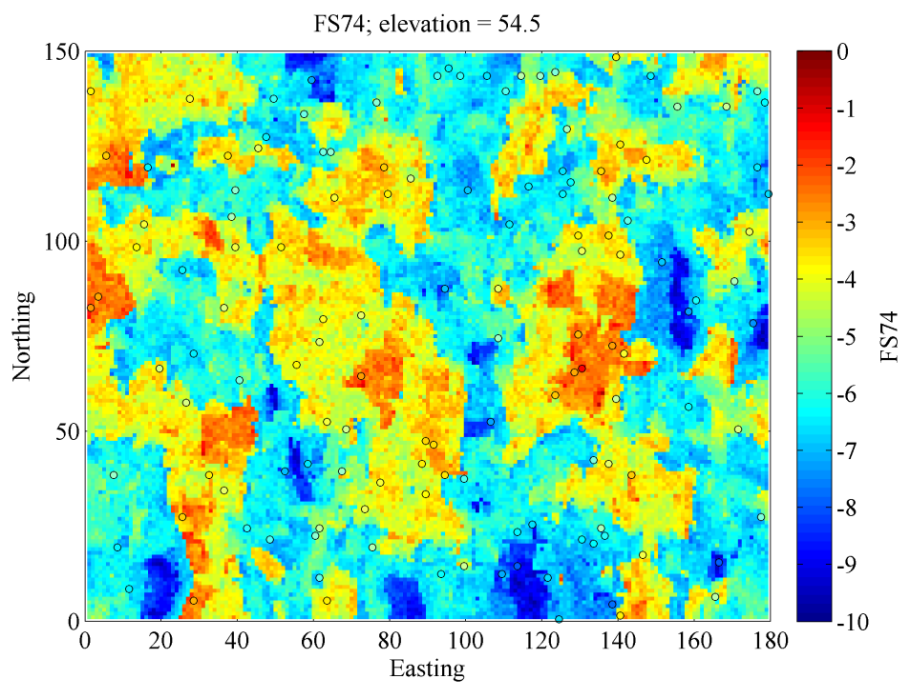
##### 4.1.1.1. Nivel 1. Condicionamiento de Datos

Un primer análisis consiste en verificar si los resultados de las simulaciones reflejan los datos reales en los resultados de las realizaciones, toda vez que las simulaciones fueron efectuadas condicionando a los datos originales o muestras empleadas. Para ello un primer análisis es el visual, en el cual se puede observar si las realizaciones están o no acorde a las muestras empleadas. En las

Figura 4-8 y Figura 4-9 se puede observar las muestras (en círculos rellenos) y la realización obtenida.

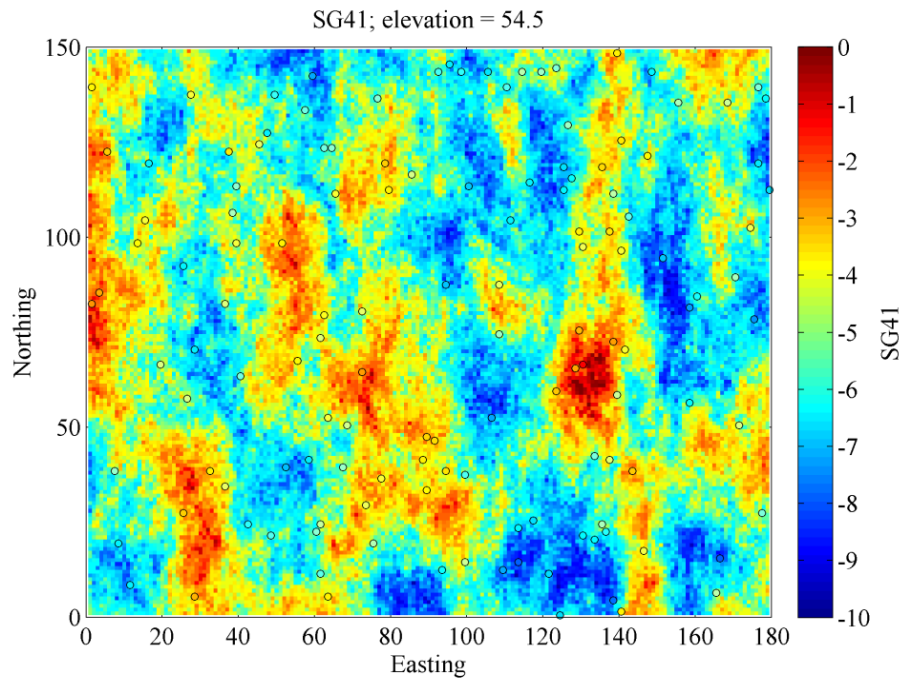


a. Realización FILTERSIM 41

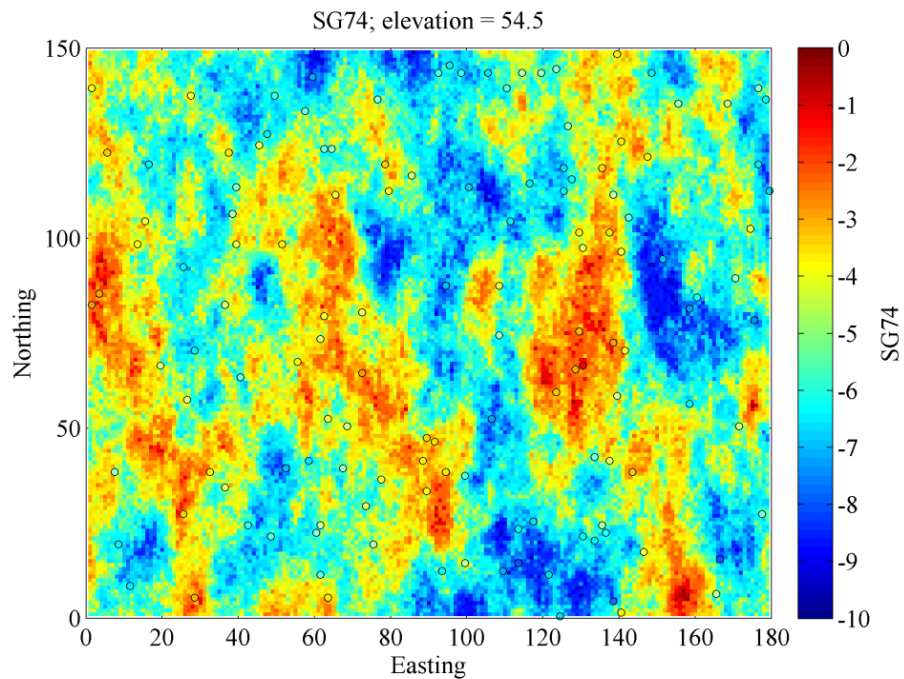


b. Realización FILTERSIM 74

**Figura 4-8:** Inspección visual de muestras y realizaciones para FILTERSIM



a. Realización SGSIM 41



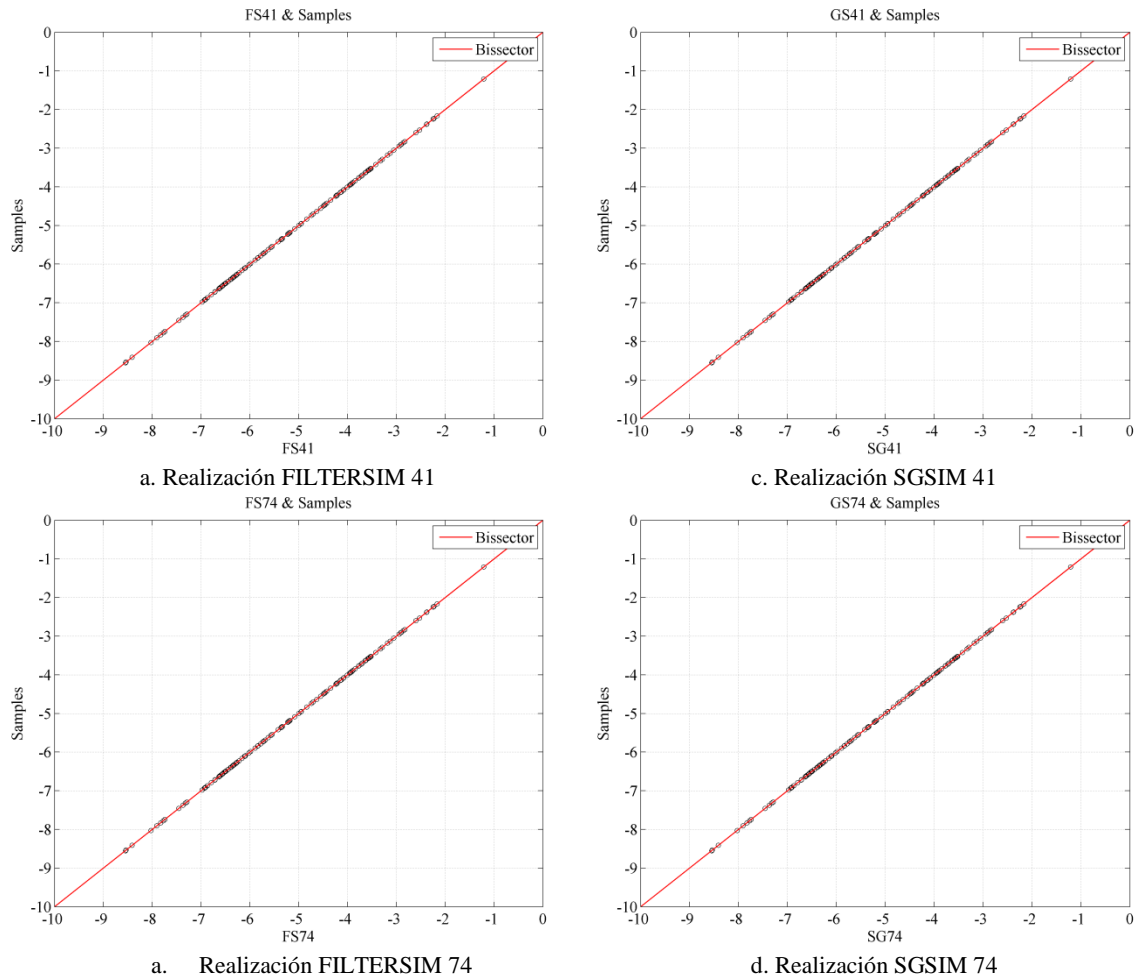
b. Realización SGSIM 74

**Figura 4-9:** Inspección visual de muestras y realizaciones para SGSIM

Como se pudo observar en las Figura 4-8 y Figura 4-9, los valores de los datos originales son reproducidos en las realizaciones. Esto se puede ver más claramente en un gráfico de dispersión (scatter plot), donde se presentan los valores simulados vs. los valores originales (datos de muestras) para las realizaciones generadas con FILTERSIM y SGSIM. Es importante considerar que para este caso los nodos de los sitios con muestra coinciden exactamente con nodos de las simulaciones generadas, situación que no siempre podría presentarse conforme como se realice la



simulación, motivo por el cual se hace importante este tipo de validación. En la Figura 4-10 se pueden observar dicho comportamientos.



**Figura 4-10:** Gráfico de dispersión de valores simulados vs. valores originales

#### 4.1.1.2. Nivel 2. Validación estadística

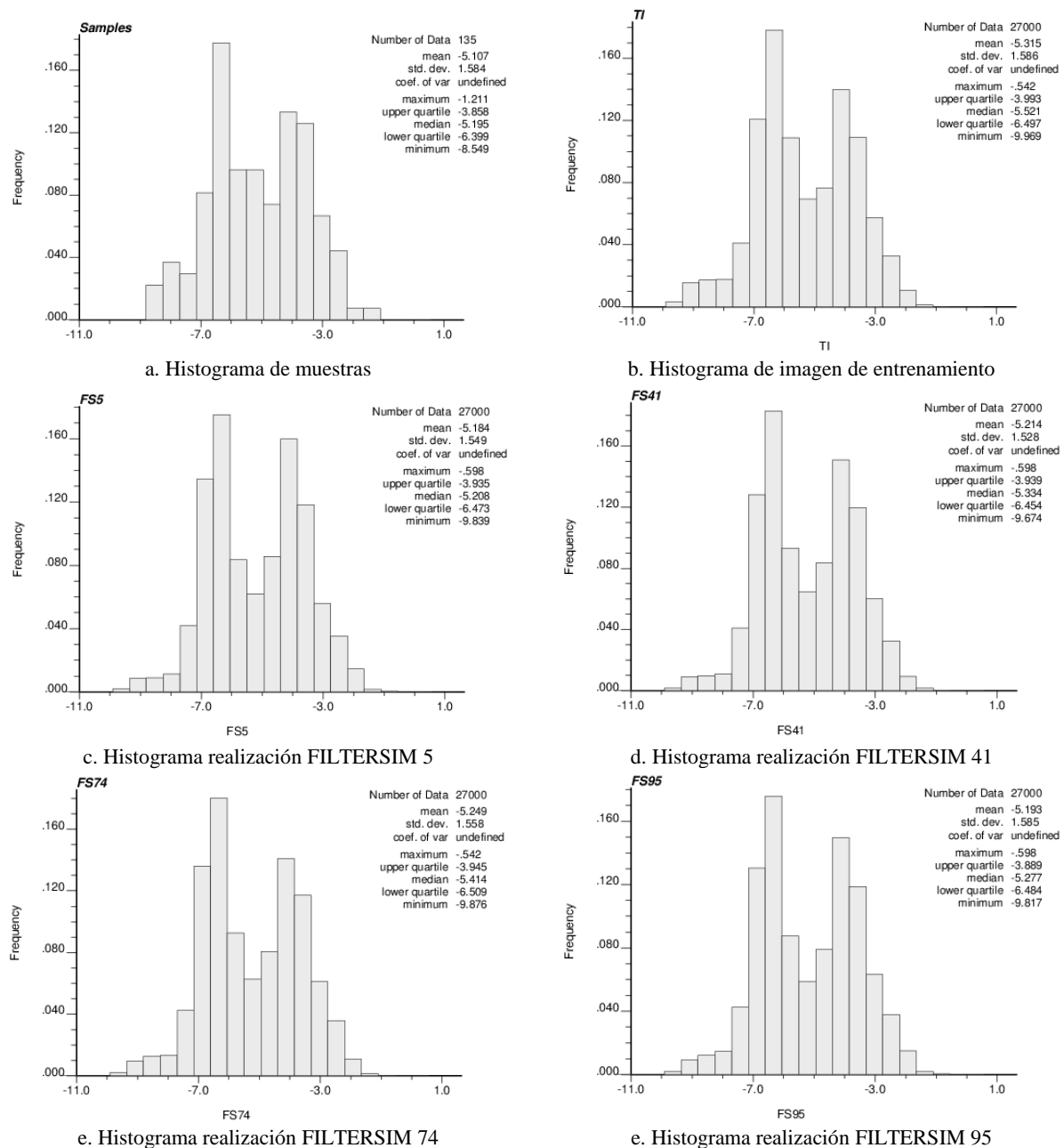
En este nivel se evalúa la reproducción de las principales estadísticas. Para ello se comparan los histogramas de los datos originales, con los de las realizaciones. Igualmente, es posible comparar para el caso de FILTERSIM, la reproducción de las estadísticas de la imagen de entrenamiento, toda vez que este algoritmo trata de reproducir dicho comportamiento.

En primer lugar (Figura 4-11) se presentan los histogramas de los datos originales (muestras) y la imagen de entrenamiento, así como los histogramas de las realizaciones 5, 41, 74 y 95 generadas con FILTERSIM. Posteriormente se presentan los histogramas igualmente de las muestras los de las realizaciones 5, 41, 74 y 95 obtenidas con el algoritmo SGSIM (ver Figura 4-12). Visualmente se pueden apreciar, para el caso de FILTERSIM, semejanzas (comportamiento bi-modal) entre el histograma de las muestras y las realizaciones, situación que se reproduce, al menos desde el punto de vista visual, de mejor manera cuando se compara con la imagen de entrenamiento. Esto se debe principalmente a que el algoritmo usa la imagen de entrenamiento como un patrón a reproducir lográndose buenos resultados (según se apreciaba visualmente).

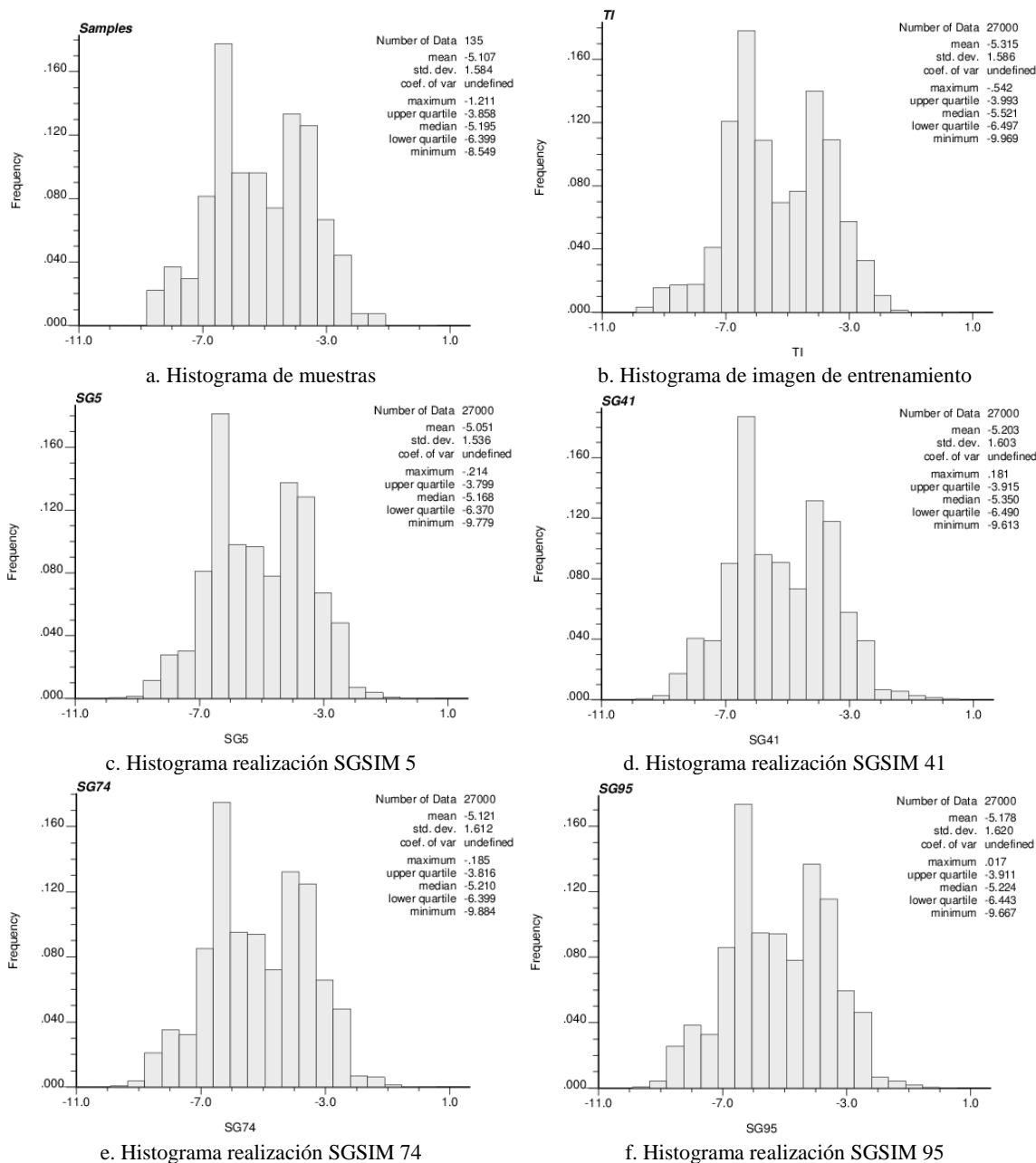


Para el caso de las simulaciones con SGSIM, se observa como las realizaciones presentan igualmente un comportamiento bi-modal pero visualmente parecen ajustarse un poco más a la forma de la distribución de las muestras que a las de las de la imagen de entrenamiento; lo cual es lo esperado dado que la imagen de entrenamiento no interviene dentro del algoritmo.

Sin embargo, para efectuar una comparación más precisa de las distribuciones se presentan posteriormente gráficos cuantil-cuantil (Q-Q plot); los cuales permiten comparar cuantil por cuantil las distribuciones de dos conjuntos de datos.



**Figura 4-11:** Histogramas para comparar reproducción de estadísticas - FILTERSIM

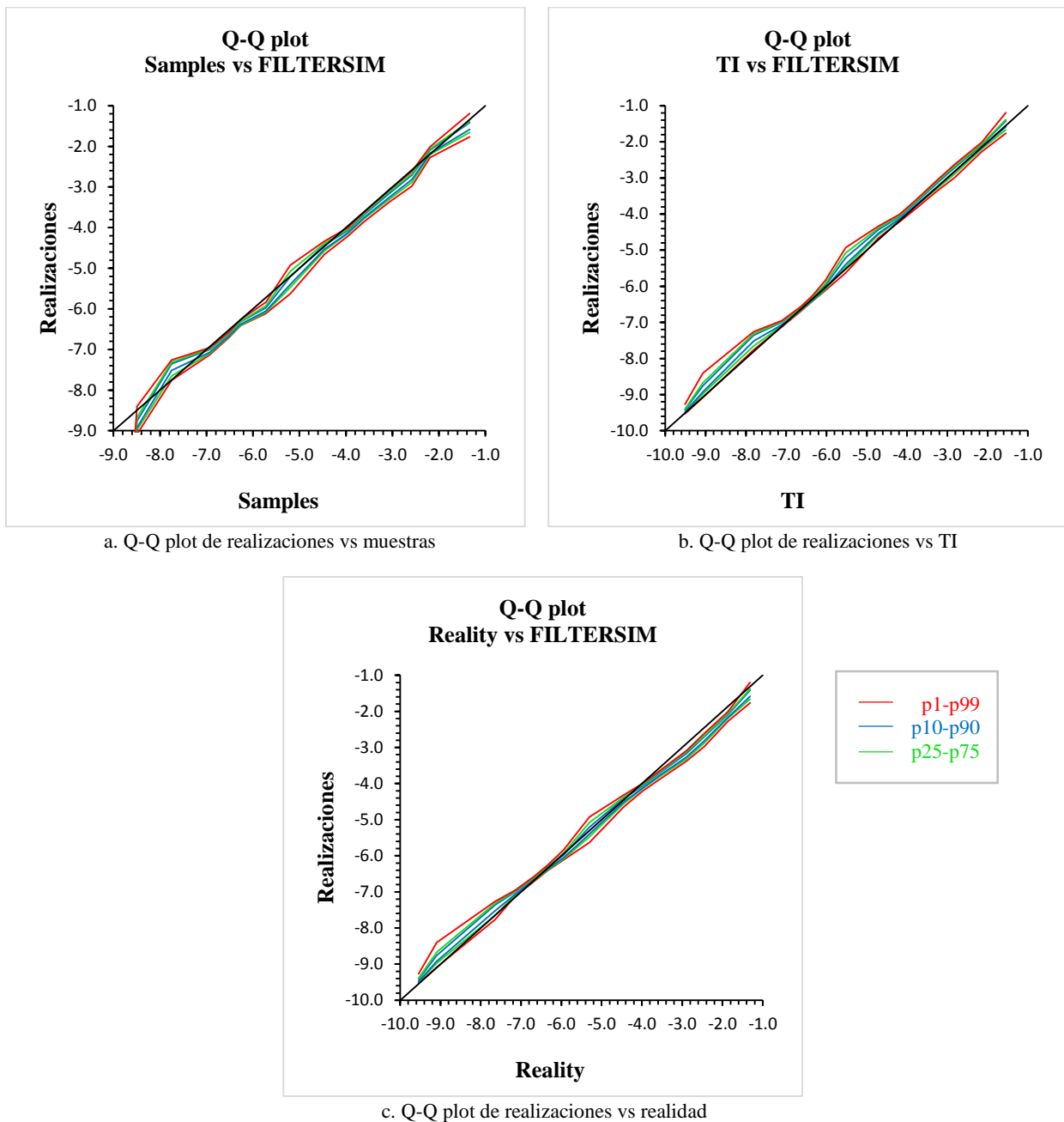


**Figura 4-12:** Histogramas para comparar reproducción de estadísticas - SGSIM

En la Figura 4-13 se presentan los Q-Q plot para el algoritmo FILTERSIM comparándolo contra las distribuciones de las muestras, de la imagen de entrenamiento y de la imagen exhaustiva (realidad). En el anexo D se presentan igualmente comparaciones para algunas realizaciones específicas de FILTERSIM.

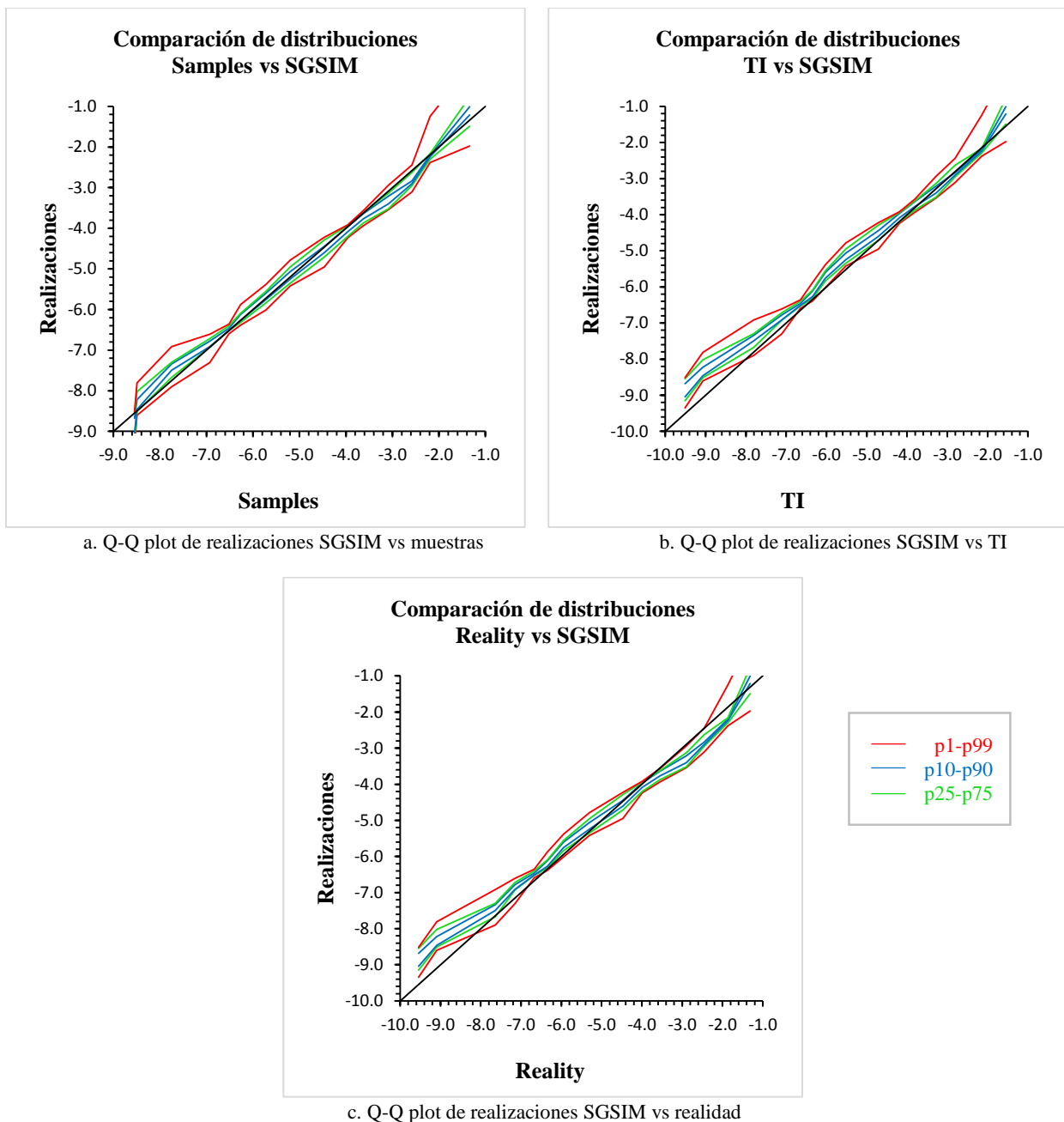
Como se puede observar, en términos generales se presenta una adecuada reproducción de las distribuciones por parte de las realizaciones de FILTERSIM (Figura 4-13), tanto en el caso de la imagen de entrenamiento, donde se presentan unas fluctuaciones en los valores inferiores; así como respecto a los datos originales; aunque en este último caso se presenta mayor oscilación en la parte inferior de la distribución, con una media levemente inferior (comportamiento similar al observado al comparar distribuciones entre las muestras y la imagen de entrenamiento). Se observa igualmente

un comportamiento entorno a la diagonal entre las distribuciones de las realizaciones con la imagen exhaustiva; lo que podría indicar una adecuada predicción del modelo de lo que sería la realidad, al menos desde el punto de la distribución de los datos.



**Figura 4-13:** Q-Q plot para análisis de distribuciones FILTERSIM

A continuación se presentan los Q-Q plot para el algoritmo SGSIM (Figura 4-14). En este caso se efectúa la comparación solamente respecto a las muestras y la realidad puesto que la imagen de entrenamiento no interviene en la obtención de las realizaciones. En el anexo E se presentan igualmente comparaciones para algunas realizaciones específicas de SGSIM.



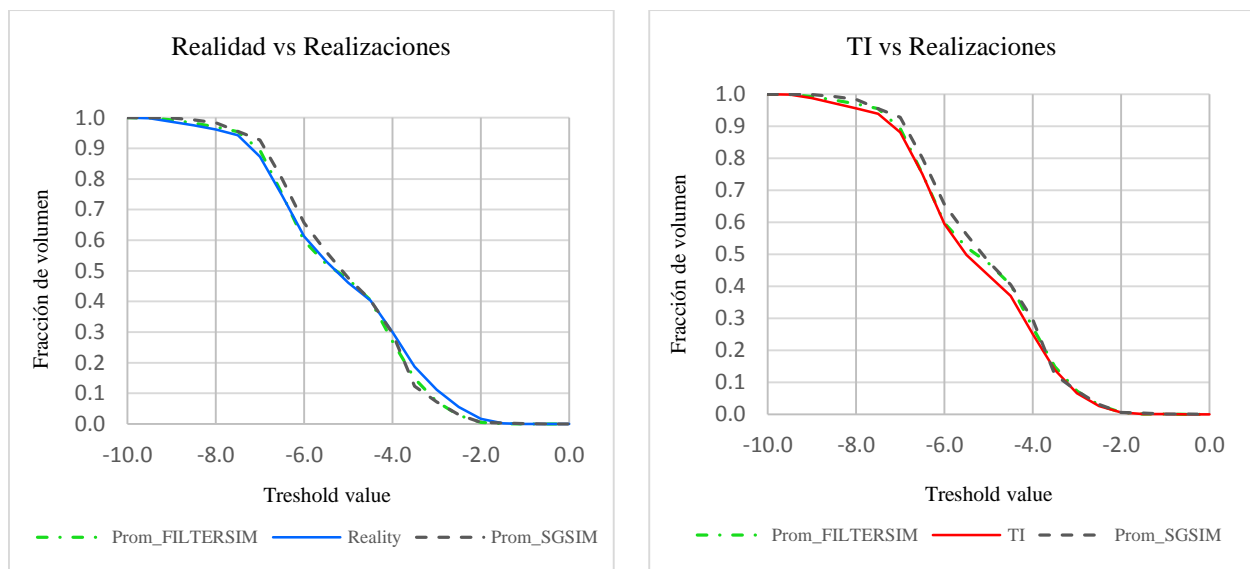
**Figura 4-14:** Q-Q plot para análisis de distribuciones SGSIM

En la figura anterior se puede apreciar igualmente distribuciones alrededor de la diagonal (equivalencia a igualdad entre las distribuciones) para el conjunto de datos pertenecientes a las muestras y las realizaciones analizadas obtenidas con SGSIM. Adicionalmente, si se revisan los QQ plot de la TI contra las realizaciones se ven igualmente ciertas diferencias entre las medias de las distribuciones, lo cual podría explicarse por la menor media que presentaba igualmente la TI respecto a la imagen exhaustiva y las muestras, que son las que trata de reproducir el algoritmo SGSIM. Por esto, algunos autores han propuesto efectuar test estadísticos que permitan verificar de manera más acertada si las simulaciones obtenidas fluctúan en rangos considerados aceptables; teniendo en cuenta eso sí que dichos test son bastante exigentes en cuanto a los supuestos para implementarlos y poder utilizar sus resultados de manera confiable.

Para este caso particular los datos no cumplen los supuestos exigidos por los test, pero se presentan a manera de ejercicio académico para mostrar las inconsistencias que podrían observarse al no tener en cuenta dichos supuestos y aplicarlos sin considerar sus supuestos fundamentales.

