UNIVERSIDAD DE TALCA

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA CIVIL DE MINAS

PREDICCIÓN Y DIAGNÓSTICO EN SUELOS CONTAMINADOS POR DAM USANDO MACHINE LEARNING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL DE MINAS

JOAQUÍN EDUARDO RETAMAL CEPEDA

PROFESOR GUÍA

Dr. MANUEL REYES JARA

MIEMBROS DE LA COMISIÓN

Mag. PABLO ZENTENO HIDALGO

Mag. SARA GODOY DEL OLMO

CURICÓ-CHILE

2020

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su encargado Biblioteca Campus Curicó certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Curicó, 2022

RESUMEN

El suelo es un recurso natural que juega un papel importante en la sostenibilidad de los ecosistemas, ya que sirve de soporte para todos los seres vivos. Además, de suministrarles el agua y los nutrientes que necesitan para su completo desarrollo. Una contaminación de este podría generar una alteración desfavorable, pudiendo disminuir la calidad de las funciones que desempeña y eventualmente suponer un riesgo a la salud humana o medio ambiente.

En Chile existen varias actividades económicas con potencial de contaminar el suelo; actividad forestal, agrícola y minería. Particularmente, la actividad minera puede desencadenar una de las problemáticas ambientales más relevantes, el drenaje ácido minero (DAM) que se genera a partir de la oxidación de sulfuros como la pirita, en presencia de oxígeno atmosférico y agua, produciéndose agua ácida cargada con sulfatos, metales y metaloides que pueden alcanzar altas concentraciones, generando un riesgo para el medio ambiente. Por lo tanto, esta memoria tiene como objetivo predecir y diagnosticar la contaminación de suelo causada por DAM. Para esto, se requiere la caracterización del emplazamiento contaminado y posteriormente, efectuar la técnica de estimación que se considere adecuada para determinar concentración de un elemento contaminante, tal como: métodos deterministas o geoestadístico. Aunque, dentro de las motivaciones de esta memoria, se busca considerar el machine learning como una nueva alternativa de estimación mediante el uso de redes neuronales. Para esto, se realiza un análisis de datos a través de métodos estadísticos para averiguar comportamiento y relaciones que existen entre variables. Posteriormente, se divide la base de datos: conjunto de entrenamiento, aquel que proporciona datos a las redes que le permiten aprender y conjunto de prueba, aquellos que serán utilizados para evaluar capacidad predictiva del modelo. Finalmente, se compara modelo machine learning con modelo geoestadístico por medio de matriz de confusión.

Los resultados obtenidos en esta memoria, muestran que cuando modelo geoestadístico y machine learning trabajan con variables continuas existen mínimas diferencias. Sin embargo, cuando modelos trabajan con variables categóricas, el método de kriging ordinario obtiene un 10% de zonas de incertidumbre. En cambio, las redes neuronales predicen entre un 3% y 7% zonas de incertidumbre. Además, de obtener mayor porcentaje de acierto en los puntos altos de contaminación, entregando como resultado 6.7 hectáreas contaminadas, lo que equivale al 20% del área de estudio.

ABSTRACT

Soil is a natural resource that plays an important role in the sustainability of ecosystems, since it serves as a support for all living things. In addition, it provides water and nutrients needed for their development. Soil contamination could generate an unfavorable alteration, being able to reduce the quality of the functions it performs and eventually being a risk to human health or the environment.

In Chile, there are several economic activities with the potential to contaminate the soil; forestry, agricultural and mining activity. Particularly, mining activity can trigger one of the most relevant environmental problems, acid mine drainage (DAM) that is generated from the oxidation of sulfides such as pyrite, in the presence of oxygen and water, producing acidic water loaded with sulfates, metals and metalloids that can reach high concentrations, creating a risk for the environment. Therefore, this report is intended to predict and diagnose soil contamination caused by DAM. For this, the characterization of the contaminated site is required and subsequently, perform the estimation technique considered appropriate to determine the concentration of a contaminating element, such as: deterministic or geostatistical methods. Although, within the motivations of this report, I seek to consider machine learning as a new alternative of estimation through the use of neural networks. For this, a data analysis is performed through statistical methods to find out the behavior and relationships that exist between variables. Subsequently, the database is divided: training set, the one that provides data to the networks that allow it to learn, and the test set, those that will be used to evaluate the predictive capacity of the model. Finally, a machine learning model is compared with a geostatistical model using a confusion matrix.

The results obtained in this report show that when geostatistical model and machine learning work with continuous variables there are minimal differences. However, when models work with categorical variables, the ordinary kriging method gets 10% uncertainty zones. In contrast, neural networks predict between 3% and 7% uncertainty zones. In addition, neural networks to obtain a higher percentage of success in the high points of contamination, giving as a result 6.7 contaminated hectares, which is equivalent to 20% of the study area.

AGRADECIMIENTOS

Quiero agradecer a mis padres por su constante apoyo, dedicación y preocupación. Sin ellos nada de esto sería posible.

A mis hermanos por estar conmigo en el momento que más lo necesite.

A Sandra por su amor incondicional y ser mi compañera de aventuras favoritas.

A Antonia por llegar a dar luz a mi vida.

Finalmente, a mis amigos y profesor guía que me ayudaron y colaboraron para poder desarrollar esta memoria.

Dedicado a

mamá y papá.

Contenidos

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

CAPÍTULO 1: INTRODUCCIÓN

1.1 Descripción del problema

La presente memoria tiene como principal interés abordar la contaminación de suelo asociado a la actividad minera.

La Organización de las Naciones Unidas para la Alimentación y la Agricultura, más conocida como FAO, define el suelo como la capa superior de la corteza terrestre que es transformada mediante procesos físicos, químicos y biológicos. En cuanto a composición, el suelo se forma de diversas partículas minerales, agua, aire, materia orgánica y organismos vivientes que interactúan en él (FAO,2015). Dado que, el suelo es considerado un recurso natural no renovable y esencial para el desarrollo de la vida, es fundamental prevenir su contaminación. Es en este contexto que nace este proyecto.

Por otra parte, el término contaminación de suelo, se refiere a la degradación de la calidad del suelo producto de la presencia de sustancias químicas que presentan una concentración más alta de lo normal, pudiendo generar consecuencias adversas sobre cualquier organismo al que no está destinado (FAO,2015). Por ejemplo, en el año 2016, durante la evaluación de desertificación hecha por CONAF, reveló que el 22% de la superficie del país presentaba riesgo de desertificación. Esto implica que un cuarto del suelo fértil y productivo del país puede perder parcialmente o total su potencial de producción.

Respecto al origen de contaminación de suelos nacionales, investigaciones reportan causas antropogénicas , siendo la minería y la industria-manufacturera los principales sitios con potencial presencia de contaminantes, con un 30,9% y 24,2% respectivamente (Ministerio del Medio Ambiente,2018). Es en este escenario que hemos puesto interés en los suelos contaminados producto de la actividad minera. Específicamente centraremos este proyecto en suelos contaminados por la generación de drenaje ácido minero (DAM). Abordar el tema de suelos contaminados por DAM es fundamental, puesto que es considerado por la Agencia de Protección del Ambiente (EPA) como una de las tres principales amenazas del ecosistema y primordial problema ambiental de las mineras.

La gravedad de la contaminación provocada por DAM, radica en que cuando se efectúan labores relacionadas con esta actividad, los sulfuros que se encuentran presentes en las rocas quedan

expuestos a la intemperie y al entrar en contacto con el agua y el aire, reaccionan formando el drenaje ácido, que a su vez puede precipitar otros metales tóxicos que también se encuentran en las rocas, pudiendo alcanzar cuerpos de agua superficiales y subterráneos, degradando la calidad del agua y suelo, además de afectar la vida de las personas y diferentes especies de animales. Por otra parte, conviene destacar que, si el DAM no se controla en el mismo sitio donde se genera, este podría perdurar durante años o décadas, hasta que uno de los principales agentes que lo causan sea eliminado (Chaparro,2015).

Ejemplo de las consecuencias provocados por los suelos contaminados por DAM, es lo ocurrido en la mina Summitville en Colorado, donde el DAM mató a todos los organismos biológicos en una franja de 27,36 kilómetros del Río Alamosa, actualmente la EPA gasta US$ 30,000 dólares al día para tratar el agua contaminada por DAM. Por otro lado, se encuentra el caso de la mina Zortman-Landusky en EE.UU, que fue abandonada en 1998 y las autoridades determinaron que la descarga de DAM tendrá que ser capturada y tratada a perpetuidad (Jepson,2002).

Los ejemplos anteriores, nos indican que la contaminación por DAM es un tema relevante, ya que las consecuencias son irreversibles y permanentes.

En el contexto nacional, las compañías mineras están concentradas principalmente en el norte del país, por lo tanto, estas zonas son las que están más expuestas a posible generación de DAM. Además, hay diferentes labores que pueden aumentar la probabilidad de generación, tales como: botaderos, depósitos de relaves, pilas de lixiviación, entre otros. En el caso de los depósitos de relaves, existen en Chile 742 depósitos, de los cuales 173 de ellos se encuentran en estado de abandono (Sernageomin,2019). Esto genera preocupación si es que no se han tomado las medidas necesarias de prevención y monitoreo.

Quiero recalcar, que lo que provoca contaminación es un mal manejo de las decisiones que se toman entorno a cualquier actividad, ya sea minería, agricultura o industrial.

Avanzando en nuestro razonamiento, cabe destacar que Chile no cuenta con una normativa que permita definir cuando nos encontramos con un sitio contaminado, lo que es un problema, porque los valores de referencia internacional no son acordes a nuestra realidad, dado que, nuestro territorio nacional es abundante en minerales de forma natural.

Por lo antes mencionado, este proyecto busca proponer herramientas o técnicas de predicción que ayudan a la prevención de este grave problema.

1.2 Objetivo general

Desarrollar modelo de machine learning para predecir y diagnosticar la contaminación de suelos por drenaje ácido minero.

1.3 Objetivos específicos

- Analizar variables que influyen en suelos contaminados por DAM.
- Generar modelo de machine learning capaz de estimar y predecir variables que influyen en suelos contaminados por DAM, a fin de determinar una nueva alternativa de estimación.
- Comparar modelo de machine learning con modelo geoestadístico para determinar capacidad predictiva.

1.4 Alcances

- La base de datos utilizada para efectos de esta memoria, no proviene de un estudio de contaminación por DAM, sino que, por otro tipo de contaminación correspondiente a un caso 2D.
- Dado que la base de datos fue compartida, se desconoce topografía del lugar, tipo de suelo, permeabilidad, porosidad y criterios que se utilizaron para implementar estrategia de muestreo de contaminantes.
- Se asume estacionariedad de la variable en estudio.
- Para comparar métodos, se utilizan solo algunos modelos de machine learning explicados en esta memoria.
- Determinada la distribución de los contaminantes en los suelos, no considera esta memoria las fases de remediación o descontaminación.

CAPÍTULO 2: MARCO TEÓRICO

En el presente capítulo, se mencionan los antecedentes teóricos que permiten fundamentar el trabajo hecho en esta memoria, que tiene como principal interés los siguientes conceptos: contaminación de suelos asociados a la actividad minera y métodos de estimación.

2.1 Contaminación de suelos en minería

Antes de explicar que se entiende por contaminación de suelos, es importante comprender que es un suelo. Guerrero (2016) define el suelo como la capa más delgada de la tierra, donde participan diversos factores en su formación: minerales, agua, aire, materia orgánica y organismos vivos. Debido a su lenta formación, se le considera un recurso natural no renovable que tiene gran importancia en la salud de las personas, las especies animales y plantas. Además, del posible desarrollo económico y social de la comunidad a causa de sus beneficios.

Respecto al término contaminación de suelo, se entiende como a la alteración negativa de la superficie terrestre por la presencia de sustancias químicas que presentan una concentración más alta de lo normal, pudiendo generar un riesgo al medio ambiente o en los diferentes organismos vivos que habitan en la tierra (Rodríguez, McLaughlin, Pennock,2019).

Esta alteración puede ocurrir de forman natural, pero el principal origen de la contaminación de suelos es por actividad antropogénica, es decir, aquellas relacionadas por la influencia del hombre en la naturaleza, por ejemplo: actividades industriales, minería, aguas residuales, agroquímicos, derrames accidentales de petróleo, desechos domésticos y municipales (Rodríguez et al.,2019).

En la minería se produce una serie de contaminantes: líquidos, sólidos y gaseosos, que de una u otra manera terminarán acumulados en los suelos generando algún tipo de impacto. Por ejemplo, el proceso de fundición genera una serie de elementos tóxicos, que al no ser manejados de manera adecuada se pueden introducir en el suelo y medio ambiente, al igual que, el escurrimiento de relaves mineros o la infiltración de productos usados en pilas de lixiviación (Rodríguez et al.,2019). Aunque, esta memoria está orientada en el drenaje ácido de minas.

2.2 Drenaje ácido de minas

El drenaje ácido de minas (DAM) consiste en la aparición de aguas con rangos de pH menor a 4 y alta concentración de elementos contaminantes, provocado por la oxidación de minerales sulfurados cuando son expuestos al aire y la humedad atmosférica en labores mineras (Pérez,2008).

De acuerdo con la Guía global del drenaje ácido de roca (Gard) publicada en el año 2009, dice que el desarrollo del DAM es un fenómeno complejo, dependiente del tiempo y que involucra procesos físicos, químicos y biológicos, que serán fundamentales en la producción, liberación, movilidad y atenuación de los contaminantes. Además, una vez iniciado el proceso de formación del DAM es difícil de detener, dado que si no se controla podría continuar e incluso verse acelerado, hasta que uno o más de sus reactantes (minerales sulfurados, oxígeno, agua) sean agotados o no estén disponibles para reaccionar.

2.2.1 Fuentes potencialmente generadoras de drenaje ácido minero

La generación de drenaje ácido producto de la oxidación de minerales sulfurosos, es un proceso que puede ocurrir en forma natural. Prueba de esto es la existencia de los gossan, rocas formadas por la oxidación progresiva de sulfuros cercanos a la superficie. Incluso, se descubrieron muchos yacimientos minerales por la coloración rojiza de las aguas, indicando presencia de minerales sulfurosos. Sin embargo, este proceso puede incrementar bruscamente su velocidad de generación por presencia de actividad minera (Pérez,2008). Por esta razón, la Guía Metodológica para la Estabilidad Química de Faenas e Instalaciones Mineras (2015) denominada para efectos de memoria como “Guía MEQFIM”, establece cuales son las instalaciones mineras con potencial de generación de DAM.

A continuación, se presenta aquellas instalaciones mineras que constituyen fuentes potencialmente generadoras de DAM.

Tabla 1: *Fuentes potencialmente generadoras de DAM durante la operación y al cierre de una faena minera (SERNAGEOMIN,2015).*

Residuos Mineros Masivos			Mina
Botaderos	**Depósitos de relaves (DS 248/06)**	**Depósitos de lixiviación**	
Estériles Baja ley Marinas y desmontes Escorias	Embalses de relaves Tranques de relaves Relaves filtrados Relaves en pasta Relaves espesados	Ripios de lixiviación Pilas permanentes Pilas dinámicas Pilas ROM	Rajo abierto Subterránea

2.2.2 Formación del drenaje ácido de minas

El drenaje ácido es originado cuando los minerales sulfurosos entran en contacto con el oxígeno y la humedad atmosférica, iniciando un complejo mecanismo de oxidación. De todos los sulfuros, la pirita (FeS_2) es el mineral más relevante asociado a la generación de DAM, pudiéndose encontrar residuos de minas con contenidos superiores al 95% de este mineral (Pérez,Schwarz,Urrutia,2017). Sin embargo, existen otros sulfuros que pueden participar de este proceso, por ejemplo: calcopirita ($CuFeS_2$), arsenopirita (FeAsS), esfalerita o blenda (ZnS), galena (PbS), entre otros (Belmonte & Tito,2012). Este proceso de oxidación puede ocurrir por dos mecanismos: oxidación abiótica, cuando la oxidación ocurre de manera directa, es decir, en ausencia de microorganismos y oxidación biótica, proceso que se lleva a cabo por la presencia de microorganismos que son capaces de acelerar el proceso de oxidación. En este caso, las bacterias que actúan como catalizadoras son: Leptospirillum ferrooxidans, Acidithiobacillus thiooxidans y Thiobacillus ferrooxidans (Life-Etad,2012).

Para entender el proceso de oxidación de sulfuros y posterior generación de acidez, se presenta la evolución de oxidación de la pirita (FeS_2).

Etapa 1: La acidez que se va produciendo, rápidamente es neutralizada por minerales carbonatados, ya que la velocidad de oxidación es baja. Por lo tanto, se mantienen condiciones de alcalinidad (pH>7) (Pérez,2008).

Etapa 2: La acidez acumulada supera la capacidad de neutralización del medio, debido a que se agotan los minerales carbonatados. Por lo tanto, el pH del agua es capaz de disminuir a valores

inferiores a 4.5, ocurriendo reacciones no solo por oxidación abiótica, sino que también biótica (Pérez,2008).

Etapa 3: Las condiciones de pH favorecen a la oxidación biótica, generando aumento en la velocidad de oxidación. En consecuencia, el agua de drenaje es ácida, presentando sulfatos y elevadas concentraciones de metales que fueron disueltos (Pérez,2008).

Respecto a las reacciones químicas que intervienen en la oxidación de la pirita, son presentadas y explicadas a continuación.

Se inicia la oxidación química de la pirita (FeS_2) al entrar en contacto con el oxígeno (O_2) y el agua (H_2O), produciendo hierro ferroso (Fe^{2+}), sulfato (SO_4^{2-}) e iones de hidrogeno (H^+) (Pérez,2008).

$$FeS_2 + \frac{7}{2}O_2 \ + \ H_2O \rightarrow Fe^{2+} + 2\ {SO_4}^{2-} + 2\ H^+ \qquad \textbf{(2.1)}$$

Posterior a esto, el hierro ferroso (Fe^{2+}) reacciona con el oxígeno para hierro férrico (Fe^{3+})

$$Fe^{2+} + \frac{1}{4}O_2 \ + \ H^+ \rightarrow Fe^{3+} + \frac{1}{2}H_2O \qquad \textbf{(2.2)}$$

Por lo general, por encima de un pH de alrededor de 3, el hierro férrico (Fe^{3+}) pierde solubilidad precipitándose como hidróxido ($Fe(OH)_3$, un precipitado de color anaranjado característico de las aguas de drenaje ácido. Durante esta reacción, también se liberan iones de hidrógeno (H^+) (Pérez,2008).

$$Fe^{3+} + \ 3H_2O \rightarrow Fe(OH)_3 + H^+ \qquad \textbf{(2.3)}$$

Finalmente, a medida que se va desarrollando la generación de ácido y se consume la alcalinidad disponible, algunos cationes férricos (Fe^{3+}) que se mantienen en solución cuando el pH es inferior a 3, pueden seguir oxidando adicionalmente a la pirita y formar nuevamente iones ferrosos , sulfato e hidrógeno (Pérez,2008).

$$FeS_2 + 14Fe^{3+} + \ 8H_2O \rightarrow 15Fe^{2+} + 2\ {SO_4}^{2-} + \ 16H^+ \qquad \textbf{(2.4)}$$

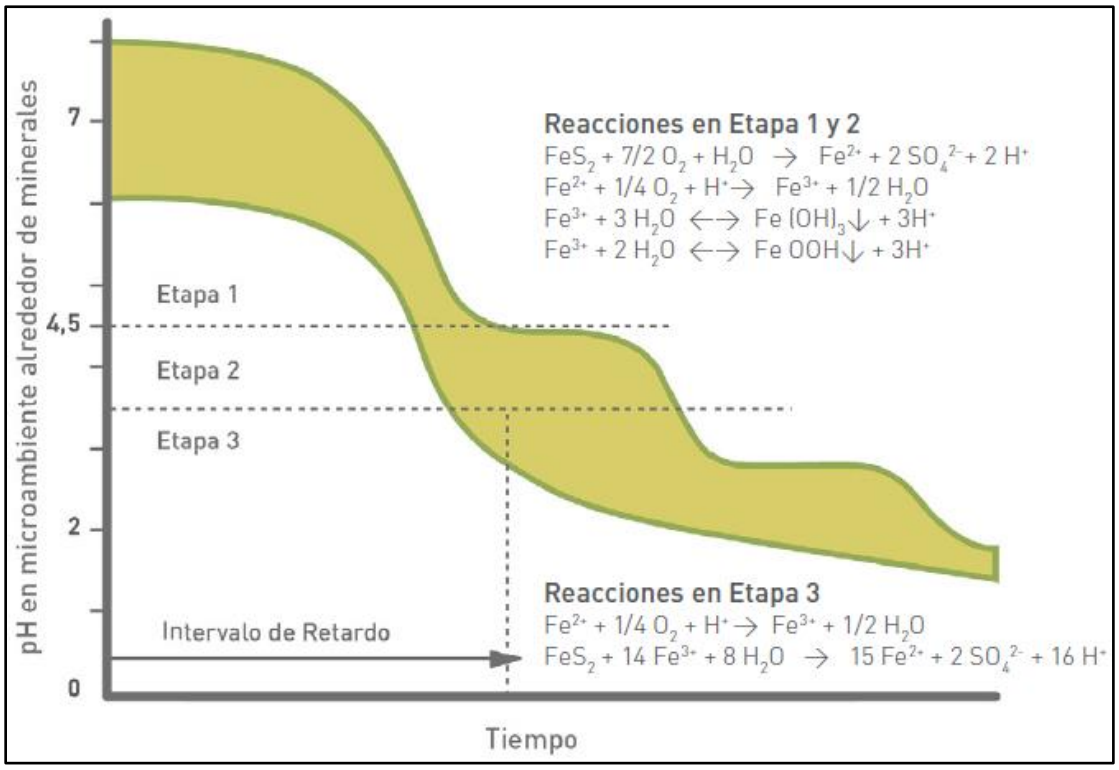

Ilustración 1: *Etapas en la generación de DAM, según la oxidación de la pirita (SERNAGEOMIN,2015).*

2.2.3 Minerales y elementos asociados al drenaje ácido minero

Como se mencionó anteriormente, en la formación de drenaje ácido minero existen diversos procesos físicos , químicos y biológicos, provocando la generación de aguas ácidas. Aunque, esto no es lo único que se produce, puesto que algunos elementos en su interacción con el agua o solución, pueden precipitar como minerales secundarios (SERNAGEOMIN,2015).

Los principales minerales secundarios formados en procesos de generación de DAM, se presentan en la siguiente tabla:

Tabla 2*: Principales minerales secundarios formados en procesos de generación de DAM (SERNAGEOMIN,2015).*

Minerales Secundarios	
Schwertmanita	$Fe_8O_8(OH)_6SO_4 – Fe_{16}O_{16}(OH)10(SO_4)_3$
Ferrihidrita	$5Fe_2O_3 \cdot 9H_20$
Jarosita	$KFe_3(SO_4)_2(OH)_6$
Alunita	$KAl_3(SO_4)_2(OH)_6$
Goethita	$FeO(OH)$
Hematita	Fe_2O_3
Siderita	$FeCO_3$

Melanterita	$FeSO_4 \cdot 7H_2O$
Epsomita	$MgSO_4 \cdot 7H_2O$
Roemerita	$Fe^{2+}Fe^{3+}{}_2(SO_4)_2 \cdot 14H_2O$
Coquimbita	$Fe_2{}^{3+}{}_2(SO_4) \cdot 9H_2O$
Bonatita	$CuSO_4 \cdot 3H_2O$
Chalcantita	$CuSO_4 \cdot 5H_2O$
Hexahydrita	$MgSO_4 \cdot 6H_2O$
Yeso	$CaSO_4 \cdot 2H_2O$
Azurita	$Cu_3(CO_3)_2(OH)_2$
Malaquita	$CU_2(CO_3)(OH)_2$
Crisocola	$CuSiO_3 \cdot 2H_2O$

Aunque, la principal preocupación aparte de la disminución de pH y la generación de acidez, es la liberación de metales y metaloides. Por ejemplo, el hierro y el aluminio son los metales disueltos que alcanzan mayores concentraciones con rangos entre 1000 a 10.000 mg/L. Mientras que otros elementos como cobre (Cu), plomo (Pb), zinc (Zn), cadmio (Cd), manganeso (Mn), cobalto (Co) y níquel (Ni) pueden llegar a concentraciones entre 100 a 1000 mg/L (SERNAGEOMIN,2015).

***Tabla 3**: Metales típicos presentes en DAM,DMN y DMAL (SERNAGEOMIN,2015).*

Metales típicos presentes en DAM	Fe^{2+},Fe^{3+},Al^{3+},Zn^{2+},Mn^{2+} Cu^{2+},Cd^{2+},Pb^{2+},Ni^{2+},As,etc
Metales típicos presentes en DMN/DMAL	Fe^{2+} ,Zn^{2+},Mn^{2+},Sb,As Cd,Cu^{2+},Mo,Se,etc.

2.2.4 Consecuencias del drenaje ácido minero

Si las actividades mineras no implementan trabajos de prevención, pronóstico y programas de monitoreo, el DAM puede ser una de las amenazas más serias que este rubro podría enfrentarse, debido a sus consecuencias ambientales, sociales y económicas, ya que su desarrollo puede continuar durante años, incluso después de que las actividades mineras han cesado (Gard,2009).

La principal preocupación que gira en torno al DAM es llegar los cursos de agua, ya sean superficiales o subterráneas, teniendo como consecuencia la modificación de las propiedades físicas y químicas del agua, provocando contaminación de esta y posterior alteración del ecosistema. Por ejemplo, los ríos afectados por este tipo de contaminación pueden causar daños letales en algunos organismos, inclusive hasta la desaparición de la fauna fluvial. Además, de la pérdida de los recursos hídricos por tornarse inservible el agua para el consumo humano, agrícola

o industrial. Por otra parte, la extensión y el grado de contaminación por metales pesados es variada, porque depende de las características geoquímicas de los sitios en que se desenvuelva (Life-Etad,2012).

En términos económicos, los costos de remediación son elevados. Sólo en Norte América los costos en minas sin operación, fueron estimados en diez mil millones de dólares aproximadamente (Gard,2009).

La siguiente tabla, resume los principales impactos y riesgos asociados para la salud de las personas y el medio ambiente dada la instalación correspondiente.