En cuanto a la comparación de las distribuciones de la imagen exhaustiva y las realizaciones, se observan claramente algunas colas en los cuantiles inferiores y superiores; así como un leve sesgo; situación observada igualmente al comparar la distribución original de las muestras y la imagen exhaustiva; lo cual muestra que el algoritmo conserva la forma de las distribuciones al generar las realizaciones.

Otro análisis que permitiría examinar el comportamiento de las distribuciones sería mediante el uso de curvas de selectividad, lo cual incluye el análisis del comportamiento de la distribución ante la aplicación de umbrales, que para el caso minero correspondería a una ley de corte. En la figura 4-15 se puede efectuar una comparación entre los resultados de las realizaciones de FILTERSIM y SGSIM respecto a la imagen exhaustiva y la imagen de entrenamiento.



**Figura 4-15:** Curvas de selectividad (Fracción de volumen vs umbral)

Como se puede apreciar, los resultados obtenidos para el promedio de las realizaciones de FILTERSIM (línea punteada verde) se ajusten de mejor manera a ambas curvas (imagen exhaustiva: línea continua azul y TI: línea continua roja). Esto se puede comprobar si se calcula el error medio cuadrático para comparar los resultados obtenidos de las realizaciones vs la imagen exhaustiva obteniéndose un EMC de  $2.84 \times 10^{-4}$  para FILTERSIM y un EMC de  $7.92 \times 10^{-4}$  para SGSIM. Si el análisis se repite respecto a la TI los EMC para FILTERSIM y SGSIM fueron de  $2.18 \times 10^{-4}$  y  $9.32 \times 10^{-4}$  respectivamente.

- **Test estadísticos**

A continuación se presentan los resultados de los test estadísticos efectuados para este análisis. Ellos corresponden a los test de reproducción de la media y la varianza; para el cual se aplica el test de Student, descrito en el marco teórico. Estos test fueron efectuados mediante algoritmos

escritos en código MatLab desarrollados por Emery (2008)<sup>3</sup> y los resultados pueden observarse en el anexo F.

En el caso de los resultados obtenidos con FILTERSIM, se efectúa el análisis para verificar la reproducción de la media de los datos originales (muestras); así como para la media de la imagen de entrenamiento. Los resultados se pueden observar en la Tabla 4.2.

**Tabla 4.2:** Resultados del test de reproducción de media para FILTERSIM

Test de media	FILTERSIM vs. muestras	FILTERSIM vs. TI
Media a comparar	-5.10669	-5.31489
Estadístico Student	22.9577	21.4271
Número grados de libertad	99	99
Nivel de significancia	0.05	0.05
Valor crítico	1.9842	1.9842
Resultado del test	Hipótesis nula rechazada.	Hipótesis nula rechazada.

La hipótesis nula plantea que existe una reproducción exacta del modelo de campo aleatorio. En dicho caso, si se rechaza la hipótesis nula significaría que las fluctuaciones observadas en las simulaciones no generan un modelo con una reproducción de la media, como ocurre con los resultados de las realizaciones de FILTERSIM. Sin embargo, este análisis puede no ser excluyente de la validez del modelo que se pretende modelar, tal como lo plantea Emery (2008); puesto que podría ser aún considerado válido según los objetivos planteados al momento de considerar el fenómeno de interés. Para este caso en particular, los resultados del test estadístico sugieren que el algoritmo FILTERSIM no reproduce exactamente el modelo del campo aleatorio. En la Tabla 4.3 se presentan los resultados para SGSIM.

**Tabla 4.3:** Resultados del test de reproducción de media para SGSIM

Test de media	SGSIM vs. muestras
Media a comparar	-5.10669
Estadístico Student	2.5702
Número grados de libertad	99
Nivel de significancia	0.05
Valor crítico	2.2622
Resultado del test	Hipótesis nula rechazada

De manera similar a lo que ocurre con el algoritmo FILTERSIM, se rechaza la hipótesis nula y por lo tanto no se reproduce exactamente el modelo del campo aleatorio. De forma similar se efectúa un análisis para verificar la reproducción de la varianza. Los resultados obtenidos, tanto para FILTERSIM, como para SGSIM son presentados en la Tabla 4.4.

**Tabla 4.4:** Resultados del test de reproducción de varianza para FILTERSIM/SGSIM

Test de varianza	FILTERSIM vs Muestras	FILTERSIM vs. TI	SGSIM vs. Muestras
Varianza	2.52742	2.51415	2.52742
Estadístico Student	7.0759	5.783	7.3818
Número grados de libertad	99	99	99
Nivel de significancia	0.05	0.05	0.05
Valor crítico	1.9842	1.9842	1.9842
Resultado del test	Hipótesis nula rechazada	Hipótesis nula rechazada	Hipótesis nula rechazada

<sup>3</sup> Código MatLab disponible en <http://www.iamg.org/images/File/documents/oldftp/VOL34/v34-11-18.zip>

Como se puede observar, en ambos casos el test igualmente rechaza la hipótesis nula. Aquí es importante considerar que si bien se llevaron a cabo estos test, los datos empleados para efectuarlos no cumplen los supuestos requeridos para efectuar estos. Por ejemplo, de acuerdo con Emery (2008) los test asumen que el dominio de simulación es mucho más grande que el dominio de integración y que por tanto las estadísticas experimentales calculadas sobre una realización son casi normalmente distribuidas; supuesto que no se cumple en este caso en particular, tal como se observó anteriormente al comparar las distribuciones.

Adicionalmente, se asume que el campo aleatorio que se quiere reproducir con las realizaciones obtenidas de las simulaciones cumplen los supuestos de Estacionariedad, de la existencia de sus momentos al menos hasta orden cuatro y que presenta la propiedad de fuerte mezcla, es decir, que el promedio espacial calculado sobre un dominio grande tiene una distribución normal. En este caso particular tampoco se cumple el supuesto de Estacionariedad del campo aleatorio puesto que como se observó la variable presenta claros comportamientos no estacionarios. Finalmente, al tratarse de una simulación condicional se presenta cierto grado de correlación entre las realizaciones obtenidas mediante las simulaciones, lo cual viola el supuesto de independencia.

De acuerdo a lo expuesto anteriormente podría decirse que no es de mucha ayuda este tipo de validación en este caso particular, dado que la variable a simular presenta patrones más complejos que no garantizan los comportamientos o asunciones requeridos para efectuar este tipo de test, cómo son Estacionariedad, que claramente no se observa ni en la imagen de entrenamiento, ni en la imagen exhaustiva de donde fueron obtenidos los datos. Esto representa igualmente desafíos importantes para el algoritmo SGSIM que requiere este tipo de asunciones igualmente para poder obtener resultados adecuados.

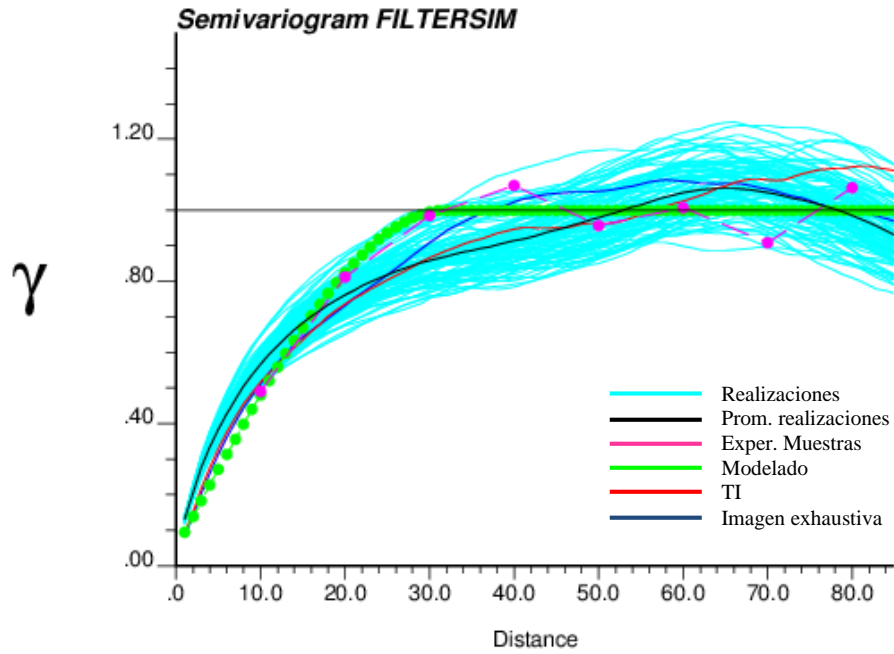
Es por eso que las modelaciones de fenómenos como los encontrados comúnmente en la naturaleza donde se observan comportamientos no estacionarios y asociados principalmente a las geociencias (en temas como la geología o hidrogeología, de vital importancia en la industria minera); deberían incluirse métricas de orden superior que permitan determinar o no la validez de dichos modelos en función de los objetivos planteados; teniendo en cuenta, tal como lo plantean Pyrcz y Deutsch (2014) que dichas validaciones presentan mayores niveles de complejidad.

#### **4.1.1.3. Nivel 3. Validación geoestadística**

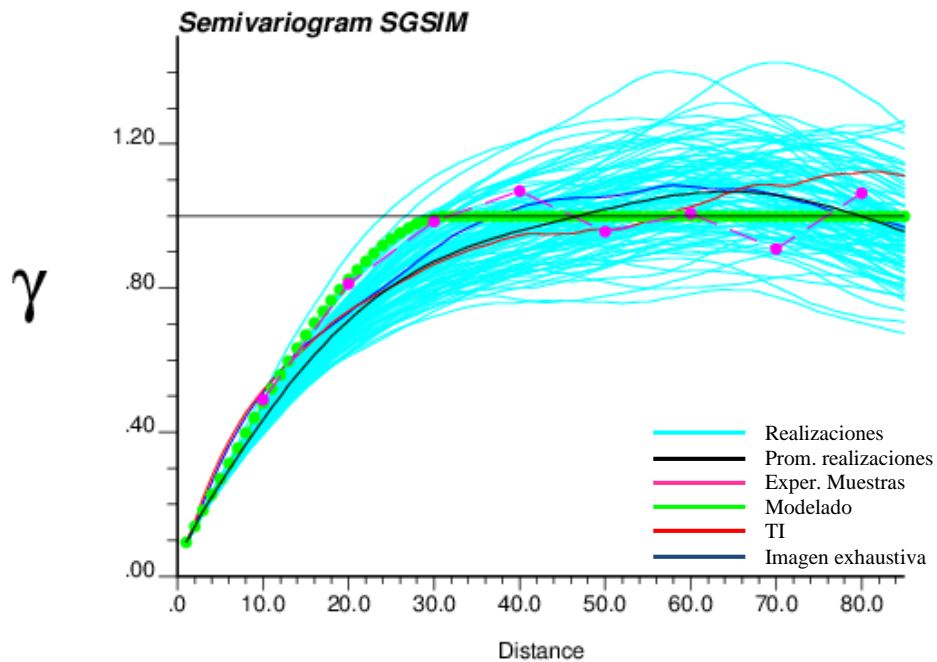
En este nivel se efectúan validaciones respecto a estadísticas de más de 1 punto. Es por eso que se validarán acá estadísticas de 2 puntos (variogramas) y multipunto.

- **Variograma**

Como parte de las validaciones de estadísticas de 2 puntos, se procede a evaluar la reproducción de los variogramas. Esta reproducción se efectúa sobre datos gaussianos, de acuerdo a lo planteado por Pyrcz y Deutsch (2014) quienes sugieren efectuar la validación en el campo gaussiano si las simulaciones son gaussianas (como es el caso de SGSIM); motivo por el cual se construyeron los variogramas experimentales de las realizaciones con datos transformados a gaussianos de media 0 y varianza 1 para el caso del FILTERSIM. Para el caso del SGSIM, se construyeron los variogramas de las realizaciones antes de efectuar la transformación de vuelta de datos gaussianos simulados a datos originales. De esta forma es posible comparar el variograma modelado que fue empleado en la simulación secuencial gaussiana a partir de las muestras convertidas en datos gaussianos. Los resultados obtenidos pueden observarse en las Figura 4-16 y Figura 4-17.



**Figura 4-16:** Reproducción de variogramas para FILTERSIM.



**Figura 4-17:** Reproducción de variogramas para SGSIM.

Como se aprecia en las figuras anteriores, visualmente se observa que los variogramas experimentales de las realizaciones fluctúan entorno al variograma modelado, aun cuando se observa una discrepancia mayor entorno a distancias entre 15 y 30 metros para el caso del SGSIM y de los 0 a los 30 metros en el FILTERSIM. En este último algoritmo, los variogramas de las realizaciones presentan un mejor ajuste con el variograma de la imagen de entrenamiento hasta alrededor de una distancia de 40 metros, observándose una desviación importante a distancias mayores a 65 metros. Por el contrario, los variogramas de las realizaciones de SGSIM parecieran presentar visualmente un mejor ajuste respecto al variograma modelado y empleado en las simulaciones secuenciales gaussianas (aunque con diferencias para distancias menores a 25 m).



Para mejorar este análisis, se procedió a calcular el error medio cuadrático (EMC) entre el variograma promedio de las realizaciones y los variogramas de la imagen exhaustiva, la imagen de entrenamiento y el variograma modelado para el SGSIM (para todos los pasos). Los resultados obtenidos se pueden apreciar en la tabla 5.

**Tabla 4.5:** Error medio cuadrático entre variogramas (FILTERSIM/SGSIM) para 85 m

VARIOGRAMA	Imagen Exhaustiva (Realidad)	Imagen de entrenamiento (TI)	Modelado (SGSIM)
Media SGSIM	0.00156	0.00341	0.00450
Media FILTERSIM	0.00325	0.00337	0.00533

Tal como se observa en la tabla anterior, el menor EMC está asociado al variograma de la imagen exhaustiva para el caso de SGSIM y para FILTERSIM la imagen de entrenamiento; lo cual podría interpretarse como un buen desempeño del algoritmo para reproducir la realidad y la imagen de entrenamiento respectivamente. Como era de esperarse FILTERSIM presenta un mejor desempeño en la reproducción del variograma de la imagen de entrenamiento y la misma situación ocurre con el variograma modelado para el caso del SGSIM dado que emplea este en la generación de las realizaciones.

Si este mismo análisis se repite para una distancia inferior (80 m), se obtienen EMC muy similares para la imagen exhaustiva y para el variograma modelado; sin embargo, para la imagen de entrenamiento el EMC se reduce sustancialmente a 0.0019 para el caso de FILTERSIM y a 0.0023 para el SGSIM; en cuyo caso ya se podría hablar de un mejor desempeño de FILTERSIM para reproducir el variograma de la imagen de entrenamiento respecto al SGSIM.

- Test estadístico para variogramas

Dadas las fluctuaciones presentes en los variogramas de las realizaciones, es posible efectuar un test estadístico para determinar si la media del vector de variograma para p pasos seleccionados es igual a la media del vector basado en el variograma modelado para las mismas p distancias seleccionadas como paso mediante el cálculo de un estadístico  $T^2$  de Hotelling<sup>4</sup> (los resultados pueden observarse en el anexo F); tal como fue descrito en el marco teórico. A continuación se presentan los resultados obtenidos para este test (ver Tabla 4.6). Es importante mencionar que este test exige una distribución normal de los datos; motivo por el cual se desarrolla el análisis con los datos gaussianos.

**Tabla 4.6:** Test de Hotelling para validación de variogramas FILTERSIM/SGSIM

Test de media y varianza	FILTERSIM	SGSIM
Número de lags	5	5
Lags	10 20 30 40 50	10 20 30 40 50
Estadístico $T^2$ Hotelling	1053.2385	1270.0729
Número grados de libertad	5 y 95	5,5
Nivel de significancia	0.05	0.05
Valor crítico	12.0375	12.0375
Resultado del test	Hipótesis nula es rechazada	Hipótesis nula es rechazada

<sup>4</sup> Código MatLab disponible en <http://www.iamg.org/images/File/documents/oldftp/VOL34/v34-11-18.zip>

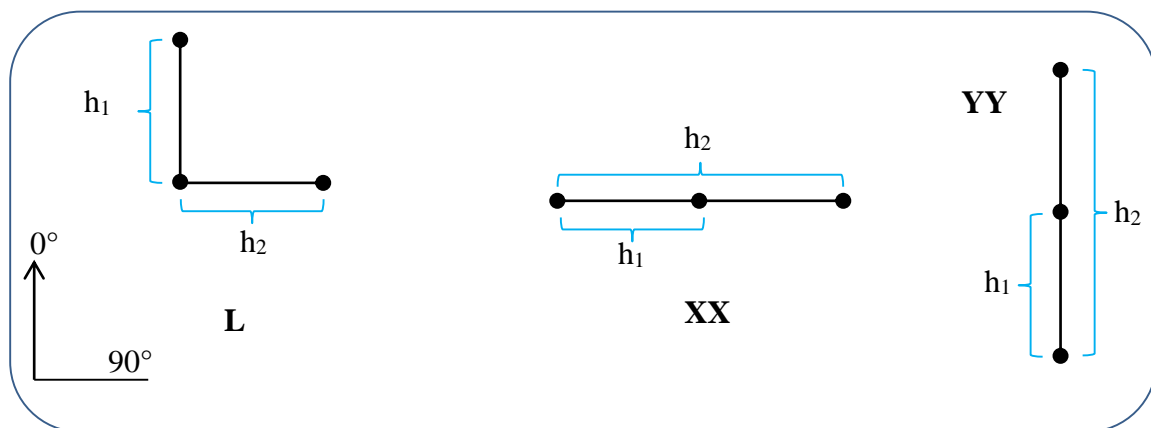
Como se puede observar, el test para ambos algoritmos, rechaza la hipótesis nula. Es importante entonces volver a retomar la validez de este test para los datos que se vienen analizando, dado los supuestos o asunciones como Estacionariedad intrínseca que exigen este test. Tal como se ha presentado con anterioridad, el comportamiento de la variable de interés, particularmente para este caso, carece de esta propiedad. En tal sentido, este test no debería ser aplicado para este caso en particular puesto que no cumple los supuestos exigidos. Se hace necesario entonces definir los objetivos que se pretenden obtener con la modelación y examinar otro tipo de métricas a emplear para determinar si las simulaciones reproducen o no las características de la variable de interés.

- **Cumulantes de orden superior**

Adicionalmente a las estadísticas de dos puntos, se obtuvieron estadísticas de orden superior mediante cumulantes espaciales. Para ello se empleó el software HOSC (High Order Spatial Cumulants) desarrollado por Dimitrakopoulos y Musthapa (2010)<sup>5</sup>, cuyos parámetros se presentan en el anexo G. Con este se obtuvieron cumulantes de orden 3 para tres configuraciones distintas:

- Configuración 1:  $\{0^\circ - 90^\circ\}$  → Denominada “L” (5 x 5 m)
- Configuración 2:  $\{90^\circ - 90^\circ\}$  → Denominada “XX” (5 x 10 m)
- Configuración 3:  $\{0^\circ - 0^\circ\}$  → Denominada “YY” (5 x 10 m)

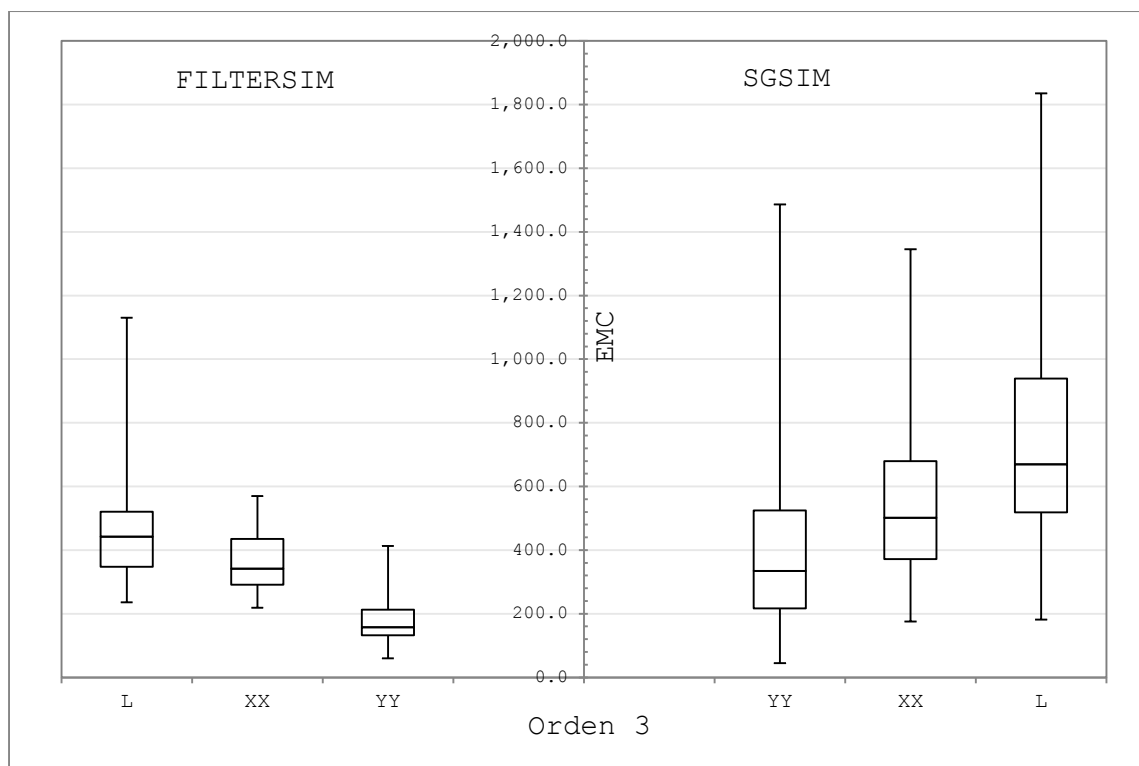
En la Figura 4-18 se presenta un esquema de estas configuraciones.



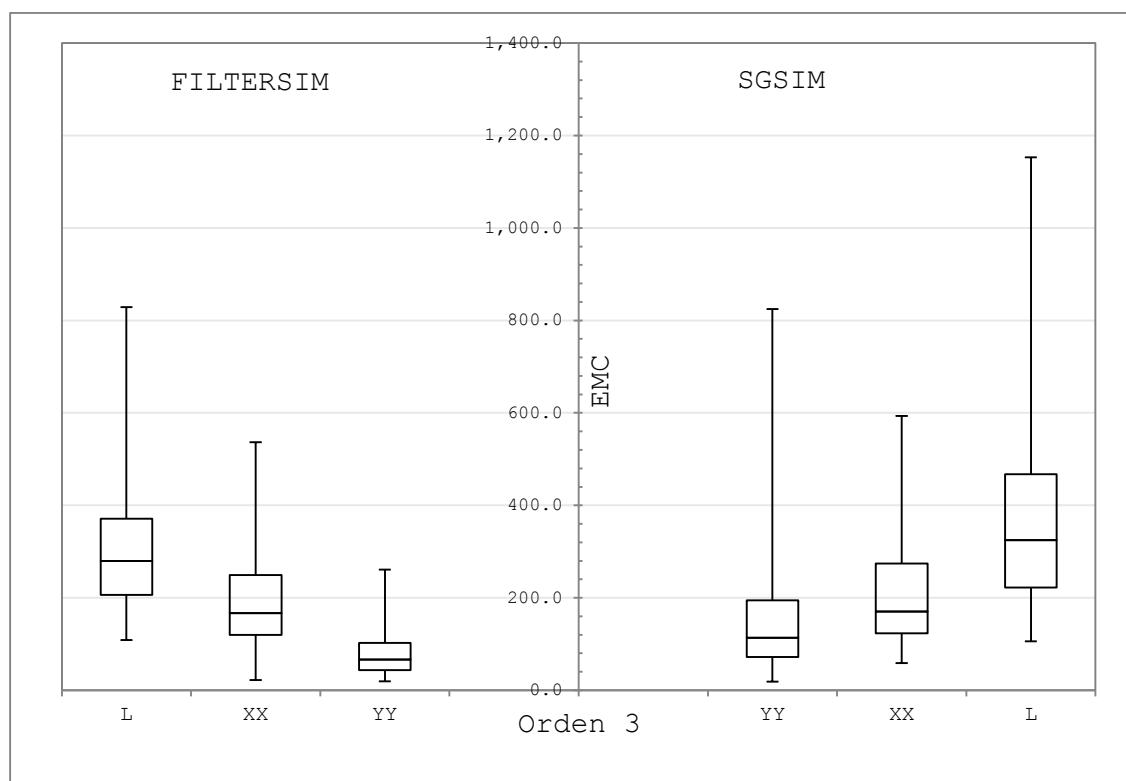
**Figura 4-18:** Esquema de configuración para cálculo de cumulantes de orden 3

De forma similar al trabajo desarrollado por De Iaco y Maggio (2011) se calculó el error medio cuadrático (EMC) entre cada una de las realizaciones y la imagen de entrenamiento y las realizaciones con la imagen exhaustiva. Los resultados obtenidos fueron graficados mediante box-plot que permiten comparar los dos tipos de algoritmos empleados. En la Figura 4-19 y Figura 4-20 se pueden observar los resultados obtenidos.

<sup>5</sup> Disponible en <http://www.iamg.org/documents/oldftp/VOL36/v36-03-06.zip>



**Figura 4-19:** Cumulantes de orden 3 (Realizaciones vs. TI).



**Figura 4-20:** Cumulantes de orden 3 (Realizaciones vs. Imagen exhaustiva).

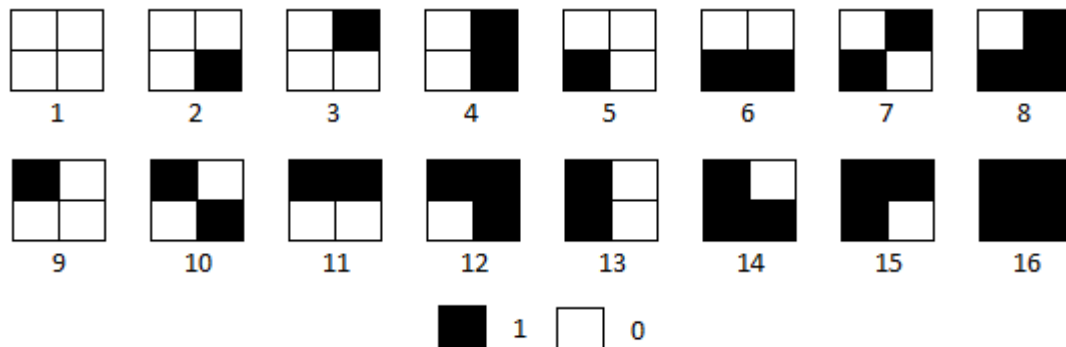
Como se puede observar en las figuras anteriores, se obtuvo un EMC inferior en el algoritmo FILTERSIM respecto al SGSIM. Estos mismos resultados fueron obtenidos en el trabajo desarrollado por De Iaco y Maggio (2011), con lo cual podría decirse que las realizaciones

obtenidas con FILTERSIM presentan un comportamiento más similar a lo observado en la imagen de entrenamiento que SGSIM. Esto sugiere que FILTERSIM captura comportamientos más complejos que los que podría reproducir SGSIM. Puede por tanto concluirse que los patrones de la distribución espacial capturados por cumulantes de orden 3 de la variable de interés, son mejor reproducidos (respecto a la imagen de entrenamiento) por las realizaciones de FILTERSIM, lo cual tiene sentido dado que el algoritmo FILTERSIM impone estadísticas de orden superior.

Al efectuar este mismo análisis para la imagen exhaustiva, se observa como mejora el desempeño del algoritmo SGSIM, toda vez que este emplea la información obtenida de esta imagen (muestras) como parámetro de partida y no emplea la imagen de entrenamiento. Sin embargo, las realizaciones de FILTERSIM parecen presentar un leve mejor desempeño respecto a SGSIM.

- **Función de densidad para múltiples puntos**

La función de densidad de múltiples puntos aplicada a la validación de las simulaciones, correspondió a determinar la frecuencia de ocurrencia de un patrón binario de 2x2 en cada una de las 100 simulaciones en ambos algoritmos (con lo cual se obtuvo un promedio de las frecuencias encontradas en las 100 simulaciones), al igual que en la imagen de entrenamiento y la imagen exhaustiva. Con esto se construyó la función de densidad de múltiples puntos planteada por Boisvert. et al (2007). En la Figura 4-21 se presentan los patrones definidos.



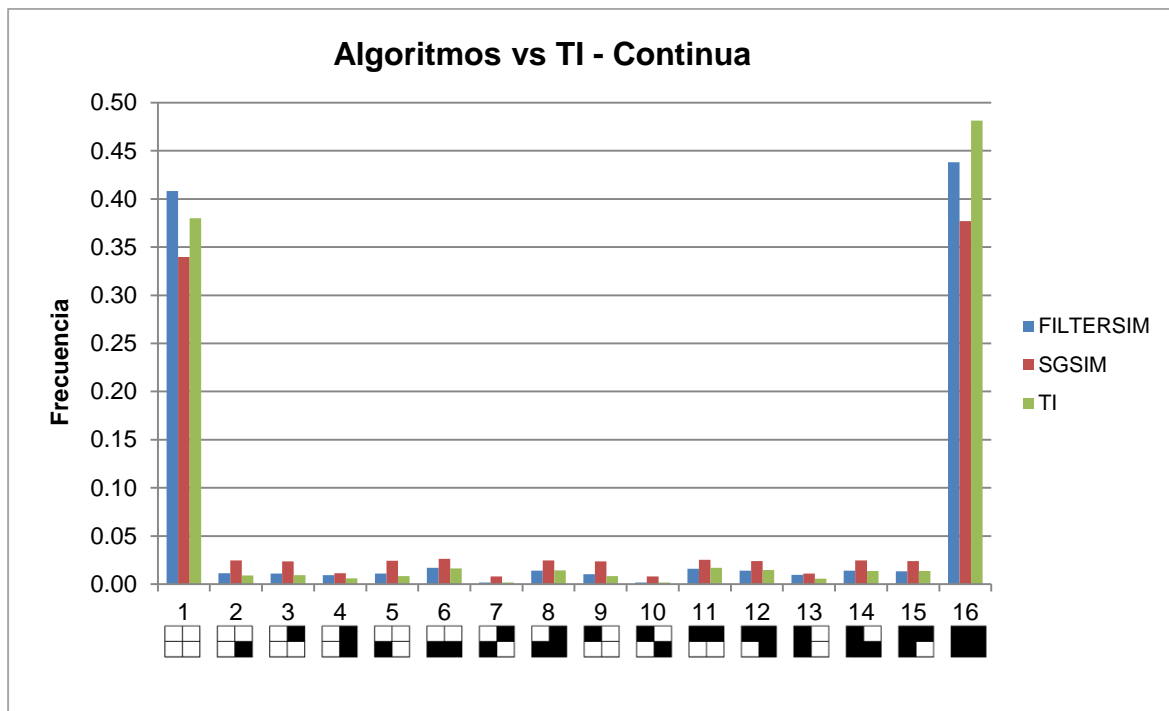
**Figura 4-21:** Patrones definidos para construcción de función de densidad de múltiples puntos

Pero para poder efectuar dicho análisis en una variable continua, se definió primero una variable indicador mediante un valor de corte. Para este caso particular, el valor de corte seleccionado fue la media de las muestras. La variable indicador será entonces:

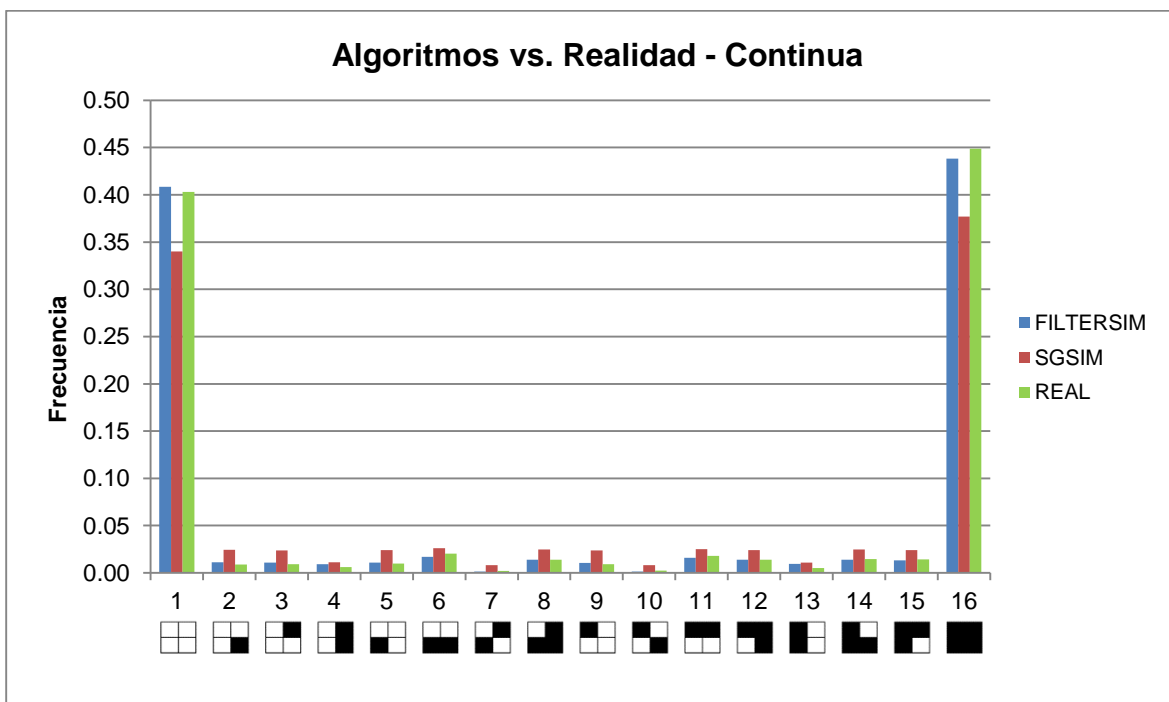
$$Variable\ Indicator = \begin{cases} 1 & Si\ u_i \leq -5.10669 \\ 0 & en\ caso\ contrario \end{cases} \quad (4.1)$$

Donde  $u_i$  corresponde al valor de la imagen (entiéndase realizaciones, imagen exhaustiva o imagen de entrenamiento) en el nodo  $i$ .