Tabla 4: *Riesgos e impactos asociados en cada fuente potencialmente generadora de DAM (SERNAGEOMIN,2015).*

Instalación	Impactos	Riesgos
Botaderos	Alteración significativa de la calidad del agua subterránea producto de la infiltración de DAM.	Afectación de la salud de las personas por el uso (agrícola, residencial, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con, y el uso del agua en lugares de afloramiento de este.
	Alteración significativa de la calidad del agua subterránea producto de la generación de DAM a causa de lluvias o crecidas, o a causa de la interacción de cursos de agua superficial con la instalación.	
	Alteración significativa de la calidad del agua superficial producto de la generación de DAM a causa de lluvias o crecidas, o producto de la interacción de cursos superficiales con la instalación.	Afectación de la salud de las personas por el uso (agrícola, residencial, recreacional, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con, y el uso del agua.
Depósitos de Relaves	Alteración significativa de la calidad del agua subterránea producto de la infiltración de DAM.	Afectación de la salud de las personas por el uso (agrícola, residencial, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con, y el uso del agua en lugares de afloramiento de este.
	Alteración significativa de la calidad del agua superficial producto de la generación de DAM.	Afectación de la salud de las personas por el uso (agrícola, residencial, recreacional, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con, y el uso del agua.

Depósitos de lixiviación	Alteración significativa de la calidad del agua subterránea producto de la generación de DAM a causa de lluvias o crecidas.	Afectación de la salud de las personas por el uso (agrícola, residencial, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con y el uso del agua en lugares de afloramiento de este.
	Alteración significativa de la calidad del agua superficial producto de la generación de DAM a causa de lluvias o crecidas.	Afectación de la salud de las personas por el uso (agrícola, residencial, recreacional, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con y el uso del agua.
Mina a Rajo Abierto	Alteración significativa de la calidad del agua del pit lake.	Afectación de la salud de las personas por el uso (agrícola, residencial, recreacional, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con, y el uso del agua.
	Alteración significativa de la calidad del agua subterránea producto de la infiltración de DAM.	Afectación de la salud de las personas por el uso (agrícola, residencial, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con, y el uso del agua en lugares de afloramiento de este.
Mina Subterránea	Alteración significativa de la calidad del agua subterránea producto de la infiltración de DAM.	Afectación de la salud de las personas por el uso (agrícola, residencial, otros) del agua. Afectación de especies y/o ecosistemas por el contacto con, y el uso del agua en lugares de afloramiento de este.

2.2.5 Predicción de drenaje ácido minero

La predicción en el DAM, está enfocada en determinar los posibles factores de una mina que son capaces de generar acidez. Para esto, se utilizan diferentes test a escala de laboratorio: test estáticos y cinéticos. Los primeros buscan determinar las características de los tipos de rocas en las áreas mineras, con el objetivo de detectar aquellos componentes que podrían generar acidez y aquellos que podrían neutralizar el potencial ácido en los desechos mineros. En cambio, los test cinéticos intentan simular condiciones reales que afectan a la generación de ácido. De esta manera, se podría conocer como un mineral se comporta a través del tiempo (Chaparro,2015).

2.3 Legislación ambiental

Las leyes o normas medioambientales tienen como objetivo preservar y proteger el medio ambiente de la contaminación o en caso de que este afectado mejorarlo (SEA,2020).

En Chile, a principios de los años noventa no existía una normativa que regulará este aspecto. Debido a esto, surgió la necesidad de crear una ley que garantizara el derecho a vivir en un medio ambiente libre de contaminación, que proteja el medio ambiente, preserve la naturaleza y conserve todo patrimonio ambiental. Como resultado, en el año 1994 se promulgó la ley 19.300 que sentó las bases generales del medio ambiente y permitió llenar el vacío existente en materia ambiental. Sin embargo, a pesar de que esta ley establece en su artículo 39 que velará que el uso del suelo se haga en forma racional, a fin de evitar su pérdida y degradación, aún en Chile no se cuenta con un marco regulatorio para la protección y calidad de uso de suelos (Mallea,s/f). Para Llona (2019) esto es un problema, porque sin una normativa es difícil calificar un terreno como contaminado, dado que no existen los parámetros mínimos o máximos de metales que debe contener un suelo y aunque, se pueda usar de referencia legislación internacional, el suelo chileno es abundante en minerales de forma natural, por lo tanto, requiere de una legislación acorde a nuestra realidad.

A pesar de todo, hoy son pocos los países que cuentan con una legislación específica sobre contaminación en suelos, destacándose: Estados Unidos, Canadá y Australia, teniendo como origen común desastres asociados a explosiones químicas y derrames que provocaron gran impacto mediático y ciudadano (Haro,2007).

En virtud de que Chile no posee normas de calidad de suelos, no es posible establecer límites máximos permisibles para contaminantes, por lo tanto, se presenta en la siguiente tabla valores de referencia de normas internacionales.

Tabla 5: *Límites máximos permisibles de normas mexicana y canadiense (Elaboración propia).*

Parámetros	Nom-147 Semarnat/SSA1-2004 (mg/kg)		Canadian Soil Quality Guidelines (mg/kg)			
	Agrícola/Residencial/ Comercial	Industrial	Agrícola	Residencial	Comercial	Industrial
Arsénico	22	260	12	12	12	12
Bario	5400	67000	750	500	2000	2000
Berilio	150	1900	4	4	8	8
Cadmio	37	450	1,4	10	22	22
Cromo	280	510	64	64	87	87
Cobre	-	-	63	63	91	91
Mercurio	23	310	6,6	6,6	24	50
Níquel	1600	20000	50	50	50	50
Plata	390	5100	20	20	40	40
Plomo	400	800	70	140	260	600
Selenio	390	5100	1	1	2,9	2,9
Talio	5,2	67	1	1	1	1
Vanadio	78	10000	130	130	130	130
Fracción de HC : Ligera	200	500	-	-	-	-
Fracción de HC : Media	1200	5000	-	-	-	-
Fracción de HC : Pesada	3000	6000	-	-	-	-

Nota: En caso de que se presenten diversos usos del suelo, debe considerarse el uso que predomine.

Cuando en los programas de ordenamiento ecológico y de desarrollo urbano no estén establecidos los usos del suelo, se usará el valor residencial.

2.4 Métodos de estimación

De acuerdo con Emery (2007) predecir o estimar "consiste en evaluar, de la manera más precisa posible, un valor que no ha sido medido, a partir de los datos disponibles" (P.5). Por lo tanto, una adecuada estimación, permitirá mejorar la comprensión y divulgación de los datos y resultados relativos al emplazamiento en estudio. Además, de disminuir el riesgo en la toma de decisiones.

Dependiendo del contexto del problema que se desea solucionar, podemos distinguir entre estimación global y local.

- **Estimación global:** Busca caracterizar una zona completa o grande por un único valor o por una distribución estadística. Se considera poco común que una estimación global sea suficiente, ya que para completarla se requieren estimaciones locales. Por ejemplo, en estudios de contaminación de suelos, la concentración promedio en toda la zona de un elemento contaminante no es suficiente, sino que se requiere distinguir entre sectores fuertemente contaminados y levemente contaminados (Emery,2007).
- **Estimación local:** Busca caracterizar los diferentes sectores de la zona de estudio, es decir, estimar el valor de sitios no muestreados o valor promedio de un bloque (Emery,2007).

Respecto a los métodos de estimación, podemos distinguir: geoestadística y métodos deterministas, los más utilizados para estimar leyes de mineral o concentración de un elemento contaminante en el suelo (Godoy,2017). Aunque, para Weeberb, Requia, Coull y Koutrakis (2019) concluyeron que machine learning también es un método que permite estimar leyes de mineral, contaminantes u otras propiedades.

2.5 Método determinista

Un método determinista es aquel que tendrá siempre los mismos resultados si sus datos de entrada no cambian, es decir, no hay existencia del azar o alguna incertidumbre asociada al proceso modelado (Villalba,2000).

Para Alfaro (2007), los métodos deterministas más tradicionales e importantes son los siguientes: media aritmética, los polígonos y el método del inverso de la distancia.

2.6 Geoestadística

En la literatura podemos encontrar diferentes definiciones sobre el término geoestadístico. Por ejemplo, algunos autores dicen que, la geoestadística es una rama de la estadística que trata fenómenos espaciales (Journel & Huiijbregts,1978). Para Petitgas (1996), es una aplicación de la teoría de probabilidades a la estimación estadística de variables espaciales.

Otras definiciones más actuales como las de Cely, Siabato, Sánchez, Rangel (2002) y Moral (2004), nos hablan de que es una técnica de la estadística que sirve para analizar, predecir y simular valores de una variable que se encuentra distribuida en el espacio o en el tiempo de forma continua, ósea, es un método que permite estimar a partir de puntos conocidos, la distribución de las características del suelo o variable en estudio, aun cuando no se disponga de muestras en algunos puntos (Godoy,2017).

La geoestadística, principalmente se ha desarrollado en el estudio de fenómenos en ciencias de la tierra (Emery,2007), aunque con el paso de los años, esta técnica se fue depurando y ampliando, tanto en sus herramientas como en sus campos de aplicación gracias a los trabajos realizados por Georges Matheron en la Escuela de Minas de París (Godoy,2017).

Los campos de aplicación actuales son variados según Godoy (2017), donde podemos destacar:

- **Minería**: Para evaluación de recursos y reservas de un yacimiento. Además, de la modelación geológica del mismo.
- **Industria petrolera**: Para modelación geológica y análisis de permeabilidad.
- **Hidrogeología**: Estimaciones de los niveles piezométricos.
- **Ecología:** Estudio de la distribución espacial y la afectación de plagas.
- **Medio Ambiente:** Estimación de suelos contaminados, cartografías geoquímicas y estudios de impacto ambiental.

A pesar de ser una técnica que surgió en el ámbito de la minería, con el propósito de predecir las reservas de oro en las minas sudafricanas, no se debe interpretar como una técnica exclusiva de la geología y la estadística. Si bien es cierto que, se integran conceptos estadísticos que se requieren conocer, no es necesario ser un experto estadístico para aplicar esta metodología, es decir, un matemático sin conocimiento alguno de un emplazamiento contaminado, no conseguirá realizar una interpretación adecuada de los resultados si no dispone de la experiencia necesaria en este tipo

de proyectos. Por lo tanto, el valor añadido de la geoestadística aparece cuando un especialista de un campo de aplicación consigue integrar y aplicar conceptos geoestadísticos a sus proyectos (Godoy,2017).

2.6.1 Etapas claves en un estudio de contaminación

Para realizar un adecuado análisis geoestadístico, se sugiere realizar una serie de etapas para poder obtener resultados óptimos (Cely et al.,2002). En la ilustración 2, se muestra el esquema metodológico para realizar este tipo de análisis, donde se aprecia que existen etapas previas y a pesar de no ser propiamente geoestadísticas, son fundamentales para un buen desarrollo del método: el muestreo y el análisis exploratorio de datos.

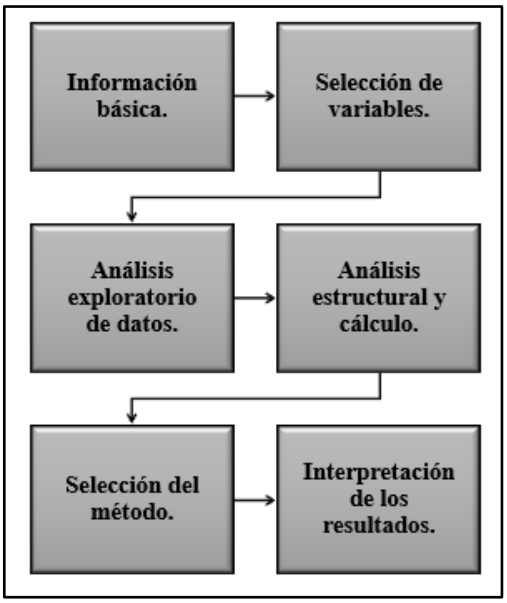

Ilustración 2: *Diagrama resumen para implementación del proceso geoestadístico (Cely et al.,2002).*

2.6.1.1 Información básica y selección de variables

Para realizar un buen análisis geoestadístico, se requiere utilizar una muestra representativa de la variable que se desea investigar (Cely et al.,2002). Por lo tanto, la estrategia de muestreo es fundamental para obtener unos datos de partida de calidad y será la base del modelo geoestadístico (Godoy,2017).

Algunos tipos de muestreo que existen son: juicio y conveniencia, catalogados como no probabilísticos. Mientras que, los muestreos aleatorios, sistemáticos y estratificado, se les considera como probabilísticos (Díaz,2006).

En la siguiente ilustración, se observan dos tipos de muestreo comúnmente utilizados: muestreo sistemático regular y muestreo aleatorio.

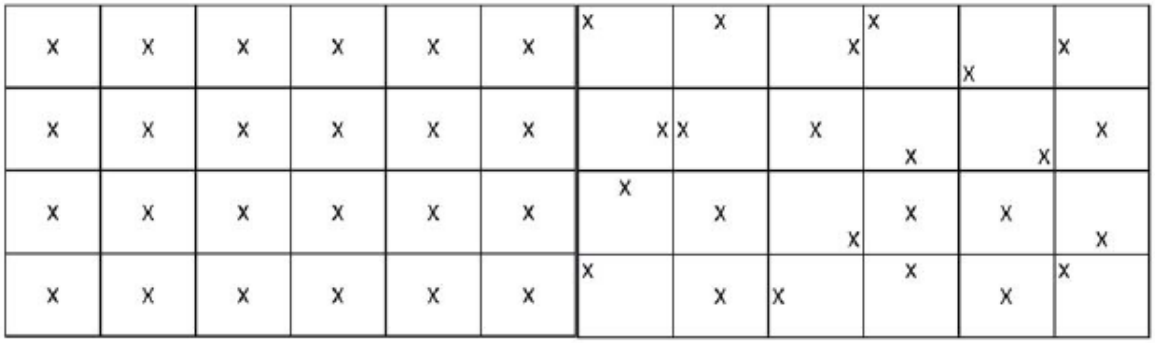

Ilustración 3: *Esquema de muestreo sistemático regular (izquierda) y muestreo aleatorio (derecha) (Godoy,2017).*

2.6.1.2 Análisis exploratorio de datos

Antes de aplicar geoestadística como tal, la realización de un análisis exploratorio de datos (EDA) ya sea descriptivo o gráfico es de suma importancia, ya que en función de las conclusiones que se extraigan de él, se procederá a realizar análisis variográfico, fundamental para definir variabilidad espacial de las concentraciones de contaminantes (Emery,2007).

Los principales objetivos de un análisis exploratorio de datos son los siguientes: análisis de cantidad, calidad y ubicación geográfica de datos disponibles, verificar la presencia de valores anómalos y distribución de la variable estudiada; además, se podría comprobar si existen correlaciones entre las principales variables de interés y otras auxiliares como: geología y topografía (Giraldo,s/f).

Las estadísticas básicas que son convenientes calcular para distinguir la distribución de valores son: medidas de posición y medidas de dispersión (Emery,2007).

Medidas de posición

- **Media aritmética:** Suma de valores de la distribución dividido por el número total de muestras.
- **Moda:** Valor que se repite más veces en una distribución.
- **Mediana:** Representa el valor de la variable de posición central en un conjunto de datos ordenados.

- **Cuantiles o percentiles:** Medidas de posición que dividen las muestras en partes de igual número de datos.
- **Mínimo y máximo:** Valores extremos de una distribución y establece el rango en que se distribuyen los mismos.

Medidas de dispersión

- **Varianza:** Es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística. Representa que tan dispersos están los datos alrededor de la media.
- **Desviación estándar:** Corresponde a la raíz cuadrada de la varianza.
- **Rango intercuartil:** Corresponde a la diferencia entre el tercer y primer cuartil de una distribución.

Respecto a las herramientas gráficas que permiten realizar un análisis exploratorio, se mencionan a continuación.

Histograma

Es una gráfica que representa la distribución de un conjunto de datos. Sirve para identificar tendencias en una población (Emery,2007).

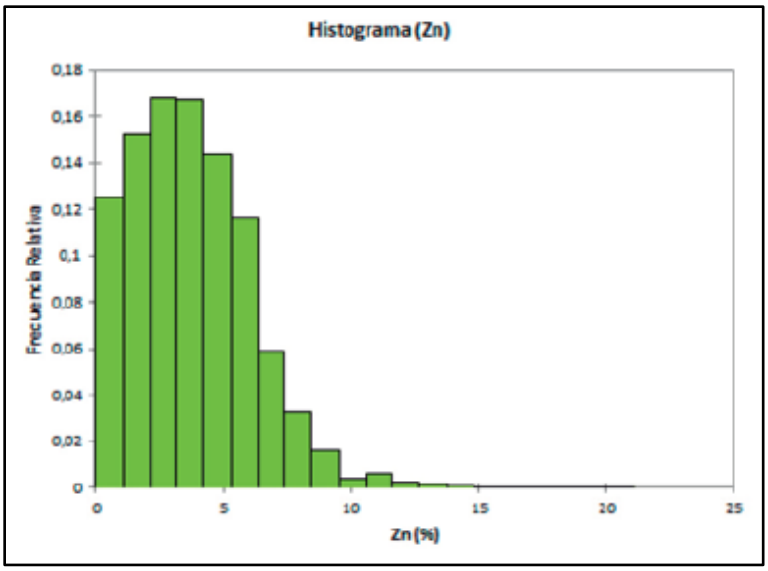

Ilustración 4: *Representación gráfica del histograma (Godoy,2017).*

Diagrama de cajas y bigotes (box-plots)

Consiste en un rectángulo, donde sus extremos están definidos por el primer y tercer cuartil. La mediana se representa como una línea que divide el rectángulo en dos partes iguales. Generalmente estos gráficos se utilizan para representar valores atípicos de una distribución (Godoy,2017).

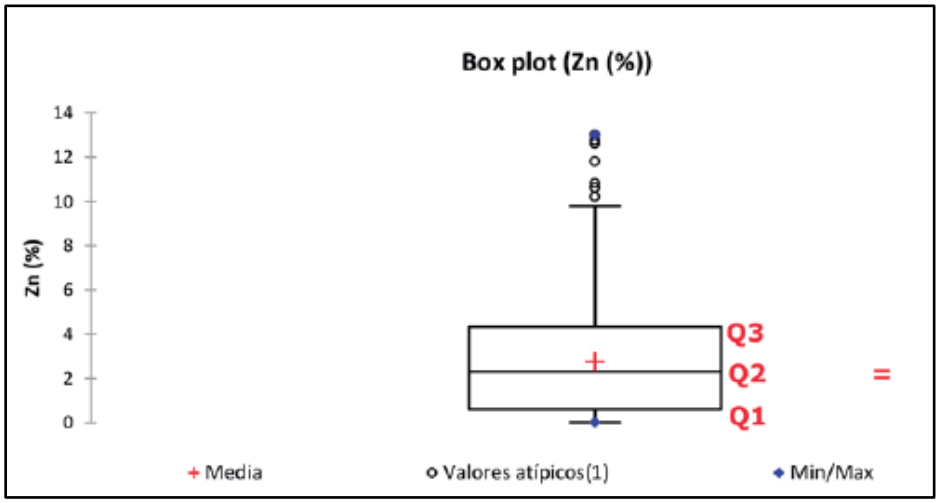

Ilustración 5: *Ejemplo de un gráfico de cajas y bigotes (Godoy,2017).*

Gráficos cuantil- cuantil (Q-Q Plot)

Este tipo de gráfico se utiliza para comparar la distribución de un conjunto de datos con otro conjunto de datos o una distribución ideal teórica (Godoy,20017).

Si los dos conjuntos de datos tienen la misma distribución, el gráfico cuantil-cuantil es lineal. En la ilustración 6, se puede apreciar un ejemplo de gráfico cuantil-cuantil para verificar normalidad de una distribución.

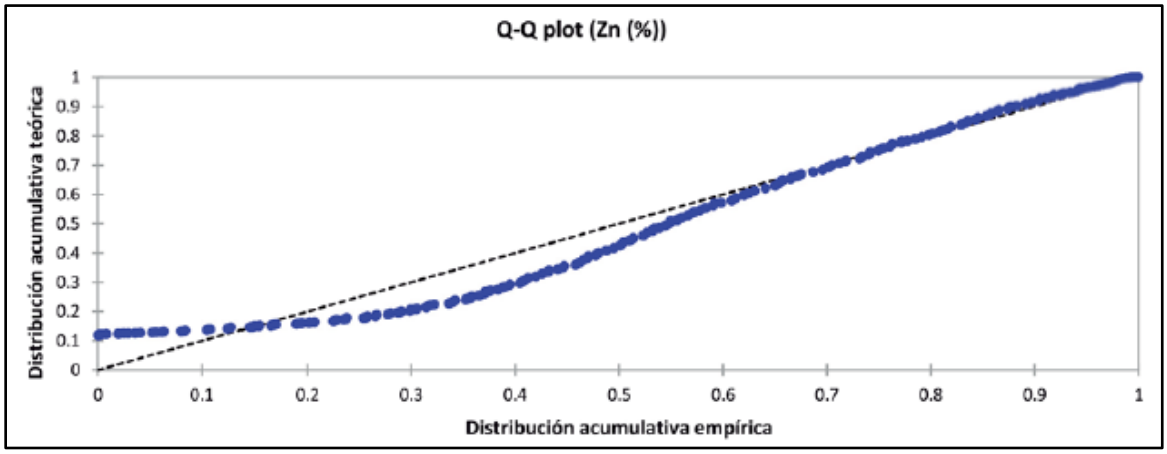

Ilustración 6: *Ejemplo de gráfico Q-Q normal (Godoy,2017).*

Diagrama de dispersión o de correlación

Es una herramienta gráfica que permite identificar la relación entre dos variables y detectar datos atípicos (aquellos puntos que se alejan de la nube). Para que esto sea posible, se requiere que ambas variables hayan sido medidas en el mismo sitio (Emery,2017).

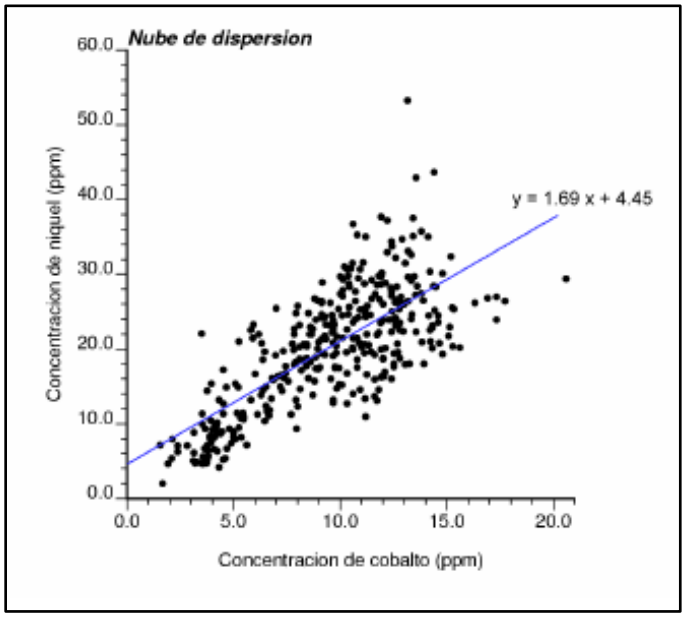

Ilustración 7: *Ejemplo de diagrama de dispersión (Emery,2007).*

2.6.1.3 Análisis estructural y cálculo

Finalizado el análisis exploratorio de datos y en función de las conclusiones que se extraigan de él, se procede a estudiar la correlación espacial del fenómeno estudiado, el que consiste en: estimar la dependencia espacial entre los datos medidos de una variable, proceso que se conoce como análisis estructural y en el cual se utilizan variogramas (Cely et al.,2002).

A partir de la información proporcionada por el variograma y los diferentes sitios muestreados, se puede generar una estimación por kriging, a fin de determinar contaminantes que se encuentran en el sitio de estudio (Vergara,2013).

2.6.1.3.1 Variograma

Según (Cely et al.,2002), los variogramas son "estimadores de la varianza poblacional relacionados con la dirección y la distancia, e indican como varían las dependencias espaciales que existen entre un punto de origen y otro punto a una determinada distancia, independientemente de su posición" (P.32).

Para modelar la continuidad espacial de una variable, se requieren de dos etapas previas, una que corresponde a la construcción del variograma experimental y la segunda, ajustar una curva teórica al variograma previamente construido (Cely et al.,2002).

La construcción del variograma experimental, se realiza a partir de muestras puntuales en puntos conocidos (Godoy,2017). Una vez que se han definido los puntos del variograma experimental, se debe pasar de un conocimiento espacial puntual a continuo, el que se realiza gracias a la modelización del variograma experimental, ajustando una curva teórica a dichos puntos de la primera curva experimental y que se denomina variograma teórico, que posteriormente, se utiliza en los diferentes métodos de interpolación como el kriging (Moral,2004).

2.6.1.3.1.1 Variograma experimental

Considerando $Z_{(x)}$ el valor de una variable regionalizada en un sitio X y $Z_{(x+h)}$ el valor de la misma variable en un punto distante h del anterior (Moral,2004), se define el variograma experimental como:

$$\hat{y}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(xi) - Z(xi+h)]^2 \qquad \textbf{(2.5)}$$

Donde N(h) corresponde al número de pares de datos separados por una distancia h.

Para caso multivariable se considera un modelo cruzado para el variograma entre las distintas especies involucradas (Z_i y z_j):

$$\hat{y}_{ij}(h) = \frac{1}{2|Nij(h)|} \sum_{Nij(h)} [Zi(x_\alpha) - Zi(x_\alpha + h) \cdot [Z_j(x_{\alpha)} - Z_j\,(x_\alpha + h)] \qquad \textbf{(2.6)}$$

2.6.1.3.1.2 Variograma modelado

Luego de haber desarrollado el cálculo del variograma experimental, es necesario ajustar un modelo teórico que permita obtener los parámetros que serán utilizados posteriormente en el método de interpolación (Sommer, Fernández, Rivas, Gutiérrez,2000). Existen varias opciones de modelos para realizar ajuste, pero deben cumplir con ciertas propiedades (Alfaro,2007).

- Función positiva: $y(h) \geq 0$
- Función par: $y(h) = y(-h)$

- Nulidad en el origen: $y(0) = 0$
- Función del tipo de negativo condicional

Debido a esto, autores como Sommer et al (2000), Cely et al (2002), Moral (2004), Emery (2007) y Alfaro (2007) mencionan los modelos que cumplen con estas propiedades y los que son más utilizados; esférico, exponencial, gaussiano y el efecto pepita puro, pudiéndose estos modelos combinarse linealmente.

Todos estos modelos tienen parámetros comunes que se describen a continuación:

- **Efecto pepita:** Representa una discontinuidad puntual del variograma en el origen. Este valor se puede atribuir a errores de medición o que parte de la estructura espacial se concentra a distancias inferiores de las observadas (Giraldo,s/f).
- **Meseta:** Corresponde al límite superior de cualquier modelo de variograma, valor donde se alcanza el rango (Giraldo,s/f).
- **Rango:** Corresponde a la distancia hasta donde hay correlación entre los datos o a partir donde dos observaciones son independientes (Giraldo,s/f).