Para determinar la frecuencia de los patrones en cada una de las mallas de las realizaciones y las imágenes, se generó un algoritmo MatLab, cuyo código se presenta en el anexo H. En las figuras Figura 4-22 y Figura 4-23 se pueden observar las funciones de densidad de múltiples puntos, comparando los resultados del promedio de las 100 realizaciones para FILTERSIM y SGSIM, con la imagen de entrenamiento y la imagen exhaustiva (realidad).



**Figura 4-22:** Función de densidad de múltiples puntos – realizaciones vs. TI



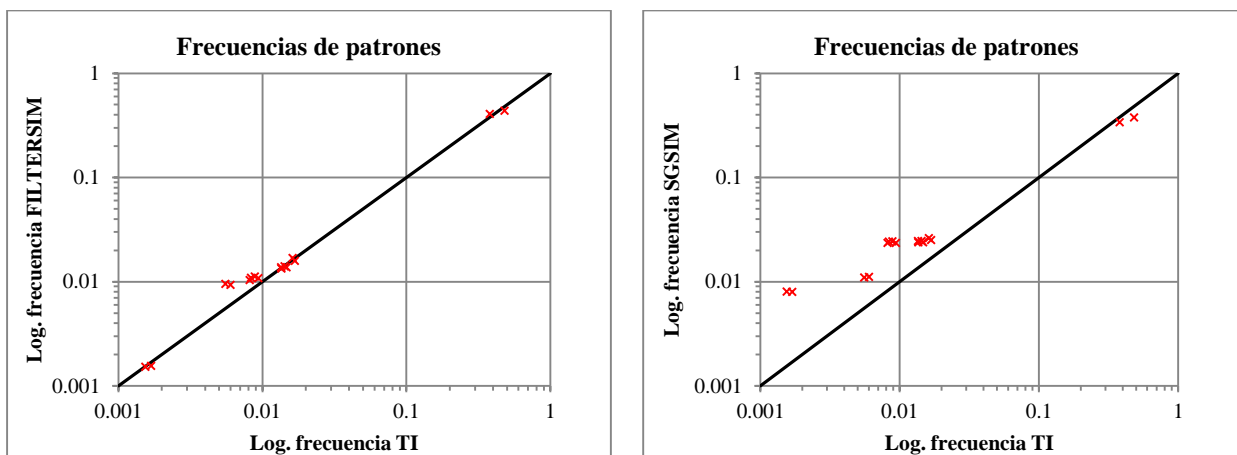
**Figura 4-23:** Función de densidad de múltiples puntos – realizaciones vs. realidad

Tal como se aprecia en las figuras anteriores, el algoritmo FILTERSIM presenta frecuencias de patrones más similares que las encontradas con los resultados del algoritmo SGSIM; tanto en el caso de la imagen de entrenamiento, como en la imagen exhaustiva.

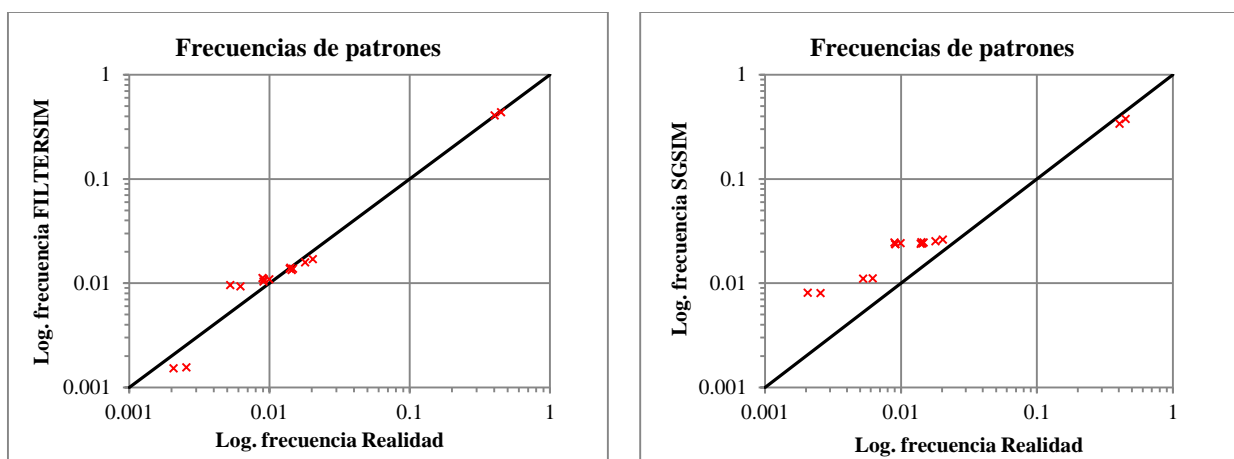
Es importante resaltar, que los resultados obtenidos al evaluar la función de densidad de múltiples puntos es consistente con los resultados obtenidos para los cumulantes de orden superior (de orden

3 calculados en este trabajo) y la comparación de variogramas; motivo por el cual se reafirma la necesidad de efectuar validaciones de diverso orden en las simulaciones a fin de determinar la consistencia del modelo y tener diferentes referencias o métricas de comparación a diferentes niveles.

Las frecuencias anteriores también se pueden analizar mediante un gráfico de dispersión, graficando las frecuencias de cada patrón entre la imagen de entrenamiento o imagen exhaustiva y las realizaciones de los algoritmos, tal como se observa a continuación.

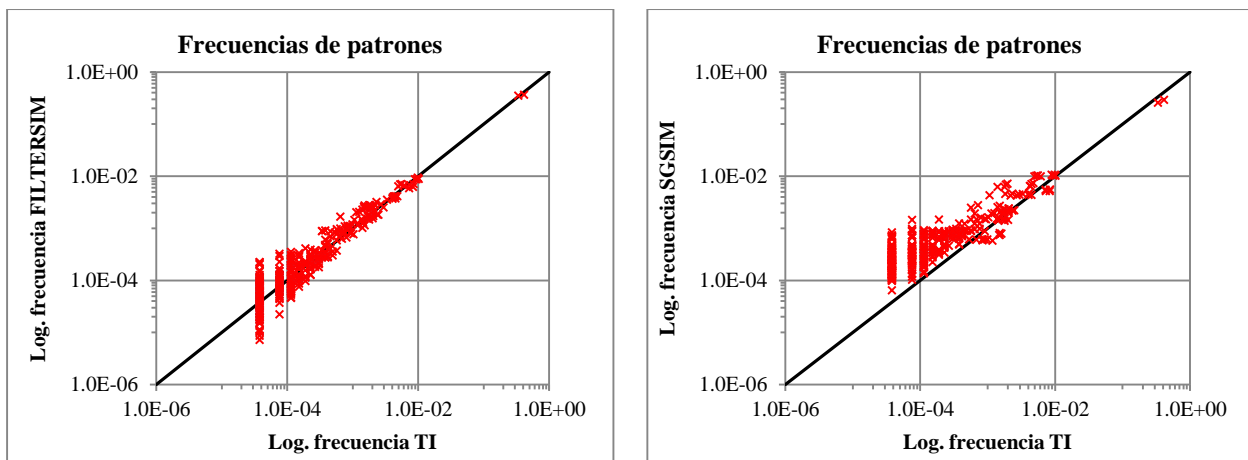


**Figura 4-24:** Gráfico de dispersión de frecuencias de patrones 2x2 TI vs. algoritmos FILTERSIM/SGSIM

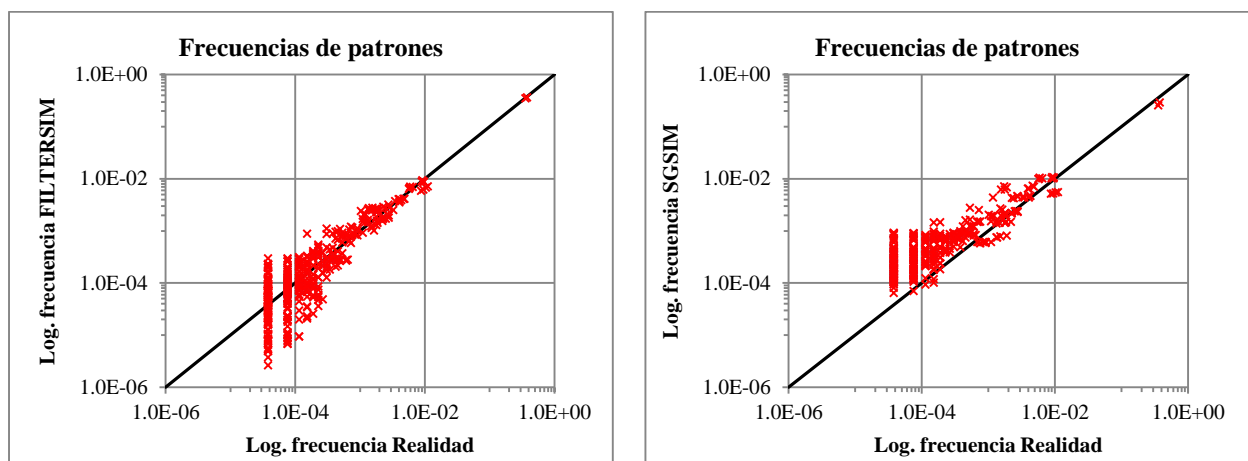


**Figura 4-25:** Gráfico de dispersión de frecuencias de patrones 2x2 realidad vs. algoritmos FILTERSIM/SGSIM

Como se puede ver, el algoritmo FILTERSIM reproduce de mejor forma los patrones evaluados, puesto que se ajustan mejor a la diagonal; tanto en el caso de TI como de la imagen exhaustiva. Al efectuar este mismo análisis pero para configuración 3x3, se obtienen 512 patrones diferentes. Dado que este número de posibles patrones es impráctico de visualizar en un gráfico de barras (tal como se presentó para una configuración 2x2), surge la relevancia del gráfico de dispersión de frecuencias. A continuación se presentan dichos gráficos para la configuración 3x3.



**Figura 4-26:** Gráfico de dispersión de frecuencias de patrones 3x3 TI vs. algoritmos FILTERSIM/SGSIM



**Figura 4-27:** Gráfico de dispersión de frecuencias de patrones 3x3 realidad vs. algoritmos FILTERSIM/SGSIM

Al observar las gráficas anteriores, se puede identificar un buen desempeño del algoritmo FILTERSIM respecto a la imagen de entrenamiento, al quedar representadas las frecuencias sobre la diagonal. Al comparar su desempeño frente a la imagen exhaustiva se ve una mayor dispersión de valores, pero sigue estando concentrado sobre la diagonal con menor coincidencia en las frecuencias menores y medias. En cambio el algoritmo SGSIM presenta un sesgo en las frecuencias. Es claro como las frecuencias pequeñas son menores a las frecuencias observadas en la imagen exhaustiva y la imagen de entrenamiento. Eso indicaría que al comparar patrones no se logran reproducir adecuadamente este tipo de métricas por parte de SGSIM para este caso particular.

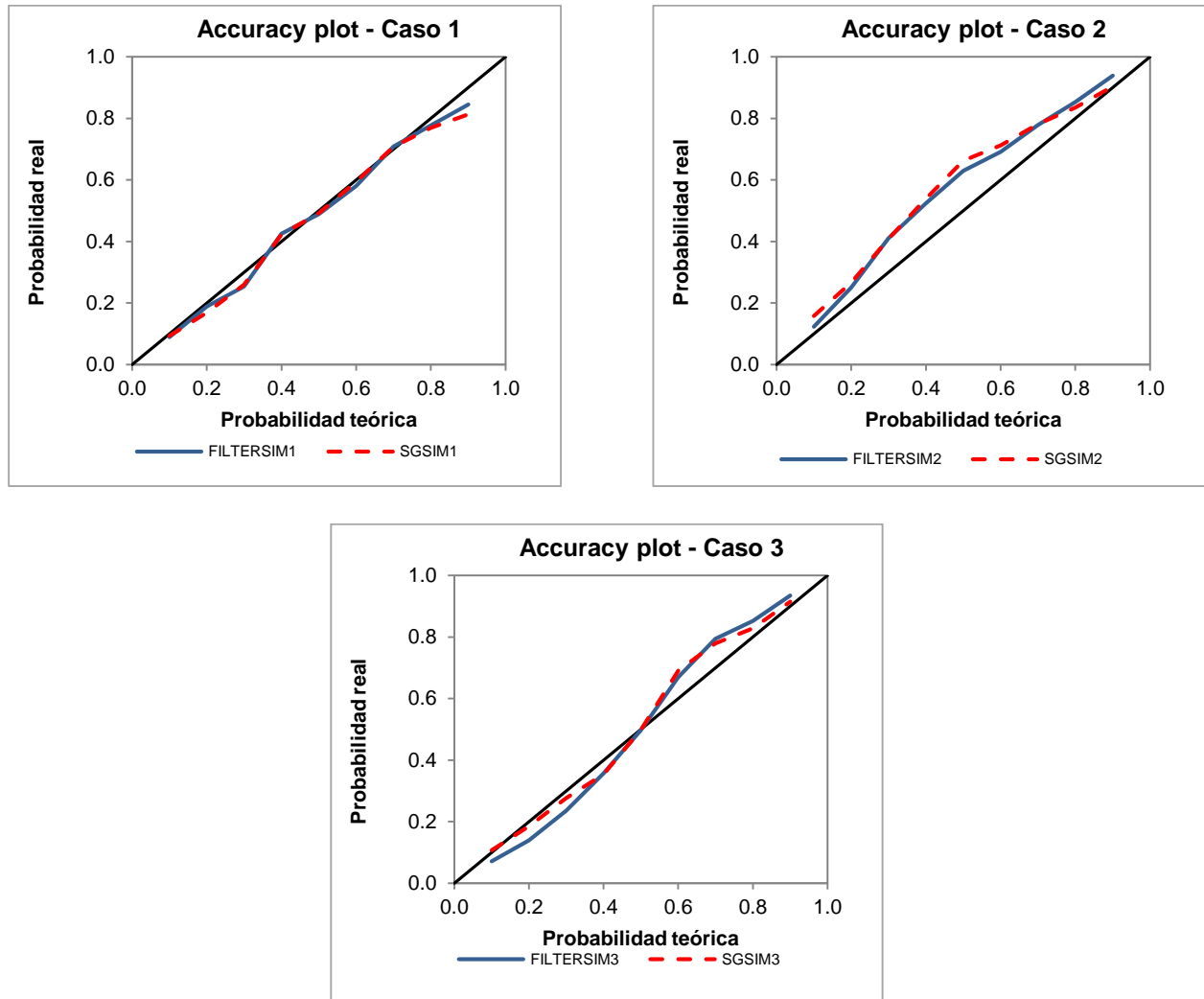
- **Validación de la distribución de incertidumbre**

Finalmente, y tal como se mencionó en el marco teórico; de manera similar a como se desarrollan técnicas de validación cruzada o Jack-knife para verificar la calidad de una estimación, en las simulaciones es necesario efectuar la comparación entre un valor conocido y distribuciones de incertidumbre generados por los modelos sobre múltiples realizaciones.

En tal sentido, se presentan a continuación los resultados obtenidos para este análisis mediante un gráfico denominado de precisión (accuracy plot). Para este caso en particular, se seleccionaron aleatoriamente otras 135 muestras dentro de la imagen exhaustiva (realidad) diferentes a las 135

muestras o datos empleados en las simulaciones. Con estas muestras se determinaron las probabilidades reales para 9 probabilidades teóricas (definidas como los deciles 1 al 9). Las probabilidades reales se obtuvieron mediante un algoritmo programado en MatLab, cuyo código se presenta en el anexo I.

En figura Figura 4-28 se presentan los accuracy plot para FILTERSIM y SGSIM.

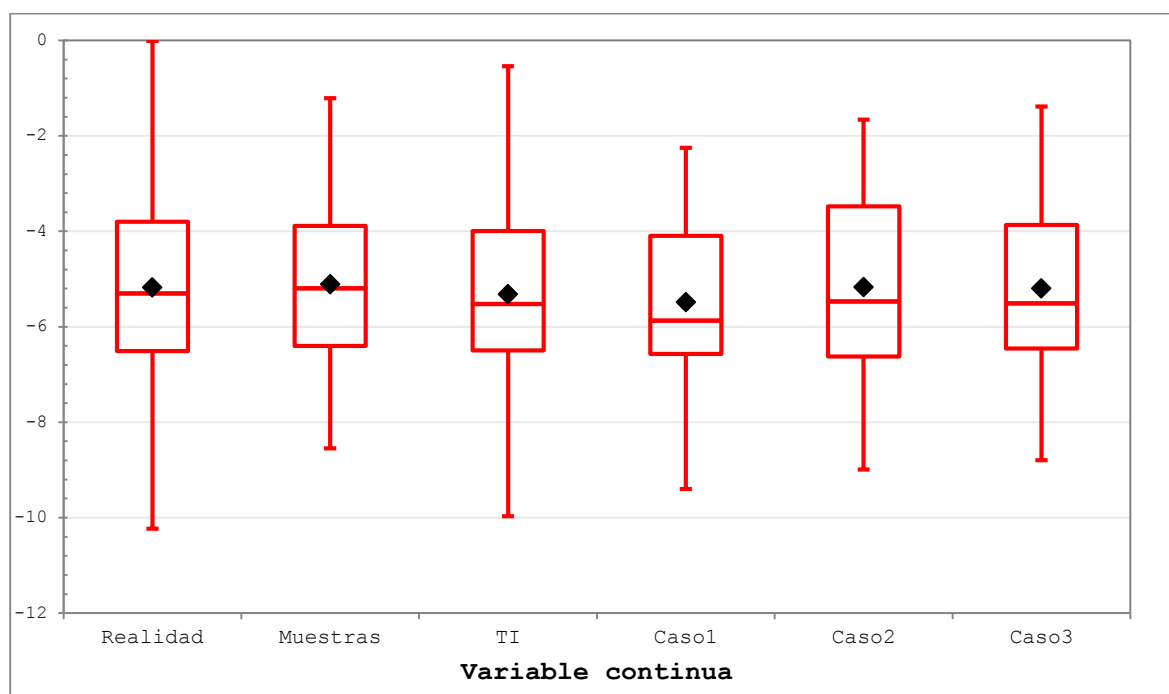


**Figura 4-28:** Accuracy plot realizaciones FILTERSIM/SGSIM para diferentes muestreos

Como se puede ver, los resultados del accuracy plot para los 3 casos son bien diferentes. Se observa que el caso 1 presenta un mejor ajuste a la diagonal; sin embargo, estando levemente por debajo de ésta en buena parte del rango analizado y al comparar las distribuciones del caso 1 con el de las muestras originales, se observa como las del caso 1 presentan una distribución levemente desplazada hacia valores inferiores (ver Figura 4-29). Por su parte, en el caso 2, se ve como las probabilidades reales estuvieron por encima de las probabilidades teóricas y si se observa la distribución del conjunto de muestras de este caso se aprecia como el rango intercuartil es más amplio respecto al conjunto de muestras originales (ver Figura 4-29). Finalmente, el caso 3 oscila entre menores probabilidades reales (cuando la probabilidad teórica es inferior al 50%), muy similar cuando es del 50% y probabilidades superiores cuando las teóricas son mayores al 50%; que al comparar con la distribución del conjunto de datos del caso 3 se observa una rango



intercuartil similar al del conjunto de muestras originales, pero con un menor rango entre primer cuartil y mediana y por ende un mayor rango entre mediana y tercer cuartil (ver Figura 4-29).

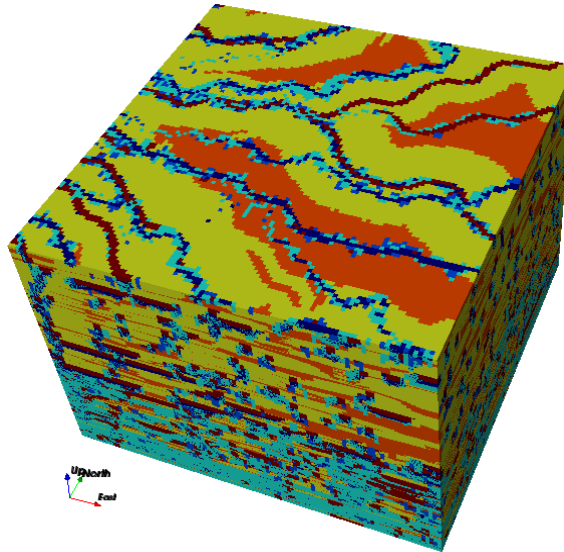


**Figura 4-29:** Box-plot datos originales vs muestreos alternativos

# CAPÍTULO 5. CASO PRÁCTICO VARIABLE CATEGÓRICA

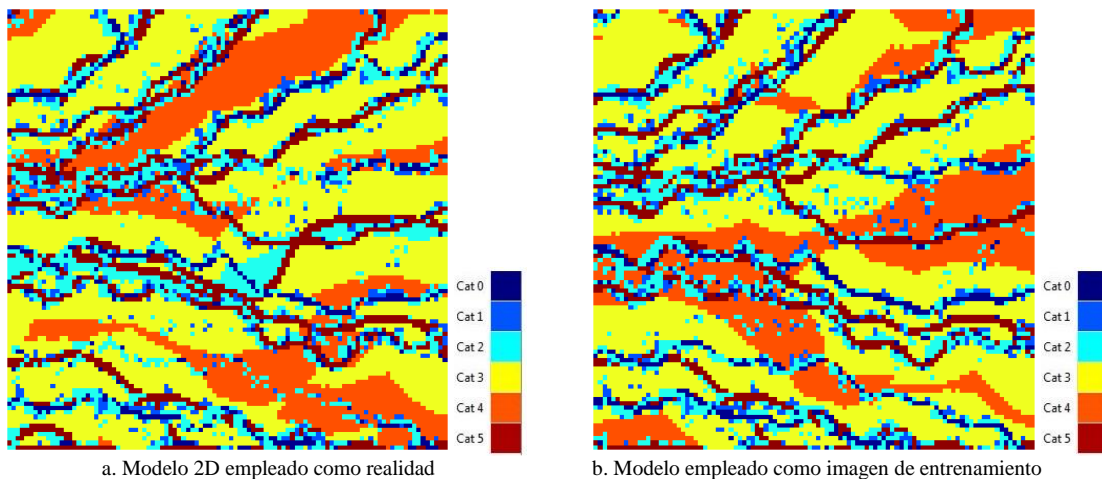
## 5.1. MODELO Y ALGORITMOS EMPLEADOS

De forma similar al caso de una variable continua, se efectuó una simulación de una variable categórica obteniendo datos desde un modelo de un acuífero con formaciones aluviales realísticas (estudio desarrollado por López (2003); Chugunova y Hu (2006)). Este modelo está compuesto por 1385 capas de 1 metro de espesor y un área transversal 10201 m<sup>2</sup> (101 m x 101 m). El modelo completo 3D puede apreciarse en la Figura 5-1.

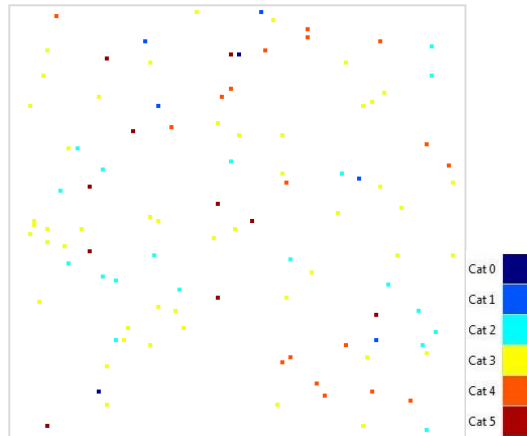


**Figura 5-1:** Modelo categórico 3D del acuífero.

De este modelo se seleccionó un nivel el cual fue considerado como imagen de entrenamiento (nivel 600) y un nivel levemente inferior (nivel 595), el cual fue empleado como imagen exhaustiva, asumida como realidad. De este último se seleccionaron 102 muestras; las cuales son empleadas para efectuar las posteriores simulaciones. En la Figura 5-2 se presentan la imagen de la realidad, la imagen de entrenamiento y las muestras seleccionadas aleatoriamente.



**Figura 5-2:** Imágenes extraídas del modelo categórico y empleadas para las simulaciones

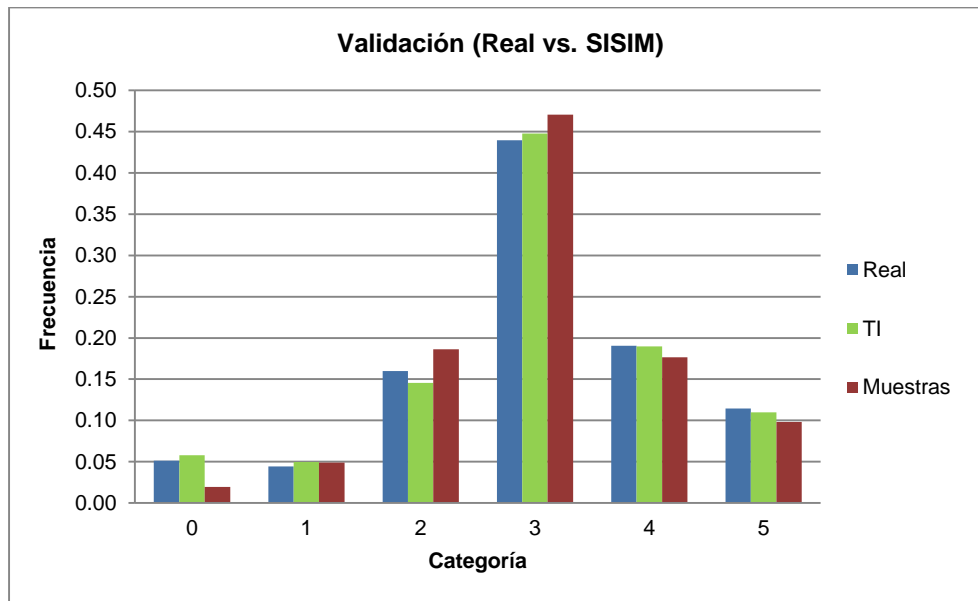


**Figura 5-3:** 102 muestras extraídas aleatoriamente del modelo de la realidad

Las estadísticas descriptivas de los datos empleados (imagen de entrenamiento, imagen exhaustiva – realidad) y las muestras se presentan en la Tabla 5.1. En la Figura 5-4 se presentan los histogramas con las frecuencias para cada categoría que describen la distribución de estos conjuntos de datos.

**Tabla 5.1:** Proporciones de categorías variable categórica

Categorías	Realidad		Muestras		TI	
	Datos	Fracción	Datos	Fracción	Datos	Fracción
	27000	1.0	102	1.0	27000	1.0
<b>Categoría 0</b>	525	0.0515	2	0.0196	591	0.0579
<b>Categoría 1</b>	451	0.0442	5	0.0490	507	0.0497
<b>Categoría 2</b>	1629	0.1597	19	0.1863	1484	0.1455
<b>Categoría 3</b>	4485	0.4397	48	0.4706	4565	0.4475
<b>Categoría 4</b>	1943	0.1905	18	0.1765	1934	0.1896
<b>Categoría 5</b>	1168	0.1145	10	0.0980	1120	0.1098



**Figura 5-4:** Histogramas con frecuencias de categorías

Como se puede ver en la gráfica y tabla anterior, existe una diferencia evidente en la proporción de datos en la categoría 0 para las muestras seleccionadas, con menos de la mitad de la proporción observada en las demás imágenes (imagen de entrenamiento a imagen exhaustiva - realidad) y

adicionalmente una mayor proporción en la categoría 3 (categoría más abundante en el modelo); lo cual podría afectar los resultados de las simulaciones. También se observan diferencias en la categoría 2. Sin embargo, esto podría corresponder a un fenómeno común en una modelación, cuando se realiza un muestreo selectivo, dado por ejemplo el no interés en determinar cierta categoría en una modelación geológica, porque no es portadora de un elemento de interés y por ende se encuentre menos muestreada.

Respecto a la imagen de entrenamiento y la imagen exhaustiva, se observa en términos generales, cierta correspondencia entre la imagen de entrenamiento y la imagen exhaustiva; sin ser exactamente iguales lo cual ocurrirá probablemente al emplear una imagen de entrenamiento como modelo a replicar.

Si bien en este caso en particular, dado que se efectuó un muestreo aleatorio a partir de la imagen exhaustiva (que generó cierto sesgo), no se tienen en cuenta proporciones desagrupadas; es importante considerar que en los casos donde se desarrollen muestreos preferenciales o dirigidos se debería efectuar la técnica de desagrupamiento para obtener estadísticas representativas de los datos, las cuales deberían ser empleadas para efectuar las validaciones respectivas.

- **Algoritmos a emplear en las simulaciones**

Para el caso de la variable categórica, se seleccionaron como algoritmos de simulación SNESIM y SISIM, cuya breve descripción se presentó en el capítulo 2 de este informe. Ambos algoritmos están disponibles en el software SGeMS de la Universidad de Stanford<sup>6</sup>. Los parámetros empleados para efectuar las simulaciones de cada uno de estos algoritmos puede verse en el anexo A.

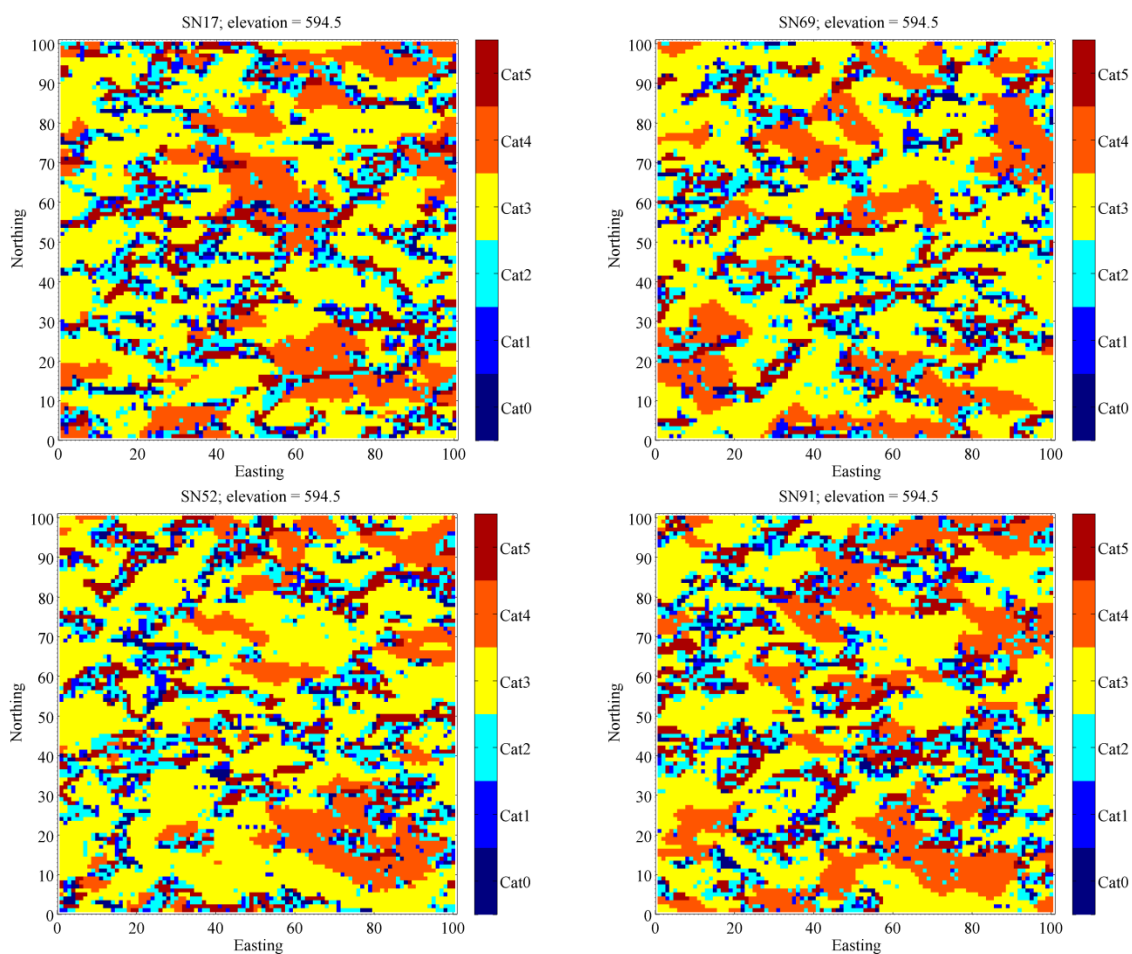
En el caso del algoritmo SISIM se obtuvo una variable indicador para cada una de las categorías encontradas en los datos; para cada una de las cuales se modeló un variograma independiente basado en la información de la imagen exhaustiva, cuyos modelos se pueden observar en el Anexo J.