Estos tres parámetros mencionados se describen gráficamente en la siguiente ilustración.

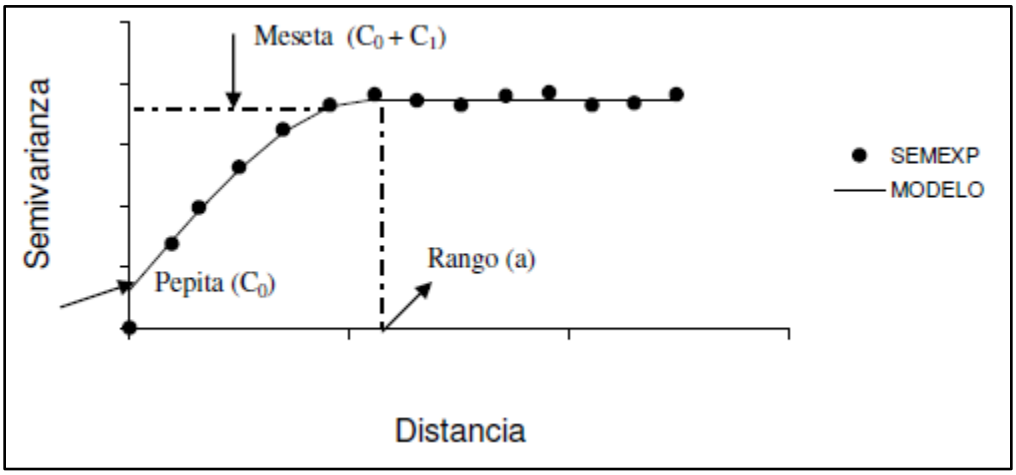

Ilustración 8: *Parámetros del variograma (Giraldo,s/f).*

Moral (2004), señala que el objetivo no es lograr un ajuste de una función a una serie de puntos, sino que, el modelo seleccionado debe explicar de manera clara el patrón de variabilidad espacial de la variable investigada y cuando se identifica el modelo, se habla que el variograma ha sido calibrado o validado.

2.6.1.3.1.2.1 Efecto pepita

El variograma pepitico de meseta C está definido como:

$$y(h) = \begin{cases} 0 \; si \; h = 0 \\ C \; caso \; contrario \end{cases} \quad \textbf{(2.7)}$$

Este modelo se traduce en que no hay correlación espacial entre los datos (Emery,2007).

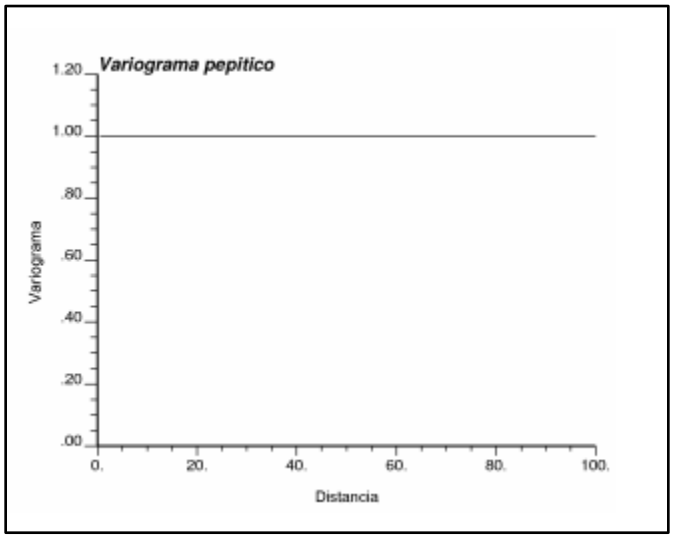

Ilustración 9: *Modelo de variograma teórico para variable sin correlación espacial (Emery,2007).*

2.6.1.3.1.2.2 Modelo esférico

El variograma esférico de alcance *a* y meseta *C* se define como:

$$y(h) = \begin{cases} C\left\{\frac{3|h|}{2a} - \frac{1}{2}\left(\frac{|h|}{a}\right)^3\right\} \; si \; |h| \leq a \\ C \; en \; caso \; contrario \end{cases} \quad \textbf{(2.8)}$$

Tiene un crecimiento rápido y lineal en el origen. Representa fenómenos continuos, pero no diferenciables (Emery,2007).

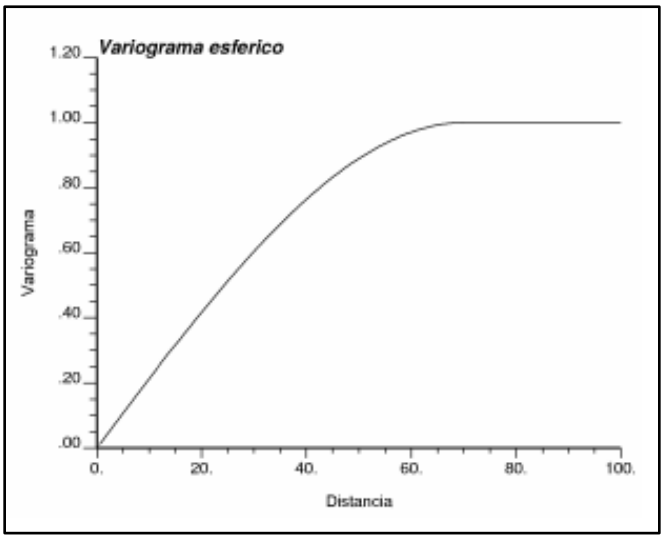

Ilustración 10: *Variograma esférico (Emery,2007).*

2.6.1.3.1.2.3 Modelo exponencial

El variograma exponencial de parámetro *a* y meseta *C* se define como:

$$y(h) = C\left\{1 - \exp\left(-\frac{|h|}{a}\right)\right\} \qquad \textbf{(2.9)}$$

A diferencia del modelo esférico que llega a la meseta exacta para |h| = *a*, el modelo exponencial alcanza la meseta *C* solo en forma asintótica (Emery,2007).

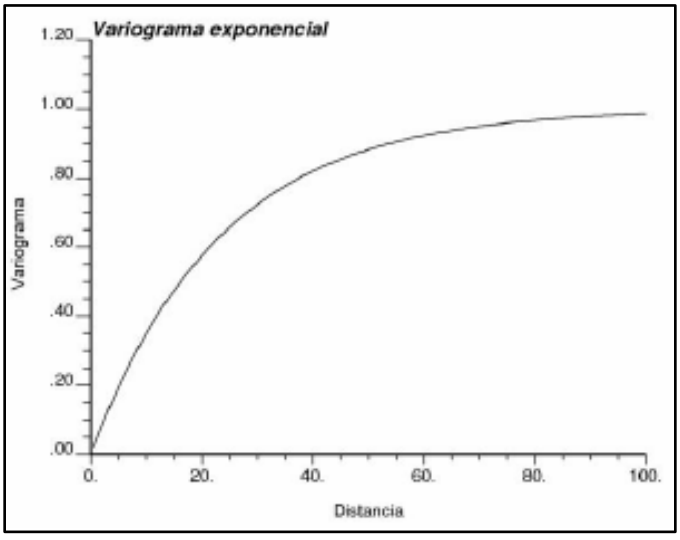

Ilustración 11: *Variograma exponencial (Emery,2007).*

2.6.1.3.1.2.4 Modelo gaussiano

El variograma gaussiano de parámetro *a* y meseta *C* se define como:

$$y(h) = C\left\{1 - \exp\left(-\frac{|h|^2}{a^2}\right)\right\} \quad \textbf{(2.10)}$$

Presenta forma parabólica cerca al origen. Su meseta se alcanza asintóticamente y el alcance práctico se puede considerar igual a $a\sqrt{3}$ (Emery,2007).

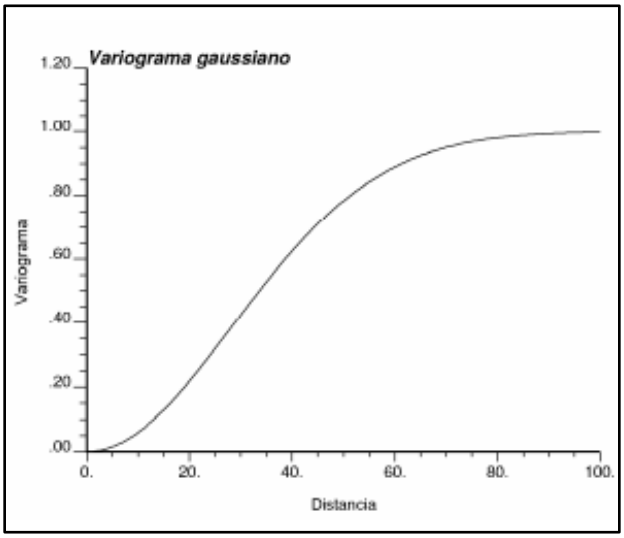

Ilustración 12: *Variograma gaussiano (Emery,2007).*

2.6.1.3.1.2.5 Modelo anidados

En reiteradas ocasiones, los variogramas experimentales presentan cambios de pendiente, lo que señala una variación en la continuidad espacial a partir de determinadas distancias, dejando de manifiesto diferentes escalas de variación en la variable regionalizada. Los diferentes tipos de modelos pueden combinarse linealmente y se les denomina modelos o estructuras anidadas, los que permite obtener modelos más complejos y modelar dichos cambios de pendiente que presentan los variogramas experimentales (Emery,2007).

$$Y(h) = Y_1(h) + 2(h) + \cdots + Y_n(h) \quad \textbf{(2.11)}$$

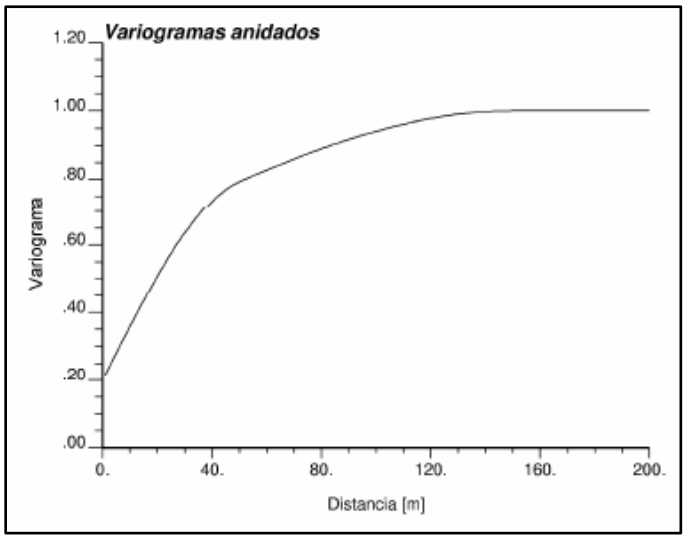

Ilustración 13: *Variograma anidado obtenido por suma de efecto pepita y dos modelos esféricos (Emery,2007).*

2.6.1.3.2 Comportamiento direccional

En general, se dispone de una serie de variogramas $Y_1(h), Y_2(h), \dots, Y_k(h)$ correspondiente a diferentes direcciones $a_1, a_2, \dots, a_k$ (Alfaro,2007).

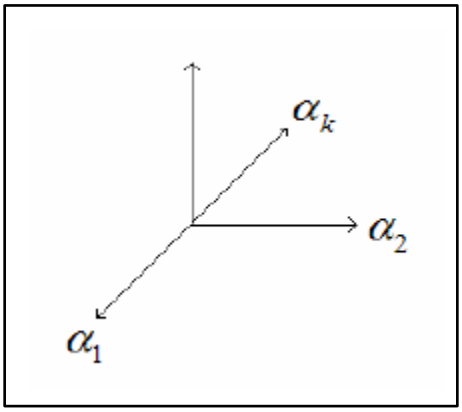

Ilustración 14: *Variogramas calculado en diferentes direcciones (Alfaro,2007).*

En el caso, que se tenga un variograma $Y(h)$ idéntico en todas las direcciones del espacio, es decir, depende solo de su modulo |h| y no de la orientación del vector h, se dice que es isótropo (Alfaro,2007). En casos como este, los variogramas calculados en las diferentes direcciones se superponen (Emery,2007).

En caso contrario, cuando los variogramas difieren según las direcciones del espacio, se manifiesta anisotropía, donde se puede distinguir la anisotropía geométrica y la anisotropía zonal (Emery,2007).

- **Anisotropía geométrica:** Es aquella en que el variograma calculado en distintas direcciones presenta la misma meseta, pero con rangos diferentes (Emery,2007).
- **Anisotropía zonal:** Es aquella en que el variograma calculado en distintas direcciones presenta el mismo rango, pero con mesetas diferentes (Emery,2007).

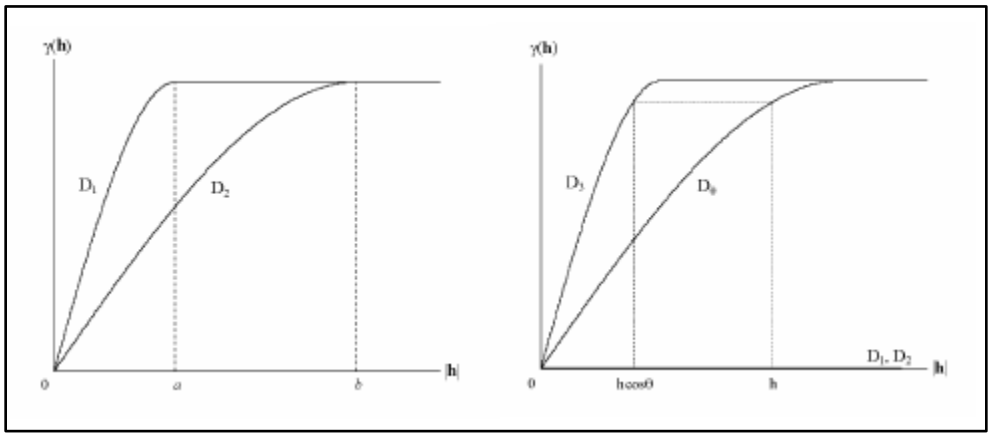

Ilustración 15: *Anisotropía geométrica (izquierda) y anisotropía zonal (derecha) (Emery,2007).*

2.6.1.4 Selección del método

Existen numerosos métodos de interpolación, los cuales se seleccionan en función de las características de la distribución o la cantidad de variables que se quieran estudiar (Godoy,2017). En geoestadística, el estimador por excelencia corresponde al kriging, un método que permite realizar cartografías de error, en función del variograma realizado y de la posición de los puntos de muestreo (Godoy, 2017).

2.6.1.4.1 Kriging

Por definición, el kriging es, un estimador lineal, insesgado que minimiza la varianza de la estimación (Godoy,2017). Su objetivo es encontrar el valor de la variable de interés en sitios donde no se dispone de información, a partir del conocimiento de los valores en los puntos próximos. Dentro de los más importantes, podemos encontrar el kriging simple y el kriging ordinario (Vergara,2013).

2.6.1.4.2 Kriging simple

Este tipo de kriging considera la siguiente hipótesis:

- La media *(m)* de la variable regionalizada estacionaria y covarianza es conocida.

En este modelo se busca estimar los valores de la variable en función de los datos medidos, utilizando la condición de insesgo y la condición de varianza mínima (Vergara,2013). El estimador tiene la siguiente forma:

$$Z*(x) = a_{ks} + \sum_{i=1}^{n} \lambda^{ks} \cdot Z(xi) \quad \textbf{(2.12)}$$

Donde:

- Z*(x) representa el valor estimado en el sitio x.
- a_{ks} un factor de adición.
- λ^{ks} son los ponderadores de los sitios con muestras $x_{i.}$

De la condición de insesgo, el valor esperado del error de estimación es:

$$E\{Z*(x) - Z(x)\} = a_{ks} + \sum_{i=1}^{n} \lambda i^{ks} \cdot E\{Z(xi)\} - E\{Z(x)\} \quad \textbf{(2.13)}$$

$$= a_{ks} + \sum_{i=1}^{n} \lambda i^{ks} \cdot m - m \quad \textbf{(2.14)}$$

Y es nulo si:

$$a_{ks} = (1 - \sum_{i=1}^{n} \lambda i^{ks}) \cdot m \quad \textbf{(2.15)}$$

De la condición de varianza de error mínima se tiene:

$$Var\{Z*(x) - Z(x)\} = Var\big(Z*(x)\big) - 2*Cov\big(Z*(x), Z(x)\big) + Var(Z(x)) \quad \textbf{(2.16)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda i \lambda j \, Cov\big(Z(xi), Z(xj)\big) - 2\sum_{1=1}^{n} \lambda i^{ks} Cov\big(Z(x), Z(xi)\big) + C(0) \quad \textbf{(2.17)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda i^{ks} \lambda j^{ks} C(xi - xj) - 2\sum_{1=1}^{n} \lambda i^{ks} C(x - xi) + C(0) \quad \textbf{(2.18)}$$

Para buscar los ponderadores que minimizan la varianza del error, es necesario derivar parcialmente con respecto a los ponderadores e igualar a cero (Vergara,2013), obteniendo que:

$$\sum_{j=1}^{n} \lambda i^{ks}\, C(xi - xj) = C(x - xi)\ \forall\ \text{i: 1 ... n} \qquad \textbf{(2.19)}$$

Las ecuaciones anteriores matricialmente se pueden ver:

$$\begin{pmatrix} C(x1-x1) & \cdots & C(x1-xn) \\ \vdots & \ddots & \vdots \\ C(xn-x1) & \cdots & C(xn-xn) \end{pmatrix} \begin{pmatrix} \lambda i^{ks} \\ \vdots \\ \lambda n^{ks} \end{pmatrix} = \begin{pmatrix} C(x1-x) \\ \vdots \\ C(xn-x) \end{pmatrix} \qquad \textbf{(2.20)}$$

El valor de la varianza de kriging tiene la siguiente expresión:

$$\sigma ks^2 = \sigma^2 - \sum_{i=1}^{n} \lambda i^{ks} C(xi - x) \qquad \textbf{(2.21)}$$

2.6.1.4.3 Kriging ordinario

Este tipo de kriging considera la siguiente hipótesis:

- La media *(m)* de la variable regionalizada no es conocida.

Vergara (2013) dice que, en el kriging ordinario se imponen condiciones para asegurar que el estimador sea insesgado y de varianza mínima. El estimador tiene la siguiente forma:

$$Z*(x) = a_{ko} + \sum_{i=1}^{n} \lambda i^{ko} \cdot Z(xi) \qquad \textbf{(2.22)}$$

Donde:

- Z*(x) representa el valor estimado en el sitio x.
- a_{ko}un factor de adición.
- λi^{ko} son los ponderadores de los sitios con muestras xi.

De la condición de insesgo, el valor esperado del error de estimación es:

$$E\{Z*(x) - Z(x)\} = a_{ko} + \sum_{i=1}^{n} \lambda i \cdot E\{Z(xi)\} - E\{Z(x)\} \qquad \textbf{(2.23)}$$

$$= a_{ko} + \sum_{i=1}^{n} \lambda i \cdot m - m \qquad \textbf{(2.24)}$$

Y es nulo si:

$$a_{ko} = 0 \; y \sum_{i=1}^{n} \lambda i^{ko} \; = 1 \tag{2.25}$$

De la condición de varianza de error mínima se tiene:

$$Var\{Z*(x) - Z(x)\} = Var\big(Z*(x)\big) - 2*Cov\big(Z*(x), Z(x)\big) + Var(Z(x)) \tag{2.26}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda i^{ko}\lambda j^{ko}\, Cov\big(Z(xi), Z(xj)\big) - 2\sum_{1=1}^{n} \lambda i^{ko} Cov\big(Z(x), Z(xi)\big) + C(0) \tag{2.27}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda i^{ko}\lambda j^{ko}\, C(xi - xj) - 2\sum_{1=1}^{n} \lambda i^{ko} C(x - xi) + C(0) \tag{2.28}$$

Para el cálculo de los ponderadores, es necesario introducir un multiplicador de Lagrange antes de calcular las derivadas parciales con respecto a los ponderadores e igualar a cero, ya que es necesario incorporar la restricción que la suma de los ponderadores debe ser uno (Vergara, 2013).

$$Var\{Z*(x) - Z(x)\} \tag{2.29}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda i^{ko}\lambda j^{ko}\, C(xi - xj) - 2\sum_{1=1}^{n} \lambda i^{ko} C(x - xi) + C(0) + 2\mu(\sum_{i=1}^{n} \lambda i^{ko} - 1) \tag{2.30}$$

Se obtiene el siguiente sistema de ecuaciones:

$$= \sum_{1=1}^{n} \lambda j^{ko} C(xi - xj) + \mu = C(x - xi) \; \forall \; \text{i: 1 ... n} \tag{2.31}$$

$$\sum_{1=1}^{n} \lambda i^{ko} = 1 \tag{2.32}$$

Expresado matricialmente tiene la siguiente forma:

$$\begin{pmatrix} C(x1-x1) & \dots & C(x1-xn) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(xn-x1) & \cdots & C(xn-xn) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda i^{ko} \\ \vdots \\ \lambda n^{k0} \\ \mu \end{pmatrix} = \begin{pmatrix} C(x1-x) \\ \vdots \\ C(xn-x) \\ 1 \end{pmatrix} \tag{2.33}$$

La varianza de kriging tiene la siguiente expresión:

$$\sigma k^2 o = \sigma^2 - \sum_{i=1}^{n} \lambda i^{ko} C(xi - x) - \mu \tag{2.34}$$

2.6.1.4.4 Kriging de indicadores

Corresponde a una variante no paramétrica del kriging ordinario. Consiste en transformar los valores iniciales de los datos de 0 a 1 en función de un criterio de contaminación establecido, es decir, todos los datos que se encuentran por debajo del criterio son reemplazados por un 0, mientras que, aquellos datos que se encuentran por encima del criterio de contaminación definido se les asigna un valor de 1 (Godoy,2017).

Los resultados obtenidos no expresan la concentración de un contaminante en un punto determinado, sino que, representan la probabilidad de superar o no el límite de contaminación establecido (Godoy,2017).

2.6.1.4.5 Validación del kriging

Existen diferentes métodos que permiten evaluar la bondad de ajuste de un modelo. Uno de los más empleados es el de validación cruzada, que usa todos los datos para predecir la autocorrelación del modelo. Este procedimiento omite un valor de la variable y predice dicho valor con los demás datos, para posteriormente, comparar predicción y el valor medido de la variable. Este proceso se debe repetir para cada uno de los valores muestreados (Cely et al.,2002).

De este procedimiento se obtiene el valor estimado y la varianza mínima de estimación para cada posición. Si el variograma previamente calculado fue modelado adecuadamente, el valor estimado debe ser muy similar al valor observado (Sommer et al.,2000).

Una forma descriptiva de hacer la validación cruzada, puede ser mediante un gráfico de dispersión de los valores observados contra los valores predichos. Por ejemplo, si la nube de puntos se ajusta más a una línea recta que pasa por el origen, el modelo de variograma que se utilizó para realizar kriging será mucho mejor (Giraldo,s/f).

2.6.1.4.6 Simulación condicional gaussiana

Una simulación es un modelo numérico que se asemeja a la variable regionalizada en estudio, es decir, que reproduce sus características espaciales y estadísticas. Se dice que es condicional, cuando en cada simulación, se asegura que los valores estimados respeten los valores observados en los puntos de muestreo. Este método exige que los datos sigan una distribución gaussiana. (Godoy,2017).

La ventaja de las simulaciones condicionales gaussianas es, permitir realizar múltiples simulaciones equiprobables de la contaminación a partir de los datos de inicio y el modelo de variograma realizado. Gracias a esto, es posible calcular la probabilidad de que en un punto dado haya o no contaminación. Además, de poder estimar la incertidumbre a las posibles fuentes de contaminación (Godoy,2017).

2.6.1.5 Interpretación de los resultados

Última etapa del proceso, que se realiza confrontando los resultados obtenidos del análisis de los datos con las hipótesis que se formularon, relacionándolas con la teoría y los procedimientos de la investigación (Alva,s/f).

En la realización de un modelo geoestadístico, se incluyen ciertas incertidumbres que son importantes mencionar para una adecuada interpretación de resultados (Godoy,2017).

- **Muestreo:** Primera fuente de incertidumbre y una de las más importantes, pues es la información fundamental que servirá de base para las distintas etapas de una estimación de contaminación de suelos (Godoy,2017):
- **Factor humano:** La manera de describir el suelo y de realizar el muestro, puede variar en función de las personas que intervengan durante la fase de caracterización. Sin embargo, es una fuente de incertidumbre difícil de cuantificar (Godoy,2017).
- **Precisión analítica:** Los resultados de laboratorio no incluyen la precisión de los aparatos con los que se realiza el análisis químico (Godoy,2017):

Además de los factores de incertidumbre nombrados anteriormente, Godoy (2017) nos dice que, hay otros que son propios del método geoestadístico:

- Realizar variograma experimental y su modelización.
- Aplicación de métodos de interpolación.

2.7 Machine learning

De acuerdo con lo que exponen Barberena, Gnoza (2018) y Alpaydin (2014), el aprendizaje automático o machine learning se engloba dentro de las disciplinas de la inteligencia artificial, que permite usar los ordenadores y otros dispositivos con capacidad computacional para que aprendan identificar patrones y relaciones que hay en datos por sí solos, es decir, dota la capacidad de

aprender a los ordenadores sin ser explícitamente programados, mejorando de forma autónoma a partir de la experiencia, a fin de predecir comportamientos y tomar futuras decisiones.

Para que las máquinas puedan aprender, los principales ingredientes que se necesitan son: datos, modelos y algoritmos (Zamora,2013).

El machine learning utiliza dos tipos de técnicas: aprendizaje supervisado y aprendizaje no supervisado (Suárez, Jiménez, Castro-Franco, Cruz-Roa, 2017). Aunque Rodríguez-Sahagún (2018), dice que es conveniente agregar una tercera técnica, el aprendizaje semisupervisado y que es una combinación entre las dos técnicas anteriormente mencionadas.

2.7.1 Aprendizaje supervisado

Son aquellos problemas en que existe una variable a predecir, donde el algoritmo aprende de un conjunto de datos que contienen determinadas características previamente etiquetadas por un experto o de forma semi-automática basándose en los datos, por lo tanto, necesita de una supervisión. En este tipo de aprendizaje, el objetivo es que el algoritmo pueda aprender de los ejemplos proporcionados las reglas que permitirán predecir etiquetas de los nuevos casos que aparezcan, es decir, dadas las propiedades de un caso del que no conocemos su valor, el algoritmo sea capaz de predecir lo más correctamente posible (Rodriguéz-Sahagún,2018).

Los problemas de aprendizaje supervisado son: clasificación y regresión. En la clasificación, se pretende predecir que categoría le corresponde a una instancia dentro de una enumeración de posibles categorías, donde la variable puede ser nominal o discreta. Mientras que, dada una cierta cantidad de datos que no se clasifican y la variable a predecir es continua, se habla de un problema de regresión (Rodriguéz-Sahagún,2018).

Algunos modelos que se basan en aprendizaje supervisado son: arboles de decisión, random forest, máquinas de vectores de soporte (Support vector machine o SVM) y redes neuronales artificiales (Zamora,2013).

2.7.1.2 Redes neuronales artificiales

Las redes neuronales artificiales (RNA) es un modelo matemático, que tiene como principio emular el funcionamiento del sistema nervioso humano. Está formado por un conjunto de neuronas o nodos

altamente interconectadas en paralelo y que transmiten señales entre sí, transformando un conjunto de datos de entrada en un conjunto de datos de salida deseado (Pedraza,s/f).

Rath (1999) señala que una RNA está compuesta por capas de información: capa de entrada, correspondiente a neuronas que reciben datos o señales que proceden del entorno, capa oculta, aquella que determina la relación entre las variables de entrada y salida, capa de salida, que proporciona la respuesta de la red a los estímulos de los datos de entrada.

A continuación, se puede observar un esquema de una red neuronal.

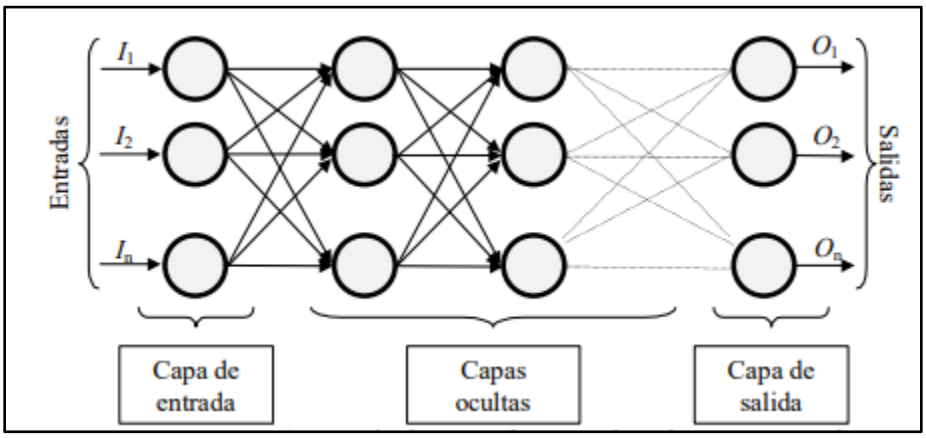

Ilustración 16: *Ejemplo red neuronal totalmente conectada (Matich,2001).*

2.7.1.2.1 Función de activación

Como se mencionó, el funcionamiento de las RNA se asemeja al del sistema nervioso humano, por lo tanto, si una neurona biológica posee estados de activación como activo o inactivo, una neurona artificial también tendrá diferentes estados de activación, cuyo rango normalmente va de 0 a 1 o de -1 a 1. Para que esto ocurra, los valores de entradas que son recibidos por las neuronas, tienen asociado un peso, permitiendo realizar una suma ponderada de los datos, que determinará con que intensidad cada variable de entrada afecta a la neurona. Este valor, se procesa al interior de la neurona mediante una función de activación, que decide si este valor numérico es capaz de propagarse a las siguientes capas intermedias hasta llegar al final de la red neuronal (Rodriguéz-Sahagún,2018).

Según Rodriguéz-Sahagún (2018), las funciones de activación más utilizadas son: función sigmoide y tangente hiperbólica, las cuales llevan las siguientes expresiones.

$$Sigmoide\ (x) = \frac{e^x}{e^x + 1} \quad \textbf{(2.35)}$$

$$Tanh\ (x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \textbf{(2.36)}$$

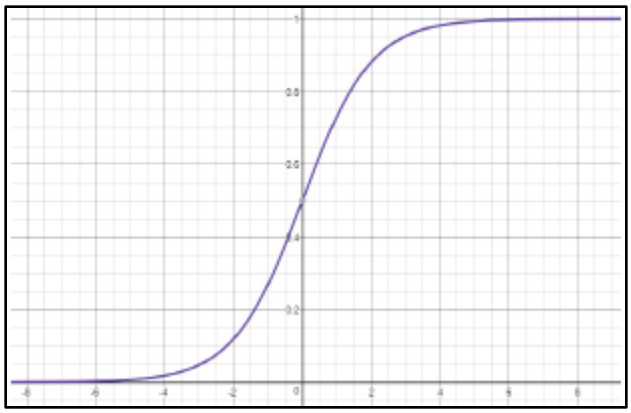

Ilustración 17: *Función sigmoide (Rodriguéz-Sahagún,2018).*

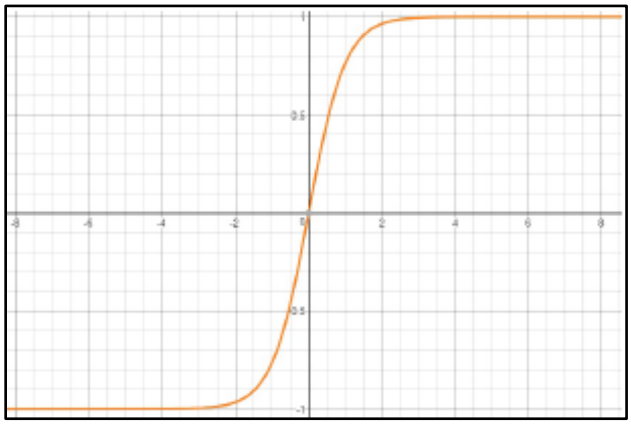

Ilustración 18: *Función tangente hiperbólica (Rodriguéz-Sahagún,2018).*

2.7.1.2.2 Entrenar redes neuronales

Para entrenar las redes neuronales, se utiliza el algoritmo propagación hacia atrás o backpropagation. Este método trabaja desde la capa de salida hasta la capa de entrada, modificando previamente los pesos de la red neuronal, para comparar datos de salida con valores deseados y calculando el error de cada una de las respuestas de salida. A partir de esto, es posible estimar el peso que mejor se ajusta dentro de las neuronas reduciendo el error. Este proceso se repite capa por capa, hasta poder averiguar la contribución de cada neurona al error total (Sánchez,2014)

2.7.2 Evaluación de modelos supervisados

Para medir la bondad de un modelo en problemas de clasificación, no es bueno tomar como referencia únicamente una única métrica. Por ejemplo, un modelo puede tener mucha confianza en una predicción y equivocarse. Esto puede ocurrir cuando los modelos sufren sobreajustes (*overfitting*), por lo tanto, no están generalizando adecuadamente. Por el contrario, con *underfitting* también es un problema, ya que el modelo es tan general que no es capaz de extraer los patrones de los datos. Por estas razones, un modelo debe evitar ambos problemas y deducir patrones existentes en los datos de una manera general como para nuevos datos (Alonso,2019).

Para asegurar que se evalúa adecuadamente se dividen los datos: conjunto de entrenamiento que contiene el 80% de los datos y conjunto de prueba con el 20% de los datos restantes. La idea es comparar los datos en que se conoce su resultado con los valores que predice el modelo. Si ambos resultados coinciden, el modelo está acertando en la predicción, sino estaría cometiendo un error (Alonso,2019).

Alonso (2019) dice que una manera gráfica de encontrar precisión y exactitud del modelo, es utilizar la matriz de confusión, donde podemos desprender 4 escenarios:

- Verdaderos positivos (TP): Es la cantidad de datos positivos que fueron clasificados correctamente como positivos por el modelo.
- Verdaderos negativos (TN): Es la cantidad de datos negativos que fueron clasificados correctamente como negativos por el modelo.
- Falsos negativos (FN): Es la cantidad de datos positivos que fueron clasificados incorrectamente como negativos por el modelo
- Falsos positivos (FP): Es la cantidad de datos negativos que fueron clasificados incorrectamente como positivos por el modelo.

Tabla 6: *Matriz de confusión (Elaboración propia).*

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (TP)	Falsos Negativos (FN)
	Negativos	Falsos positivos (FP)	Verdaderos Negativos (TN)

En el caso de modelos de regresión, autores como Comesaña, Dago-Morales, Talavera, Núñez & Hernández (2010), utilizan la raíz del error del cuadrático medio (RMSD) y R al cuadrado (R^2).

2.7.3 Aprendizaje no supervisado

En los algoritmos de aprendizaje no supervisado, el modelo de entrenamiento está formado por entradas y no se tiene información a que clase pertenecen los datos, por lo tanto, su objetivo es agrupar los datos por características o patrones similares en un número definido de clases (Suárez et al.,2017).

Los siguientes modelos que se catalogan bajo este concepto son:

- Segmentación (clusters).
- Detectores de anomalías.
- Reglas de asociación.

CAPÍTULO 3: METODOLOGÍA

En el presente capítulo, se mencionan los procedimientos adoptados que permitieron relacionar los objetivos planteados con la problemática de investigación expuesta al inicio de esta memoria.

Para esto, se utilizó como referencia la metodología planteada por Cely et al. (2002), quienes plantean las etapas claves que se deben realizar en un estudio de contaminación de suelos usando geoestadística.

1. **Información básica:** al momento de realizar un estudio geoestadístico, la obtención de datos del campo de estudio es imprescindible. Es por esto, que la estrategia de muestreo es fundamental para obtener datos de calidad, dado que las muestras recopiladas deben ser representativas de la variable que se desea investigar.

   En la presente memoria, los datos utilizados fueron facilitados por Dña. Sara Godoy del Olmo, docente del curso "Geoestadística aplicada a la contaminación de suelos" de la plataforma Ingeoexpert.

   Es importante recalcar, que los datos obtenidos no provienen de un estudio de contaminación por DAM, sino que, la contaminación proviene del derrame de hidrocarburos. Aunque, la metodología que se requeriría para estimar contaminación por DAM, sería esencialmente la misma que se utiliza en esta memoria para contaminación de suelos. Incluso, las estimaciones de DAM podrían ser mejores si a un conjunto de muestras se incluye un modelo cinético químico, en conjunto con muestreos temporales. Sin embargo, en esta memoria no se desarrolló esta idea en forma numérica, debido a que no se dispone de tales datos de DAM en forma temporal, pero se plantea y comenta la metodología en base a modelos teóricos.
2. **Selección de variables:** obtenida la base de datos, se debió seleccionar aquellas variables predictoras que influyeron en mayor proporción en el fenómeno de contaminación.
3. **Análisis exploratorio de datos:** antes de aplicar geoestadística como tal, se hizo un análisis exhaustivo de la base de datos, tales como: cantidad, calidad y ubicación geográfica de datos disponibles, verificación de valores anómalos, distribución de las variables estudiadas. También, se comprobó si había errores en la base datos o existía correlación entre las principales variables de interés. Por último, se calcularon algunas estadísticas básicas como: medidas de posición y dispersión.

Este procedimiento fue de gran importancia, ya que en función de las conclusiones que se extrajeron, permitió entender la variabilidad espacial de las concentraciones contaminantes.

4. **Selección del método:** existen numerosos métodos de interpolación, los cuales se seleccionan en función de las características de la distribución o cantidad de variables que se quieran estudiar. Dicho esto, los métodos que se consideraron pertinentes para alcanzar los objetivos planteados fueron los siguientes: kriging ordinario, kriging de indicadores y simulación condicional gaussiana.
5. **Análisis estructural y cálculo:** se debió estudiar la correlación espacial del fenómeno de contaminación. Para esto, se calcularon variogramas experimentales para las distintas variables que fueron sometidas al estudio y para los tres métodos de estimación seleccionados previamente. Luego, cada variograma experimental fue ajustado a una curva teórica, proceso que se conoce como modelación del variograma. Mencionar, que antes de realizar los variogramas para el método de kriging de indicadores y simulaciones condicionales gaussianas, los datos originales fueron transformado de acuerdo con la distribución que exige el método para aplicarse. Por último, se realizó validación cruzada para predecir autocorrelación del modelo.
6. **Interpretación de resultados:** se obtuvo la estimación de concentración del o los elementos contaminantes que se encontraban en la zona de estudio delimitada. Con estas estimaciones, se calculó superficies y volúmenes de contaminación, además de definir como estaba distribuida la contaminación en el área de estudio: zonas bajas, medianas y altas de contaminación.

   En el caso de las simulaciones condicionales gaussianas, se realizó un post-tratamiento que permitió recopilar el conjunto de simulaciones ejecutadas e interpretarlas, tal como: estimación de la concentración media de simulaciones hechas y estimación de probabilidad de superar criterio de contaminación establecido.

   Finalmente, los resultados de los distintos métodos de estimación empleados, se presentan como tablas, mapas de predicción y probabilidad para facilitar su interpretación.

Los softwares empleados fueron los siguientes: Past en el análisis estadístico de los datos y SGeMS para aplicar los distintos métodos de estimación geoestadístico mencionados. La particularidad que presentan ambos softwares es que son de uso gratuito, intuitivos, por lo que no requieren de gran tiempo de capacitación y no se necesitan equipos computacionales de alto nivel para que funcionen

bien. Aunque, el software SGeMS al exportar sus resultados no entrega los datos georreferenciados y tampoco genera validación cruzada. En su remplazo, se utilizó un código en Python y el software gvSIG para realizar validaciones cruzadas.

Respecto al proceso aplicado para construir modelos de machine learning, se tomó como referencia las etapas que mencionan Gnoza y Barberena (2018) con algunas variables.

1. **Entendimiento y preparación de datos:** proceso centrado en la extracción de conocimiento; datos que existen y combinación o creación de nuevas variables a partir de las existentes, transformación de datos en caso de ser requerido, limpieza, entre otros.
2. **Análisis exploratorio:** tratamiento estadístico que se someten los datos para descubrir comportamiento y relaciones que pueden existir entre los mismos.
3. **Selección de técnica y construcción de modelo:** en base al problema que se pretende resolver y los datos disponibles, se seleccionó una técnica y se construyó un modelo de referencia que sirvió para comparación. Para evitar que el modelo construido sufriera *overfitting* o *underfitting*, se dividió la base de datos: conjunto de entrenamiento que contiene el 80% de los datos y que proporciona al algoritmo datos que le permiten aprender, mientras que el 20% restante conocido como conjunto de prueba se utilizó para probar que tan bien predice el modelo entrenado.
4. **Construir modelos iterativos:** puede que el resultado obtenido del primer modelo construido no sea adecuado e inclusive completamente erróneo, por lo que se debe volver al punto anterior, entrenar el modelo y cambiar los ajustes. Esta etapa fue un ciclo de retroalimentación, donde los modelos construidos se fueron comparando con uno de referencia con criterios objetivos, para así definir el mejor modelo.
5. **Predicción:** con el modelo que se obtuvieron resultados satisfactorios, se ingresaron datos nuevos para realizar predicciones.
6. **Comparación e interpretación de resultados:** finalmente se comparó el modelo de machine learning con modelo geoestadístico para determinar capacidad predictiva. Se utilizó una matriz de confusión o tabla de clasificación y se interpretaron sus resultados.

Para implementar machine learning, se utilizó algunas librerías de Python como: Jupyter Notebook y Orange Canvas.

CAPÍTULO 4: DESARROLLO

4.1 Caso estudio

Como se ha mencionado anteriormente, en la oxidación de minerales sulfurosos ocurren una serie de reacciones que terminan por generar ácido sulfúrico, compuesto muy reactivo y altamente corrosivo. A pesar de esto, es muy utilizado en la minería, principalmente en el proceso de lixiviación, siendo el ferrocarril su principal vía de transporte. Aunque, durante los últimos años se han registrado descarrilamientos de trenes con cargamento de ácido sulfúrico en el norte del país, pudiendo hacer que el proceso de acidez se vea acelerado producto del derrame mismo. Para resolver el problema, se cuenta con datos provenientes de un estudio de contaminación por derrame de hidrocarburos y entendiéndose que el contexto es el mismo, la metodología mencionada aplica para ambos casos u otro tipo de elementos contaminantes que se encuentren en el suelo.

Respecto al emplazamiento, la contaminación fue originada por el descarrilamiento de un tren en una zona urbana y que transportaba vagones cisternas cargados con miles de litros de petróleo. Posterior al descarrilamiento, ocurrieron alrededor de cuatro y seis explosiones que arrasó con numerosos inmuebles.

La ciudad en cuestión se encuentra a orillas de un río y de un lago en el que desemboca.

Las muestras recopiladas fueron tomadas entre 0 y 2 metros de profundidad en rellenos antrópicos. Además, al momento de realizar muestro, se detectaron concentraciones significativas de níquel en zonas puntuales.

4.2 Análisis exploratorio de datos

Antes de aplicar cualquier técnica, es importante realizar un EDA para tener una primera aproximación acerca del entendimiento básico de los datos y las relaciones que puede haber entre las diferentes variables sometidas a estudio.

La base de datos cuenta con la siguiente información: identificación del sondaje o muestra, coordenadas UTM correspondiente para caso 2D, profundidad media, espesor de la capa donde se aloja la contaminación y concentración de elementos contaminantes; hidrocarburo y níquel.

De las variables previamente mencionadas, las que fueron sometidas a estudio y permitieron establecer zonas prioritarias de descontaminación son: hidrocarburos (ppm), níquel (ppm) y espesor de la muestra (m).

Para comprender la distribución de sondajes o muestras en el área de estudio, se recomienda observar la siguiente ilustración.

Ilustración 19: *Muestreo en área de estudio (Elaboración propia).*

De la ilustración anterior, podemos observar que el método utilizado para tomar muestras correspondió a un muestreo asistemático, es decir, muestreo al azar sin que la selección de puntos de la toma de muestras se haya realizado a distancias uniformes. Aun así, los puntos son representativos del área de estudio.

Se infiere que la toma de muestras se realizó de dicha forma, para poder comprender hasta que zonas fueron alcanzadas producto del derrame de hidrocarburos.

Es importante mencionar, que las muestras de hidrocarburos y níquel comparten la misma ubicación. Por otra parte, las muestras que contienen mayor concentración de contaminantes se representan con puntos más grandes y se observan en las siguientes ilustraciones.

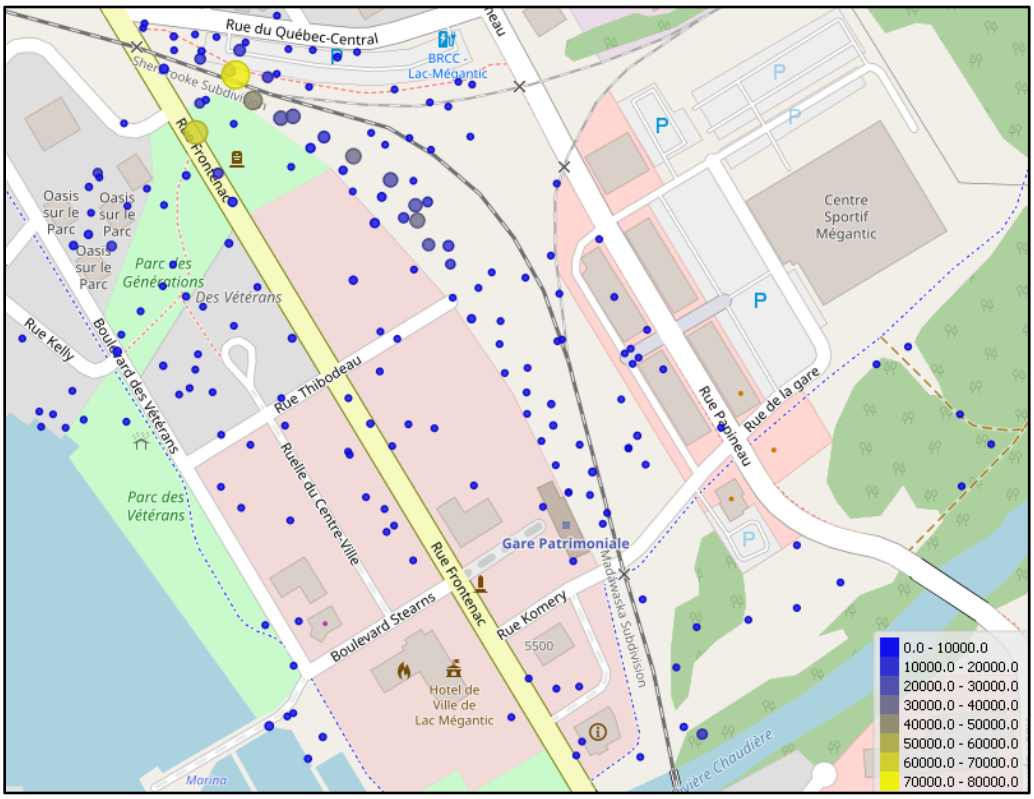

Ilustración 20: *Distribución espacial HC (Elaboración propia).*

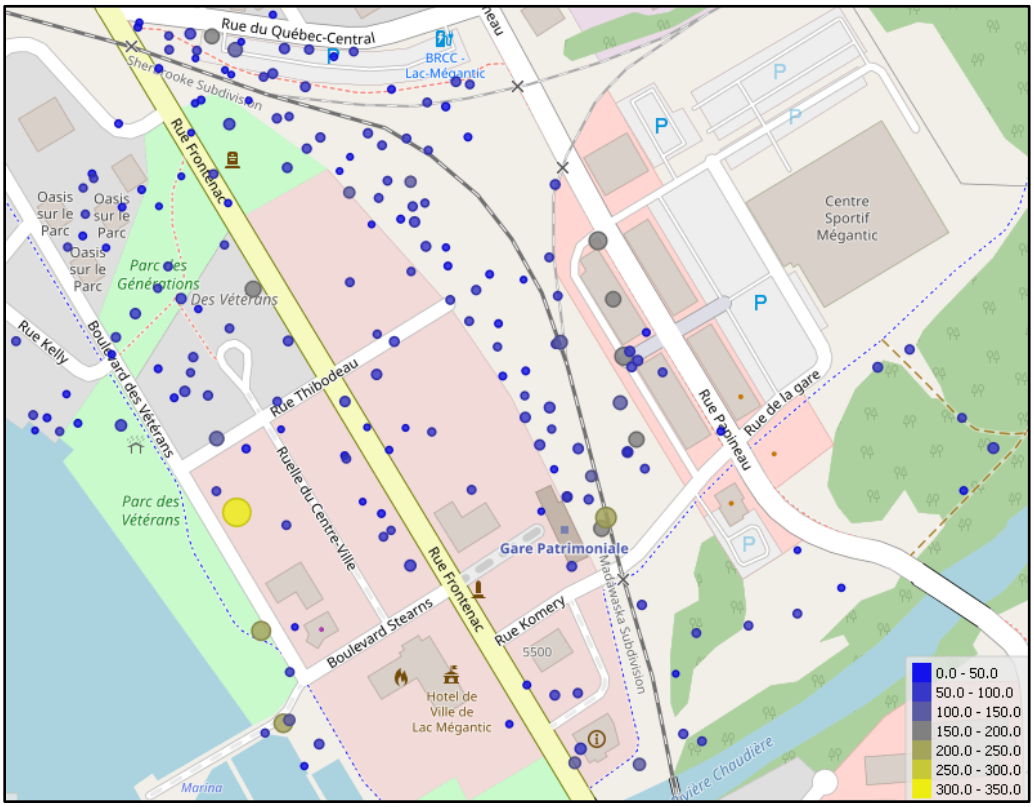

Ilustración 21: *Distribución espacial Ni (Elaboración propia).*

De las evidencias anteriores, se observa que precisamente en la zona donde ocurre el descarrilamiento (Sherbrooke Subdivisión), existen muestras con mayor concentración de hidrocarburos. Además, en zonas cercanas al lago producto del escurrimiento de petróleo al mezclarse con el agua utilizada para atacar los incendios generados por las explosiones.

Para el caso del níquel, las altas concentraciones de este metal se observan en caso puntuales del área de estudio.

Con este análisis general, se tiene una primera aproximación acerca de cuáles serán las zonas más bajas o altas respecto a contaminación.

Finalmente, se obtienen estadísticas descriptivas de las variables en estudio y se emplean algunas herramientas gráficas.

Tabla 7: *Estadísticas descriptivas elementos contaminantes (Elaboración propia).*

	Hidrocarburos (ppm)	**Níquel (ppm)**
Cantidad de datos	192	192
Mínimo	100	30
Máximo	79000	340
Media	3842.3	66.8
Mediana	100	57.5
Varianza	9.75E+07	1615.5
Error estándar	712.5	2.9
Desviación estándar	9872.5	40.2
Q1	100	45
Q3	1475	73
Skewness	4.5	3.1
Curtosis	25.1	13.6
Coeficiente de variación	256.9	60.2

De la tabla 7, las conclusiones más importantes que se pueden extraer son:

- En total se recogieron 192 muestras con las cuales se realiza la estimación. Además, se conoce los valores mínimos y máximos de los distintos contaminantes.
- Se deduce que ambos contaminantes no siguen una distribución normal, dado que los valores de la media y la mediana no coinciden. Esto ocurre por la presencia de valores extremos. Por ejemplo, en el caso de los hidrocarburos el 75% de los datos tienen una concentración inferior a 1475 ppm y el promedio para aquella variable es alrededor de 3840

ppm. De esta manera, se espera que ambos casos presenten una forma logarítmica, donde la frecuencia relativa más significativa corresponda a concentraciones bajas, próximas al límite de detección de laboratorio. Este tipo de distribuciones es muy normal cuando se trata de suelos contaminados.

- La varianza en el caso de los hidrocarburos es extremadamente alta, por lo que podría presentar inconvenientes al momento de realizar algún método de estimación.
- Los resultados de curtosis y skewness, indican que la distribución de ambas variables presenta asimetría positiva con apuntamiento leptocúrtica.
- Finalmente, las estadísticas descriptivas como las herramientas gráficas utilizadas para analizar la variable espesor, se pueden observar en apéndice A.

Respecto a la distribución de los datos de hidrocarburos y níquel, se presentan a continuación.