La decisión de emplear la imagen exhaustiva en vez de los datos seleccionados, se tomó para poder efectuar un mejor cálculo de los variogramas experimentales de la variable indicador, dada la baja representatividad de la categoría 0 (la cual en la muestra seleccionada aleatoriamente, solo presentaba 2 registros en esta categoría lo cual dificultaba en gran medida la construcción de un variograma mínimamente representativo del comportamiento para esta categoría). De igual manera, tratando de lograr un mayor equilibrio entre los dos algoritmos empleados en las simulaciones al momento de efectuar comparaciones.

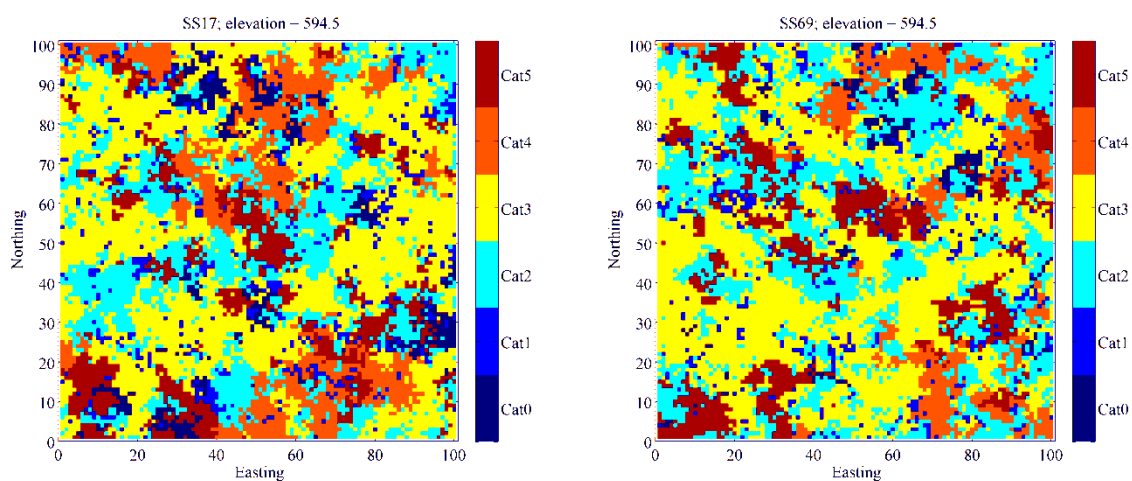
Se generaron 100 realizaciones para poner en práctica las diferentes métricas de validación. En las siguientes figuras se presentan algunas de las realizaciones obtenidas para estos algoritmos.

---

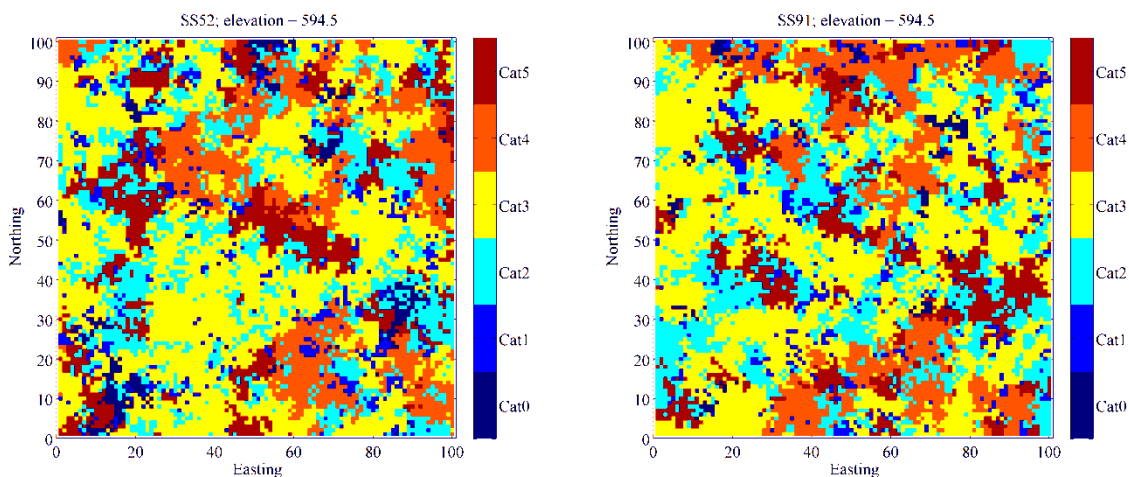
<sup>6</sup> Disponible en <http://sgems.sourceforge.net/>



**Figura 5-5:** Ejemplos de realizaciones obtenidas para SNESIM (SN)



**Figura 5-6:** Ejemplos de realizaciones 17 y 69 obtenidas para SISIM (SS)

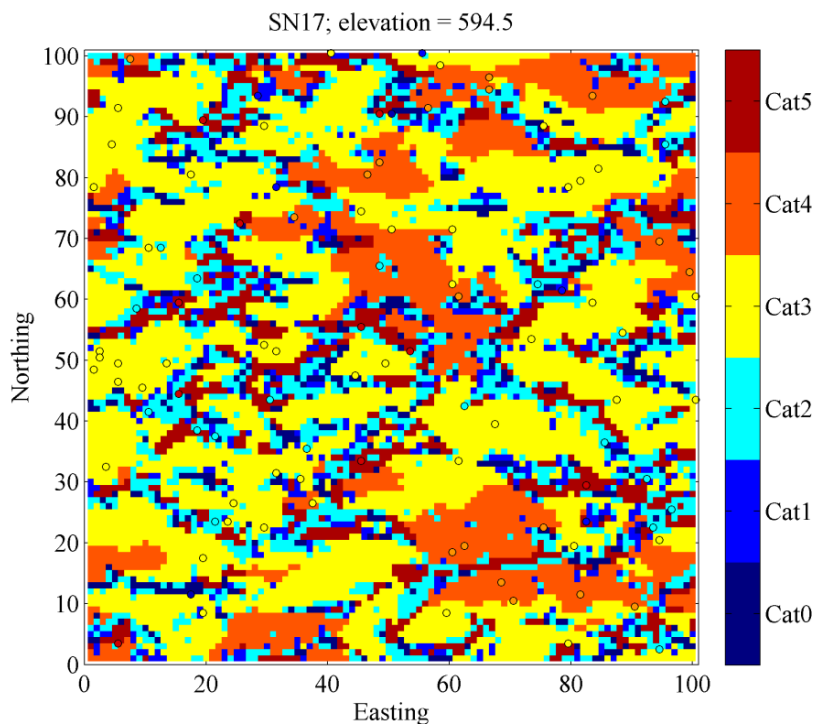


**Figura 5-7:** Ejemplo de realizaciones 52 y 91 obtenidas con SISIM (SS)

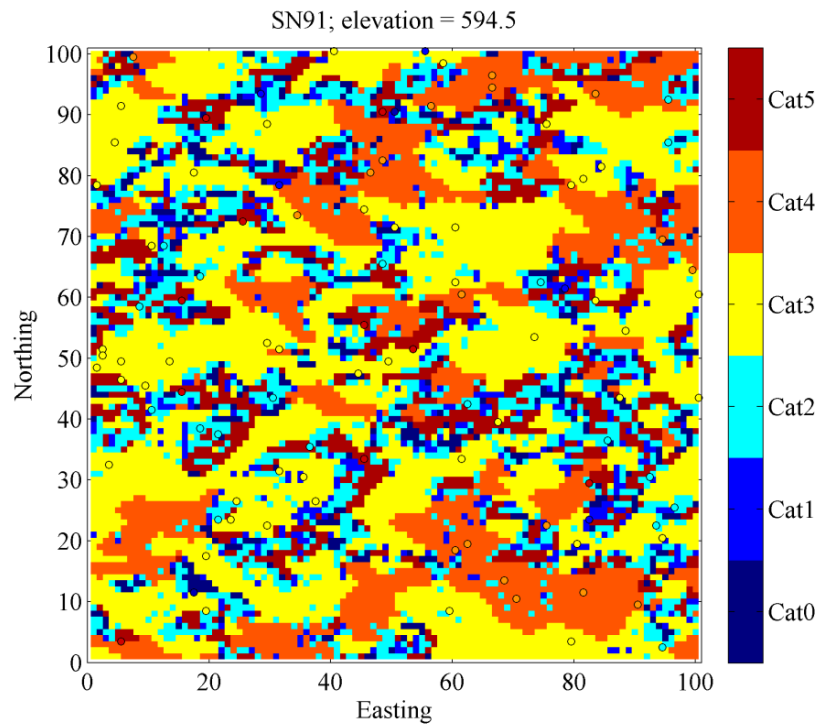
### 5.1.1. Análisis para validación de simulaciones

#### 5.1.1.1. Nivel 1. Condicionamiento de Datos

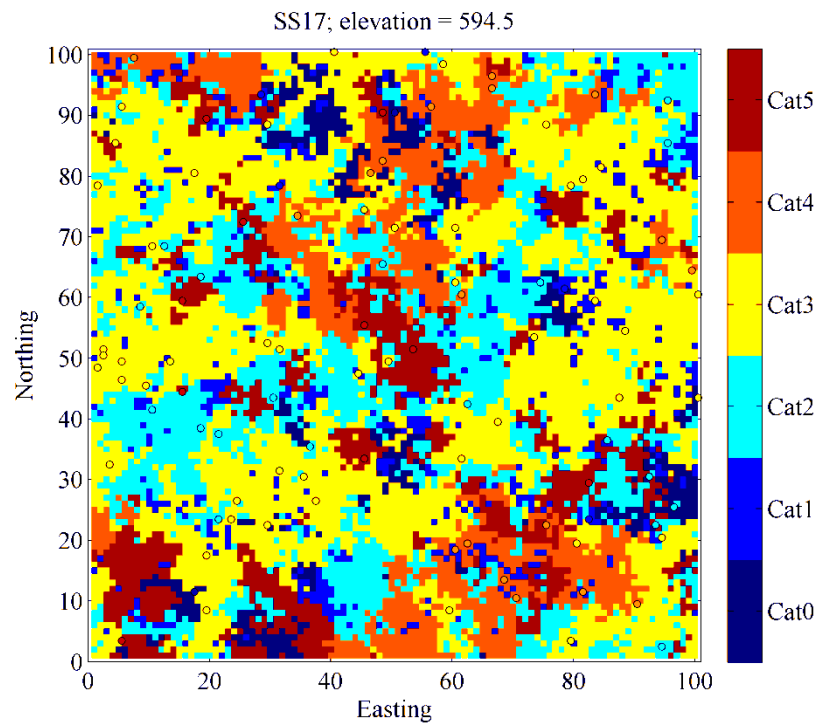
Un primer análisis consiste en verificar si los resultados de las simulaciones reflejan los datos reales en los resultados de las realizaciones, toda vez que las simulaciones fueron efectuadas condicionando a los datos originales o muestras empleadas. Para ello un primer análisis es el visual, en el cual se puede observar si las realizaciones están o no acorde a las muestras empleadas. Entre la Figura 5-8 y Figura 5-11 se puede observar las muestras (en círculos rellenos) y la realización obtenida, en donde los valores de los datos originales son reproducidos en las realizaciones.



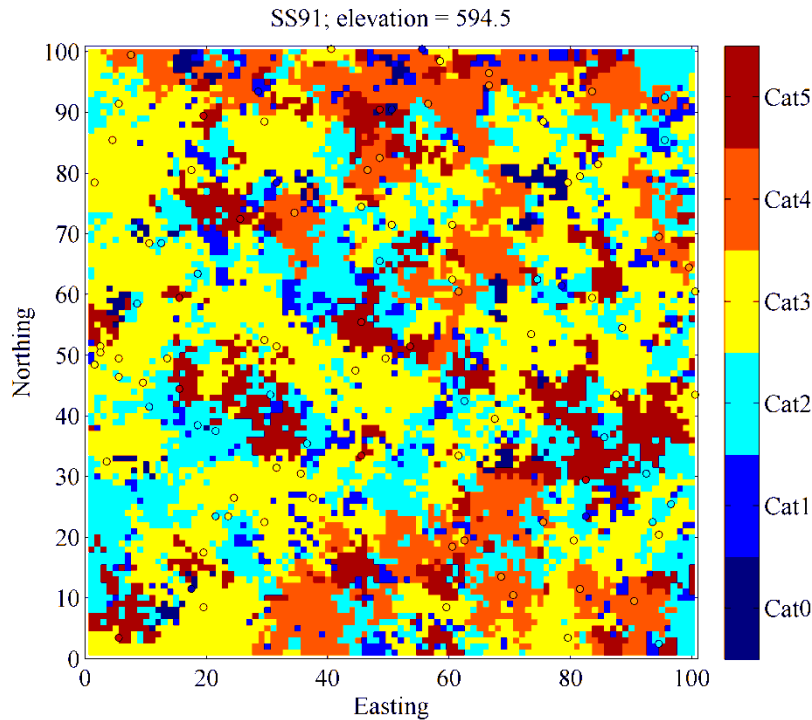
**Figura 5-8:** Inspección visual de muestras y realización 17 - SNESIM



**Figura 5-9:** Inspección visual de muestras y realización 91 - SNESIM



**Figura 5-10:** Inspección visual de muestras y realización 17 - SISIM



**Figura 5-11:** Inspección visual de muestras y realización 91 – SISIM

Esto se puede ver más claramente en un gráfico de dispersión (scatter plot), donde se presentan los valores simulados vs. los valores originales (datos de muestras) para las realizaciones generadas con SNESIM y SISIM. Para este caso en particular el algoritmo ejecuta adecuadamente el condicionamiento de los datos y asigna a los sitios con datos el valor conocido previamente

Al igual que lo acontecido en el caso de la variable continua, es importante aclarar que para este caso los nodos de los sitios con muestra coinciden exactamente con nodos de las simulaciones generadas, situación que difícilmente ocurrirá en un caso real, motivo por el cual se hace importante este tipo de validación y no asumir como obvio esta validación. Se observa en la Figura 5-12 como son reproducidos los datos condicionantes en las realizaciones obtenidas en ambos algoritmos (SNESIM y SISIM).

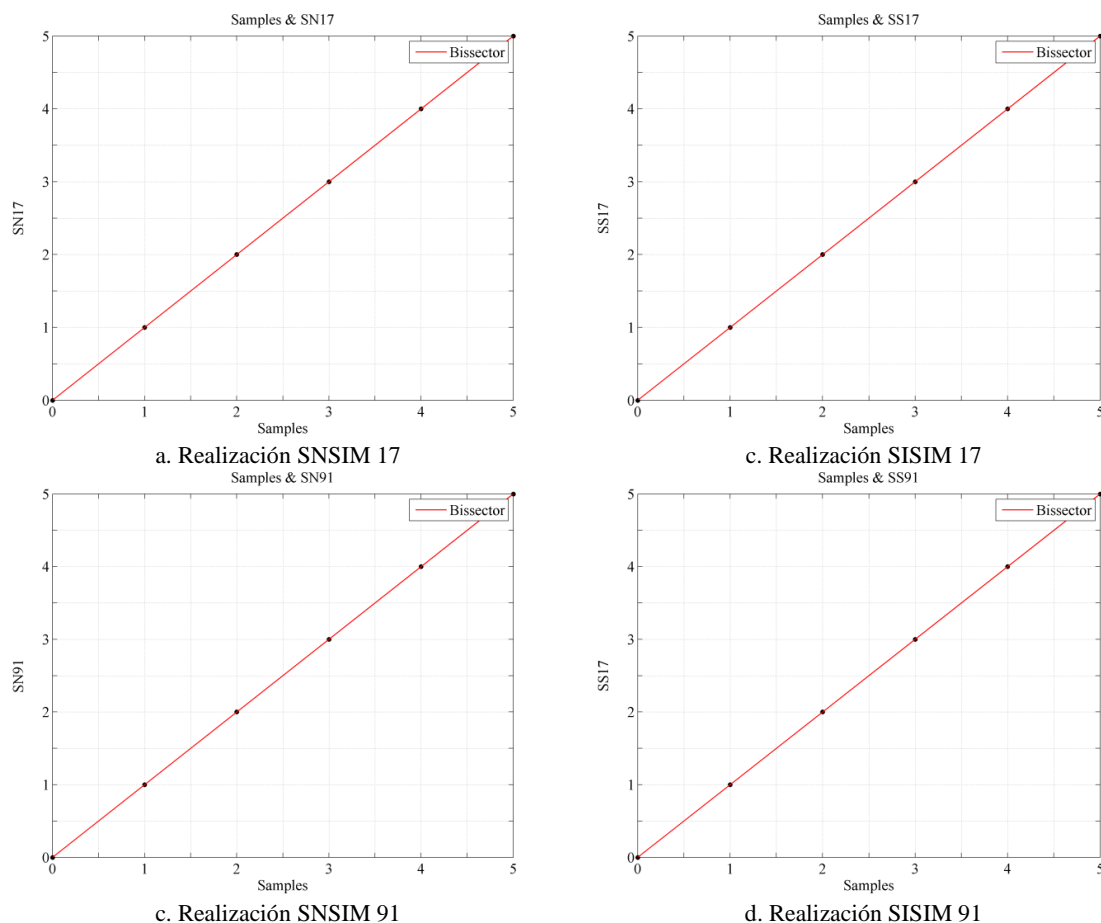
#### 5.1.1.2. Nivel 2. Validación estadística

En este nivel, dado que se trata de una variable categórica, se procederá a evaluar la validación respecto a la reproducción de las proporciones para cada una de las categorías. Dicho análisis se efectuará comparando las proporciones de las realizaciones (analizando el promedio de estas) contra las proporciones de la imagen de entrenamiento, la imagen exhaustiva y las muestras extraídas de esta.

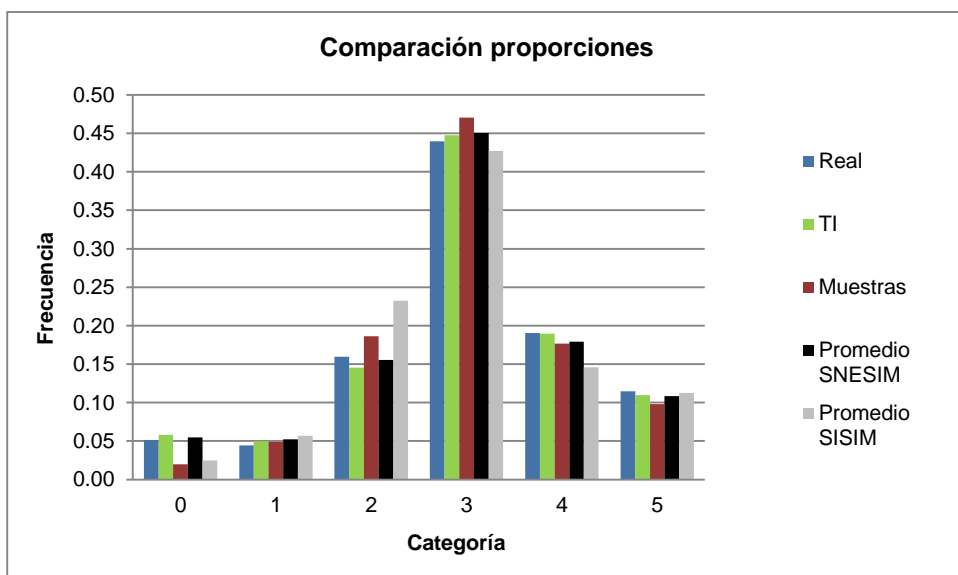
Como se puede observar en la Figura 5-13, el algoritmo SNESIM conserva mejor las proporciones respecto a la imagen de entrenamiento (barras verde y negra); lo cual era el comportamiento que se esperaba encontrar dada su construcción. Este genera unas proporciones por categoría muy similares a las observadas en la imagen exhaustiva (barras azul y negra). Por el contrario el algoritmo SISIM presenta proporciones más similares a las de las muestras (barras roja y gris), salvo en la categoría 4 donde se presenta un mayor cambio. Por ejemplo, conserva mejor la proporción de la categoría 0 de la imagen de entrenamiento y la imagen exhaustiva que corresponde



a la categoría menos representativa dentro del conjunto de datos de la muestra. Sin embargo, el algoritmo SISIM presenta una mejor reproducción de la proporción de la categoría 0 respecto a las muestras, pero con cambios importantes en las proporciones de las categorías 2, 3 y 4 (las más abundantes) con lo cual podría decirse que SISIM no reproduce proporciones de las muestras en general.



**Figura 5-12:** Gráfico de dispersión de valores simulados vs. valores originales



**Figura 5-13:** Reproducción de proporciones SNESIM/SISIM

- **Test estadísticos**

Para efectuar estos test, se empleó una variable indicador. Se analizaron las categorías con menor y mayor representatividad, es decir las categorías 0 y 3 respectivamente. Se analizó inicialmente la reproducción de la media y posteriormente la de la varianza. Se aplica en primer lugar el test de Student para la media para para al algoritmo SNESIM comparándolo con las medias de las muestras y la imagen de entrenamiento; luego se analiza SISIM respecto a la media de las muestras y también la imagen de entrenamiento, tanto en el indicador para la categoría 0, como para el indicador de la categoría 3. Posteriormente se ejecuta el test de Student para la varianza a las mismas categorías (0 y 3) en ambos algoritmos. Estos test fueron efectuados mediante algoritmos escritos en código MatLab desarrollados por Emery (2008)<sup>7</sup> y los resultados pueden observarse en el anexo F.

En la Tabla 5.2 se pueden observar los resultados obtenidos en el test de reproducción de medias cuando se trata de evaluar la media de la función indicador de las muestras para las categorías 0 y 3. En ambos casos cuando se efectúa el test, es rechazada la hipótesis nula con lo cual podría decirse que no existe una reproducción exacta de la media de las muestras por parte de las realizaciones para el caso de SNESIM. Sin embargo, al evaluar respecto a la media de la imagen de entrenamiento, para el caso de la categoría 3 el test acepta la hipótesis nula con lo cual puede afirmarse que no existe evidencia estadística significativa para rechazar la reproducción exacta de la media por parte de las realizaciones. En el caso de la categoría 0 la hipótesis nula es rechazada. Sin embargo, dado que los test son desarrollados sin cumplirse los supuestos bajos los cuales deben desarrollarse (dado que la variable indicador no los cumple), estos resultados deben ser observados con cautela y servir solo como ejercicio académico.

**Tabla 5.2:** Resultados del test de reproducción de media de muestras para SNESIM

<b>Test de media</b>	<b>SNESIM Ind 0 vs. muestras</b>	<b>SNESIM Ind 3 vs. muestras</b>	<b>SNESIM Ind 0 vs. TI</b>	<b>SNESIM Ind 3 vs. TI</b>
Media a comparar	0.019608	0.470588	0.057935	0.447505
Estadístico Student	77.2144	8.5947	7.2973	1.2765
Número grados de libertad	99	99	99	99
Nivel de significancia	0.05	0.05	0.05	0.05
Valor crítico	1.9842	1.9842	1.9842	1.9842
Resultado del test	Hipótesis nula rechazada.	Hipótesis nula rechazada.	Hipótesis nula rechazada.	Hipótesis nula aceptada.

Con el algoritmo SISIM se rechaza la hipótesis nula en todos los casos (Tabla 5.3).

**Tabla 5.3:** Resultados del test de reproducción de media de muestras para SISIM

<b>Test de media</b>	<b>SISIM Ind 0 vs. muestras</b>	<b>SISIM Ind 3 vs. muestras</b>	<b>SISIM Ind 0 vs. TI</b>	<b>SISIM Ind 3 vs. TI</b>
Media a comparar	0.019608	0.470588	0.057935	0.447505
Estadístico Student	6.8722	14.6098	42.6289	6.8508
Número grados de libertad	99	99	99	99
Nivel de significancia	0.05	0.05	0.05	0.05
Valor crítico	1.9842	1.9842	1.9842	1.9842
Resultado del test	Hipótesis nula rechazada.	Hipótesis nula rechazada.	Hipótesis nula rechazada.	Hipótesis nula rechazada.

<sup>7</sup> Código MatLab disponible en <http://www.iamg.org/images/File/documents/oldftp/VOL34/v34-11-18.zip>

Los resultados para el test de Student de reproducción de varianza para los algoritmos SNESIM y SISIM en las categorías 0 y 3 se presentan en la Tabla 5.4 y Tabla 5.5.

**Tabla 5.4:** Resultados del test de reproducción de varianza para SNESIM

Test de varianza	SNESIM Ind 0. vs. var. muestras	SNESIM Ind 3. vs. var. muestras	SNESIM Ind 0. vs. var. TI	SNESIM Ind 3. vs. var. TI
Varianza	0.019414	0.251602	0.054584	0.247269
Estadístico Student	79.7244	18.3468	7.3335	1.0427
Número grados de libertad	99	99	99	99
Nivel de significancia	0.05	0.05	0.05	0.05
Valor crítico	1.9842	1.9842	1.9842	1.9842
Resultado del test	Hipótesis nula rechazada.	Hipótesis nula rechazada.	Hipótesis nula rechazada.	Hipótesis nula aceptada

**Tabla 5.5:** Resultados del test de reproducción de varianza para SISIM

Test de varianza	SISIM Ind 0. vs. var. muestras	SISIM Ind 3. vs. var. muestras	SISIM Ind 0. vs. var. TI	SISIM Ind 3. vs. var. TI
Varianza	0.019414	0.251602	0.054584	0.247269
Estadístico Student	6.6126	18.3599	41.4965	8.1465
Número grados de libertad	99	99	99	99
Nivel de significancia	0.05	0.05	0.05	0.05
Valor crítico	1.9842	1.9842	1.9842	1.9842
Resultado del test	Hipótesis nula rechazada.	Hipótesis nula rechazada.	Hipótesis nula rechazada.	Hipótesis nula rechazada.

Como se puede observar, el test de Student para la varianza rechaza la hipótesis nula en todos los casos para el algoritmo SISIM. Misma situación ocurre para el caso de SNESIM.

Tal como se mencionó, estos test exigen ciertos supuestos para su aplicación; que la variable indicador no cumple, como son estacionariedad; con lo cual podría decirse que no son test útiles en este caso.

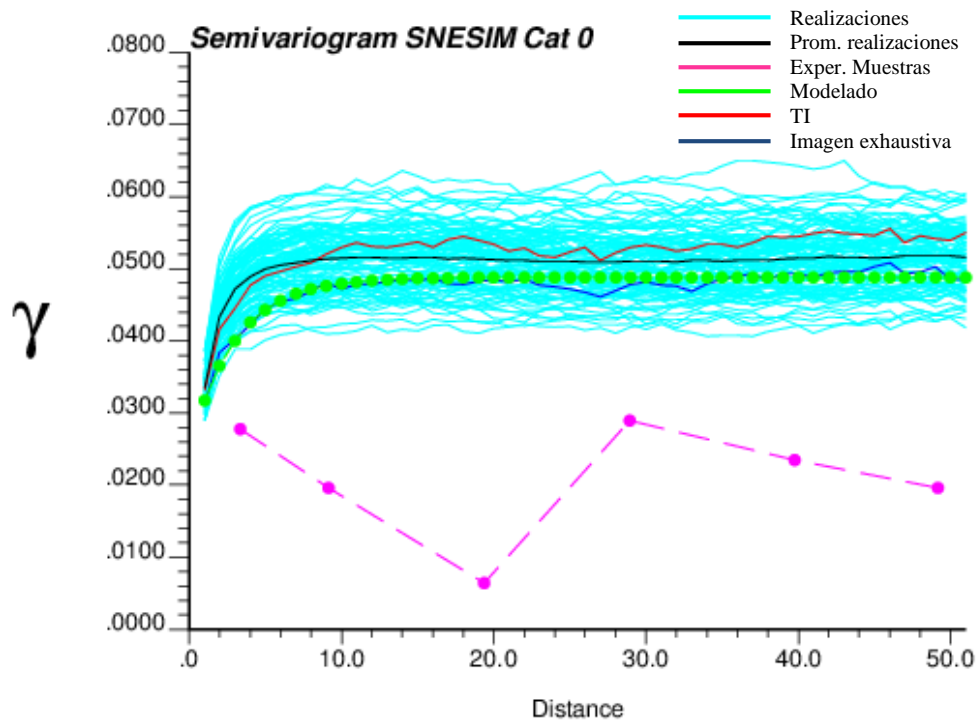
### 5.1.1.3. Nivel 3. Validación geoestadística

En un primer análisis se presenta el chequeo de variogramas (estadísticas de 2 puntos). Posteriormente se analizará la reproducción de estadísticas de orden superior, como lo son los cumulantes de orden 3 analizados para este caso.

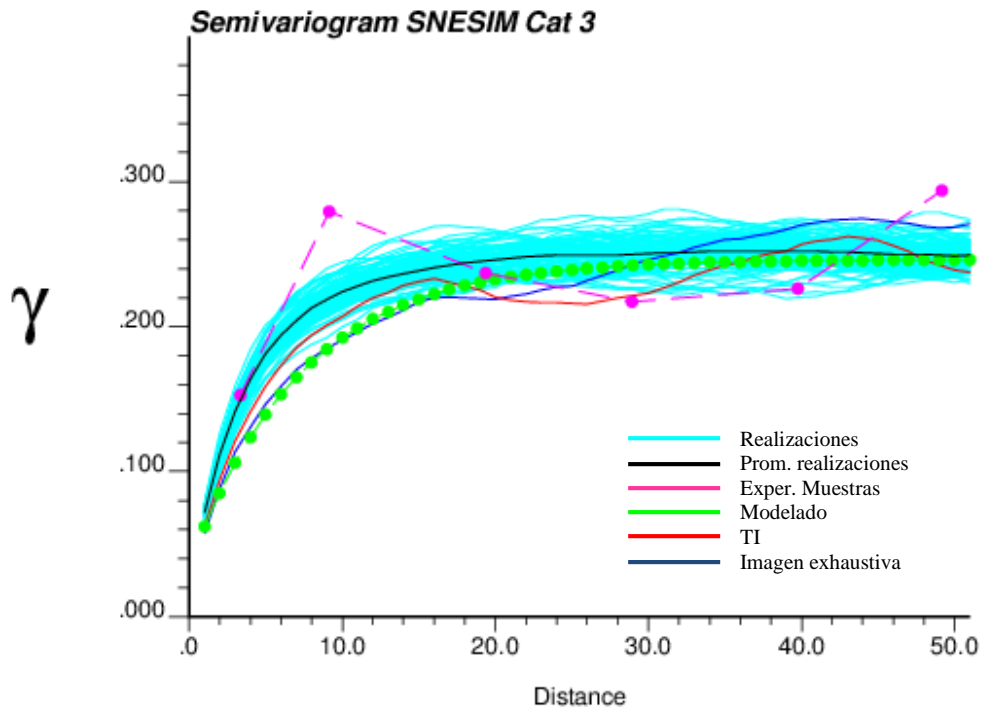
- **Variograma**

Respecto a la validación de variogramas se procederá a comparar los variogramas experimentales de las realizaciones de la categoría 0 (categoría menos representativa dentro del conjunto de datos) y los de la categoría 3 (la de mayor proporción) con los variogramas experimentales de la imagen exhaustiva, la imagen de entrenamiento y el variograma experimental de las muestras; así como el variograma modelado a partir de este último variograma experimental y el cual fue empleado en la simulación con el algoritmo SISIM (para ambas categorías 0 y 3).