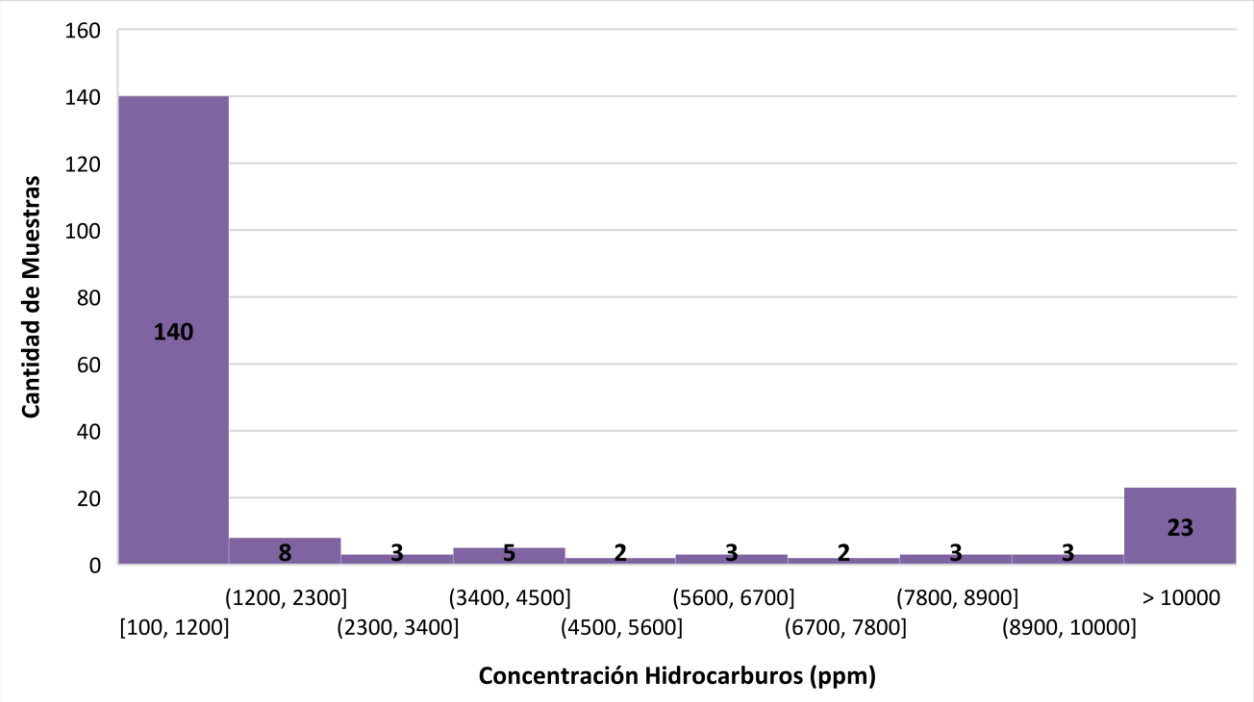

Ilustración 22: *Histograma HC (Elaboración propia).*

Como ya se había mencionado, la mayor parte de los datos se encuentra en valores bajos de contaminación, donde para el caso del hidrocarburo aproximadamente el 80% de las muestras ronda en valores de 100 y 2300 ppm. En cambio, solo un 12% de las muestras tiene concentración superior a 10000 ppm.

Este tipo de distribución alejada de la normalidad, tendrá consecuencias importantes a la hora de aplicar ciertas técnicas de estimación como las simulaciones condicionales gaussianas.

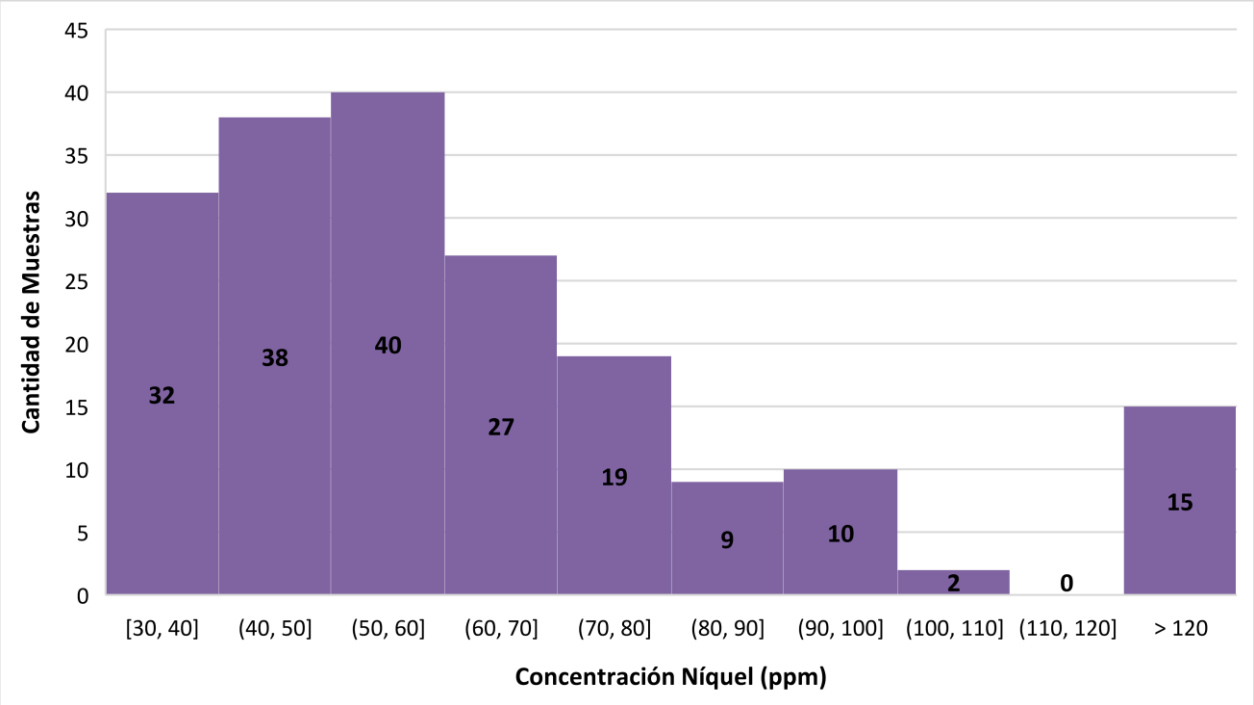

Ilustración 23: *Histograma Ni (Elaboración propia).*

En el caso del histograma del níquel, alrededor del 40% de las muestras tiene una concentración que ronda entre 40 y 60 ppm. En cambio, los datos que se escapan de la tendencia son solo un 8% de las muestras y tienen concentración superior a 120 ppm. Además, se aprecia una distribución más homogénea en comparación al histograma del hidrocarburo.

Las ilustraciones 24 y 25 corresponden a diagramas de cajas y bigotes; gráficos utilizados principalmente para detectar errores en la base de datos. Sin embargo, en este caso, los datos anómalos encontrados están intrínsecamente ligados al muestreo de partida, donde los valores máximos dan idea de contaminación.

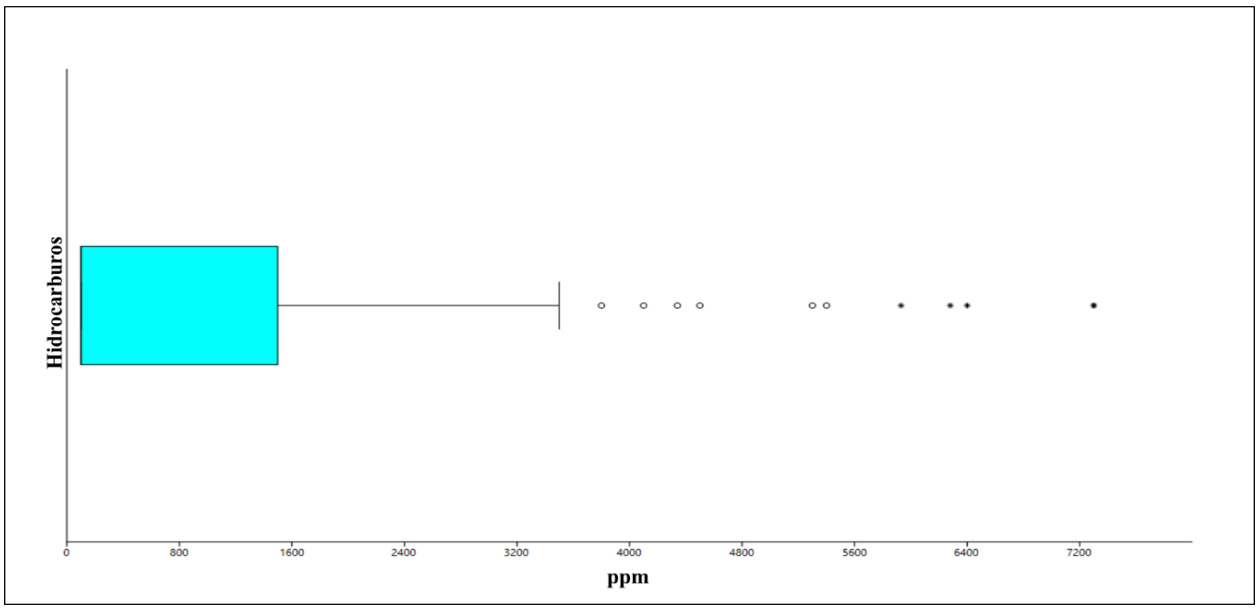

Ilustración 24: *Diagrama de cajas y bigotes HC (Elaboración propia).*

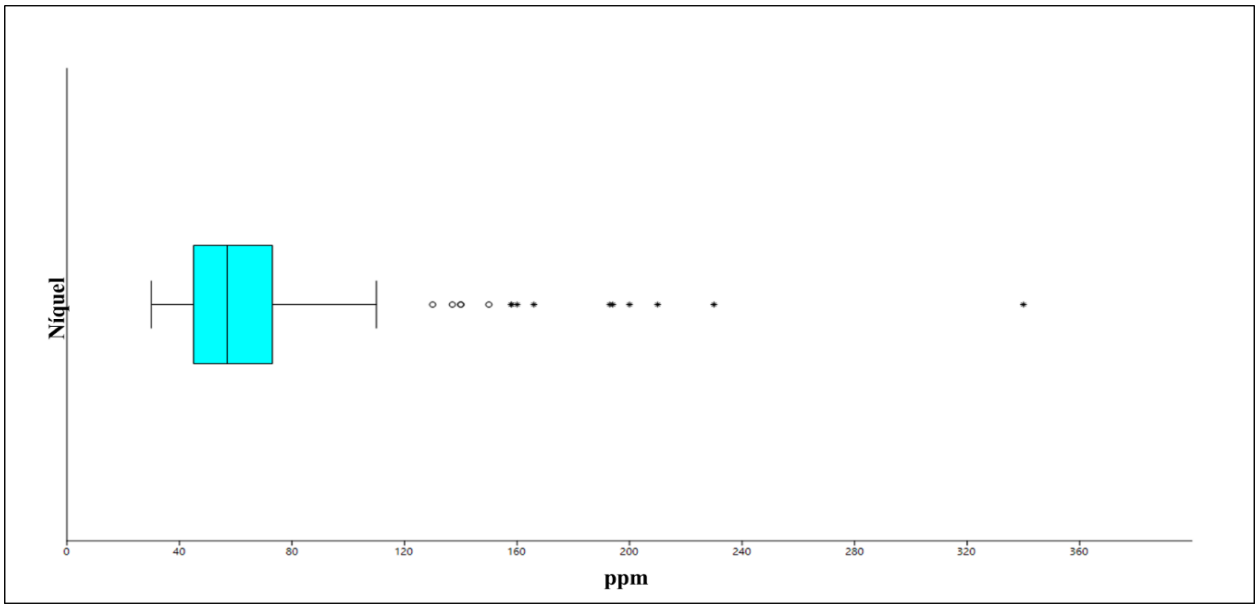

Ilustración 25: *Diagrama de cajas y bigotes Ni (Elaboración propia).*

Para el hidrocarburo, se consideran valores atípicos o anómalos aquellas muestras con concentración superior a 3537 ppm. Esto significa, que de las 192 muestras iniciales hay 40 datos considerados atípicos. En cambio, las muestras de níquel con concentraciones superiores a 115 ppm se consideran datos atípicos, teniendo un total de 15 datos. Un número bastante menor respecto a las muestras de hidrocarburos.

4.2.1 Análisis bivariado

Se generan gráficos de correlación para poder determinar dependencia o independencia de una variable respecto de la otra.

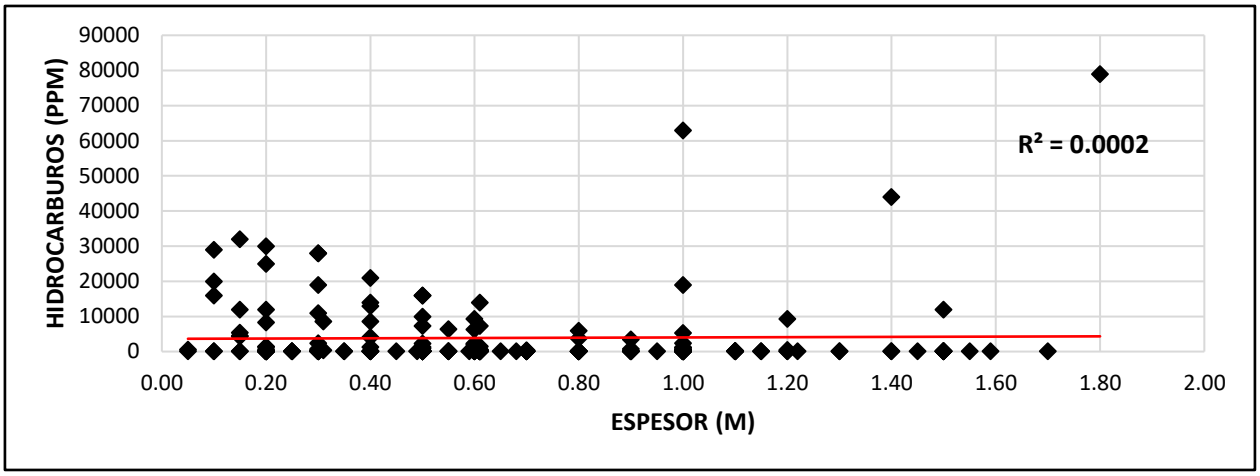

Ilustración 26: *Diagrama de dispersión HC vs espesor (Elaboración propia).*

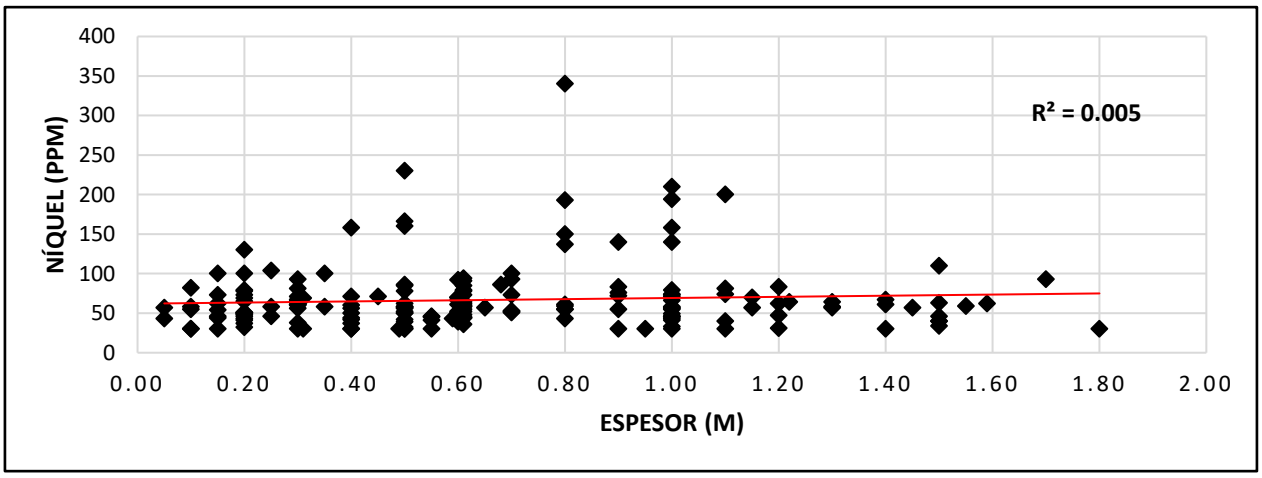

Ilustración 27: *Diagrama de dispersión Ni vs espesor (Elaboración propia).*

Con los gráficos anteriormente expuestos, se descarta que los valores de concentración obtenidos dependan del espesor de la muestra que aloja la contaminación. Además, dado que el hidrocarburo no es propio del terreno de estudio, no existe correlación entre la variable HC y Ni, por lo que es coherente que sus variogramas presenten estructuras distintas.

4.3 Selección del método de estimación

Los métodos geoestadísticos utilizados fueron los siguientes: kriging ordinario, kriging de indicadores y simulaciones condicionales gaussianas, mientras que, para machine learning se utilizaron redes neuronales. Sin embargo, antes de aplicar la técnica del kriging o cualquier otro

tipo de técnica de interpolación, es necesario definir cuál será el área de estudio. Para esto, se definió un tamaño de grilla o mallado que está condicionado por el tamaño de las muestras (valores máximos y mínimos de coordenadas Este y Norte), dado que debe permitir que todas las muestras se encuentren dentro de él. Esta grilla es el límite de las interpolaciones, es decir, todas las estimaciones de las diferentes técnicas aplicadas se generan dentro de esta área de estudio.

Tabla 8: *Estadísticas descriptivas base de datos (Elaboración propia).*

	Coordenadas Este (m)	**Coordenadas Norte (m)**
Máximo	353501	5048890
Mínimo	352867	5048400
Rango	634	490

Observando la tabla 8, se aprecia que en el eje de las abscisas hay valores entre 352867 y 353501 metros, lo que da una longitud de 634 metros, mientras que, en el eje de las ordenadas, estos varían entre 5048400 y 5048890 metros, dando una longitud de 490 metros. Con esto, se estableció un rango de 650 metros dirección Este y 512 metros dirección Norte. De esta manera, la grilla no quedó al límite de las muestras y todos los valores quedaron dentro del área de estudio, dando una superficie total de 33 hectáreas, lo que es equivalente a 43 canchas de fútbol profesional.

De manera general, en contaminación de suelos se realiza estimación de bloques, ya que se piensa en los futuros trabajos de descontaminación que se llevan a cabo en mallas de varios metros. Por esto, el número de celdas que se obtienen en el mallado dependen del tamaño de bloque con que se desea trabajar. En este caso, se trabajó con tamaños de bloques de 2x2 y de 10x10, a fin de determinar influencia en los resultados cuando se trabaja con bloques de distintos tamaños. Los parámetros ingresados para la creación del mallado se observan en la tabla 9.

Tabla 9: *Parámetros malla (Elaboración propia).*

	Mallado 2x2	**Mallado 10x10**
Número de bloques en X	325	65
Número de bloques en Y	256	52
Coordenada origen X	352860	352865
Coordenada origen en Y	5048393	50483393

4.4 Variogramas

Para poder estudiar la correlación espacial de las variables se requieren los variogramas, por lo tanto, es un proceso clave dentro del estudio geoestadístico, porque a partir de la información proporcionada por él se puede generar estimación de contaminantes en el sitio de estudio. Por lo tanto, se definieron algunos parámetros como: *numbers of lags, lag separation, lag tolerance, number of directions, azimuth, dip, tolerance y bandwith* para crear los variogramas experimentales. Conviene subrayar, que estos parámetros se mantuvieron fijos en los 3 casos geoestadísticos aplicados: kriging ordinario, kriging indicadores y simulaciones condicionales gaussianas.

Al tratarse de datos irregularmente espaciados pueden ocurrir diferentes fenómenos, tales como:

- Ocurrir que no existan valores de la variable a la distancia h.
- Ocurrir que no existan valores de la variable en una respectiva dirección.

Por esto, no se pudo utilizar un tamaño de *lag* que fuese parecido al espaciamiento de los datos. En consecuencia, entendiendo que el variograma debe ser representativo en la generalidad del fenómeno estudiado, se usó un tamaño de *lag* que permitiera hacer el cálculo con una cantidad considerable de pares de puntos y un número de *lag* que pudiera representar tanto el dominio de la zona de derrame, así como el emplazamiento general. Por lo tanto, se definió un tamaño de *lag* de 40 metros con 15 pasos, de esta manera poder abarcar casi la totalidad del sector. Mientras que, el *lag tolerance* corresponde a la mitad del tamaño del *lag.*

Por otra parte, se realizó un análisis en la horizontal (dip igual a cero), considerando 4 direcciones diferentes: Norte-Sur, Este-Oeste y sus respectivas diagonales, de esta manera poder detectar diferentes direcciones de anisotropía si es que las hubiese. En caso contrario, trabajar con variograma omnidireccional. Aunque, es altamente probable que exista alguna dirección de preferencia, donde las razones que lo justifican se exponen a continuación:

- En el caso estudio, el combustible derramado es un fluido, por lo tanto, es un proceso dinámico, es decir, un sistema cuyo estado evoluciona con el paso del tiempo y esta evolución está condicionada por la geología de la zona.
- De la geología, algunas causas que podrían influir son el tipo de suelo en que se está trabajando el fenómeno; arcilloso, arenoso, salino, limoso, calizo, los cuales determinan el

grado de permeabilidad y porosidad. También, se debe tener en consideración las canalizaciones o cavidades subterráneas. Estos factores se desconocen, ya que se está trabajando en una única capa del suelo, dado que la toma de muestras fue echa a nivel superficial.

- La orientación de las vías del tren en la zona que se produjo el descarrilamiento tiene influencia en la dirección de anisotropía.
- Finalmente, el último factor a considerar es la fuerza de gravedad.

Las tablas 10 y 11 resumen los parámetros mencionados anteriormente.

Tabla 10: *Parámetros lags variogramas experimentales (Elaboración propia).*

Numbers of lags	Lag separation	Lag tolerance
15	40	20

Tabla 11*: Parámetros direccionales variogramas experimentales (Elaboración propia).*

Azimuth (°)	Dip (°)	Tolerance	Bandwith
0	0	22.5	500000
45	0	22.5	500000
90	0	22.5	500000
135	0	22.5	500000

En el caso del kriging ordinario, se obtuvieron variogramas experimentales direccionales que presentaron comportamiento poco estructurado, más bien algo errático , con saltos entre cada punto calculado a diferentes distancias, donde los rangos obtenidos para todas las direcciones caen ligeramente en el mismo intervalo, siendo un poco mayor para 45° azimuth, lo que significa que la variable es mucho más continua en dicha dirección y corresponderá al eje mayor. En cambio, el eje menor corresponderá a la perpendicular del eje mayor, que en este caso es 135° azimuth.

Las siguientes ilustraciones corresponden a los variogramas experimentales para el caso de kriging ordinario.

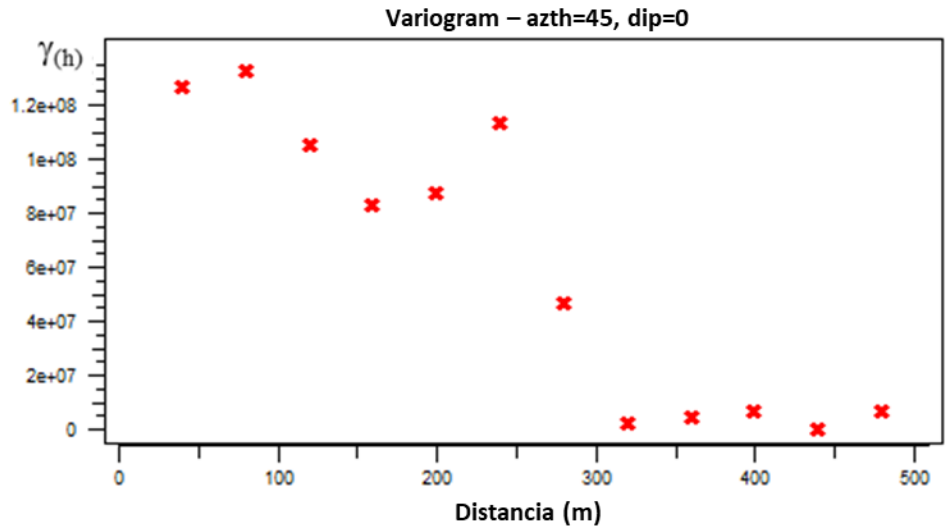

Ilustración 28: *Variograma experimental 45° HC (Elaboración propia).*

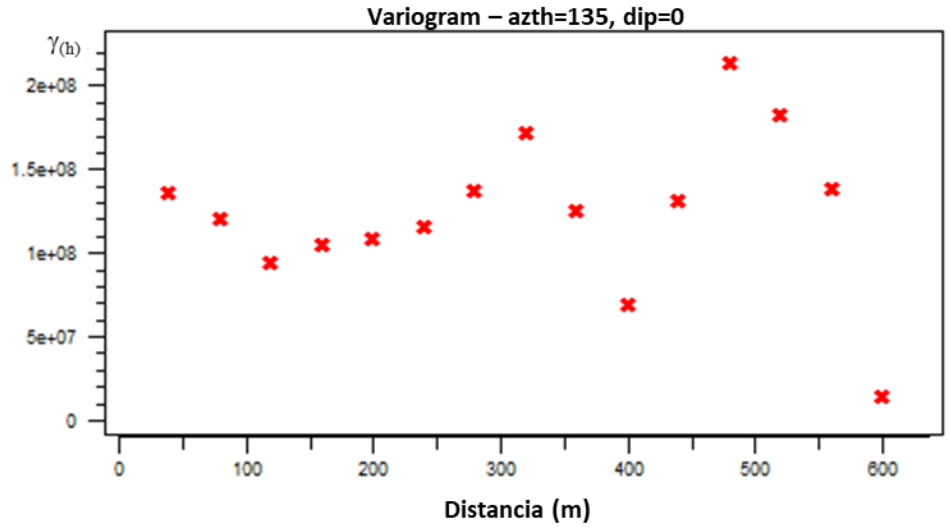

Ilustración 29: *Variograma experimental 135° HC (Elaboración propia).*

Los variogramas experimentales, presentan gran movilidad del valor gamma producto de la alta variabilidad que existen en los datos.

Respecto al modelado del variograma, se asignó un efecto pepita de 0.1 con una meseta cercana al valor de la varianza de los datos del hidrocarburo. Luego, se ajustó a un modelo esférico debido a que presenta un comportamiento lineal en el origen, expresando un rango o alcance de 96 metros en ambas direcciones. Lo anteriormente expuesto, se expresa en las ilustraciones 30 y 31.

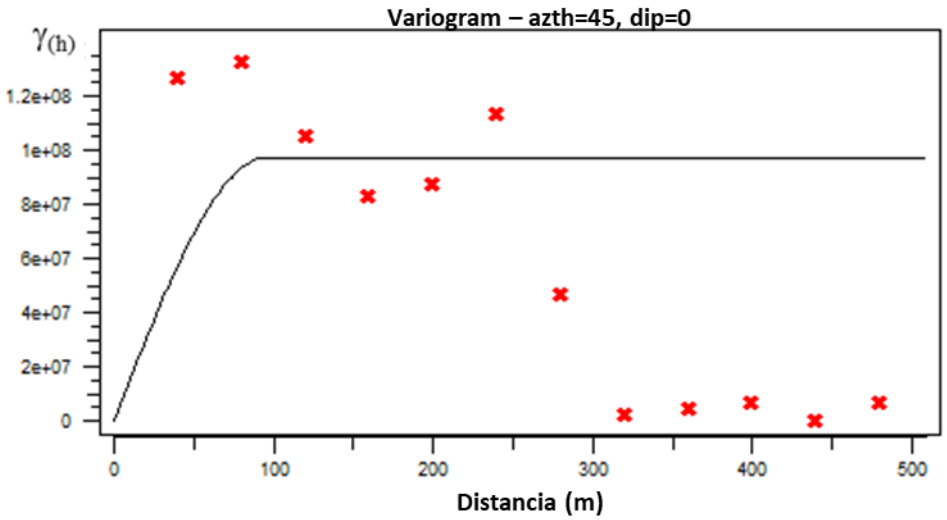

Ilustración 30: *Variograma modelado 45° HC (Elaboración propia).*

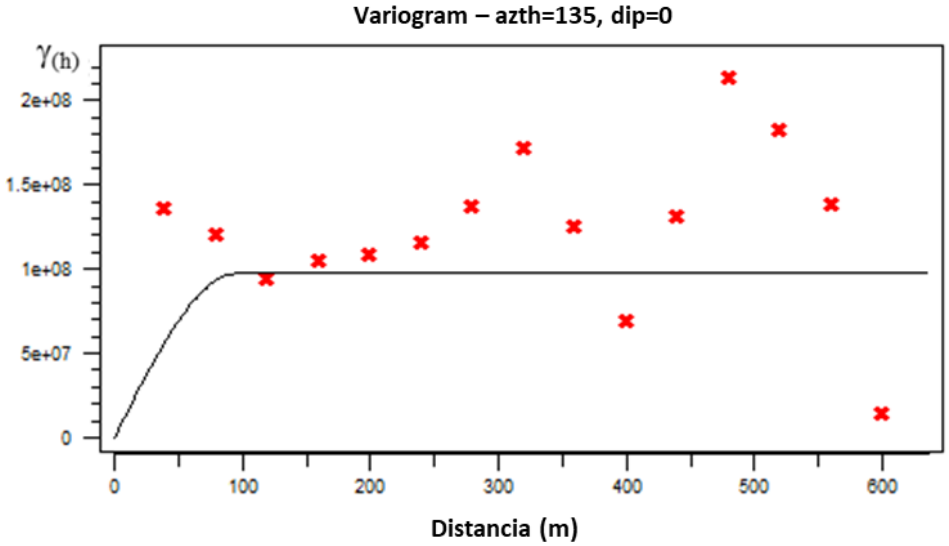

Ilustración 31: *Variograma modelado 135° HC (Elaboración propia).*

Por otra parte, los variogramas para variable níquel y espesor, se pueden observar en apéndice B.

Para el caso de kriging de indicadores, es necesario previamente transformar la base de datos en función de un criterio de contaminación establecido. Pero ¿Cuáles son los límites máximos permisibles para los elementos contaminantes estudiados? En este contexto, existe un vacío en la legislación chilena, ya que no cuenta con una normativa que permita calificar un terreno como contaminado. De este modo, se usó como referencia la normativa internacional mexicana "Nom-147 Semarnat/SSA1-2004" y la canadiense "Soil Quality Guidelines". Se hace necesario resaltar que el límite máximo permisible está definido en función del uso del suelo. Como la contaminación en este caso ocurre en una zona residencial, los límites máximos permisibles son los siguientes:

750 ppm para hidrocarburo y 100 ppm para níquel. Definido estos límites, se transforma la base de datos reemplazando con un 0 aquellos datos que estén bajo el criterio establecido y con un 1 los datos que estén sobre el criterio establecido. En consecuencia, se tiene una base de datos con variables categóricas, donde las estadísticas descriptivas para esta renovada base de datos se aprecian en apéndice A.

Los variogramas experimentales y modelados para las distintas variables con caso kriging de indicadores, se observan en apéndice B.

Finalmente, para aplicar la técnica de simulaciones condicionales, se requiere que los datos de partida sigan una distribución normal, en caso contrario, hay que transformarlos. A diferencia del caso de kriging de indicadores, solo la variable de hidrocarburo se le aplica esta técnica de estimación.

Previamente, en el análisis exploratorio de datos ya se confirmó que la variable hidrocarburo no sigue una distribución normal, sino que parece seguir una distribución logarítmica. En consecuencia, los datos fueron transformados mediante el algoritmo trans del software SGeMS a una distribución normal.

Las estadísticas descriptivas como sus variogramas para este caso, se pueden observan en apéndice A y B respectivamente.

4.5 Construcción de modelo machine learning

La técnica utilizada para construir modelo machine learning fue redes neuronales artificiales. Una técnica de aprendizaje supervisado, donde el algoritmo aprende de los ejemplos proporcionados las reglas que permiten predecir etiquetas de los nuevos casos y no se conozca su valor.

Para implementar esta técnica, se utilizó el software de aprendizaje automático y minería de datos Orange Canvas y que incluye bibliotecas comunes como numpy, scipy y scikit-learn. Aunque, previamente a generar modelo en software, se definieron parámetros propios de la red neuronal, tales como: número de neuronas por capa oculta, función de activación, solucionador para la optimización del peso, tasa de aprendizaje y número máximo de iteraciones.

Construir un modelo de machine learning puede resultar agotador, porque se deben construir modelos iterativos debido a que no se sabe con exactitud cuáles son sus parámetros óptimos.

Entonces, como no es posible definir a priori sus parámetros, se decidió construir 4 modelos de redes neuronales que mantuvieran constante las siguientes variables: función de activación, solucionador para la optimización del peso y tasa de aprendizaje, en tanto el número de neuronas por capa oculta y máximo de iteraciones sufrieron variaciones. El valor de estos parámetros, se pueden observar en tabla 12.

Tabla 12: *Parámetros redes neuronales (Elaboración propia).*

Modelo	Neuronas por capa oculta	Máximo de iteraciones	Función de activación	Optimizador	Tasa de aprendizaje
NN 100_200	100	200	ReLU	ADAM	0.0001
NN 1000_200	1000	200	ReLU	ADAM	0.0001
NN 100_2000	100	2000	ReLU	ADAM	0.0001
NN 1000_2000	1000	2000	ReLU	ADAM	0.0001

Se optó por la función de activación de unidad lineal rectificada (ReLU), porque es sencilla y no posee regiones de saturación como las funciones tangentes hiperbólica y sigmoidal, que provocan estancamiento durante el entrenamiento de la red. Por lo tanto, cuando se ocupa el optimizador descenso de gradiente con función ReLU, se obtienen resultados que convergen mucho más rápidos.

Por último, se eligieron valores bajos y altos en los parámetros que variaron, a fin de determinar su grado de influencia en las estimaciones.

La siguiente ilustración corresponde a diagrama generado para aplicar red neuronal en software Orange Canvas.

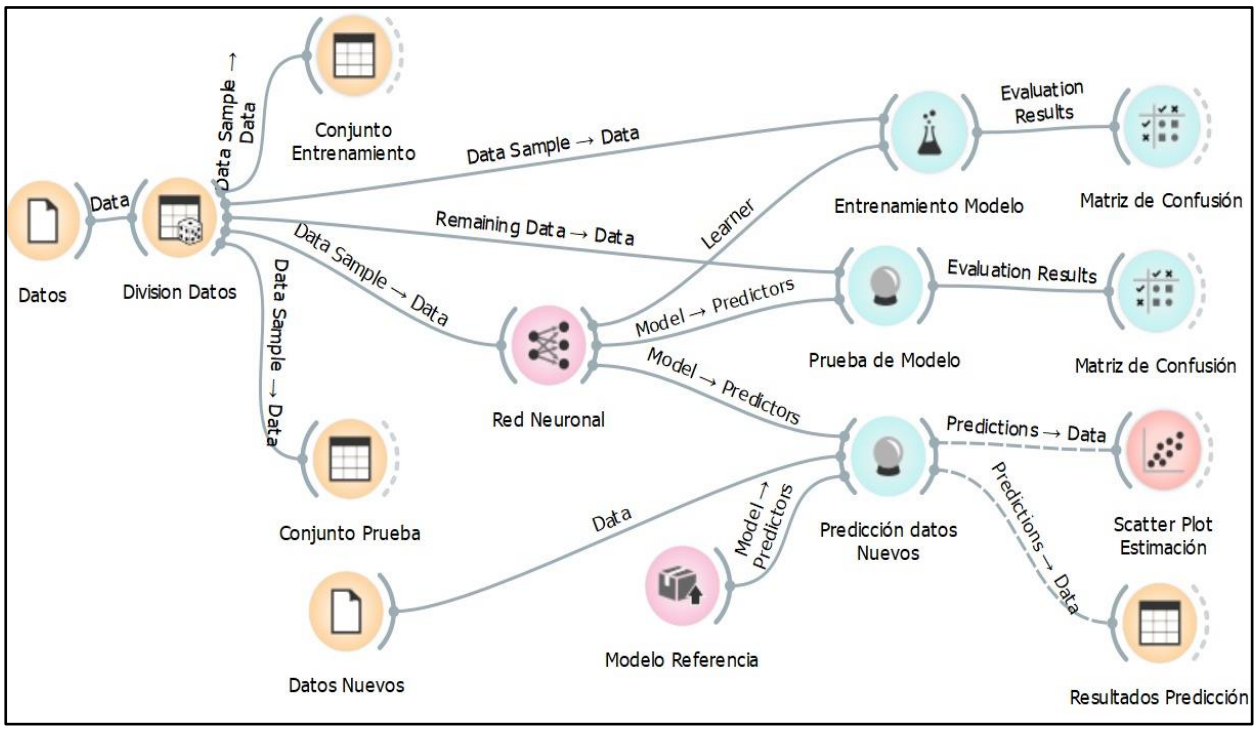

Ilustración 32: *Diagrama modelo red neuronal Orange Canvas (Elaboración propia).*

Pero, ¿Cómo entrenar redes neuronales en un caso por contaminación por DAM? Si se dispone de un modelo cinético, este sería utilizado para entrenar a la red neuronal, ya que contaría con datos adicionales a las muestras y que fueron generados de forma sintética por el modelo cinético.

CAPÍTULO 5: RESULTADOS Y DISCUSIÓN

En el presente capítulo, se exponen resultados de estimaciones realizadas por métodos geoestadísticos y machine learning.

Cabe señalar que, para realizar estimación por métodos geoestadísticos, se debe escoger un elipsoide de búsqueda para definir cantidad de datos que entran en el proceso de cálculo. Se trabajó con un mínimo de 2 y máximo de 12 datos, con un elipsoide configurado por un radio de 300 x 300 con el fin de tomar mayor número de muestras posibles.

5.1 Estimación kriging ordinario

Hidrocarburos

En la ilustración 33 se muestra el resultado de estimación por kriging ordinario para variable hidrocarburo con tamaño de bloques de 2x2.

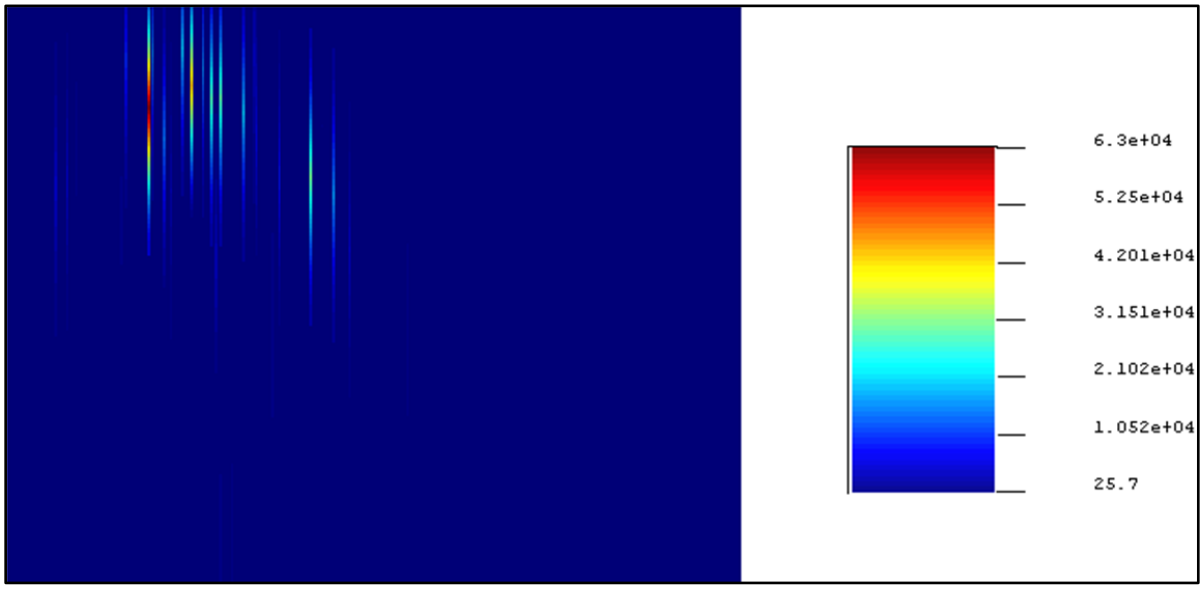

Ilustración 33: *Estimación kriging ordinario HC para bloques 2x2.*

La ilustración 33 demuestra que no se obtienen resultados que permitan definir zonas de contaminación cuando se estima mediante kriging ordinario. No obstante, no es algo que llame la atención, pues es un resultado esperable al tener datos de inicio con una varianza excesivamente alta. Atendiendo a estas consideraciones, es que se busca una nueva alternativa que permita disminuir valor de la varianza, por ejemplo: ¿Qué ocurre si se estima sin datos atípicos?

Caso estimación sin datos atípicos

El EDA de inicio nos dice que se consideran datos atípicos aquellas muestran con concentración superior a 3537 ppm. Esto significa, que de las 192 muestras iniciales hay 40 datos considerados atípicos. Por lo tanto, la estimación realizada mediante este caso se realiza con 152 muestras.

Antes de aplicar kriging ordinario, se repitieron los procedimientos de EDA y modelación variográfica. Estos resultados se pueden observar en apéndice A y B respectivamente.

La ilustración 34 y 35 muestra el resultado de estimación por kriging ordinario sin datos atípicos con tamaño de bloques de 2x2 y 10x10.

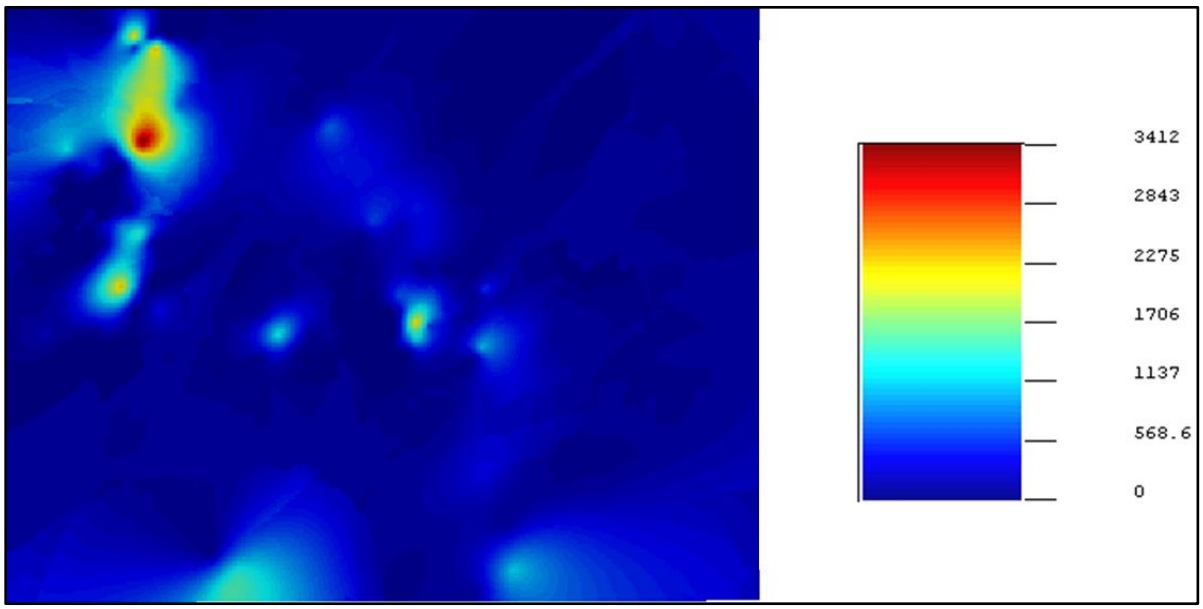

Ilustración 34: *Estimación kriging ordinario HC, caso sin datos atípicos bloques 2x2 (Elaboración propia).*

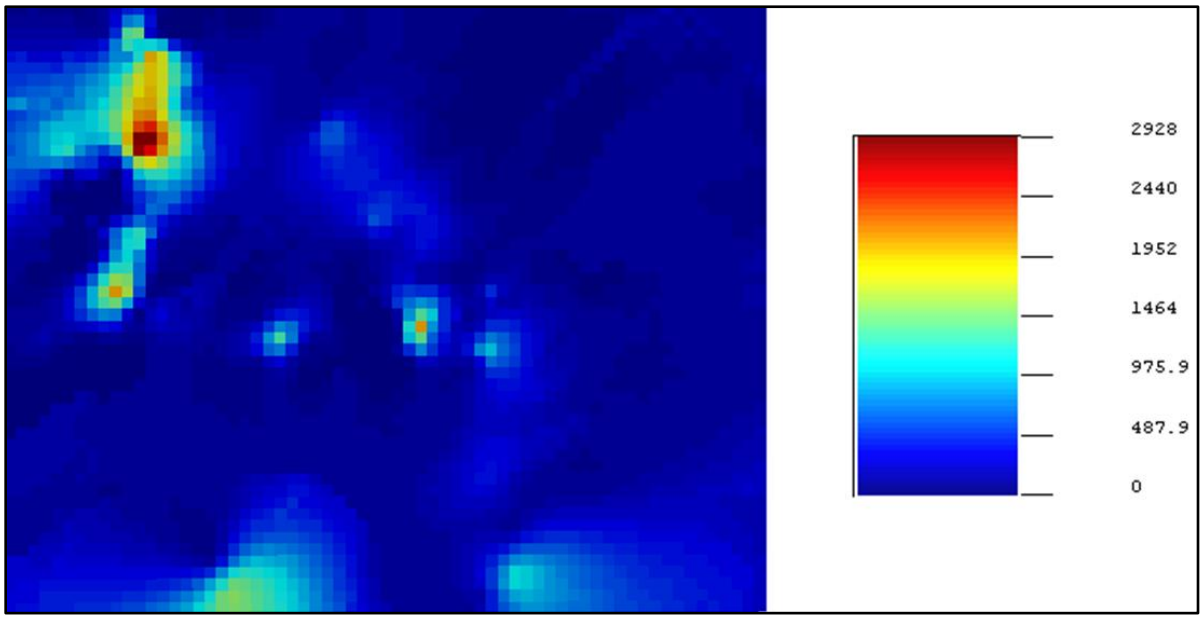

Ilustración 35: *Estimación kriging ordinario HC, caso sin datos atípicos bloques 10x10 (Elaboración propia).*

De los resultados anteriores, se aprecia que la zona noroeste presenta mayor concentración de hidrocarburo. Esto tiene mucho sentido, ya que fue la zona donde se produjo el descarrilamiento del tren con combustible. También, se aprecia que existe contaminación en zonas cercanas al lago, lo que podría significar que la contaminación sigue fuera de los límites de la zona estudiada. Sin embargo, hay que recordar que para este caso no se consideraron datos mayores a 3537 ppm, por lo que esta estimación podría estar subestimada.

En relación con las diferencias que se encuentran al utilizar distintos soportes, se puede mencionar que, al utilizar tamaño de bloques mucho más pequeños, se obtiene una mejor precisión de la dispersión de los elementos. También, se obtiene una variación en la escala de concentración de hidrocarburo, donde la estimación realizada con bloques de 2x2 entrega valores más altos en comparación a bloques de 10x10.

De modo general, si se aumenta el tamaño del bloque, se obtiene una varianza de la predicción mucho menor.

En la ilustración 36 y 37 se muestra la varianza de estimación del hidrocarburo con tamaño de bloques de 2x2 y 10x10.

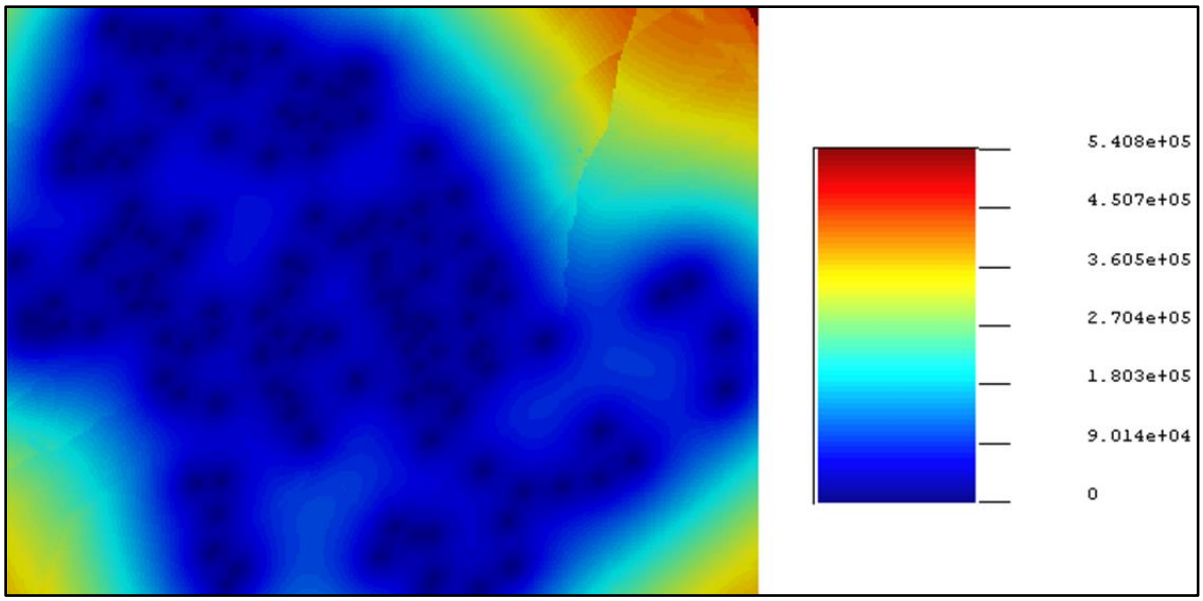

Ilustración 36: *Varianza kriging ordinario HC, caso sin datos atípicos bloques 2x2 (Elaboración propia).*

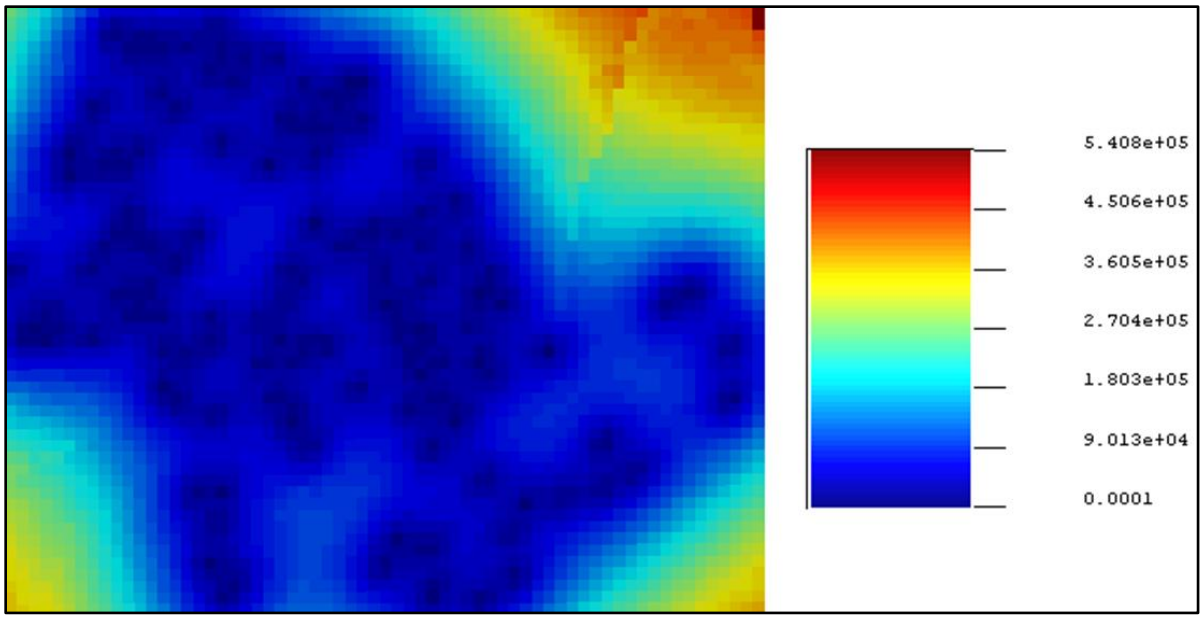

Ilustración 37: *Varianza kriging ordinario HC, caso sin datos atípicos bloques 10x10 (Elaboración propia).*

Respecto a la varianza de estimación, expresa el grado de precisión de la estimación realizada, pero no tiene en cuenta la noción de incertidumbre de la estimación.

Si se analiza los resultados de la varianza de estimación, se puede observar que los puntos azules oscuros corresponden a un dato de muestreo y representan los valores de varianza mínimo, puesto que en esa localización se dispone de un dato conocido. A medida que la estimación se aleja de esos puntos, los valores de varianza aumentan y se representan por tonalidades azules y verdes suaves, hasta alcanzar valores más significativos que vienen representados por tonalidades

amarillos y naranjas. Esto significa que, las estimaciones realizadas con varianzas alta carecen de seguridad y de confianza por no disponer de datos en dichas zonas.

Realizado el kriging ordinario, podemos visualizar el histograma del kriging y estadísticas descriptivas de la estimación.