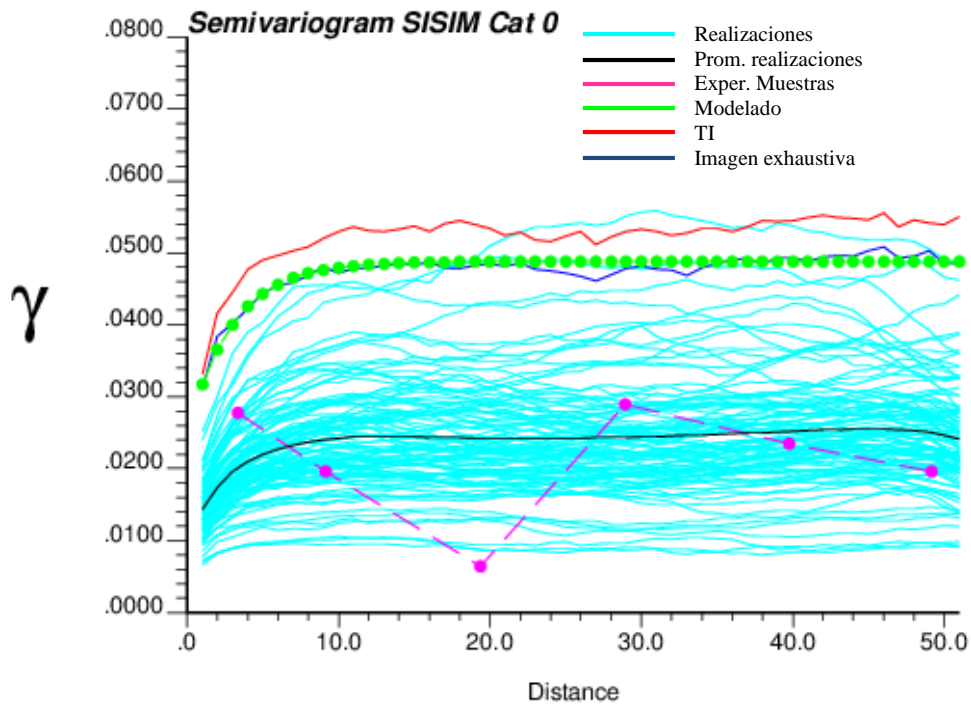
Los resultados obtenidos al graficar los diversos variogramas descritos se presentan desde la Figura 5-14 a la Figura 5-17.



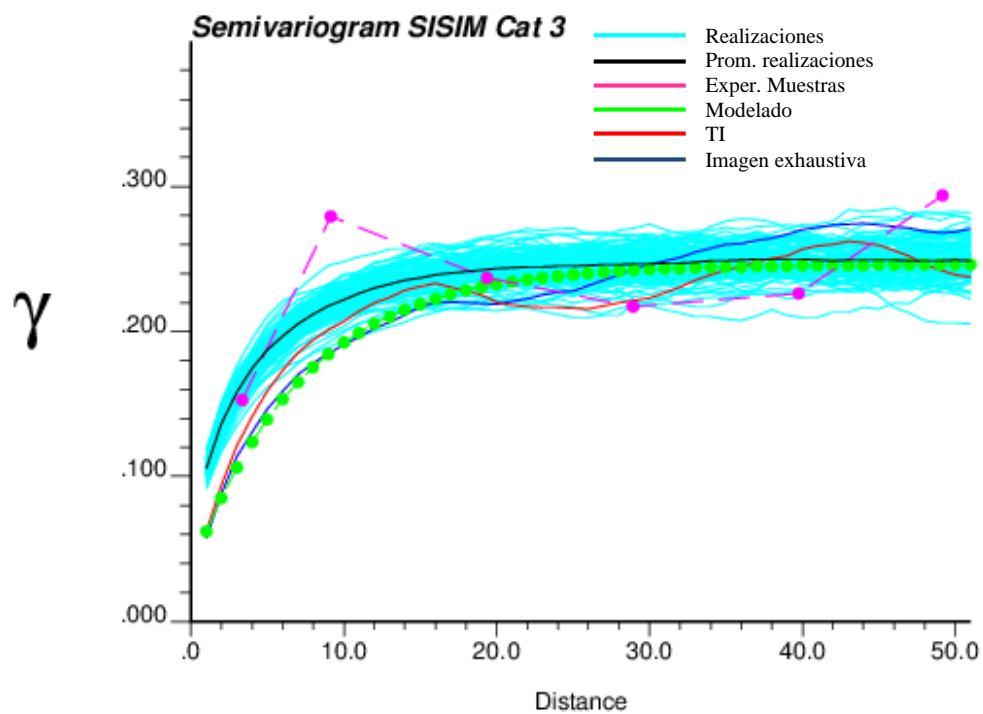
**Figura 5-14:** Gráfico comparativo de variogramas para SNESIM Categoría 0.



**Figura 5-15:** Gráfico comparativo de variogramas para SNESIM Categoría 3.



**Figura 5-16:** Gráfico comparativo de variogramas para SISIM Categoría 0.



**Figura 5-17:** Gráfico comparativo de variogramas para SISIM Categoría 0.

Como se aprecia en la Figura 5-14, el variograma promedio de las realizaciones generado por SNESIM (para la categoría 0) se aproxima mucho más al de la imagen de entrenamiento (comportamiento esperado) y presenta un mejor comportamiento que lo observado en el algoritmo SISIM (Figura 5-16). En este último existe evidencia clara del cambio de meseta entre los variogramas promedio de las realizaciones y los de la imagen exhaustiva y la imagen de entrenamiento, lo cual es debido a sesgo en las proporciones debido al muestreo, dado que la

varianza para una variable categórica es función de la proporción. El mejor desempeño de SNESIM se puede corroborar al observar el error medio cuadrático en la Tabla 5.6, lo cual ocurre también para los demás variogramas.

De igual forma, para la categoría 3 se presenta un menor EMC entre el promedio de las realizaciones de SNESIM con la imagen de entrenamiento, situación esperada. En cuanto al variograma modelado, se presenta también un mejor desempeño de SNESIM. Esta situación se debe probablemente a que las proporciones de la imagen exhaustiva son similares a las de la imagen de entrenamiento y por ende las mesetas similares, lo cual difiere de las proporciones de las muestras que influyen en la generación de las realizaciones por parte de SISIM dada la construcción del algoritmo a través de kriging de indicadores que emplean estas categorías.

Finalmente, se muestra como los variogramas de las muestras (principalmente de la categoría 0) difieren considerablemente de los variogramas de las imágenes exhaustiva y de entrenamiento (por un sesgo en las proporciones). Sin embargo, en el caso de la categoría 3, SISIM trata de reproducir ese comportamiento.

**Tabla 5.6.** EMC entre variogramas (SNESIM/SISIM) para categoría 0 (51 m)

VARIOGRAMA	Imagen Exhaustiva (Realidad)	Imagen de entrenamiento (TI)	Modelado (SGSIM)
Media SISIM	0.000550492	0.000807396	0.000562709
Media SNESIM	1.22104E-05	4.88437E-06	1.04702E-05

**Tabla 5.7.** EMC entre variogramas (SNESIM/SISIM) para categoría 3 (51 m)

VARIOGRAMA	Imagen Exhaustiva (Realidad)	Imagen de entrenamiento (TI)	Modelado (SGSIM)
Media SISIM	0.000612025	0.000385048	0.000453337
Media SNESIM	0.000509681	0.000345962	0.000356824

- **Test estadístico variogramas**

Dadas las fluctuaciones presentes en los variogramas de las realizaciones, al igual que lo efectuado con la variable continua, se llevó a cabo un test estadístico para determinar si dichas fluctuaciones están dentro de cierto grado de confianza a través del test de Hotelling. Es importante mencionar que este test se presenta solo para los variogramas de la categoría 3, que es donde se observa cierta reproducción entorno al variograma modelado (claramente la categoría 0 no cumple esta condición y por tanto no se evalúa).

**Tabla 5.8.** Test de Hotelling para validación de variogramas SNESIM/SISIM

Test de media y varianza	SNESIM	SISIM
Número de lags	5	5
Lags	10 20 30 40 50	10 20 30 40 50
Estadístico T <sup>2</sup> Hotelling	21014.2382	15702.9178
Número grados de libertad	5 y 95	5 y 95
Nivel de significancia	0.05	0.05
Valor crítico	12.0375	12.0375
Resultado del test	Hipótesis nula es rechazada	Hipótesis nula es rechazada

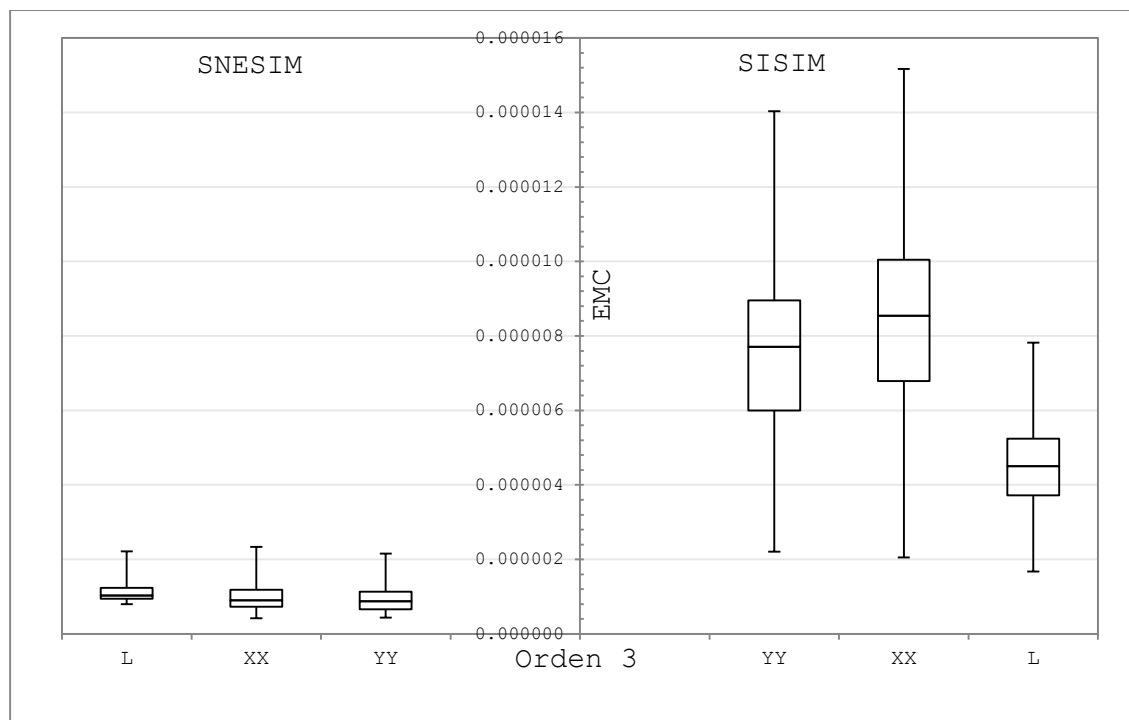
Como se puede observar, el test para ambos algoritmos, rechaza la hipótesis nula. Como se mencionó en la variable continua este test exige cumplir ciertas asunciones, tales como Estacionariedad intrínseca y la variable en análisis carece claramente de esta propiedad; adicionalmente porque las simulaciones efectuadas emplean una imagen de referencia o se caracteriza solamente un sector parcial de la distribución para explicar la variable de interés, motivo por el cual no sería aplicable para este caso en particular y debería ser tomado como un ejercicio académico.

- **Cumulantes de orden superior**

Se calcularon cumulantes de orden superior (de orden 3 en este caso) mediante el software HOSC (High Order Spatial Cumulants) desarrollado por Dimitrakopoulos y Musthapa (2010)<sup>8</sup> en código FORTRAN. Estos se calcularon para tres configuraciones distintas, las cuales corresponden a la misma configuración presentada en la variable continua (Figura 4-18):

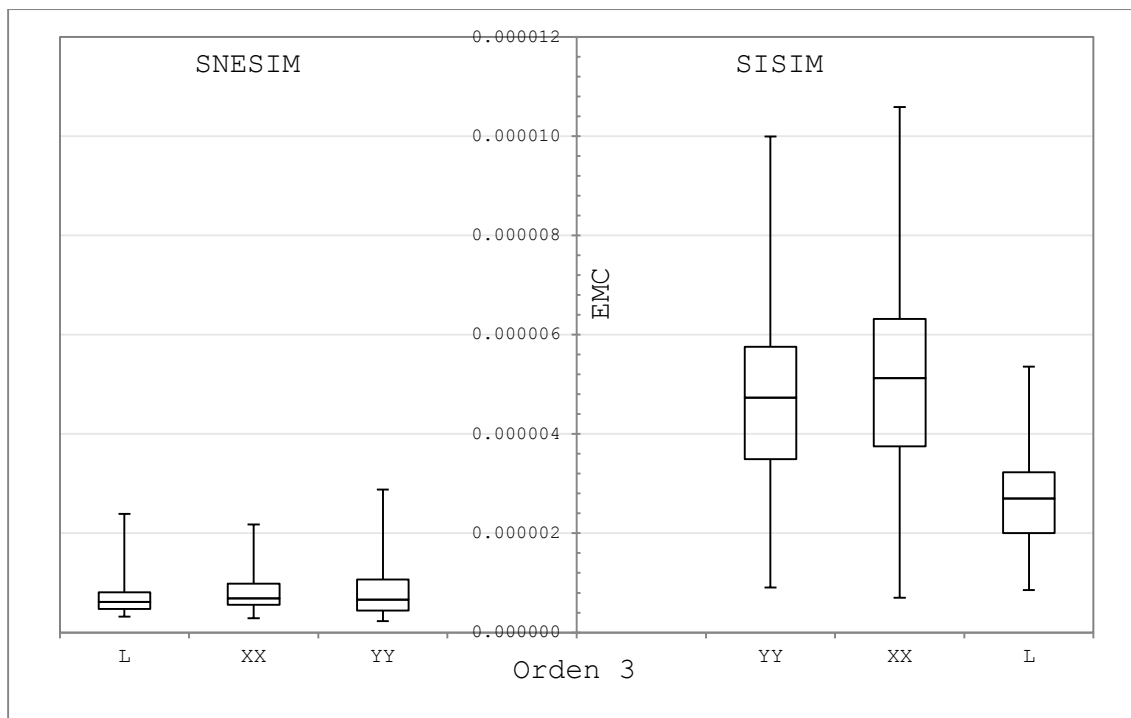
- Configuración 1:  $\{0^\circ - 90^\circ\}$  → Denominada “L”
- Configuración 2:  $\{90^\circ - 90^\circ\}$  → Denominada “XX”
- Configuración 3:  $\{0^\circ - 0^\circ\}$  → Denominada “YY”

Para cada una de las configuraciones anteriores y las 100 realizaciones (de SNESIM y SISIM) se calculó el error medio cuadrático (EMC) entre los cumulantes de orden 3 de la imagen de entrenamiento y la imagen exhaustiva vs cada una de las realizaciones en las categorías 0 y categoría 3. Obtenidos los EMC se graficaron box plot que muestran las distribuciones de estos y permiten ver el desempeño de cada uno de los algoritmos respecto a los cumulantes de orden 3.



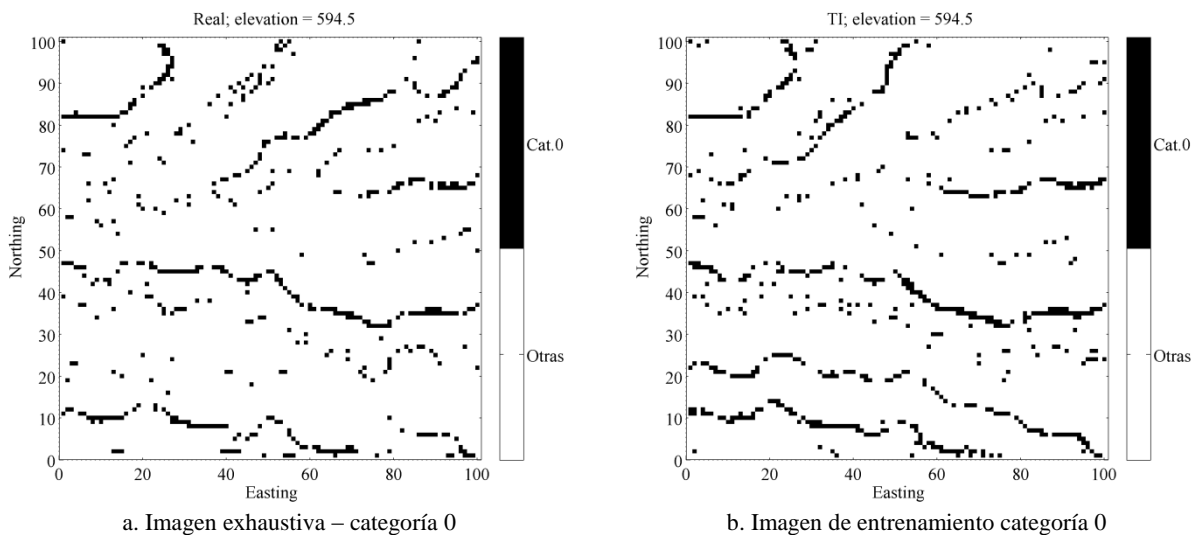
**Figura 5-18.** EMC cumulantes realizaciones vs. TI (Categoría 0)

<sup>8</sup> Código FORTRAN disponible en <http://www.iamg.org/documents/oldftp/VOL36/v36-03-06.zip>



**Figura 5-19.** EMC cumulantes realizaciones vs. realidad (Categoría 0)

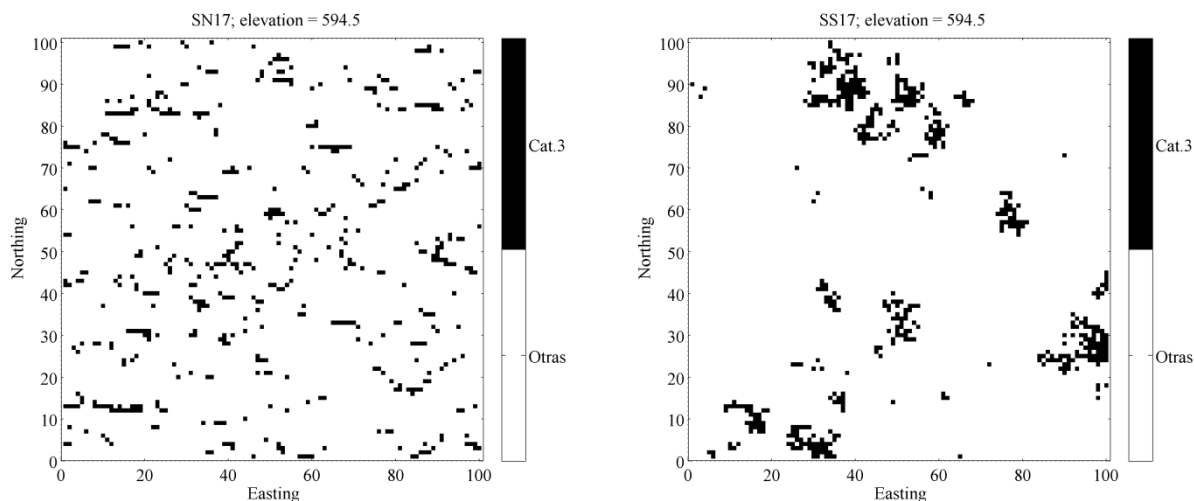
En las figuras anteriores se puede ver como para la categoría 0, el algoritmo SNESIM captura de mejor manera los patrones de distribución espaciales calculados por cumulantes de orden 3, para todas las configuraciones analizadas. Por otro lado, cuando la comparación se efectúa contra la imagen exhaustiva (realidad) mejora el desempeño del algoritmo SISIM y el algoritmo SNESIM presenta desempeños similares respecto a los cumulantes de orden 3 calculados para dicha imagen. Se puede ver igualmente diferencias significativas, las cuales pueden ser explicadas en el sesgo en la reproducción de proporciones. En la Figura 5-20 se puede observar la imagen binaria de la realidad y la imagen de entrenamiento de esta categoría.



**Figura 5-20.** Imágenes binarias de categoría 0



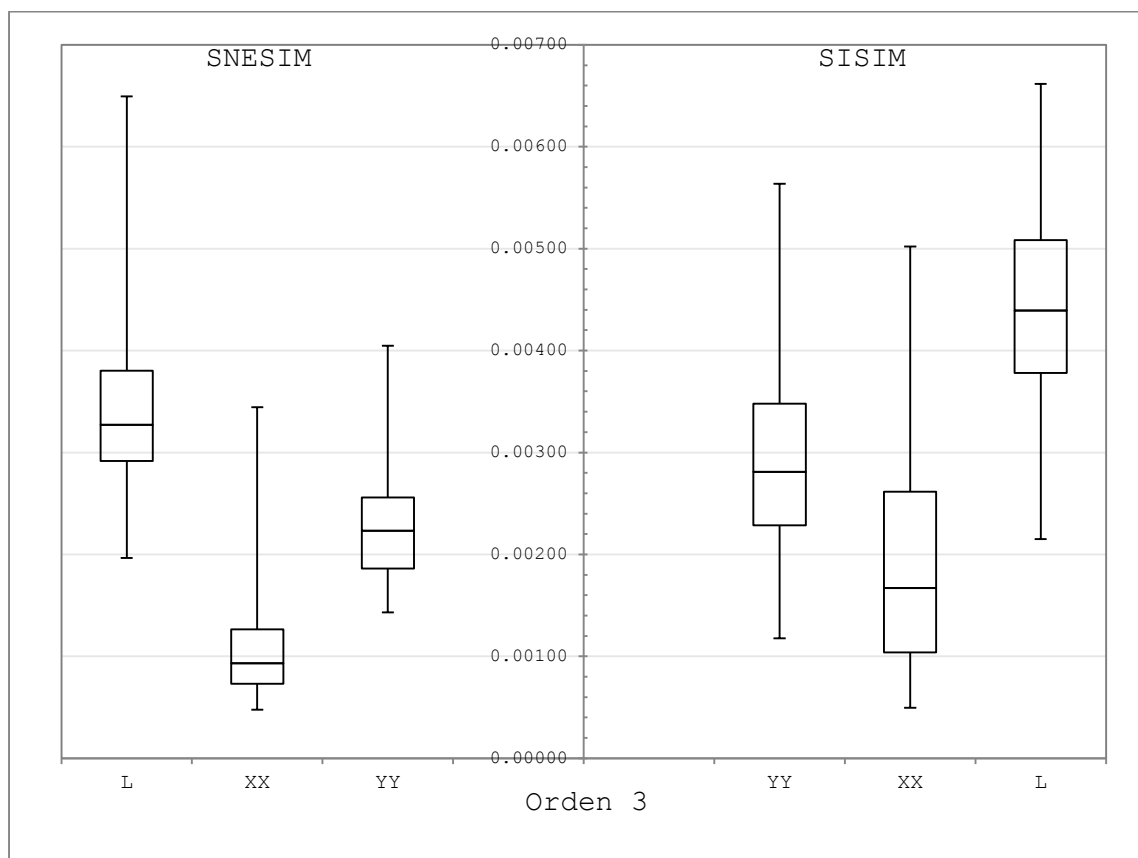
Igualmente se presentan dos imágenes binarias de la categoría 0, correspondientes a realizaciones de los algoritmos en cuestión, lo cual puede dar más claridad respecto a los resultados obtenidos.



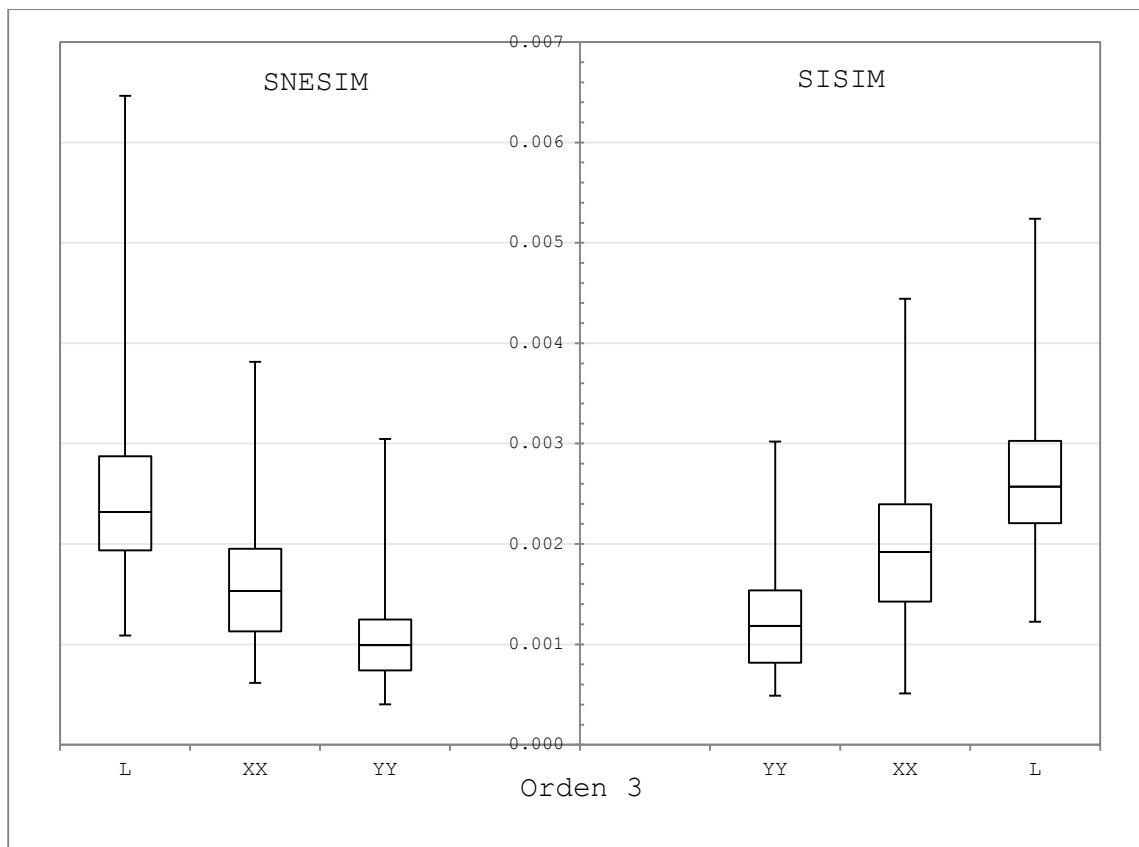
a. Imagen SNESIM – categoría 0

b. Imagen SISIM categoría 0

**Figura 5-21:** Ejemplo de imágenes binarias de realizaciones obtenidas para categoría 0



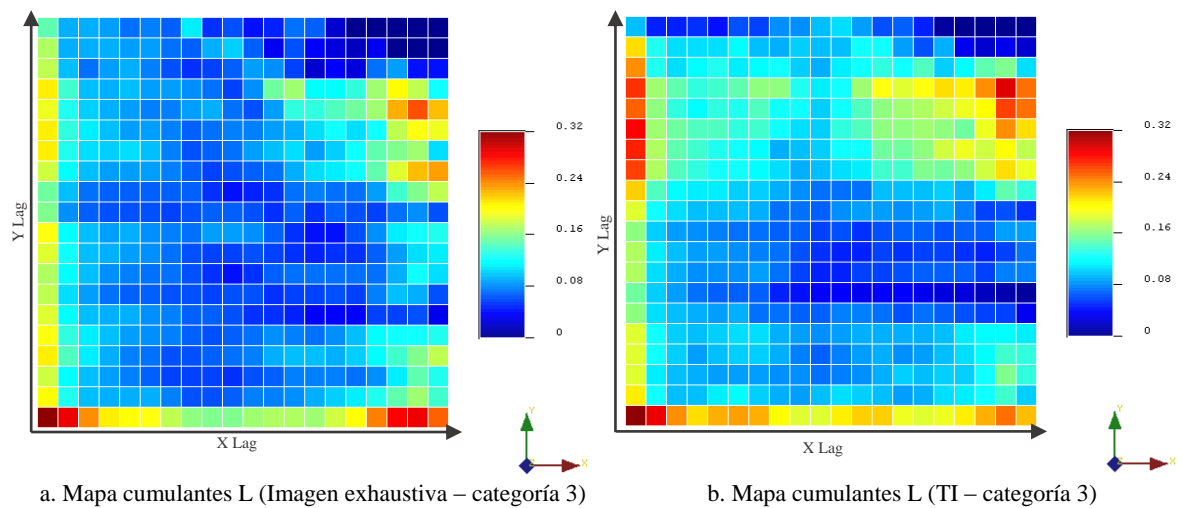
**Figura 5-22:** EMC cumulantes realizaciones vs. TI (Categoría 3)



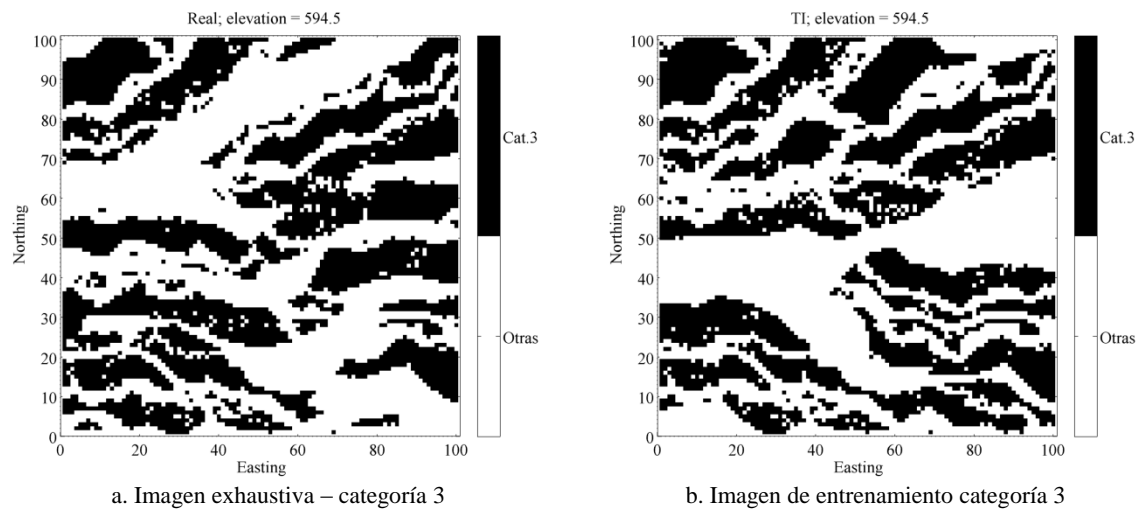
**Figura 5-23:** EMC cumulantes realizaciones vs. realidad (Categoría 3)

Respecto a la categoría 3, se puede ver igualmente una mejor captura de las distribuciones espaciales del algoritmo SNESIM respecto a las relaciones capturadas por los cumulantes de orden 3 para las realizaciones de SISIM, tanto en el caso de la imagen de entrenamiento, como en la imagen exhaustiva.

Por otro lado, se observa un comportamiento diferente en la configuración XX cuando se comparan los cumulantes de orden 3 de las realizaciones de los algoritmos con la imagen de entrenamiento; presentándose una reducción importante en el EMC para ambos algoritmos. Esto podría explicarse en la existencia de patrones horizontales en las imágenes en análisis, los cuales pueden observarse en el mapa de cumulantes de configuración L (Figura 5-24), donde se puede apreciar una franja inferior de valores altos a lo largo del eje x en ambos casos (mayor en la imagen de entrenamiento), lo que informa sobre la presencia de estructuras horizontales que cuando se calculan los cumulantes en la configuración XX se verá reflejado en menores diferencias. En estos mapas también se observan zonas de mayores valores en el costado nor-este en ambos casos (de mayor magnitud en la imagen de entrenamiento, inclusive tratándose de conectar hacia el oeste con la zona de valores altos); esto muestra zonas de mayor conectividad; tal como se puede apreciar en el mapa binario de las imágenes que se presenta en la Figura 5-25.