Tabla 13: *Estadísticas descriptivas estimación, kriging ordinario HC caso sin datos atípicos bloques 2x2 (Elaboración propia).*

Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
83200	-196.6	3411.8	238.02	112.3	103029	273.214

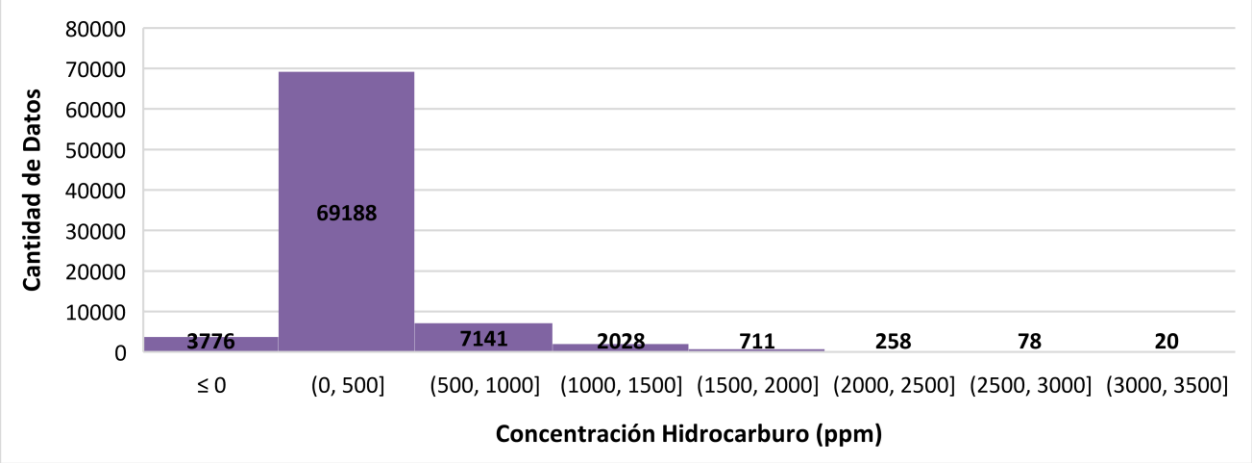

Ilustración 38: *Histograma de estimación kriging ordinario HC, caso sin datos atípicos bloques 2x2 (Elaboración propia).*

Si comparamos las estadísticas descriptivas de las estimaciones con las del análisis exploratorio de datos; caso sin datos atípicos *(ver apéndice A: caso sin datos atípicos)*, podemos ver que hay una disminución de la varianza y del promedio de los datos.

Analizando el histograma, se observa que los valores máximos puntuales de la distribución de origen desaparecen. Asimismo, la frecuencia relativa de los valores cercanos a cero también es menor en el histograma de los valores estimados. Sin duda, el kriging tiene un efecto de suavizado, ya que diluye los valores extremos, ya sean máximos o mínimos. Además, el promedio de la distribución obtenida para los valores estimados es menor que la observada para los datos de partida *(ver apéndice A: caso sin datos atípicos).* Esto podría significar que al aplicar la técnica del kriging, la contaminación del hidrocarburo en los suelos está siendo subestimada.

El histograma del kriging y estadísticas descriptivas de la estimación para bloques 10x10 se pueden observar en apéndice C.

Níquel

Para el níquel, se obtuvieron los siguientes resultados al aplicar kriging ordinario.

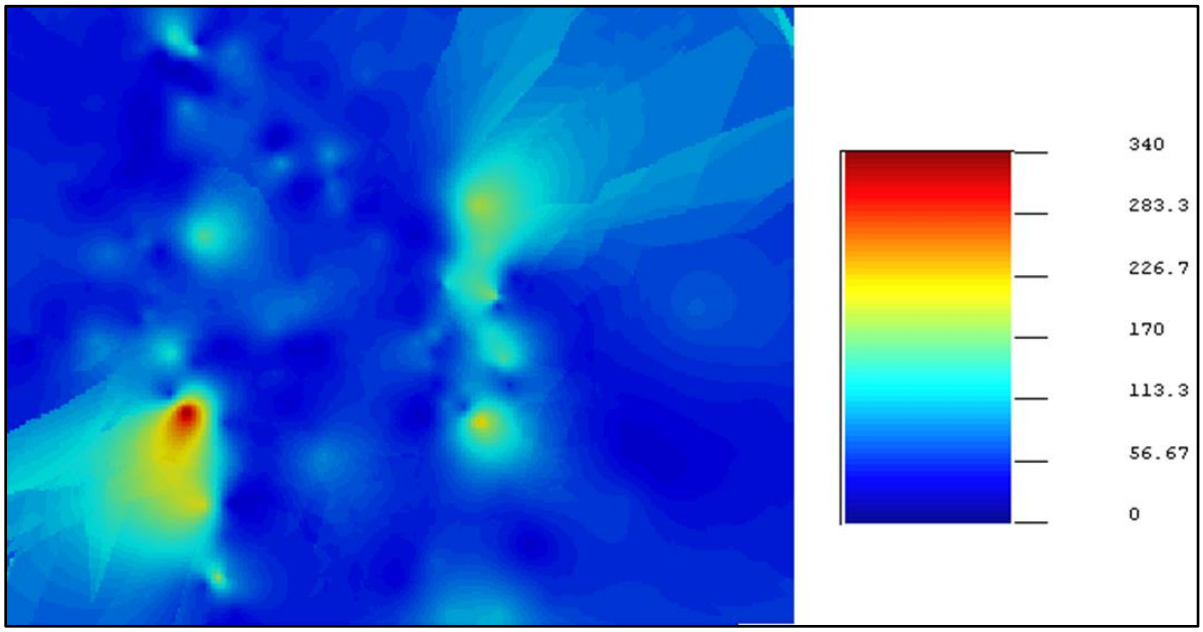

Ilustración 39: *Estimación kriging ordinario Ni bloques 2x2 (Elaboración propia).*

La alta concentración de níquel ocurre en zonas puntuales, tales como: zona cercana al lago y división de vía férrea, con un máximo de concentración de 340 ppm. Sin embargo, por el resultado de la varianza de estimación aquellas zonas carecen de seguridad y de confianza por no contar con la suficiente cantidad de muestras.

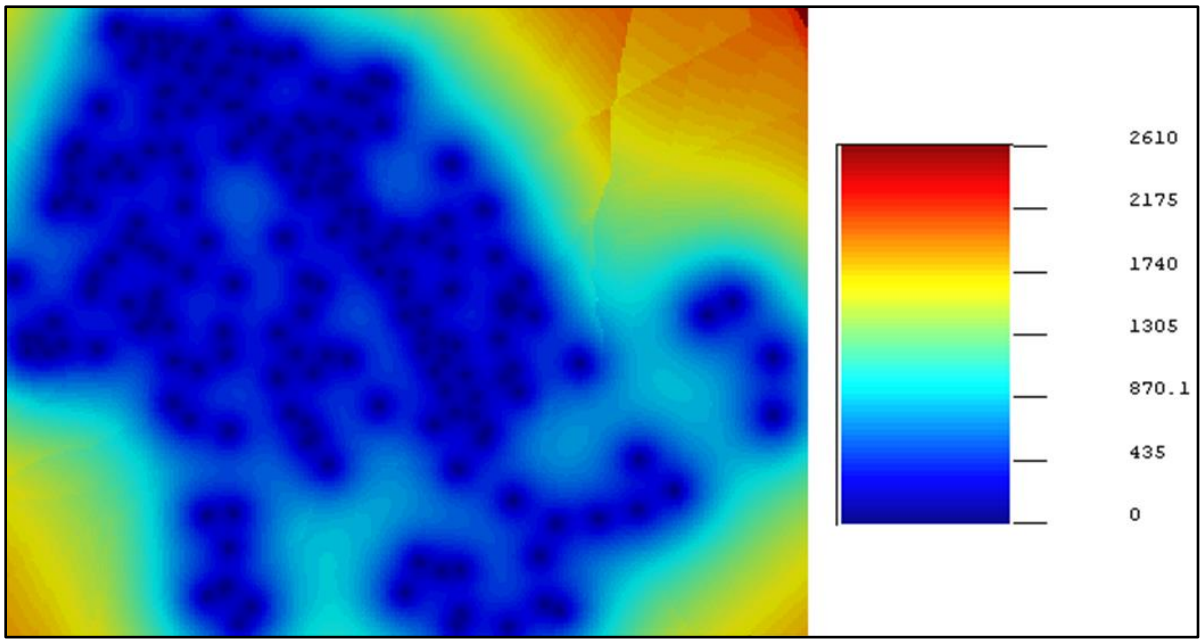

Ilustración 40: *Varianza kriging ordinario Ni bloques 2x2 (Elaboración propia).*

El histograma del kriging y estadísticas descriptivas de la estimación para níquel se pueden observar en apéndice C. En cambio, su estimación con tamaño de bloques 10x10 en apéndice E. Por último, los resultados de estimación de la variable espesor se pueden observar en apéndice C y E.

5.2 Estimación kriging de indicadores

Los resultados obtenidos mediante kriging de indicadores, representan una estimación de la probabilidad de superar o no el límite de descontaminación que se haya fijado, es decir, no expresan la concentración de un contaminante en un punto determinado. De acuerdo con la normativa se fijó un criterio de contaminación de 750 ppm para hidrocarburo y 100 ppm para níquel.

Para este caso solo se trabajó con bloques de 2x2, ya que en el caso anterior se explica la influencia que se tiene al trabajar con distintos tamaños de soporte.

Hidrocarburo

Los resultados de la estimación por kriging de indicadores se observan en la siguiente ilustración.

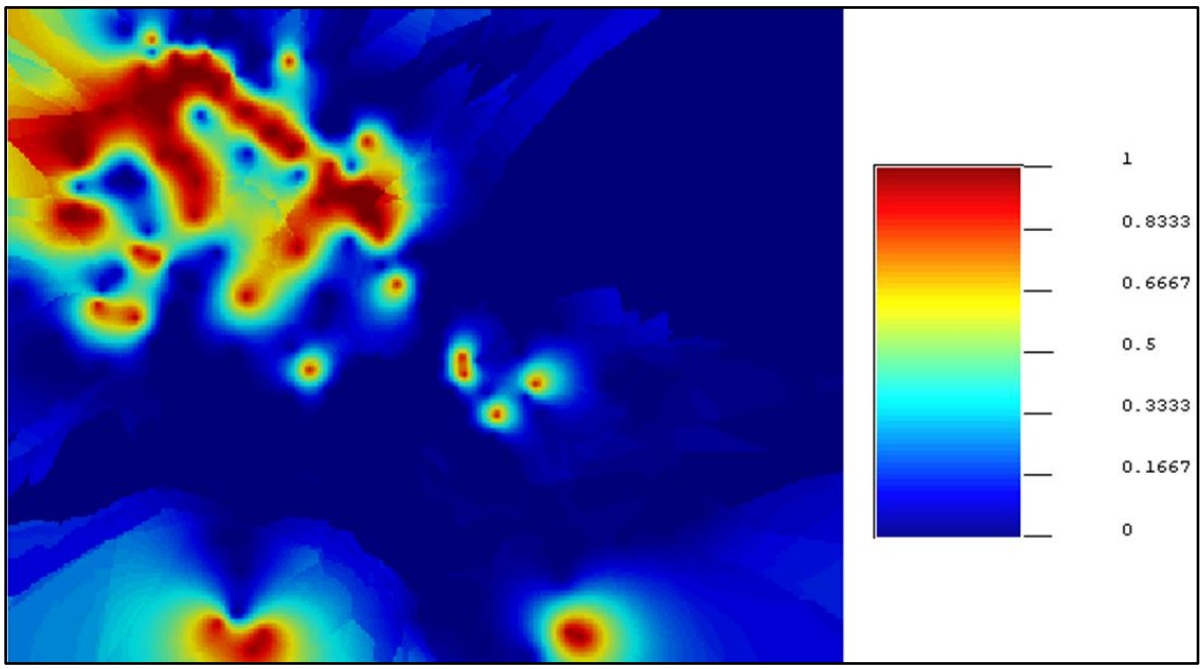

Ilustración 41: *Estimación kriging de indicadores HC bloques 2x2 (Elaboración propia).*

Si nos fijamos en la escala, podemos ver que varía entre 0 y 1, lo que significa que la estimación realizada representa ahora la probabilidad de superar 750 ppm. El inconveniente con este tipo de estimación es que no tiene en cuenta el grado de contaminación, ya que se consideran iguales todas las muestras que superen el límite, pero permite añadir una noción de riesgo que no era posible obtener con la técnica del kriging ordinario. También, como es menos sensible a valores extremos permite realizar la estimación con el total de muestras recopiladas, entregando un detalle más fino de la zona noroeste.

Níquel

El resultado de estimación para la variable níquel mediante kriging de indicadores, se observa en ilustración 42.

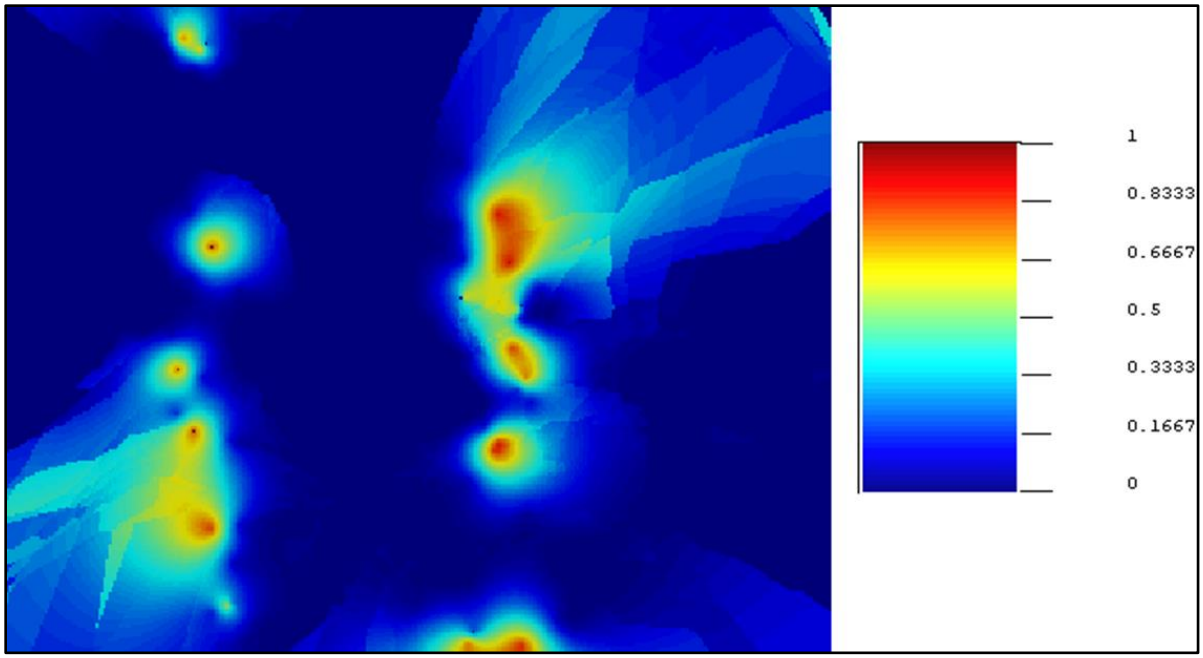

Ilustración 42: *Estimación kriging de indicadores Ni bloques 2x2 (Elaboración propia).*

En esta ocasión, se logra apreciar algunos círculos de contaminación que no se apreciaban cuando se aplicaba kriging ordinario.

Las estadísticas descriptivas de estimación por kriging de indicadores, se aprecian en la tabla 14.

Tabla 14: *Estadísticas descriptivas estimación kriging de indicadores bloques 2x2 (Elaboración propia).*

Variable	Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
HC	83200	-0.05	1.03	0.17	0.05	0.07	0.3
Ni	83200	-0.03	1	0.12	0.04	0.03	0.2

Observando la tabla 14, se aprecia que el 75% de las estimaciones de ambas variables tiene baja probabilidad de superar criterio de contaminación

5.3 Estimación simulaciones condicionales gaussianas

Los resultados que se obtienen al aplicar cualquier método geoestadístico son estimaciones de la realidad, dicho esto, no existe un único resultado que sea capaz de reproducir fehacientemente el fenómeno real de la contaminación que se está estudiando. Por esto, es que tiene un gran valor

añadido las simulaciones condicionales, porque permiten realizar múltiples estimaciones equiprobables de la contaminación a partir de los datos de partida y del modelo de variograma realizado.

En total, se generaron 50 simulaciones mediante esta técnica que se observan en apéndice E. Sin embargo, de las simulaciones realizadas ¿Qué simulación representa mejor el fenómeno de contaminación?

Para poder realizar una correcta interpretación de las simulaciones generadas, se realiza un post-tratamiento por medio de estimación de concentración media de simulaciones realizadas y probabilidad de superar criterio de contaminación establecido. Esto permite definir zonas y volúmenes de contaminación.

El resultado de la probabilidad de superar criterio de contaminación, se observa en ilustración 43.

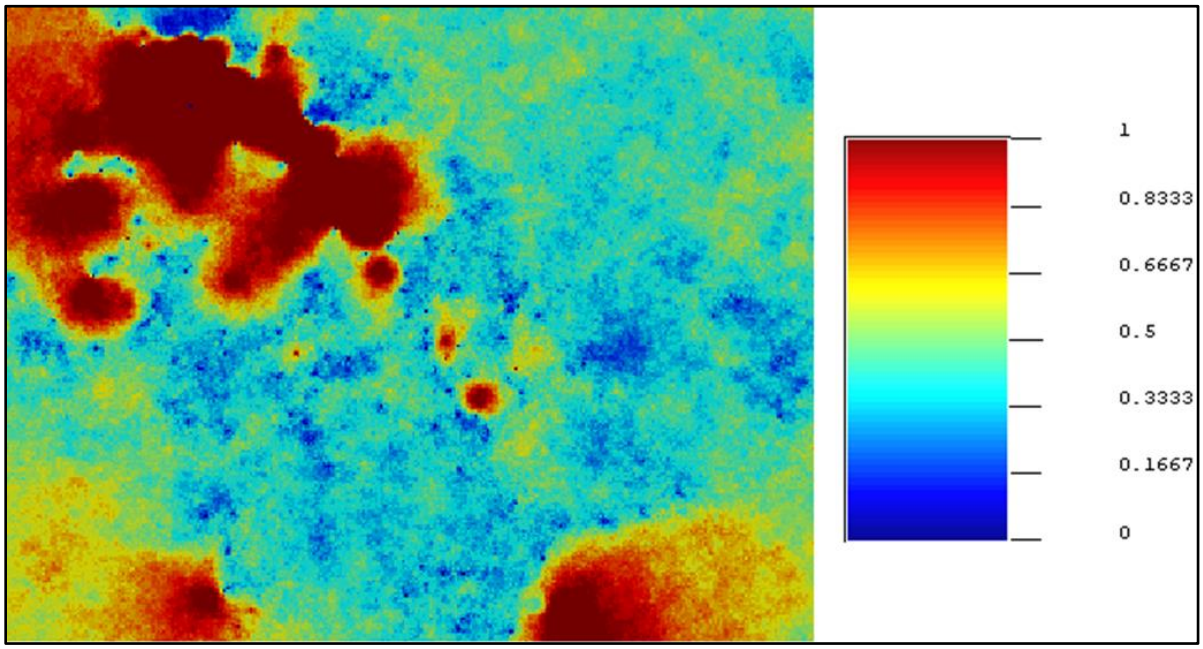

Ilustración 43: *Estimación simulación condicional gaussiana HC bloques 2x2 (Elaboración propia).*

Este resultado se obtiene mediante la combinación de las 50 simulaciones realizadas y en donde los bloques se fueron formando de acuerdo con la cantidad de veces que aparecieron como contaminados.

Se logra apreciar que hay gran cantidad de bloques con probabilidades entre un 40 y 60%. Esto se produce porque las simulaciones son capaces de salirse de su rango, es decir, encontrar valores más

chicos o altos respecto a los datos originales, por ende, aumenta la varianza. Esto es bueno, ya que hace la estimación más flexible que el kriging, que esencialmente no es capaz de estimar nada fuera de sus rangos. Esto se aprecia en las estadísticas descriptivas de estimación de la concentración media de las simulaciones realizadas, donde la simulación fue capaz de estimar un dato 106 veces más pequeño que el mínimo original.

Tabla 15: *Estadísticas descriptivas estimación de la concentración media de simulaciones realizadas bloques 2x2 (Elaboración propia).*

Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
83200	-1627	79000	1695	499.4	1.6E+07	1711

5.4 Zonas contaminadas métodos geoestadísticos

A continuación, se presentan las zonas con riesgo de contaminación. Se hace necesario resaltar que, se trabaja con variables continuas cuando se estima por kriging ordinario, mientras que, la estimación por kriging de indicadores trabaja con variables categóricas. En cambio, las simulaciones pueden obtener resultados para ambos tipos de variables, continuas o categóricas.

Estimación kriging ordinario

La ilustración 44 y 45 exponen las zonas que superan criterio de contaminación establecido para elementos contaminantes: hidrocarburo y níquel.

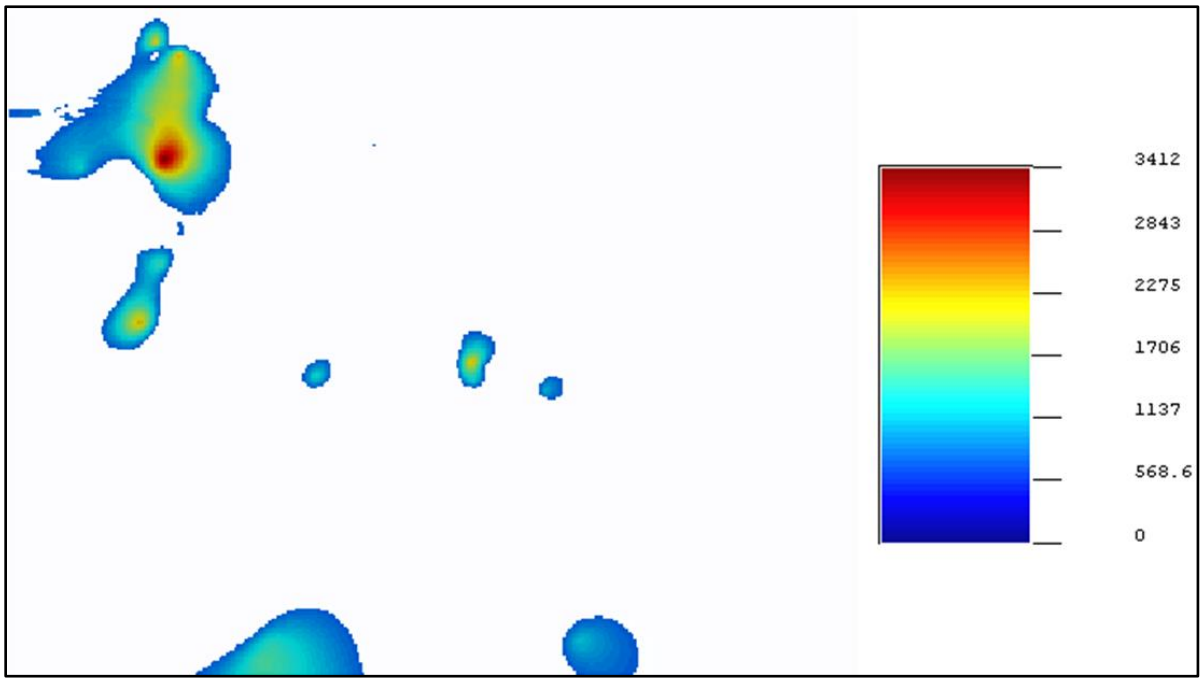

Ilustración 44: *Zonas que superan criterio de contaminación establecido, caso sin datos atípicos HC bloques 2x2 (Elaboración propia).*

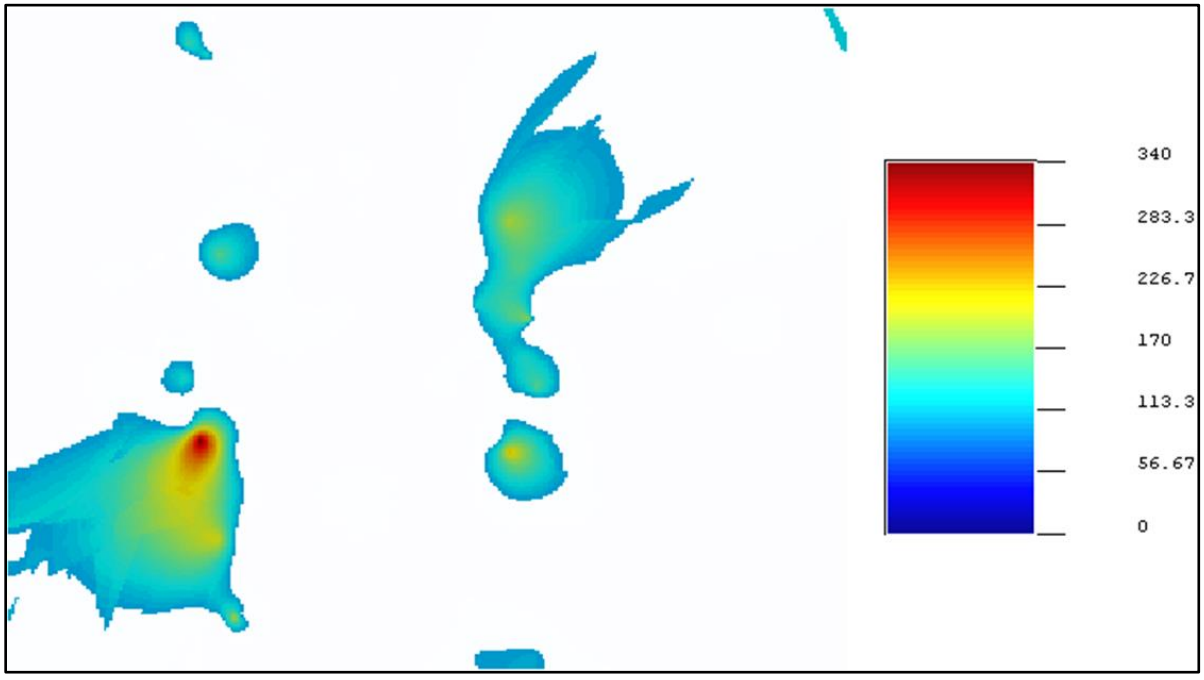

Ilustración 45: *Zonas que superan criterio de contaminación establecido, kriging ordinario Ni bloques 2x2 (Elaboración propia).*

Estimación kriging de indicadores

Los resultados obtenidos en el apartado 5.2 muestran que la escala que varía entre 0 y 1. Esto significa que las zonas azules muestran probabilidades inferiores al 30% de superar criterio de contaminación establecido, por lo tanto, corresponden a zonas con bajo riesgo de contaminación. En cambio, las zonas con tonalidades rojas muestran probabilidades mayores al 60% de superar criterio de contaminación establecido. Estas son las denominadas zonas con alto riesgo de contaminación. Sin embargo, ¿A que corresponden las zonas que están comprendidas entre 30% y 60% de probabilidad de superar criterio de contaminación establecido? Estas zonas tienen probabilidad de estar contaminada como de no estarlo, por lo tanto, en estas zonas no se está seguro de la presencia de la contaminación. En consecuencia, estas son las denominadas zonas de incertidumbre.

Los resultados de contaminación por hidrocarburo se presentan desde la ilustración 46 hasta la 48. En cambio, las zonas por contaminación de níquel se pueden observar en apéndice H.