**Figura 5-24:** Mapa de cumulantes L orden 3 (categoría 3)

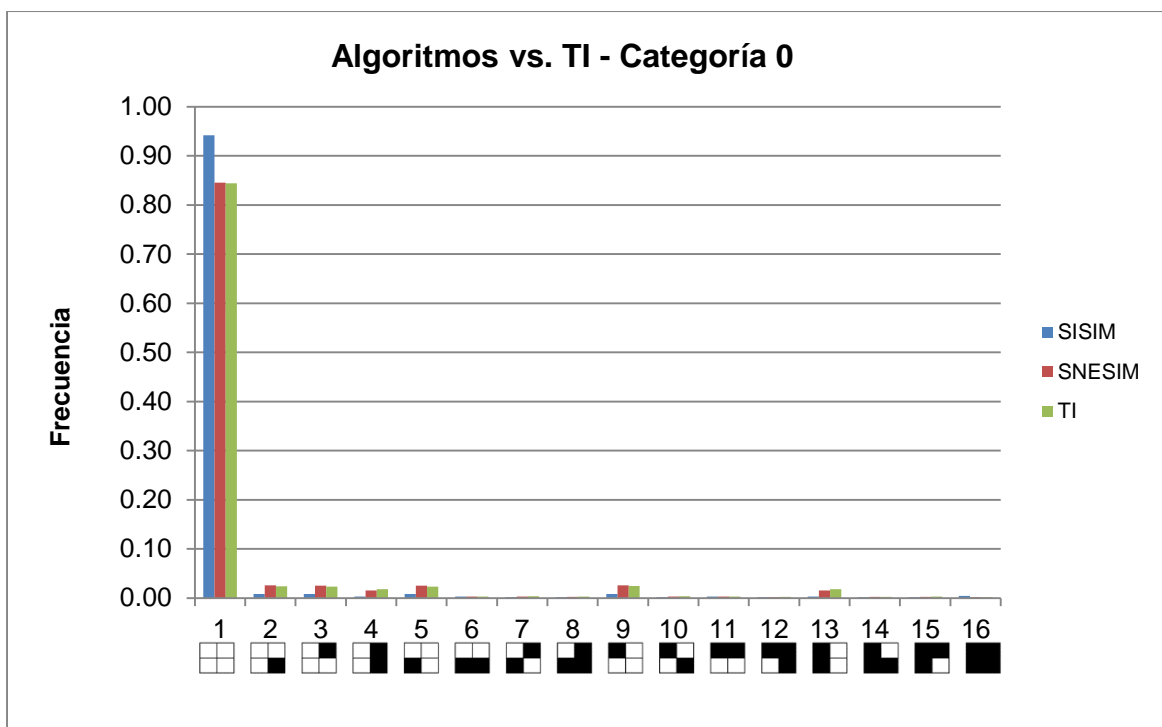


**Figura 5-25.** Imágenes binarias de categoría 3

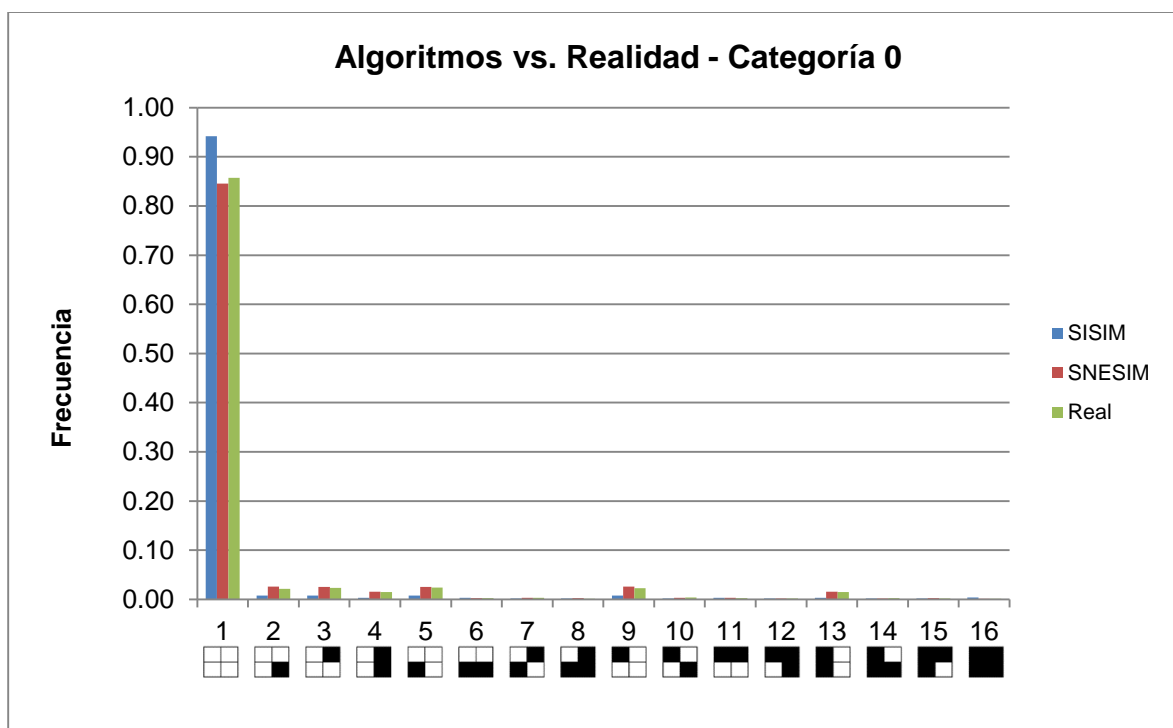
- **Función de densidad de múltiples puntos**

Para el caso de la variable categórica se usó igualmente la función de densidad de múltiples puntos, determinando las frecuencias de ocurrencia de patrones en las imágenes (de entrenamiento y exhaustiva) y el promedio de las realizaciones para complementar el análisis de validación presentado.

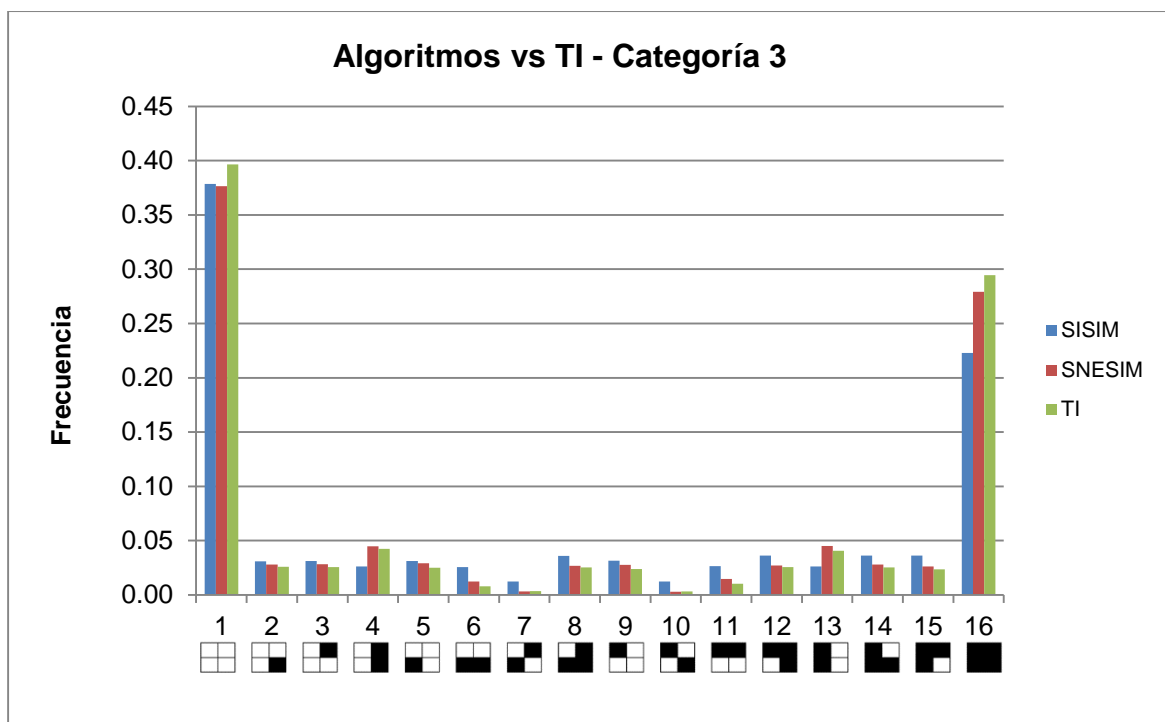
Los patrones empleados corresponden a los mismos descritos en la Figura 4-21 empleados en la variable continua, y fueron calculados tanto para la categoría 0, como para la categoría 3. En las figuras siguientes se pueden observar los resultados obtenidos. Igualmente en el anexo H se presenta el código MatLab empleado para determinar las frecuencias de ocurrencia de patrones.



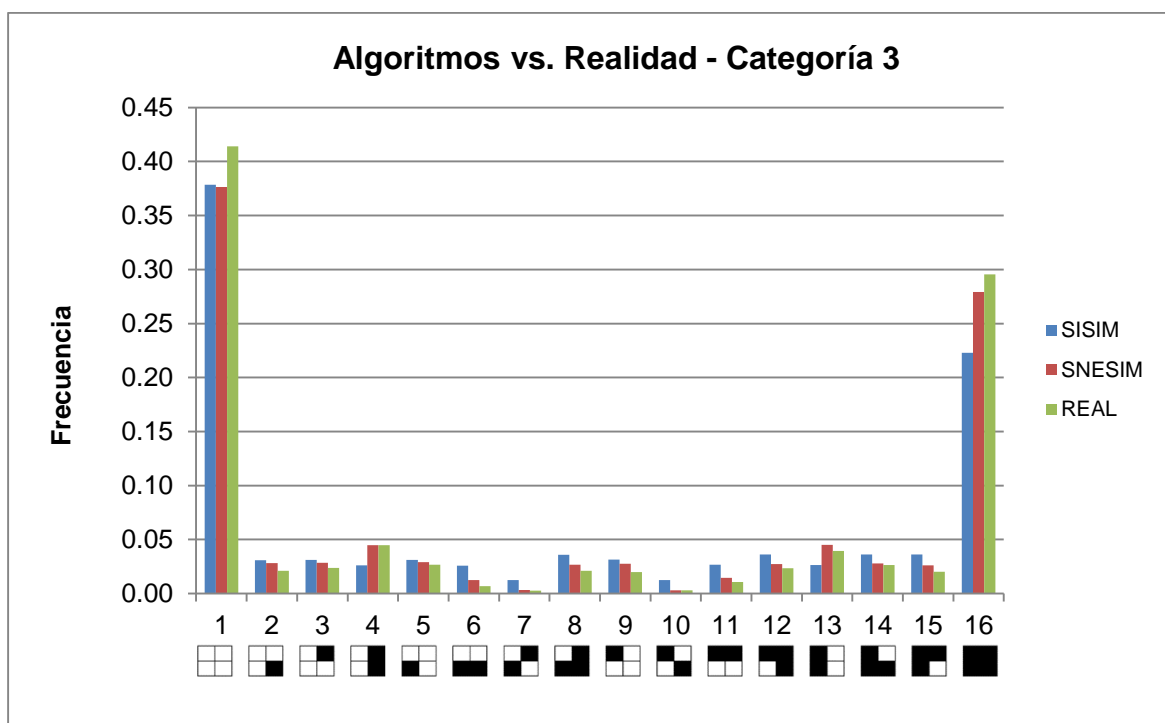
**Figura 5-26:** Función de densidad de múltiples puntos cat0 – realizaciones vs. TI



**Figura 5-27.** Función de densidad de múltiples puntos cat0 – realizaciones vs. realidad



**Figura 5-28.** Función de densidad de múltiples puntos cat3 – realizaciones vs. TI

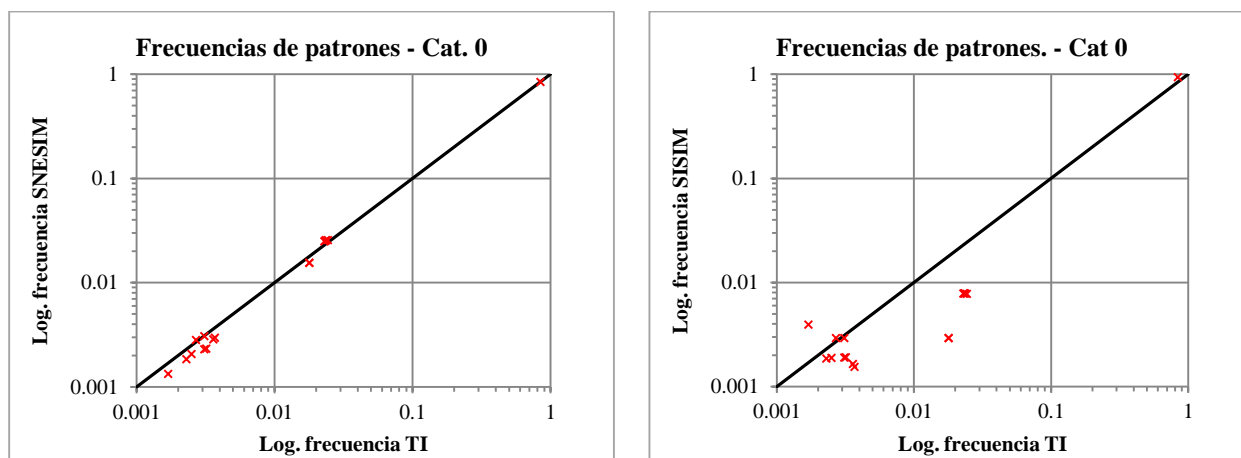


**Figura 5-29.** Función de densidad de múltiples puntos cat3 – realizaciones vs. realidad

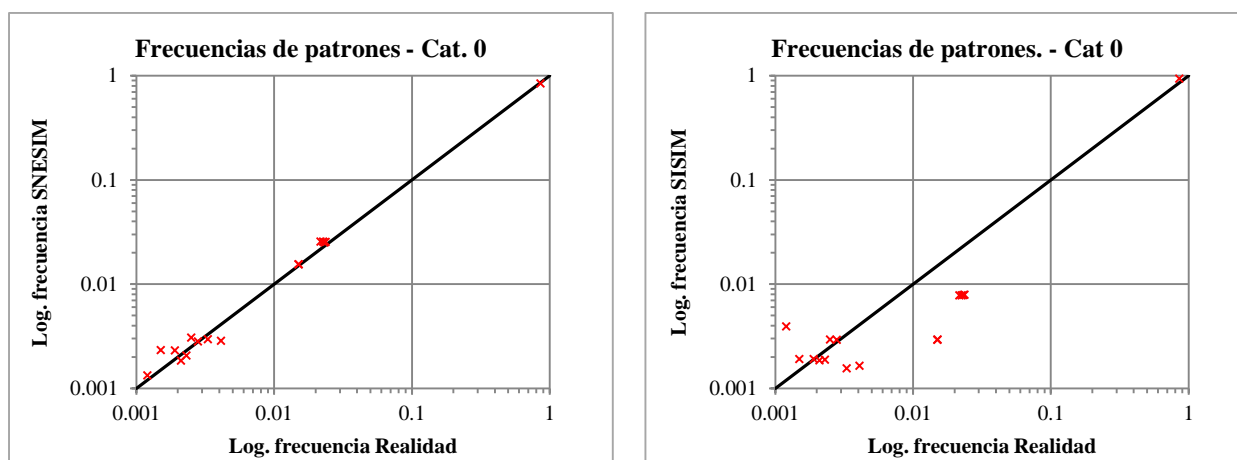
Al observar la categoría 0, se observa que el algoritmo SNESIM presenta frecuencias más parecidas para todos los patrones que las encontradas para SISIM en cuanto a la imagen de entrenamiento se refiere y que sería lo esperado al momento de efectuar una validación de este tipo de simulaciones y acorde con los resultados obtenidos en otras métricas aplicadas.

En cuanto a la categoría 3, se observa igualmente cierta correspondencia entre la imagen exhaustiva y el promedio de las realizaciones obtenidas mediante el algoritmo SNESIM. Esto no ocurre con los resultados del algoritmo SISIM donde se aprecian ciertas discrepancias importantes en ciertos patrones. Esto permite concluir, al menos desde el punto de vista de la función de densidad de múltiples puntos, que el algoritmo SISIM no logra reproducir adecuadamente ciertos comportamientos encontrados en la variable original, lo cual resalta la importancia de adelantar este tipo de validaciones, cuando de variables no estacionarias se trata y que son simuladas mediante algoritmos no paramétricos; puesto que permiten determinar mayores comprobaciones a las obtenidas con métricas de los niveles 1 y 2.

Se construyeron igualmente gráficos de dispersión para observar de mejor manera las frecuencias de patrones 2x2 presentadas en las anteriores figuras, y como complemento también se construyeron gráficos de frecuencia de patrones 3x3, las cuales se presenta a continuación.

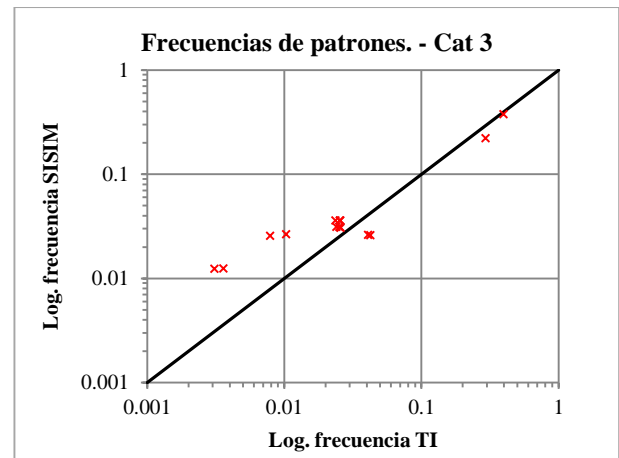
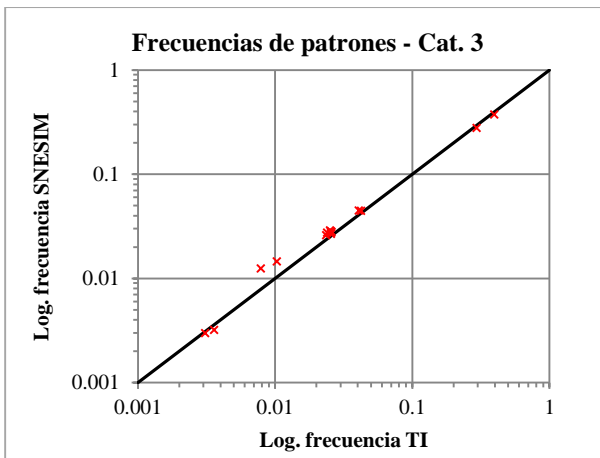


**Figura 5-30:** Gráfico de dispersión de frecuencias de patrones 2x2 TI cat 0 vs. algoritmos SNESIM/SISIM

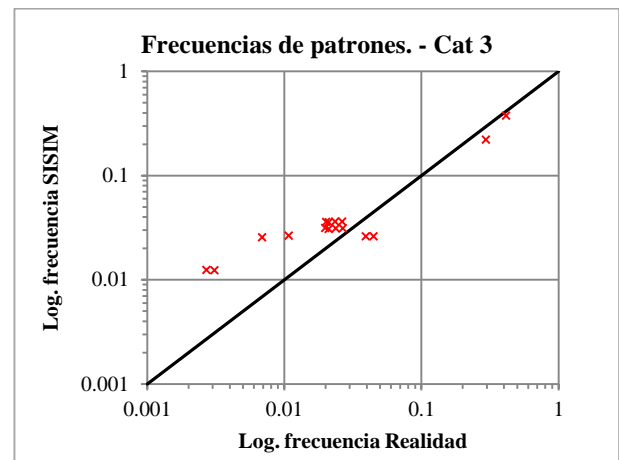
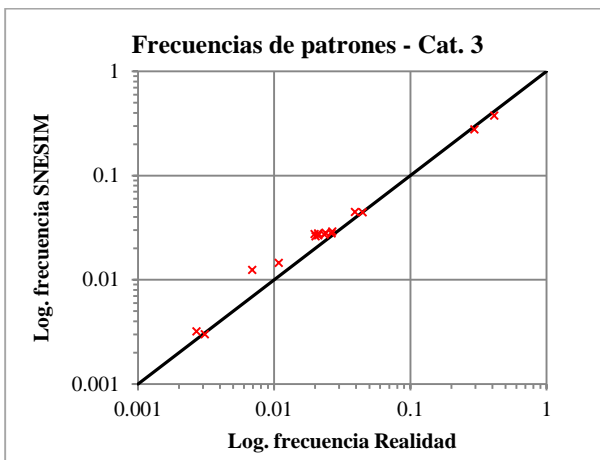


**Figura 5-31:** Gráfico de dispersión de frecuencias de patrones 2x2 realidad cat 0 vs. algoritmos SNESIM/SISIM

Como se aprecia en las gráficas anteriores, existe una mejor reproducción de los patrones por parte de SNESIM, que lo que genera SISIM, tanto para la imagen de entrenamiento como para la imagen exhaustiva. El mismo comportamiento se observa en las siguientes dos figuras que corresponden a la categoría 3.

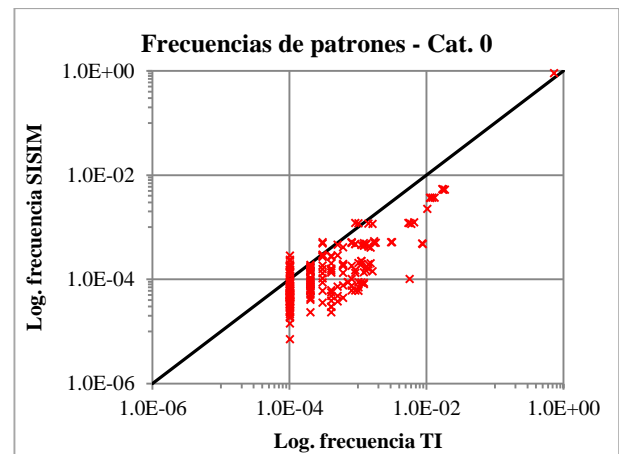
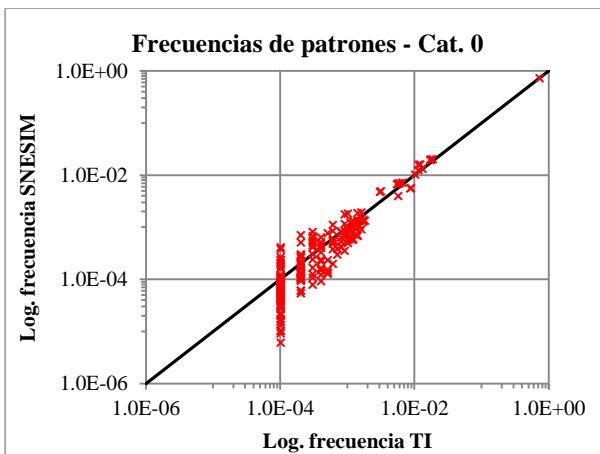


**Figura 5-32:** Gráfico de dispersión de frecuencias de patrones 2x2 TI cat 0 vs. algoritmos SNESIM/SISIM

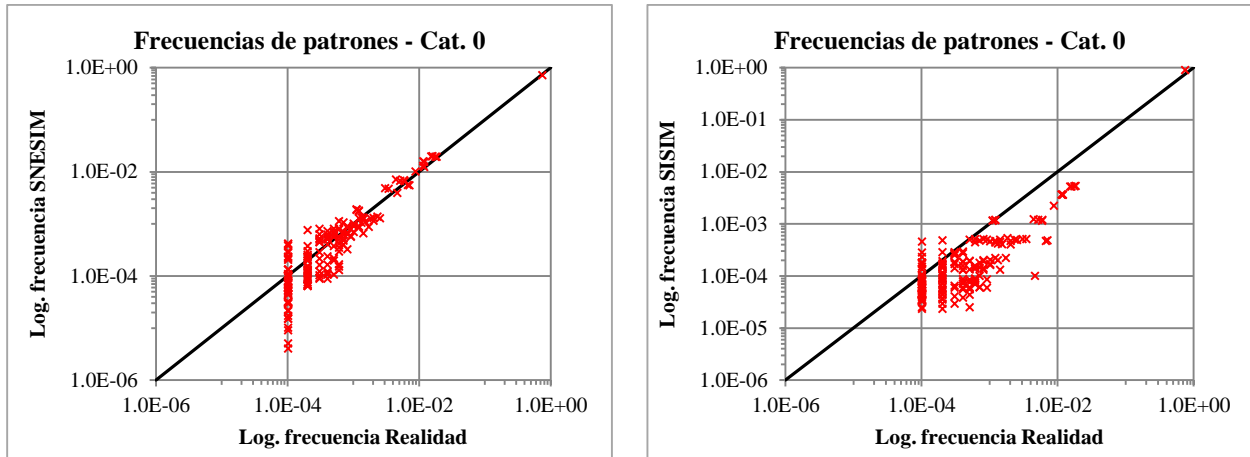


**Figura 5-33:** Gráfico de dispersión de frecuencias de patrones 2x2 realidad cat 0 vs. algoritmos SNESIM/SISIM

En cuanto a los patrones de 3x3, se puede apreciar como para las frecuencias más bajas se presenta una mayor dispersión (debido a la escala logarítmica, en donde un pequeño cambio de frecuencia se traduce en una representación más dispersa). En este caso, se tiene igualmente que SNESIM genera patrones con frecuencias más similares a las de la TI y la realidad; tanto para la categoría 0, como para la 3.

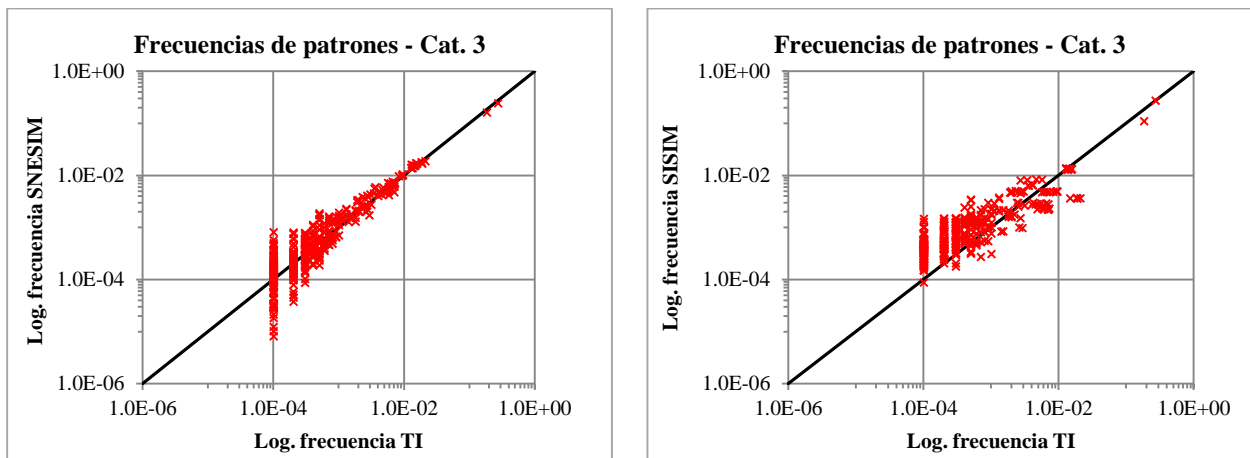


**Figura 5-34:** Gráfico de dispersión de frecuencias de patrones 3x3 TI cat 0 vs. algoritmos SNESIM/SISIM

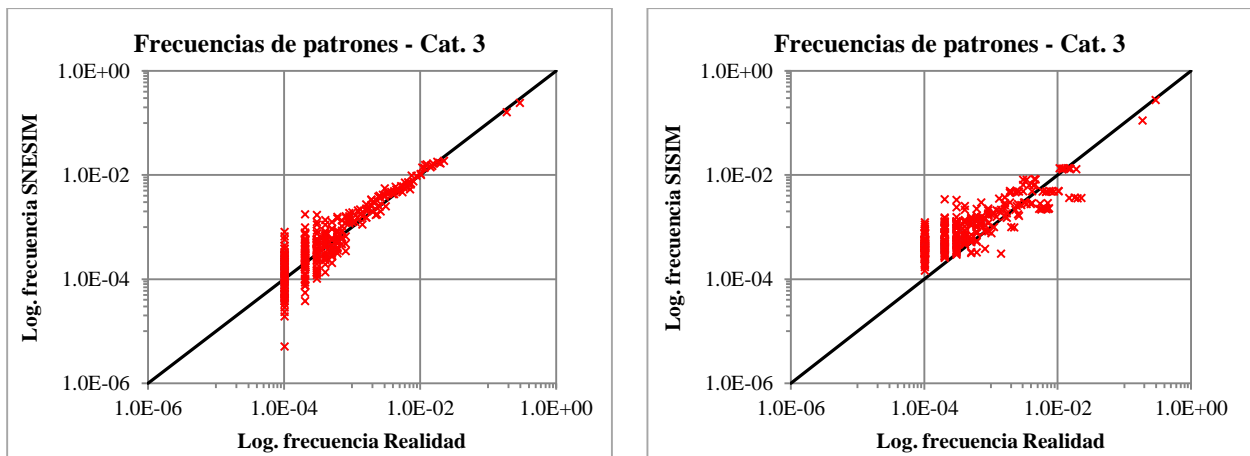


**Figura 5-35:** Gráfico de dispersión de frecuencias de patrones 3x3 realidad cat 0 vs. algoritmos SNESIM/SISIM

Se destaca en las figuras anteriores que las frecuencias de patrones obtenidas del algoritmo SISIM están en gran medida por encima de las frecuencias de la imagen de entrenamiento y la imagen exhaustiva. Un comportamiento contrario se presenta al analizar la categoría 3, en donde se puede apreciar que las frecuencias obtenidas con este algoritmo son inferiores a las que se reportan para la imagen de entrenamiento y la imagen exhaustiva.



**Figura 5-36:** Gráfico de dispersión de frecuencias de patrones 3x3 TI cat 3 vs. algoritmos SNESIM/SISIM



**Figura 5-37:** Gráfico de dispersión de frecuencias de patrones 3x3 realidad cat 3 vs. algoritmos SNESIM/SISIM



## CAPÍTULO 6. CONCLUSIONES

### 6.1. SUPUESTOS FUNDAMENTALES DE LAS VALIDACIONES

Al momento de efectuar las validaciones de simulaciones geoestadísticas, es importante considerar los supuestos fundamentales sobre los que se basan en ocasiones los algoritmos o las métricas mismas para aplicarlas. En este sentido, surgen dos temas de gran relevancia a considerar: por un lado la Estacionariedad, principio básico de la geoestadística, inclusive con inferencia en las simulaciones multi-punto y por otro lado la inferencia.

#### Estacionariedad

- Tal como se mencionó en el estado del arte, algunas técnicas geoestadísticas se basan en principio de estacionariedad para poder obtener resultados, a fin de lograr inferir la función de distribución dada la imposibilidad de conocer la variable aleatoria en todo el dominio. Este concepto tiene implicancias al momento de efectuar la inferencia (tanto en modelos paramétricos, como en los no paramétricos) e igualmente al realizar las validaciones de estos modelos.
- Si bien las técnicas multi-punto pretenden generar modelos con patrones en muchos casos no estacionarios, es importante tener en cuenta que cuando las simulaciones multi-punto fueron desarrolladas, se efectuó una asunción importante para inferir los momentos estadísticos multi-punto desde la imagen de entrenamiento, la cual estaba relacionada con la estacionariedad del campo de estudio (Strebelle, 2012). Esto exige que los patrones, geometrías, proporciones de una categoría sean relativamente homogéneos a lo largo de la imagen de entrenamiento para poder reproducir un patrón mediante el modelo multi-punto. Dado esta restricción inclusive se han desarrollado algunas variantes de algoritmos como SNESIM y FILTERSIM, mediante la inclusión de regiones y conceptos de rotación y afinidad incluidos en SGeMS (Wu. et al, 2007). Estos temas no son abordados en el presente estudio pero podrían ser tenidos en cuenta para futuros complementos de la propuesta que acá se presenta para efectuar validaciones acerca de que la imagen de entrenamiento cumpla esta propiedad (como paso previo a las modelaciones).
- Dado entonces la propiedad subyacente de estacionariedad requerida en el campo aleatorio a ser reproducido, algunos test exigen que se cumpla esta condición para su correcta aplicación. Es el caso de los test estadísticos, que los hace bien exigentes entonces al momento de tratar de ser empleados como métricas de validación en casos donde las variables claramente no presentan comportamientos estacionarios. Esto hace prácticamente inviables de aplicar este tipo de test en las geociencias dado que sus resultados no serán acordes a su formalismo.
- Dado que el tema de la estacionariedad de la imagen de entrenamiento no es abordado en el presente estudio puesto que el enfoque del trabajo está relacionado a la validación de los resultados obtenidos producto de una modelación y no a la selección de la imagen de entrenamiento más adecuada o que cumpla ciertos requisitos (por ejemplo compatibilidad con eventos de datos condicionantes o estacionariedad); es importante considerar que se han efectuado algunas propuestas relacionadas a validar la estacionariedad de la imagen de entrenamiento. Mirowski. et al (2009) propone verificar la estacionariedad de la imagen de

entrenamiento para efectuar las simulaciones en términos de la orientación, escala y distribución de categorías.

### **Inferencia.**

- La inferencia es empleada para caracterizar la función aleatoria de manera que se pueda reproducir mediante simulaciones ciertos parámetros. En el caso de los modelos paramétricos (tales como el modelo gaussiano, del cual hace parte el SGSIM) se infiere la distribución por medio de estadísticas de dos puntos como lo es el variograma; y este a su vez se puede inferir a partir de datos discretos obtenidos de muestreos dispersos en el espacio. Es por eso fundamental para este tipo de simulaciones realizar las validaciones de las estadísticas básicas y del variograma mismo.
- La inferencia en las simulaciones multi-punto requieren de una fuente exhaustiva de información, que suministre la frecuencia de patrones que permiten definir la función de distribución condicional, dado que el variograma no permite modelar algunas formas o comportamientos que pretenden simular los algoritmos multi-punto. Dicha fuente exhaustiva permite reproducir las relaciones de más alto orden logrando modelar patrones más complejos. Esta fuente de información corresponde a la imagen de entrenamiento. La construcción entonces del algoritmo mismo, exige validar sus resultados con métricas de más alto orden para verificar estas relaciones espaciales, como lo son cumulantes espaciales de orden superior, funciones de densidad de múltiples puntos, funciones de conectividad.
- Es importante considerar que para poder inferir adecuadamente el campo aleatorio (es decir, tener una inferencia representativa) mediante una fuente exhaustiva (imagen de entrenamiento) se requeriría que esta fuera lo suficientemente grande. Sin embargo, dado que difícilmente se tendrá una imagen exhaustiva disponible tan grande para generar el campo aleatorio, las simulaciones multi-punto difícilmente podrán generar un modelo de campo aleatorio bien definido y los resultados de las simulaciones con estos algoritmos dependerán en gran medida de una adecuada implementación de sus parámetros (Emery y Lantuéjoul, 2013).

## **6.2. DISCUSIÓN SOBRE MÉTRICAS DE VALIDACIÓN**

Una vez puestas en práctica las diferentes métricas de validación seleccionadas, es importante sacar conclusiones sobre su desempeño para validar las simulaciones; con el fin de aportar ideas para una propuesta metodológica o guía de buenas prácticas a implementarse en el ámbito de simulaciones geoestadísticas.

En este orden de ideas, se presentan a continuación algunos temas a destacar como resultado de los casos prácticos descritos.

- Es fundamental considerar siempre la verificación de los datos condicionantes; pues son el punto de partida para las simulaciones, dado que es lo que se conoce de la realidad. Independiente si los algoritmos son paramétricos (basados en variogramas, es decir, basados en un modelo); o si son no paramétricos (basados en datos o modelos conceptuales como el caso de la imagen de entrenamiento o una función indicador) es obligatorio efectuar este chequeo puesto que son la única información veraz que se conoce y el modelo debería

reproducirla. El no reproducir los datos duros exigiría rechazar el modelo (siempre que la simulación sea condicional).

Para efectuar dicha validación se deberá tener en cuenta si los datos duros son desplazados a nodos de la grilla de simulación como parte del algoritmo empleado, para que al momento de efectuar el chequeo no se obtengan conclusiones erradas.

- En los casos que el campo aleatorio de la variable en estudio sea estacionario y ergódico, debería propenderse por desarrollar test estadísticos que determinen si las fluctuaciones encontradas en el modelo simulado son consistentes con las estadísticas del modelo del campo aleatorio. Sin embargo, deberá tenerse claro que los resultados son concluyentes en cuanto a que el modelo no reproduce adecuadamente las estadísticas de interés; pero que ello no indica que la aceptación del test sugiera que el modelo es correcto, si no que no existe evidencia significativa para rechazarlo.

Adicionalmente, es importante tener en cuenta los supuestos sobre los cuales se desarrollan los test estadístico, lo cual los hacen difíciles de aplicar en muchas circunstancias prácticas como medios de validación. En términos generales podría decirse, que serían de mayor utilidad bajo el análisis de simulaciones no condicionales, donde al menos se respete el supuesto de independencia para las realizaciones obtenidas en la simulación; sin embargo esta situación sería factible para la simulación secuencial gaussiana (alterando el proceso de simulación para posterior a los test efectuar el condicionamiento), no así para las simulaciones multi-punto o simulación secuencial de indicadores.

Podría decirse que a medida que se vayan incluyendo datos condicionantes, se irá perdiendo mayor validez del test estadístico.

- Aunque este estudio se analiza la reproducción de una curva de selectividad (fracción de volumen vs umbral de corte), bien podrían efectuarse análisis adicionales con curvas de selectividad que incluyan otras funciones de transferencia. Igualmente a dicho análisis se podría adicionar el cambio de soporte y analizar los resultados en procura de lograr un mejor entendimiento de la variable a modelar y poder mejorar los resultados que se pretenden lograr con el modelamiento.
- Si bien es cierto algunos modelos de simulación, como es el caso de los algoritmos de las simulaciones multi-punto, no consideran el variograma en su construcción; es importante considerar su chequeo dado que este mide la relación espacial de la variable entre dos puntos, o en su defecto chequear parámetros relacionados como la covarianza o el correlograma.

Esta métrica de validación tiene la ventaja que puede ser calculada tanto para simulaciones paramétricas, así como para validaciones no paramétricas; permitiendo en los primeros comparar la reproducción de variogramas modelados y en los segundos reproducción del variograma de la imagen de entrenamiento o el variograma de indicadores. Para el caso particular de este estudio se tuvo la oportunidad de comparar igualmente con una imagen exhaustiva (considerada la realidad), lo cual nunca ocurrirá en la vida diaria dado el desconocimiento que se tiene de esta y de la cual solo se conocen algunos sitios.

- Al momento de efectuar validaciones de simulaciones multi-punto es importante considerar estadísticas o métricas de orden superior, dado que este tipo de simulaciones igualmente imponen estadísticas de orden superior en su cálculo. Métricas como las mencionadas hasta el momento en este numeral no permiten validar la reproducción de ciertas características como patrones, conectividad, entre otros elementos distintivos por los cuales son empleados este tipo de algoritmos.

Ello hace entonces necesario implementar el uso de técnicas como las acá discutidas. Se sugiere entonces tener en cuenta los cumulantes de orden superior, puesto que permiten caracterizar algunos aspectos como conectividad, distinguir anisotropías o el tamaño de cuerpos dependiendo de la configuración y distancias empleadas (Dimitrakoupulus y Musthapa, 2010), lo cual es una herramienta valiosa a la hora de verificar este tipo de aspectos en un modelo simulado. De igual forma, permite efectuar una comparación entre diferentes modelos para determinar un mejor desempeño capturando comportamientos espaciales más complejos o bien para calibrar los modelos y compararlos luego de efectuados dichos cambios. Esta herramienta podría ser empleada igualmente para validar simulaciones paramétricas sobre las cuales se quiere verificar la reproducción de cierta anisotropía por ejemplo.

- De forma similar, las estadísticas multi-puntos como la función de densidad de múltiples puntos y los gráficos de frecuencia son importantes toda vez que reflejan el comportamiento de un modelo para reproducir patrones. Dado que las simulaciones multi-punto subyacen sobre el concepto de patrones obtenidos a partir de la imagen de entrenamiento es casi obligatorio determinar si los resultados obtenidos presentan patrones similares a los de la imagen de entrenamiento.

Si bien esta técnica permite comparar entre dos modelos (simulado e imagen de entrenamiento), para determinar su similaridad en cuanto reproducción de patrones definidos; bien podría ser empleada igualmente para validar una simulación paramétrica a partir de patrones obtenidos de los datos condicionantes y verificar si el modelo simulado reproduce los patrones allí encontrados. Si bien este tipo de métrica no fue aplicada a los casos prácticos, se encuentra implementada en el trabajo desarrollado por Pérez (2013). Igualmente se puede aplicar para comparar los resultados obtenidos contra otro tipo de algoritmo o modelo de interés.

- Es importante considerar que las simulaciones pretenden caracterizar la incertidumbre ocasionada por el desconocimiento de la realidad a través de múltiples escenarios ficticios de esta. Esto conlleva a la necesidad también de verificar la distribución de dicha incertidumbre y por ello similar a como se suele realizar con una validación cruzada o un Jack-knife para una estimación; es importante evaluar si la distribución de incertidumbre es acorde a probabilidades teóricas, que para el caso acá considerado estuvo relacionado a 9 deciles seleccionados. Para eso un gráfico de precisión, suele ser la respuesta y debería ser entonces considerado como parte de las validaciones de incertidumbre.
- Finalmente, conforme a los resultados obtenidos de las validaciones, debería decirse que los modelos obtenidos con los algoritmos SISIM y SGSIM presentan desempeños no muy adecuados para este caso particular conforme a los parámetros de simulación empleados. Esto no indica que sean algoritmos de peor desempeño “*per se*” que FILTERSIM o

SNESIM, lo cual dependerá del comportamiento de la variable de interés y de una buena calibración de los parámetros de simulación; motivo por el cual podría decirse que son modelos que requieren mayores ajustes que podrían efectuarse a través de análisis de sensibilidad ajustando los parámetros de modelación y determinar posteriormente si hay mejoras en su desempeño. En caso contrario a esta decisión, y de no obtenerse mejoras significativas debería decirse que estos modelos deben descartarse.

### **6.3. BUENAS PRÁCTICAS DE VALIDACIÓN**

Finalmente, como parte de los resultados del presente estudio, se tiene una propuesta encaminada a determinar unas buenas prácticas de validación. Es decir, definir qué elementos deberían ser incluidos al momento de efectuar las validaciones geoestadísticas, teniendo en cuenta que el trabajo está enfocado en gran parte a las simulaciones multi-punto y que se basa en los niveles de validación analizados a lo largo del estudio: 1: Nivel de condicionamiento de datos, 2: Nivel estadístico, 3: Nivel geoestadístico y 4: Nivel de predicción.

Esta propuesta parte de las experiencias recogidas al aplicar algunas de las métricas propuestas en los casos prácticos analizados, así como de la revisión de antecedentes. Con ella se pretende que al momento de modelar una variable a través de simulaciones geoestadísticas se pongan en práctica ciertas medidas mínimas para determinar si el modelo representa o reproduce adecuadamente la variable o fenómeno que se quiere modelar. Esto representa un avance en el estado del arte al proponer un flujo de trabajo que permite abordar de manera ordenada y sistémica el tema de las validaciones, algo que hasta el momento no se encuentra bien establecido en la práctica del modelamiento geoestadístico mediante simulaciones, porque tal como lo asegura Mariethoz y Caers (2015) durante la última década el enfoque ha estado puesto en desarrollar algoritmos y métodos para generar imágenes de entrenamiento y el campo de las validaciones es un área aún en desarrollo.

En la Figura 6-1 se presenta un flujo grama con la metodología propuesta, en el cual es importante destacar un aspecto relacionado a los test estadísticos; que como se pudo observar en los casos prácticos desarrollados en el presente estudio, carecían de validez dado los fuertes supuestos que exigen para su aplicación. Es por esto, que si bien se encuentran dentro de la línea de flujo, es necesario tener muy presente la validez de estos test conforme la variable que se esté modelando y el esfuerzo que puede significar realizar unas validaciones adicionales que permitan determinar si la información de los modelos cumple los supuestos para su aplicación. Para el caso particular del supuesto de estacionariedad es necesario verificar este comportamiento, lo cual no siempre es fácil de efectuar. En la metodología propuesta se presenta una forma subjetiva asociada a la presencia de derivas en diferentes direcciones; pero mayores esfuerzos serían necesarios para determinar la estacionariedad de una variable (dentro de las cuales podría remitirse a trabajos propuestos por Mirowski et al, (2009)). Futuras investigaciones podrían desarrollarse en dicha área a fin de complementar la propuesta acá presentada.

Es importante aclarar, que la propuesta metodológica presentada no permite establecer umbrales de aceptación o rechazo de un modelo en particular, puesto que ello obedece a los objetivos que el modelador se plantee. Adicionalmente depende también del tipo de variable, fenómeno o proceso a modelar, como se distribuye este en el espacio, su comportamiento, la información disponible con que se cuenta. Esta propuesta permite al modelador tener conocimiento de si su modelo puede tener deficiencias al no reproducir las características esperadas y por ende darle información sobre la validez del mismo (es decir, generar ciertas alertas sobre el desempeño de su modelo), sobre lo

cual tendrá que tomar la decisión si continúa empleado éste para la toma de decisiones, aun teniendo ciertas limitaciones, o si por el contrario decide rechazarlo pero todo bajo el criterio mismo del modelador. Un trabajo posterior podría enfocarse a desarrollar posibles umbrales de rechazo para determinados tipos de modelos, no tan amplio, como el caso de las geociencias.

Finalmente, se aclara que en el flujo grama se presentan (destacado en texto rojo) otro tipo de métricas no puestas en práctica en el presente trabajo, tal como se sustentó en el capítulo 5, que igualmente podrían ser aplicadas para consolidar un mayor nivel de validación (Pérez, 2013 y Mariethoz y Caers, 2014).

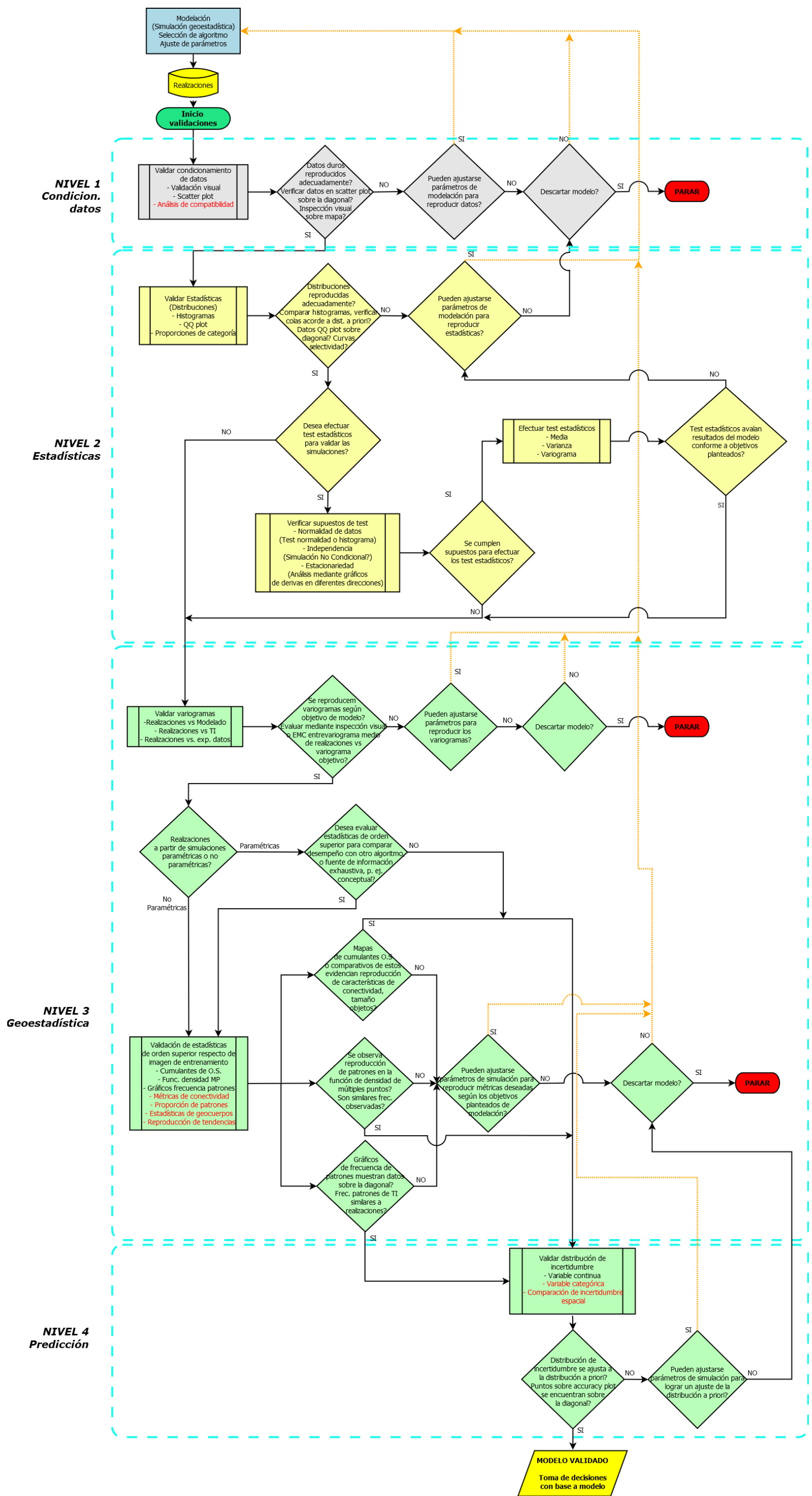


Figura 6-1: Diagrama de flujo para validar simulaciones

## BIBLIOGRAFÍA

- AHMED, Shakeel. Application of geostatistics in hydrosiences. En Groundwater. Springer Netherlands, 2007. p. 78-111.
- ASGHARI, Omid; SOLTANI, Fatemeh; HASSAN, B. A. The comparison between sequential gaussian simulation (SGS) of Choghart ore deposit and geostatistical estimation through ordinary kriging. Australian Jour. Basic Applied Sci, 2009, vol. 3, no 1, p. 330-341.
- BOISVERT, Jeff B.; PYRCZ, Michael J.; DEUTSCH, Clayton V. Multiple-point statistics for training image selection. Natural Resources Research, 2007, vol. 16, no 4, p. 313-321.
- BOISVERT, Jeff B.; PYRCZ, Michael J.; DEUTSCH, Clayton V. Multiple point metrics to assess categorical variable models. Natural resources research, 2010, vol. 19, no 3, p. 165-175.
- CAERS, Jef; SRINIVASAN, Sanjay. Statistical pattern recognition and geostatistical data integration. En Soft Computing for Reservoir Characterization and Modeling. Physica-Verlag HD, 2002. p. 355-386.
- CHILES, Jean-Paul; LANTUÉJOL, Christian. Prediction by conditional simulation: models and algorithms. En Space, Structure and Randomness. Springer New York, 2005. p. 39-68.
- DE IACO, S.; MAGGIO, S. Validation techniques for geological patterns simulations based on variogram and multiple-point statistics. Mathematical Geosciences, 2011, vol. 43, no 4, p. 483-500.
- DEUTSCH, Clayton V.; JOURNAL, André G. Geostatistical software library and user's guide. New York, 1992.
- DIMITRAKOPOULOS, Roussos; MUSTAPHA, Hussein; GLOAGUEN, Erwan. High-order statistics of spatial random fields: exploring spatial cumulants for modeling complex non-Gaussian and non-linear phenomena. Mathematical Geosciences, 2010, vol. 42, no 1, p. 65-99.
- EMERY, Xavier. Statistical tests for validating geostatistical simulation algorithms. Computers & Geosciences, 2008, vol. 34, no 11, p. 1610-1620.
- EMERY, X. Geoestadística, Universidad de Chile, Facultad de ciencias físicas y matemáticas, Universidad de Chile, Santiago, Chile, pp. 21-28, 2013.
- EMERY, Xavier; LANTUÉJOL, Christian. Can a training image be a substitute for a random field model?. Mathematical Geosciences, 2013, vol. 46, no 2, p. 133-147.
- ERSOY, Adem; YUNSEL, Tayfun Yusuf; ATICI, Ümit. Geostatistical conditional simulation for the assessment of contaminated land by abandoned heavy metal mining. Environmental toxicology, 2008, vol. 23, no 1, p. 96-109.
- GOOVAERTS, Pierre. Geostatistical modelling of uncertainty in soil science. Geoderma, 2001, vol. 103, no 1, p. 3-26.
- JOURNAL, Andre G., et al. Stochastic imaging of the Wilmington clastic sequence. SPE formation Evaluation, 1993, vol. 8, no 01, p. 33-40.



- KING, Susan L. Sequential Gaussian simulation vs. simulated annealing for locating pockets of high-value commercial trees in Pennsylvania. *Annals of Operations Research*, 2000, vol. 95, no 1-4, p. 117-203.
- KLISE, Katherine A., et al. Exploring solute transport and streamline connectivity using lidar-based outcrop images and geostatistical representations of heterogeneity. *Water resources research*, 2009, vol. 45, no 5.
- KRISHNAN, Sunderrajan; JOURNAL, A. G. Spatial connectivity: from variograms to multiple-point measures. *Mathematical Geology*, 2003, vol. 35, no 8, p. 915-925.
- LEUANGTHONG, Oy; MCLENNAN, Jason A.; DEUTSCH, Clayton V. Minimum acceptance criteria for geostatistical realizations. *Natural Resources Research*, 2004, vol. 13, no 3, p. 131-141.
- MARIETHOZ, Gregoire; CAERS, Jef. *Multiple-point Geostatistics: Stochastic Modeling with Training Images*. John Wiley & Sons, 2014.
- MIROWSKI, P., TETZLAFF, D., DAVIES, R., McCORMICK, D., WILLIAMS, N., SIGNER, C., 2009: Stationary Scores on Training Images for Multipoint Geostatistics. *Mathematical Geosciences*, vol. 41, pp. 447-474.
- MUSTAPHA, Hussein; DIMITRAKOPOULOS, Roussos. A new approach for geological pattern recognition using high-order spatial cumulants. *Computers & Geosciences*, 2010, vol. 36, no 3, p. 313-334.
- OLEA, Ricardo A. Building on crossvalidation for increasing the quality of geostatistical modeling. *Stochastic Environmental Research and Risk Assessment*, 2012, vol. 26, no 1, p. 73-82.
- ORTIZ, Julián M.; DEUTSCH, Clayton V. Indicator simulation accounting for multiple-point statistics. *Mathematical Geology*, 2004, vol. 36, no 5, p. 545-565.
- ORTIZ, Julián M.; LEUANGTHONG, Oy. *A Practical Approach to Validate the Variogram Reproduction from Geostatistical Simulation*. 2012.
- PÉREZ STRUTZ, Cristián. Training image selection and model validation using multiple point statistics [en línea]. Santiago, Chile: Universidad de Chile - Facultad de Ciencias Físicas y Matemáticas, 2013
- PYRCZ, Michael J.; DEUTSCH, Clayton V. *Geostatistical reservoir modeling*. Oxford university press, 2014.
- REMY, Nicolas; BOUCHER, Alexandre; WU, Jianbing. *Applied geostatistics with SGeMS: a user's guide*. Cambridge University Press, 2009.
- REZAEI, Hassan, et al. Multiple-point geostatistical simulation of dykes: application at Sungun porphyry copper system, Iran. *Stochastic Environmental Research and Risk Assessment*, 2014, vol. 28, no 7, p. 1913-1927.
- SARGENT, Robert G. Verification and validation of simulation models. En *Proceedings of the 37th conference on Winter simulation*. winter simulation conference, 2005. p. 130-143..
- SCHLÜTER, Steffen; VOGEL, Hans-Jörg. On the reconstruction of structural and functional properties in random heterogeneous media. *Advances in Water Resources*, 2011, vol. 34, no 2, p. 314-325..

- SOLENG, Harald H.; SYVERSVEEN, Anne Randi; KOLBJØRNSSEN, Odd. Comparing Facies Realizations–Defining Metrics on Realization Space. En 10th European Conference on the Mathematics of Oil Recovery. 2006.
- STREBELLE, Sebastien. Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, 2002, vol. 34, no 1, p. 1-21.
- STREBELLE, Sebastien; CHEVRON, E. T. C. Multiple-Point Geostatistics: from Theory to Practice. 2012. Ninth International Geostatistics Congress, Oslo, Norway.
- WACKERNAGEL, Hans. Multivariate geostatistics: an introduction with applications. Springer Science & Business Media, 2013. p 213-220.
- WESTERN, Andrew W.; BLÖSCHL, Günter; GRAYSON, Rodger B. Toward capturing hydrologically significant connectivity in spatial patterns. *Water Resources Research*, 2001, vol. 37, no 1, p. 83-97.
- ZHANG, Tuanfeng; SWITZER, Paul; JOURNAL, Andre. Filter-based classification of training image patterns for spatial simulation. *Mathematical Geology*, 2006, vol. 38, no 1, p. 63-80.
- ZHAO, Shan, et al. Thickness, porosity, and permeability prediction: comparative studies and application of the geostatistical modeling in an Oil field. *Environmental Systems Research*, 2014, vol. 3, no 1, p. 1-24.

# ANEXOS

## ANEXO A – Parámetros empleados en simulaciones

### PARAMETROS FILTERSIM

```
<parameters> <algorithm name="filtersim_cont" />
  <GridSelector_Sim value="Simulation" region="" />
  <Property_Name_Sim value="FILTERSIM" />
  <Nb_Realizations value="100" />
  <Seed value="211175" />
  <PropertySelector_Training grid="TI" property="v" region="" />
  <Patch_Template_ADVANCED value="7 7 1" />
  <Scan_Template value="11 11 1" />
  <Trans_Result value="1" />
  <Hard_Data grid="Samples" property="v" region="" />
  <Use_SoftField value="0" />
  <Region_Indicator_Prop value="SGSIM__real0" />
  <Active_Region_Code value="" />
  <Use_Previous_Simulation value="0" />
  <Previous_Simulation_Prop value="SGSIM__real0" />
  <Use_Region value="0" />
  <Nb_Multigrids_ADVANCED value="3" />
  <Debug_Level value="0" />
  <Cmin_Replicates value="10 10 10" />
  <Data_Weights value="0.5 0.3 0.2" />
  <CrossPartition value="0" />
  <KMeanPartition value="1" />
  <Nb_Clusters_ADVANCED value="200" />
  <Nb_Clusters_ADVANCED2 value="2" />
  <Use_Normal_Dist value="1" />
  <Use_Score_Dist value="0" />
  <Filter_Default value="1" />
  <Filter_User_Define value="0" />
</parameters>
```

### PARAMETROS SGSIM

```
<parameters> <algorithm name="sgsim" />
  <Grid_Name value="Simulation" region="" />
  <Property_Name value="SGSIM" />
  <Nb_Realizations value="100" />
  <Seed value="14071789" />
  <Kriging_Type value="Simple Kriging (SK)" />
  <Trend value="0 0 0 0 0 0 0 0 0" />
  <Local_Mean_Property value="" />
  <Assign_Hard_Data value="1" />
  <Hard_Data grid="Samples" property="v" region="" />
  <Max_Conditioning_Data value="12" />
  <Max_Conditioning_Simul_Data value="12" />
  <Search_Ellipsoid value="30 30 30
0 0 0" />
  <AdvancedSearch use_advanced_search="0"></AdvancedSearch>
  <Use_Target_Histogram value="1" />
  <nonParamCdf ref_in_distribution="0" ref_on_file="0" ref_on_grid="1"
```

```

break_ties = "0" distribution = "" filename = "" grid = "Samples" region = ""
property = "v">
<LTI_type function = "Power" extreme = "-10" omega = "3" />
<UTI_type function = "Power" extreme = "0.2" omega = "0.333" />
</nonParamCdf>
  <Variogram nugget = "0.05" structures_count = "1" >
    <structure_1 contribution = "0.95" type = "Spherical" >
      <ranges max = "32" medium = "16" min = "0" />
      <angles x = "0" y = "0" z = "0" />
    </structure_1>
  </Variogram>
</parameters>

```

## **PARAMETROS SNESIM**

```

<parameters> <algorithm name = "snesim_std" />
  <Hard_Data grid = "Samples" region = "" property = "facies" />
  <use_pre_simulated_gridded_data value = "0" />
  <Use_ProbField value = "0" />
  <ProbField_properties count = "0" value = "" />
  <TauModelObject value = "1 1" />
  <use_vertical_proportion value = "0" />
  <Use_Affinity value = "0" />
  <Use_Rotation value = "0" />
  <Cmin value = "1" />
  <Constraint_Marginal_ADVANCED value = "0" />
  <resimulation_criterion value = "-1" />
  <resimulation_iteration_nb value = "1" />
  <Nb_Multigrids_ADVANCED value = "3" />
  <Debug_Level value = "0" />
  <Subgrid_choice value = "0" />
  <expand_isotropic value = "1" />
  <expand_anisotropic value = "0" />
  <aniso_factor value = "
" />
  <GridSelector_Sim value = "Simulation" region = "" />
  <Property_Name_Sim value = "SNESIM" />
  <Nb_Realizations value = "100" />
  <Seed value = "211175" />
  <PropertySelector_Training grid = "TI" region = "" property = "facies" />
  <Nb_Facies value = "6" />
  <Marginal_Cdf value = "0.057935497 0.04970101 0.145475934 0.447505147
0.189589256 0.109793158
" />
  <Max_Cond value = "60" />
  <Search_Ellipsoid value = "30 30 30
0 0 0" />
</parameters>

```

## **PARAMETROS SISIM**

```

<parameters> <algorithm name = "sisim" />
  <Grid_Name value = "Simulation" region = "" />
  <Property_Name value = "NEW3SISIM" />
  <Nb_Realizations value = "100" />
  <Seed value = "14071789" />
  <Categorical_Variable_Flag value = "1" />

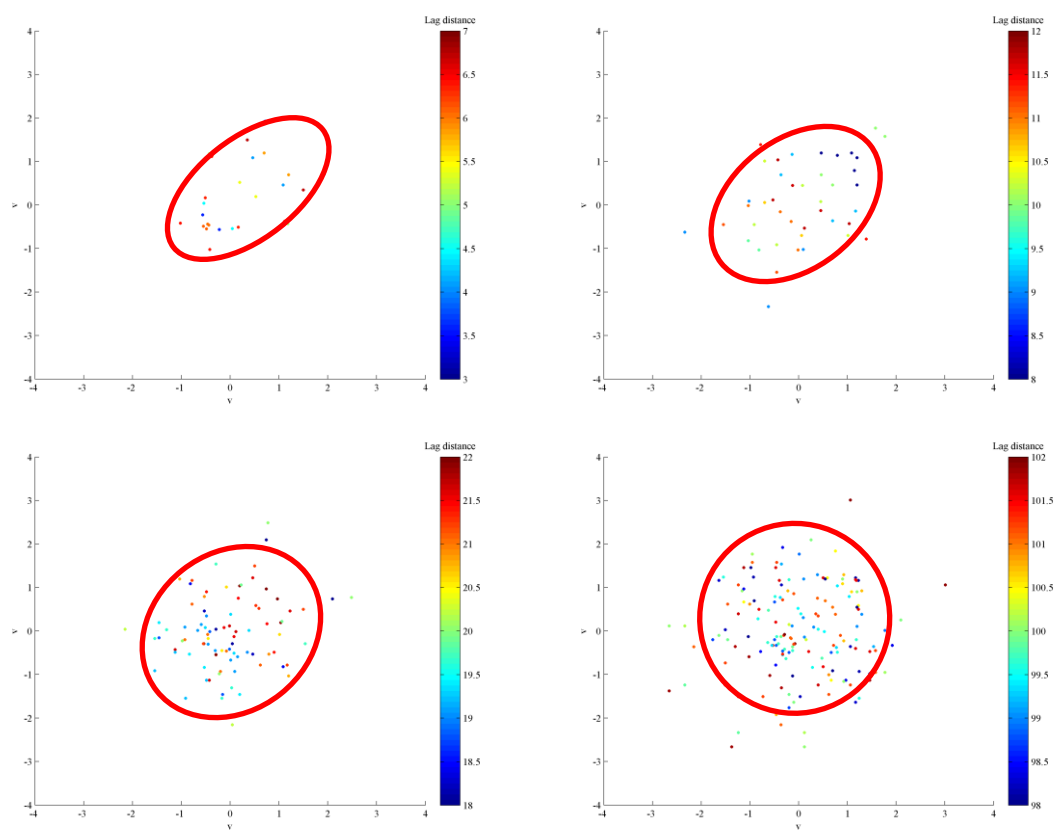
```

```

<Nb_Indicators value="6" />
<Marginal_Probabilities value="0.019607843 0.049019608 0.18627451
0.470588235 0.176470588 0.098039216
" />
<lowerTailCdf function="Power" extreme="0" omega="3" />
<upperTailCdf function="Power" extreme="0" omega="0.333" />
<Median_Ik_Flag value="0" />
<Full_Ik_Flag value="1" />
<Hard_Data_Grid value="Samples_Ind" region="" />
<Hard_Data_Property value="Samples" />
<Assign_Hard_Data value="1" />
<coded_grid value="Samples_Ind" region="" />
<coded_props count="6" value="Ind_0;Ind_1;Ind_2;Ind_3;Ind_4;Ind_5" />
<Max_Conditioning_Data value="12" />
<Search_Ellipsoid value="15 15 15
0 0 0" />
<AdvancedSearch use_advanced_search="0"></AdvancedSearch>
<Variogram_Full_Ik nugget="0.025" structures_count="1" >
<structure_1 contribution="0.0238216" type="Exponential" >
<ranges max="9" medium="9" min="0" />
<angles x="0" y="0" z="0" />
</structure_1>
</Variogram_Full_Ik>
<Variogram_Full_Ik_2 nugget="0.028" structures_count="1" >
<structure_1 contribution="0.0142609" type="Exponential" >
<ranges max="4.5" medium="4.5" min="0" />
<angles x="0" y="0" z="0" />
</structure_1>
</Variogram_Full_Ik_2>
<Variogram_Full_Ik_3 nugget="0.06" structures_count="1" >
<structure_1 contribution="0.074202" type="Exponential" >
<ranges max="7" medium="7" min="0" />
<angles x="0" y="0" z="0" />
</structure_1>
</Variogram_Full_Ik_3>
<Variogram_Full_Ik_4 nugget="0.035" structures_count="1" >
<structure_1 contribution="0.211384" type="Exponential" >
<ranges max="22" medium="22" min="0" />
<angles x="0" y="0" z="0" />
</structure_1>
</Variogram_Full_Ik_4>
<Variogram_Full_Ik_5 nugget="0.01" structures_count="1" >
<structure_1 contribution="0.144207" type="Exponential" >
<ranges max="25" medium="25" min="0" />
<angles x="0" y="0" z="0" />
</structure_1>
</Variogram_Full_Ik_5>
<Variogram_Full_Ik_6 nugget="0.02" structures_count="1" >
<structure_1 contribution="0.081399" type="Exponential" >
<ranges max="9" medium="9" min="0" />
<angles x="0" y="0" z="0" />
</structure_1>
</Variogram_Full_Ik_6>
</parameters>

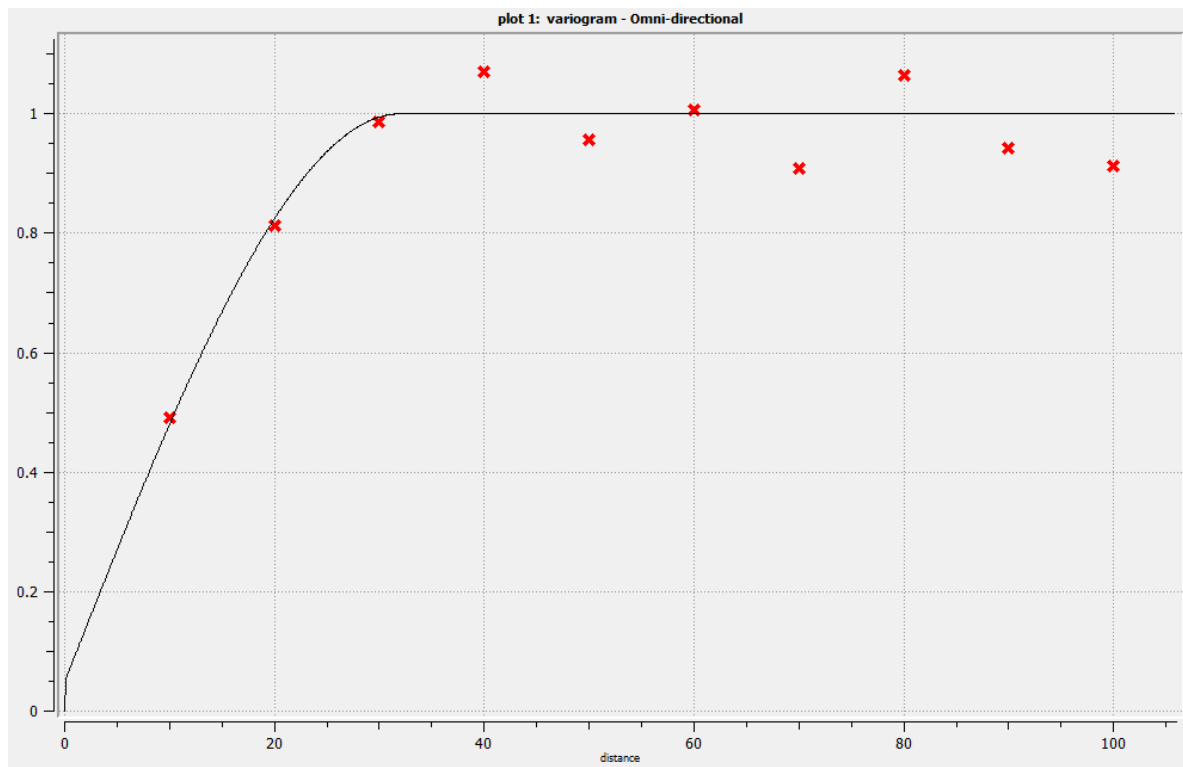
```

## ANEXO B – Gráficos de correlación diferida

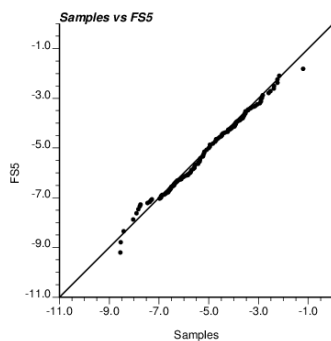


Gráficos de correlación diferida a 5, 10, 20 y 100 m.

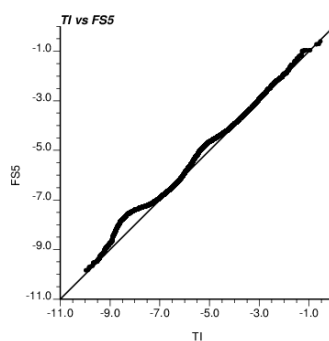
## ANEXO C – Variograma modelado SGSIM



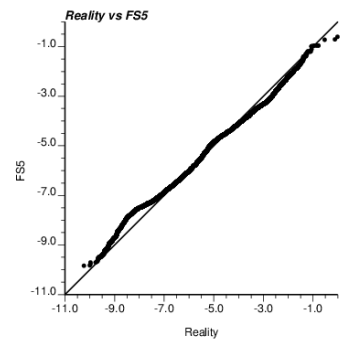
## ANEXO D – QQ Plots para diferentes realizaciones de FILTERSIM



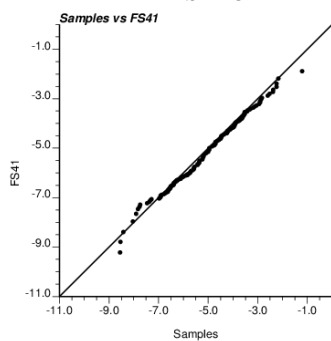
a. QQ-plot Muestras – realización FILTERSIM 5



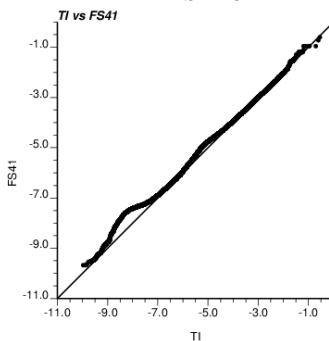
e. QQ-plot TI – realización FILTERSIM 5



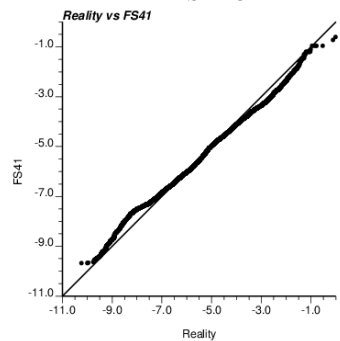
i. QQ-plot Realidad – realización FILTERSIM 5



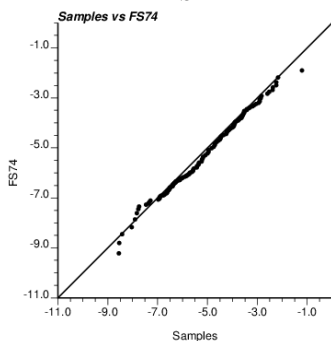
b. QQ-plot Muestras – realización FILTERSIM 41



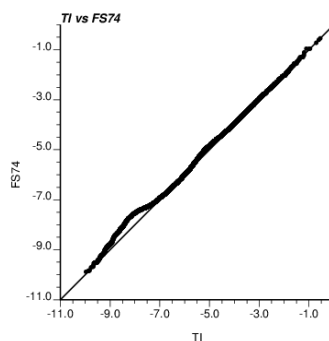
f. QQ-plot TI – realización FILTERSIM 41



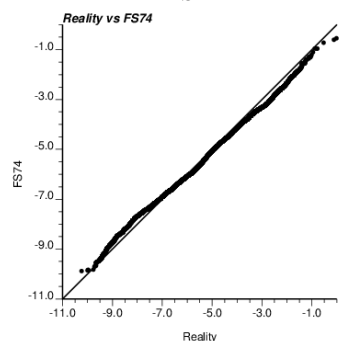
j. QQ-plot Realidad – realización FILTERSIM 41



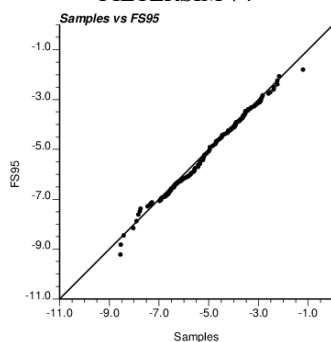
c. QQ-plot Muestras – realización FILTERSIM 74



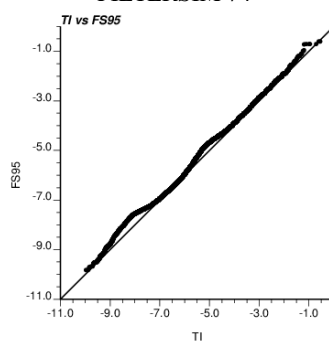
g. QQ-plot TI – realización FILTERSIM 74



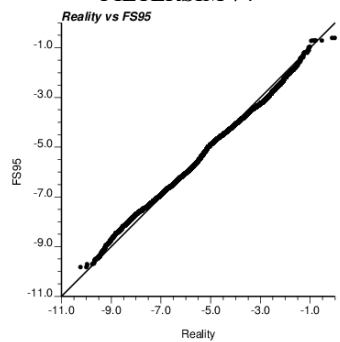
k. QQ-plot Realidad – realización FILTERSIM 74



d. QQ-plot Muestras – realización FILTERSIM 95



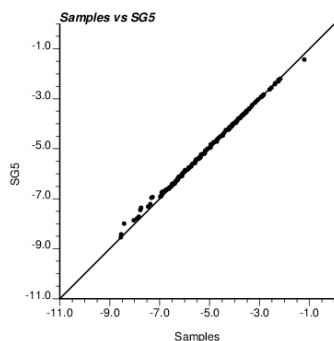
h. QQ-plot TI – realización FILTERSIM 95



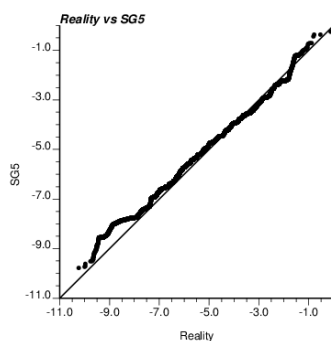
l. QQ-plot Realidad – realización FILTERSIM 95



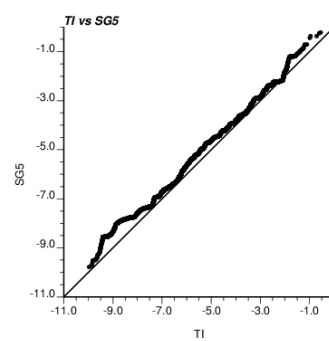
## ANEXO E – QQ Plots para diferentes realizaciones de SGSIM



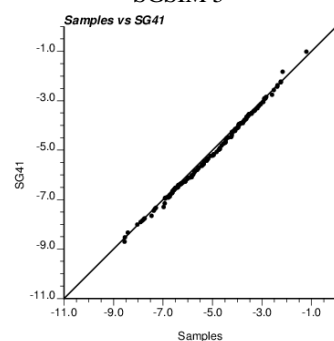
a. QQ-plot Muestras – realización SGSIM 5



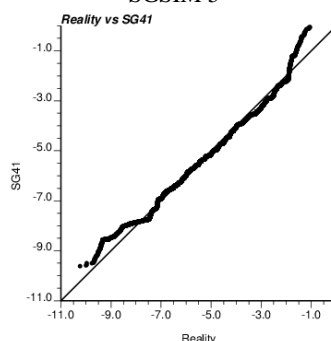
e. QQ-plot Realidad – realización SGSIM 5



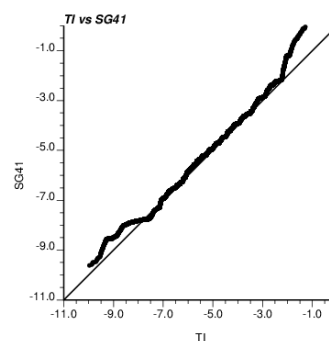
i. QQ-plot TI – realización SGSIM 5



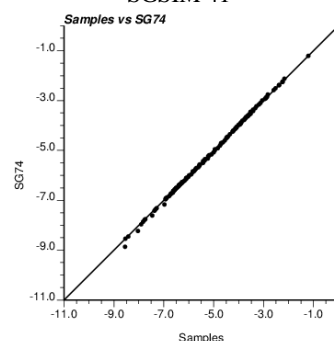
b. QQ-plot Muestras – realización SGSIM 41



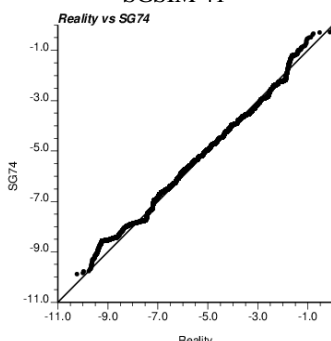
f. QQ-plot Realidad – realización SGSIM 41



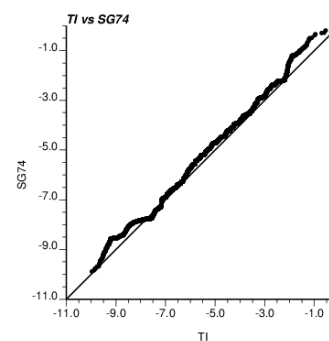
j. QQ-plot TI – realización SGSIM 41



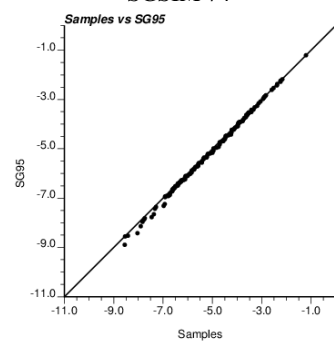
c. QQ-plot Muestras – realización SGSIM 74



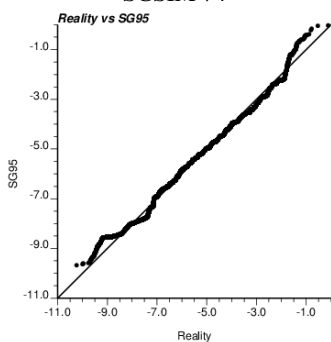
g. QQ-plot Realidad – realización SGSIM 74



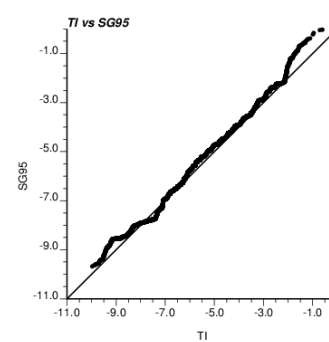
k. QQ-plot TI – realización SGSIM 74



d. QQ-plot Muestras – realización SGSIM 95



h. QQ-plot Realidad – realización SGSIM 95



l. QQ-plot TI – realización SGSIM 95

## ANEXO F – Reportes de Test estadísticos

<b>FILTERSIM vs Media Samples</b> [T,Tmax] = Student_mean(FILTERSIM,27000,100,-5.10669,0.05);  Student's test on the mean ***** Student statistics: 22.9577 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is rejected. Simulation algorithm fails	<b>SGSIM vs Media Samples</b> [T,Tmax] = Student_mean(SGSIM,27000,100,-5.10669,0.05);  Student's test on the mean ***** Student statistics: 2.5702 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is rejected. Simulation algorithm fails
<b>FILTERSIM vs Media TI</b> [T,Tmax] = Student_mean(FILTERSIM,27000,100,-5.31489,0.05);  Student's test on the mean ***** Student statistics: 21.4271 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is rejected. Simulation algorithm fails	
<b>FILTERSIM vs Varianza Samples</b> [T,Tmax] = Student_var(FILTERSIM,27000,100,2.52742,0.05);  Student's test on the variance ***** Student statistics: 7.0759 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is rejected. Simulation algorithm fails	<b>SGSIM vs Varianza Samples</b> [T,Tmax] = Student_var(SGSIM,27000,100,2.52742,0.05);  Student's test on the variance ***** Student statistics: 7.3818 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is rejected. Simulation algorithm fails
<b>FILTERSIM vs Varianza TI</b> [T,Tmax] = Student_var(FILTERSIM,27000,100,2.51415,0.05);  Student's test on the variance ***** Student statistics: 5.783 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842	

Null hypothesis is rejected. Simulation algorithm fails.	
<b>FILTERSIM (Test <math>T^2</math> Hotelling)</b>	<b>SGSIM (Test <math>T^2</math> Hotelling)</b>
<pre>[T2,T2max] = Hotelling_variog(FS_sc,180,150,1,1,1, 1,10,[1 38.4 18 0 0 0 0],[0.7],1,[0.3],[10 20 30 40 50],[0 0 0 0 0],[0 0 0 0 0],0.05);</pre>	<pre>[T2,T2max] = Hotelling_variog(GS_sc,180,150,1,1,1, 1,10,[1 38.4 18 0 0 0 0],[0.7],1,[0.3],[10 20 30 40 50],[0 0 0 0 0],[0 0 0 0 0],0.05);</pre>
<p>Hotelling's test on the semi-variogram ***** * Hotelling statistics: 131.3737 Numbers of degrees of freedom: 5, 5 Critical value: 45.453</p>	<p>Hotelling's test on the semi-variogram ***** * Hotelling statistics: 146.6062 Numbers of degrees of freedom: 5, 5 Critical value: 45.453</p>
Null hypothesis is rejected. Simulation algorithm fails	Null hypothesis is rejected. Simulation algorithm fails

<b>SNESIM Cat. 0 vs. media Muestras</b>	<b>SISIM Cat. 0 vs. media Muestras</b>
<pre>[T,Tmax] = Student_mean(SN0,10201,100,0.0196078431 372549,0.05);</pre>	<pre>[T,Tmax] = Student_mean(SS_IND0,10201,100,0.01 9608,0.05);</pre>
<p>Student's test on the mean ***** Student statistics: 77.2144 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p>	<p>Student's test on the mean ***** Student statistics: 6.8722 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p>
Null hypothesis is rejected. Simulation algorithm fails	Null hypothesis is rejected. Simulation algorithm fails
<b>SNESIM Cat. 3 vs. media Muestras</b>	<b>SISIM Cat. 3 vs. media Muestras</b>
<pre>[T,Tmax] = Student_mean(SN3,10201,100,0.4705882352 94118,0.05);</pre>	<pre>[T,Tmax] = Student_mean(SS_IND3,10201,100,0.47 0588,0.05);</pre>
<p>Student's test on the mean ***** Student statistics: 8.5947 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p>	<p>Student's test on the mean ***** Student statistics: 14.6098 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p>
Null hypothesis is rejected. Simulation algorithm fails	Null hypothesis is rejected. Simulation algorithm fails
<b>SNESIM Cat. 0 vs. media TI</b>	<b>SISIM Cat. 0 vs. media TI</b>

<p>[T,Tmax] = Student_mean(SN0,10201,100,0.0579354965 19949,0.05);</p> <p>Student's test on the mean ***** Student statistics: 7.2973 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p> <p>Null hypothesis is rejected. Simulation algorithm fails</p>	<p>[T,Tmax] = Student_mean(SS_IND0,10201,100,0.05 7935,0.05);</p> <p>Student's test on the mean ***** Student statistics: 42.6289 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p> <p>Null hypothesis is rejected. Simulation algorithm fails</p>
<p><b>SNESIM Cat. 3 vs. media TI</b></p> <p>[T,Tmax] = Student_mean(SN3,10201,100,0.4475051465 54259,0.05);</p> <p>Student's test on the mean ***** Student statistics: 1.2765 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p> <p>Null hypothesis is accepted. Simulation algorithm is validated</p>	<p><b>SISIM Cat. 3 vs. media TI</b></p> <p>[T,Tmax] = Student_mean(SS_IND3,10201,100,0.44 7505,0.05);</p> <p>Student's test on the mean ***** Student statistics: 6.8508 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p> <p>Null hypothesis is rejected. Simulation algorithm fails</p>
<p><b>SNESIM Cat. 0 vs. varianza Muestras</b></p> <p>[T,Tmax] = Student_var(SN0,10201,100,0.01941370607 649,0.05);</p> <p>Student's test on the variance ***** Student statistics: 79.7244 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p> <p>Null hypothesis is rejected. Simulation algorithm fails</p>	<p><b>SISIM Cat. 0 vs. varianza Muestras</b></p> <p>[T,Tmax] = Student_var(SS_IND0,10201,100,0.019 414,0.05);</p> <p>Student's test on the variance ***** Student statistics: 6.6126 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p> <p>Null hypothesis is rejected. Simulation algorithm fails</p>
<p><b>SNESIM Cat. 3 vs. varianza Muestras</b></p> <p>[T,Tmax] = Student_var(SN3,10201,100,0.25160163075 131,0.05);</p> <p>Student's test on the variance ***** Student statistics: 18.3468 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p> <p>Null hypothesis is rejected. Simulation algorithm fails</p>	<p><b>SISIM Cat. 3 vs. varianza Muestras</b></p> <p>[T,Tmax] = Student_var(SS_IND3,10201,100,0.251 602,0.05);</p> <p>Student's test on the variance ***** Student statistics: 18.3599 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842</p> <p>Null hypothesis is rejected. Simulation algorithm fails</p>

<b>SNESIM Cat. 0 vs. varianza TI</b>  [T,Tmax] = Student_var(SN0,10201,100,0.05458432564 28147,0.05);  Student's test on the variance ***** Student statistics: 7.3335 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is rejected. Simulation algorithm fails	<b>SISIM Cat. 0 vs. varianza TI</b>  [T,Tmax] = Student_var(SS_IND0,10201,100,0.054 584,0.05);  Student's test on the variance ***** Student statistics: 41.4965 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is rejected. Simulation algorithm fails
<b>SNESIM Cat. 3 vs. varianza TI</b>  [T,Tmax] = Student_var(SN3,10201,100,0.24726852999 802,0.05);  Student's test on the variance ***** Student statistics: 1.0427 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is accepted. Simulation algorithm is validated	<b>SISIM Cat. 3 vs. varianza TI</b>  [T,Tmax] = Student_var(SS_IND3,10201,100,0.247 269,0.05);  Student's test on the variance ***** Student statistics: 8.1465 Number of degrees of freedom: 99 Significance level: 0.05 Critical value: 1.9842  Null hypothesis is rejected. Simulation algorithm fails
<b>SNESIM (Test T<sup>2</sup> Hotelling)</b>  [T2,T2max] = Hotelling_variog(SN_IND3,101,101,1,1,1, 1,100,[2 22 22 0 0 0 0],0.211384,1,0.035,[10 20 30 40 50],[0 0 0 0 0],[0 0 0 0 0],0.05);  Hotelling's test on the semi-variogram ***** Hotelling statistics: 21014.2382 Numbers of degrees of freedom: 5, 95 Critical value: 12.0375  Null hypothesis is rejected. Simulation algorithm fails	<b>SISIM (Test T<sup>2</sup> Hotelling)</b>  [T2,T2max] = Hotelling_variog(SS_IND3,101,101,1, 1,1,1,100,[2 22 22 0 0 0 0],0.211384,1,0.035,[10 20 30 40 50],[0 0 0 0 0],[0 0 0 0 0],0.05);  Hotelling's test on the semi- variogram ***** Hotelling statistics: 15702.9178 Numbers of degrees of freedom: 5, 95 Critical value: 12.0375  Null hypothesis is rejected. Simulation algorithm fails

## ANEXO G – Parámetros de cumulantes de orden 3

### Configuración L

```
# DIMENSION OF THE PROBLEM: 1, 2 OR 3
2
# REGULAER OR IRREGULAR GRID : 1 REGULAR, 0 IRREG
1
# NUMBER OF POINTS AND DIMENSION OF THE IMAGE IN EACH DIRECTION
101 101 0 100 100 0
# ORDER OF CUMULANTS : 1, 2 ,3, 4 OR 5
3
# DIRECTIONS : (ORDER OF CUMULANTS - 1) DIRECTIONS
2
# NUMBER OF LAGS IN EACH DIRECTION: NLAG_DIR1, NLAG_DRI2, ...
20 20
# LAG SEPARATION IN EACH DIRECTION: LAG_DIR1, LAG_DIR2, ...
5 5
# LAG TOLERANCE IN EACH DIRECTION
0.0 0.0
# DIRECTIONS DEFINITION : AZIMUTH,AZI_TOL,BANDH,DIP,DTOL,BANDV (FOR EACH
DIRECTION)
90 0 0 0 0 0
0 0 0 0 0 0
```

### Configuración XX

```
# DIMENSION OF THE PROBLEM: 1, 2 OR 3
2
# REGULAER OR IRREGULAR GRID : 1 REGULAR, 0 IRREG
1
# NUMBER OF POINTS AND DIMENSION OF THE IMAGE IN EACH DIRECTION
101 101 0 100 100 0
# ORDER OF CUMULANTS : 1, 2 ,3, 4 OR 5
3
# DIRECTIONS : (ORDER OF CUMULANTS - 1) DIRECTIONS
2
# NUMBER OF LAGS IN EACH DIRECTION: NLAG_DIR1, NLAG_DRI2, ...
20 20
# LAG SEPARATION IN EACH DIRECTION: LAG_DIR1, LAG_DIR2, ...
5 10
# LAG TOLERANCE IN EACH DIRECTION
0.0 0.0
# DIRECTIONS DEFINITION : AZIMUTH,AZI_TOL,BANDH,DIP,DTOL,BANDV (FOR EACH
DIRECTION)
90 0 0 0 0 0
90 0 0 0 0 0
```

### Configuración YY

```
# DIMENSION OF THE PROBLEM: 1, 2 OR 3
2
# REGULAER OR IRREGULAR GRID : 1 REGULAR, 0 IRREG
1
# NUMBER OF POINTS AND DIMENSION OF THE IMAGE IN EACH DIRECTION
101 101 0 100 100 0
# ORDER OF CUMULANTS : 1, 2 ,3, 4 OR 5
```

```

3
# DIRECTIONS : (ORDER OF CUMULANTS - 1) DIRECTIONS
2
# NUMBER OF LAGS IN EACH DIRECTION: NLAG_DIR1, NLAG_DRI2, ...
20 20
# LAG SEPARATION IN EACH DIRECTION: LAG_DIR1, LAG_DIR2, ...
5 10
# LAG TOLERANCE IN EACH DIRECTION
0.0 0.0
# DIRECTIONS DEFINITION : AZIMUTH,AZI_TOL,BANDH,DIP,DTOL,BANDV (FOR EACH
DIRECTION)
0 0 0 0 0 0
0 0 0 0 0 0

```

## ANEXO H – Código MatLab frecuencia de patrones

### Código MatLab para calcular frecuencia de patrones NxN

```
function [combs,Count] = SearchPatternsGeneral(filename,N);

fid=fopen(filename,'r');
if fid==-1
    disp('cannot open file')
    fclose(fid);
    return
end
x=fscanf(fid,'%i',1);
y=fscanf(fid,'%i',1);
z=fscanf(fid,'%i',1);
nrealiz=fscanf(fid,'%i',1);
namevar='';
for i=1:nrealiz;
    namevar=[namevar,{fscanf(fid,'%s',1)}];
end
grid=fscanf(fid,'%f',x*y*z*nrealiz);
grid=reshape(grid,nrealiz,x*y*z);
grid=reshape(grid,x,y,z,nrealiz);

fclose(fid)

c = cell(1,N*N);
[c{end:-1:1}] = ndgrid([0 1]);           %% Generar combinación
posible de matriz N*N con 0 y 1
combs = cell2mat(cellfun(@(x) x(:), c, 'uni', 0)); %% Resultados como vector
combs = reshape(combs.',N,N,[]);         %% Generar matrices NxM:
combs(:,:,1), combs(:,:,2),...

combination = size(combs,3);

Columna = size(grid,2)
Fila = size(grid,1)

for j =1:nrealiz

    A = grid(:,:,j)
    A2 = nan(size(A)+2)
    A2(2:end-1, 2:end-1) = A;

    for i = 1:combination;
        Patterns = combs(:,:,i)
        R = nnz(nlfilter(A2, [N N], @(x) (isequal(x,Patterns))))
        Frecuencia = R/((Columna-N+1)*(Fila-N+1))
        AnswerA(i,j)= i
        AnswerB(i,j)= Frecuencia
    end
end

Count =[AnswerA(:,1) AnswerB];
```



## ANEXO I – Código MatLab Accuracy plot

### Código MatLab para selección de muestras diferentes a las empleadas en Simulación

```
function [REALITY,NEW_SAMPLES,INITIAL_SAMPLES,Msg] =  
SamplesAP(Reality,Initial_Samples,samplesize);  
  
%Cargar archivo con Realidad  
fidR=fopen(Reality,'r');  
if fidR==-1  
    disp('cannot open file')  
    fclose(fidR);  
    return  
end  
  
nbptsR=fscanf(fidR,'%i',1);  
nvariabR=fscanf(fidR,'%i',1);  
dummyR=fscanf(fidR,'%s',1);  
dummyR=fscanf(fidR,'%s',1);  
dummyR=fscanf(fidR,'%s',1);  
  
namevarR='';  
for i=1:nvariabR-3  
    namevarR=[namevarR,{fscanf(fidR,'%s',1)}];  
end  
REALITY=fscanf(fidR,'%f',nbptsR*nvariabR);  
REALITY=reshape(REALITY,nvariabR,nbptsR)';  
  
fclose(fidR);  
  
%Seleccionar muestras alternativas de la Imagen Exhaustiva para Jack-Knife  
  
NEW_SAMPLES = REALITY(randperm(size(REALITY, 1), samplesize), :);  
  
%Cargar archivo con muestras iniciales  
  
fidB=fopen(Initial_Samples,'r');  
if fidB==-1  
    disp('cannot open file')  
    fclose(fidB);  
    return  
end  
  
nbptsB=fscanf(fidB,'%i',1);  
nbvarB=fscanf(fidB,'%i',1);  
dummyB=fscanf(fidB,'%s',1);  
dummyB=fscanf(fidB,'%s',1);  
dummyB=fscanf(fidB,'%s',1);  
  
namevarB='';  
for i=1:nbvarB-3  
    namevarB=[namevarB,{fscanf(fidB,'%s',1)}];  
end  
INITIAL_SAMPLES=fscanf(fidB,'%f',nbptsB*nbvarB);  
INITIAL_SAMPLES=reshape(INITIAL_SAMPLES,nbvarB,nbptsB)';
```

```

fclose(fidB);

%Verificar que muestras alternativas no sean mismas muestras que las
%disponibles

R2 = NEW_SAMPLES(:,1:3);
Check = nnz(ismember(INITIAL_SAMPLES(:,1:3),R2,'rows'));

if Check > 0
    Msg = 'Al menos una muestra corresponde con otra muestra inicial. Intente
nuevamente'
    disp(Msg);
else
    Msg = 'Muestras seleccionadas Ok'
    disp(Msg);
end

```

### **Código MatLab para obtener probabilidad real para Accuracy Plot (se ejecuta en conjunto con anterior código)**

```

function [REALIZ,SAMP_REALIZ_SEL,p,TablaXY] =
Accuracy(Realizaciones,NEW_SAMPLES);

%Cargar archivo con realizaciones
fid=fopen(Realizaciones,'r');
if fid==-1
    disp('cannot open file')
    fclose(fid);
    return
end

nbptsA=fscanf(fid,'%i',1);
ncolumnas=fscanf(fid,'%i',1);
dummy=fscanf(fid,'%s',1);
dummy=fscanf(fid,'%s',1);
dummy=fscanf(fid,'%s',1);

namevar='';
for i=1:ncolumnas-3
    namevar=[namevar,{fscanf(fid,'%s',1)}];
end
REALIZ=fscanf(fid,'%f',nbptsA*ncolumnas);
REALIZ=reshape(REALIZ,ncolumnas,nbptsA)';

fclose(fid);

%Seleccionar los nodos de realizaciones con muestras para Validación

R2 = NEW_SAMPLES(:,1:3);

k = ismember(REALIZ(:,1:3),R2,'rows');
SAMP_REALIZ_SEL = REALIZ(k,:);

nbptsB = size(NEW_SAMPLES,1)

```

```

%Calcular los intervalos teóricos y reales
p = [10:10:90]

for i = 1:9

    zmin = [(100-p(i))/2]
    zmax = [(100+p(i))/2]

    Percentil_Inf = prctile(NEW_SAMPLES(:,4),zmin);
    Percentil_Sup = prctile(NEW_SAMPLES(:,4),zmax);

    Cantidad(i) = nnz(Percentil_Inf <= SAMP_REALIZ_SEL(:,4:end) &
SAMP_REALIZ_SEL(:,4:end) <= Percentil_Sup);

    Prob_teorica(i) = p(i)/100
    Prob_real(i) = (Cantidad(i)/(nbptsB*(ncolumnas-3)));
end

%Tabla con resultados para graficar el Accuracy_plot
TablaXY = [Prob_teorica,Prob_real];
TablaXY = reshape(TablaXY,9,2);

```

## ANEXO J – Variogramas de variable indicador SISIM