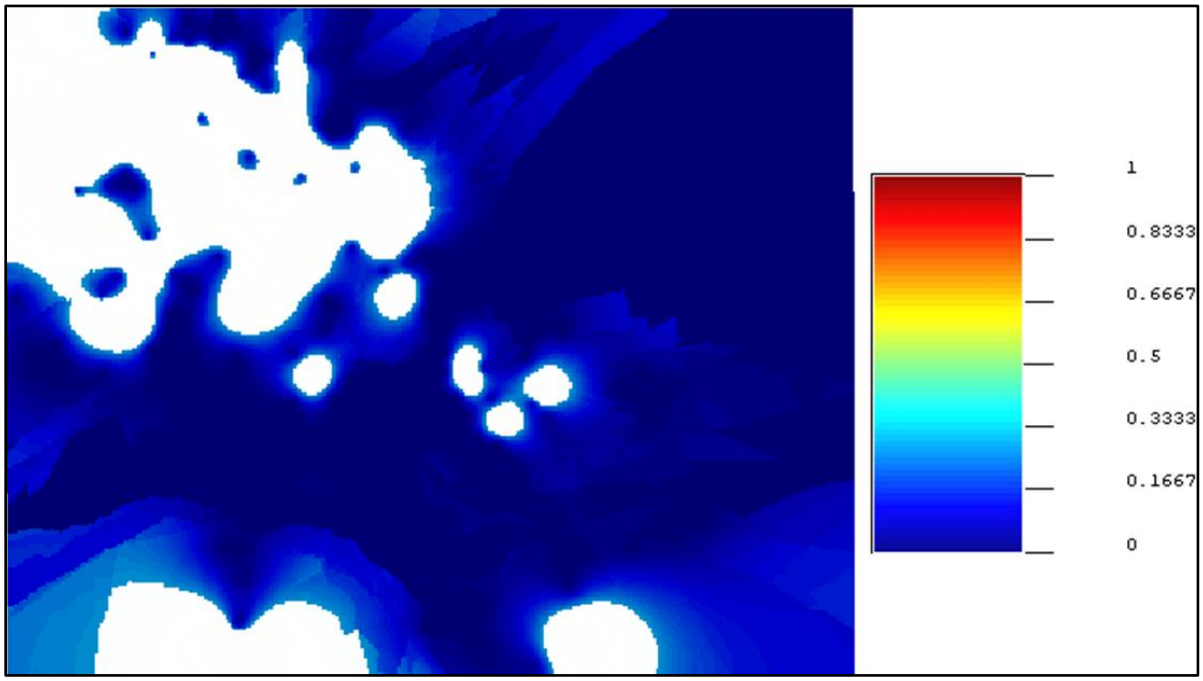

Ilustración 46: *Zona con bajo riesgo de contaminación, kriging de indicadores HC (Elaboración propia).*

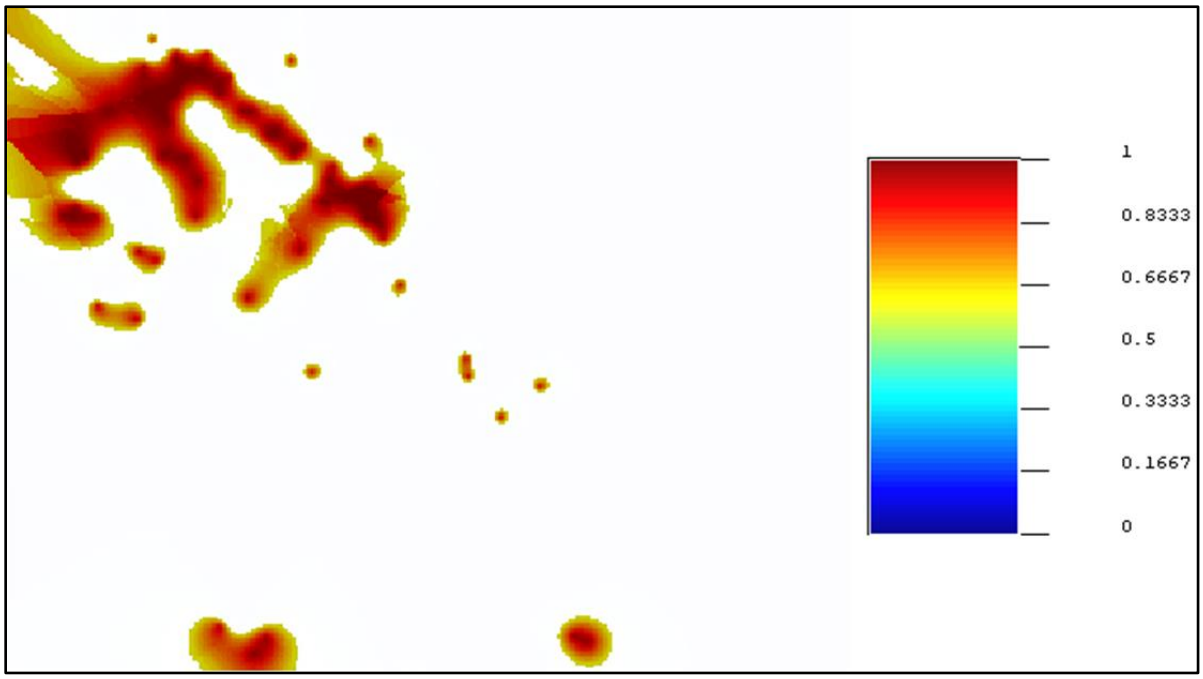

Ilustración 47: *Zonas con alto riesgo de contaminación, kriging de indicadores HC (Elaboración propia).*

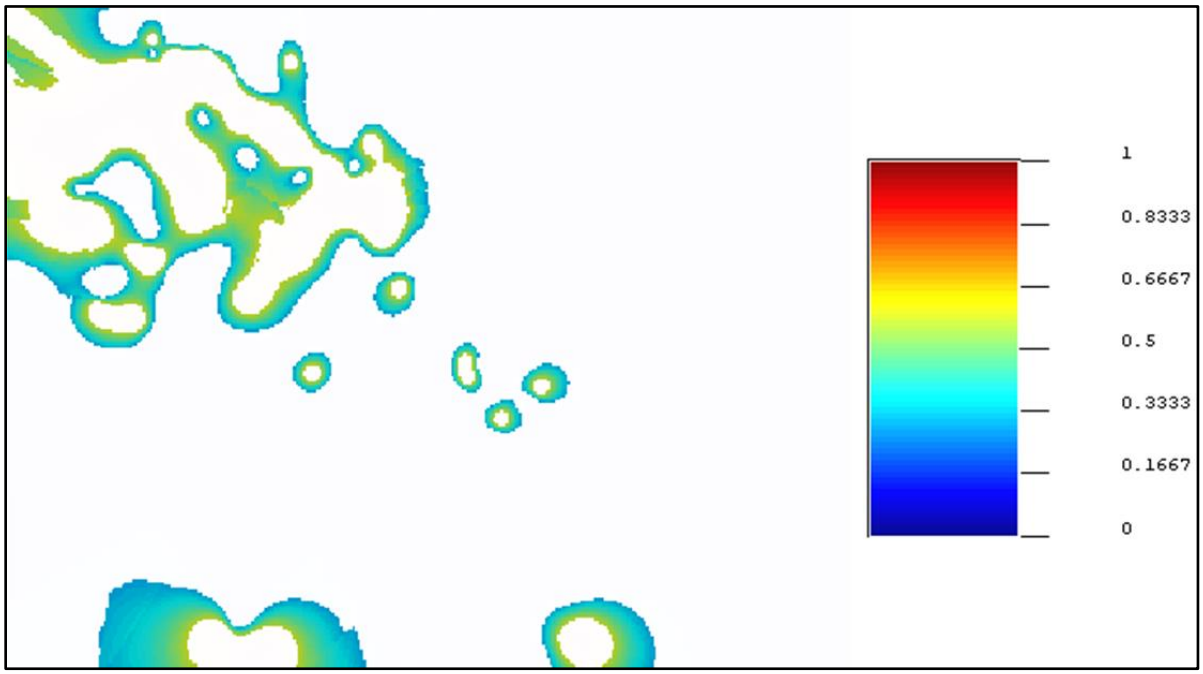

Ilustración 48: *Zonas de incertidumbre, kriging de indicadores HC (Elaboración propia).*

Simulaciones condicionales gaussianas

Por medio del post-tratamiento es posible integrar las 50 simulaciones realizadas, obteniendo como resultado la estimación de probabilidad de superar criterio de contaminación establecido. La diferencia de este caso respecto a kriging de indicadores es que, de las 50 simulaciones realizadas

los bloques azules aparecieron como contaminados en menos de un 30% de los casos. En cambio, las zonas rojas aparecieron como contaminadas en más de un 60% de los casos. Por último, los bloques comprendidos entre el 30% y 60% de los casos, aparecieron una vez cada dos de las 50 simulaciones realizadas.

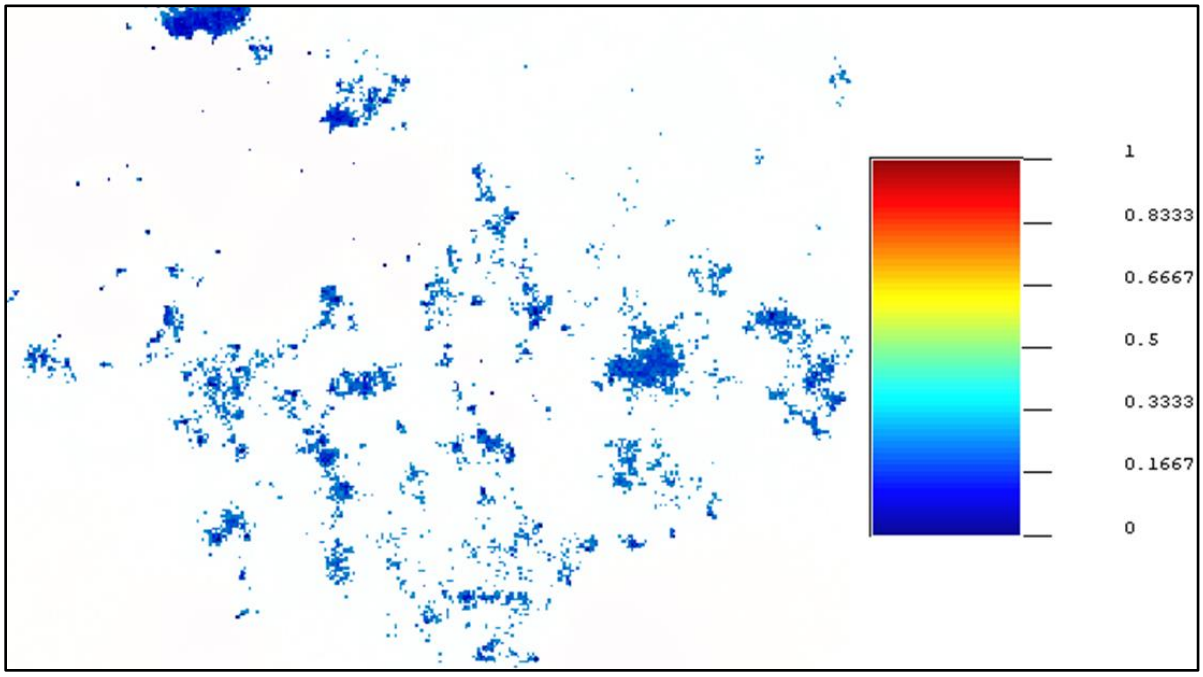

Ilustración 49: *Zonas con bajo riesgo de contaminación, simulaciones condicionales HC (Elaboración propia).*

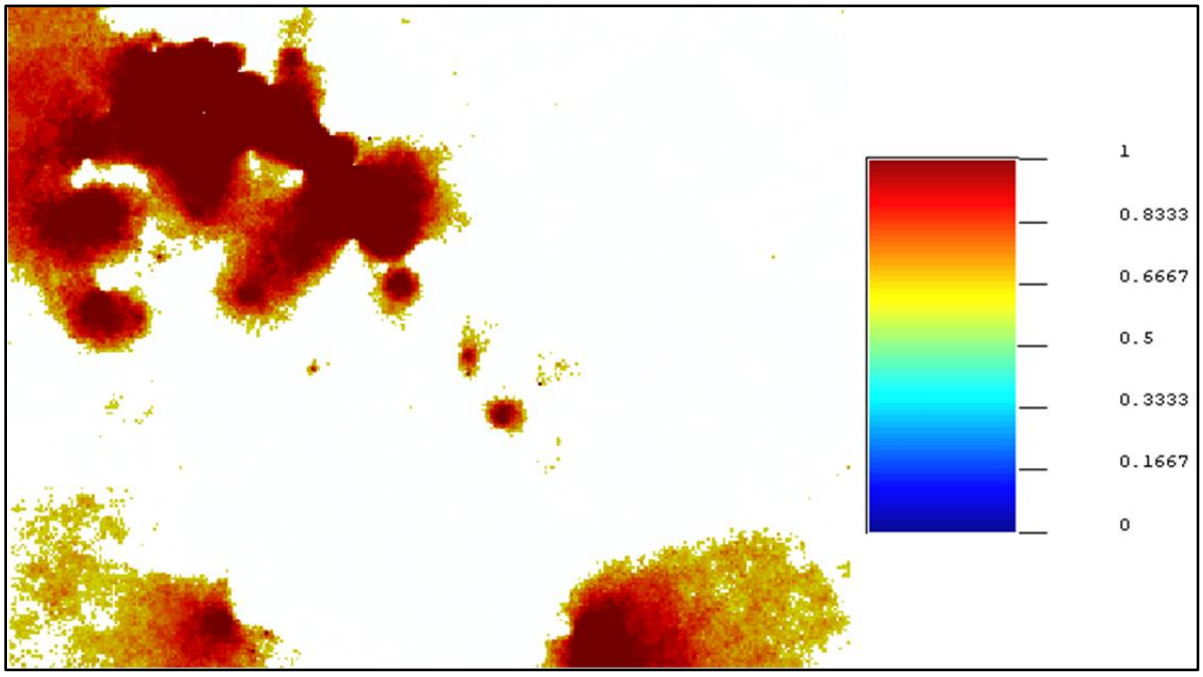

Ilustración 50: *Zonas con alto riesgo de contaminación, simulaciones condicionales HC (Elaboración propia).*

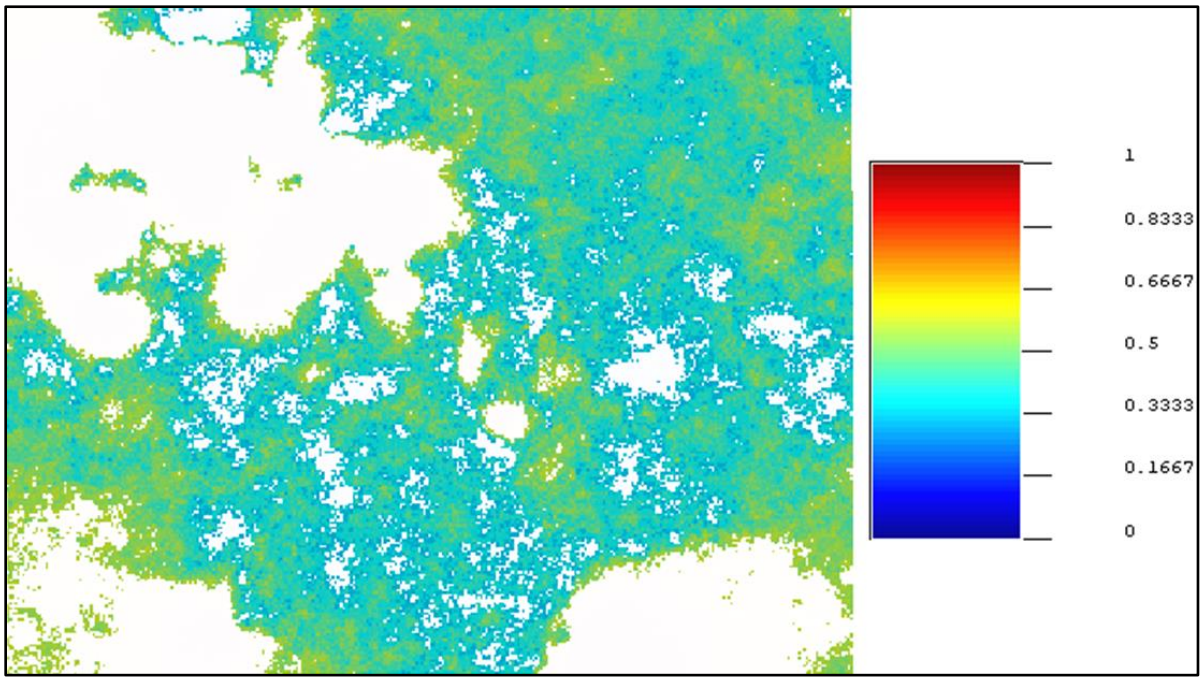

Ilustración 51: *Zonas de incertidumbre, simulaciones condicionales HC (Elaboración propia).*

Observando la ilustración 51, se logra apreciar que existen numerosas zonas de incertidumbre. En estas situaciones lo recomendable es planificar una campaña de investigación complementaria para disminuir la presencia de dichas zonas, es decir, en zonas de incertidumbre volver a muestrear.

5.4.1 Superficies de contaminación

Observando los resultados anteriores, no es posible decir con certeza porcentaje contaminado de la zona de estudio, por lo tanto, filtrando histograma por el valor de criterio de contaminación se puede calcular superficies de contaminación.

Kriging ordinario

Tabla 16: *Superficie de contaminación kriging ordinario HC (Elaboración propia).*

Bloques	Caso	Bloques que superan criterio de contaminación	Superficie de contaminación total (ha)	Superficie de contaminación total (%)
2x2	Datos inicio	2617	1.05	3.2
	Sin datos atípicos	5629	2.3	6.8
10x10	Datos inicio	-	-	-
	Sin datos atípicos	235	2.4	7

Tabla 17: *Superficie de contaminación kriging ordinario Ni (Elaboración propia).*

Bloques	Bloques que superan criterio de contaminación	Superficie de contaminación total (ha)	Superficie de contaminación total (%)
2x2	11532	4.6	13.9
10x10	461	4.6	13.6

Los resultados anteriores muestran que la contaminación por hidrocarburo abarca un 7% de la zona de estudio. Aunque, la contaminación es mucho mayor ya que el caso aplicado no consideraba las muestras atípicas. En paralelo, la contaminación por níquel abarca un 14% de la zona de estudio, lo que no es menor ya que son alrededor de 4.6 hectáreas contaminadas.

En cuanto al volumen, se utiliza el kriging ordinario de los espesores, porque esta variable corresponde al espesor de la capa donde se aloja la contaminación. Por tanto, multiplicando el espesor de la capa y el tamaño de los bloques empleados en nuestra grilla, se obtiene el volumen de contaminación que hay en cada bloque. Esto es posible solamente cuando se estiman variables continuas.

El volumen de contaminación se observa en las tablas 18 y 19.

Tabla 18: *Volumen de contaminación kriging ordinario HC (Elaboración propia).*

Bloques	Caso	Volumen total de contaminación (m^3)
2x2	Datos inicio	6294
	Sin datos atípicos	15132
10x10	Datos inicio	-
	Sin datos atípicos	15804

Tabla 19: *Volumen de contaminación kriging ordinario Ni (Elaboración propia).*

Bloques	Volumen total de contaminación (m^3)
2x2	32107
10x10	32234

Los resultados de volumen obtenidos equivalen a llenar aproximadamente 10 veces una piscina con medidas olímpicas.

Kriging de indicadores

Para el cálculo de contaminación a partir de probabilidades, se debe decidir cuál es el riesgo que se está dispuesto aceptar. Por lo tanto, es posible realizar varios cálculos en función del riesgo.

Tabla 20: *Superficie de contaminación probable, kriging de indicadores elementos contaminantes (Elaboración propia).*

Variable	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
HC	0-30	Bajo	65503	26.2	78.7
	30-60	No se sabe	8972	3.6	10.8
	60-100	Alto	8725	3.5	10.5
Ni	0-30	Bajo	72725	29.1	87.4
	30-60	No se sabe	8744	3.5	10.5
	60-100	Alto	1731	0.7	2.1

Los resultados demuestran que hay numerosas zonas con baja probabilidad de estar contaminada. Mientras que, la probabilidad de que haya zonas con alta contaminación es de un 10% para hidrocarburo y 2% para níquel.

Respecto a las zonas de incertidumbre, es importante conocerlas ya que están presentes en cualquier tipo de proyecto, lo interesante es plantearse que hacer con ellas: ¿Volver a muestrear en el caso que sea posible?, ¿Incluirlas como suelos contaminados?, ¿Centrarse en las zonas con alta probabilidad y realizar muestreo de control para verificar contactos con las zonas de incertidumbre?

En líneas generales, lo recomendable seria volver a muestrear, pero no siempre están los recursos disponibles. Sin embargo, para esta situación no existen un porcentaje elevado de zonas de incertidumbre, así que se pueden incluir como suelos contaminados.

Simulaciones condicionales gaussianas

A continuación, se presenta resultado de superficie de contaminación del post-tratamiento de simulaciones condicionales**.**

Tabla 21*: Superficie de contaminación probable, simulaciones condicionales HC (Elaboración propia).*

Variable	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
HC	0-30	Bajo	6484	2.6	7.8
	30-60	No se sabe	52258	21	62.8
	60-100	Alto	24458	9.8	29.4

De acuerdo con los resultados obtenidos, hay una alta probabilidad de que haya zonas que superen el criterio de contaminación, abarcando aproximadamente un 30% de la zona de estudio. Llama la atención, la cantidad de zonas de incertidumbre que estiman la simulación en comparación al kriging. Esto ocurre porque las simulaciones pueden estimar valores fuera de sus rangos, por tanto, aumenta la varianza, mientras que el kriging no es capaz de estimar nada fuera de sus rangos.

Finalmente, como se ha mencionado cada simulación presenta un resultado distinto, por tanto, presentan una superficie de contaminación diferente. Por ello, se presenta el listado de superficie de contaminación asociada a cada simulación.

Tabla 22: *Superficie de contaminación simulaciones condicionales HC (Elaboración propia).*

N° Sim	Bloques	Superficie (ha)	Superficie (%)	N° Sim	Bloques	Superficie (ha)	Superficie (%)
1	45543	18.2	54.7	**2**	43229	17.3	52
3	43978	17.6	52.9	**4**	36901	14.8	44.4
5	38189	15.3	45.9	**6**	42089	16.8	50.6
7	42771	17.1	51.4	**8**	43839	17.5	52.7
9	36708	14.7	44.1	**10**	39295	15.7	47.2
11	39588	15.8	47.6	**12**	49130	19.7	59.1
13	44780	17.9	53.8	**14**	47569	19	57.2
15	50944	20.4	61.2	**16**	39649	15.9	47.7
17	38844	15.5	46.7	**18**	44818	17.9	53.9
19	41857	16.7	50.3	**20**	47308	18.9	56.9
21	46531	18.6	55.9	**22**	43643	17.5	52.5
23	46859	18.7	56.3	**24**	36394	14.6	43.7
25	40753	16.3	49	**26**	49138	19.7	59.1
27	47087	18.8	56.6	**28**	49014	19.6	58.9
29	40465	16.2	48.6	**30**	45822	18.3	55.1
31	42688	17.1	51.3	**32**	49928	20	60
33	44729	17.9	53.8	**34**	46710	18.7	56.1
35	51416	20.6	61.8	**36**	54941	22	66
37	45949	18.4	55.2	**38**	42518	17	51.1
39	51374	20.5	61.7	**40**	49007	19.6	58.9
41	50526	20.2	60.7	**42**	43980	17.6	52.9
43	40976	16.4	49.3	**44**	44841	17.9	53.9
45	39223	15.7	47.1	**46**	45552	18.2	54.8
47	47967	19.2	57,7	**48**	41806	16.7	50.2
49	46230	18.5	55.6	**50**	41833	16.7	50.3

También, es posible visualizar estas superficies en una curva en función de los percentiles, con el objetivo de determinar el nivel de contaminación global. La curva obtenida se presenta en la siguiente ilustración.

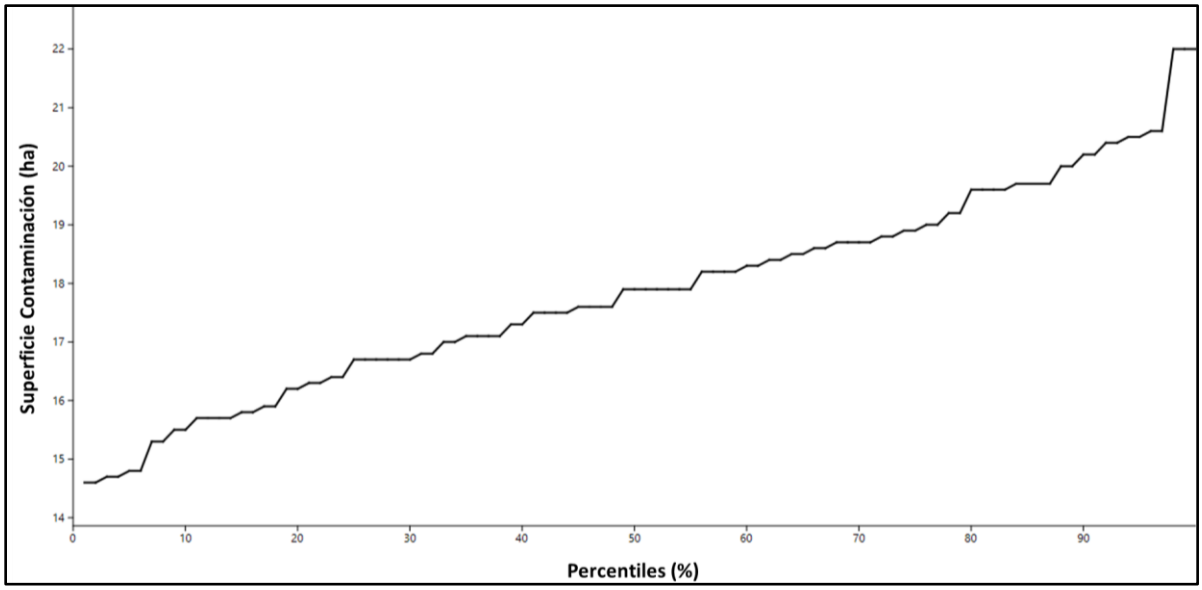

Ilustración 52: *Superficie de contaminación en percentiles (Elaboración propia).*

De la ilustración anterior, se realizan los siguientes comentarios:

- Percentil 5, representa el cálculo optimista entregando un 14.8 ha contaminadas.
- Percentil 50, representa el cálculo probable entregando un 17.9 ha contaminadas.
- Percentil 95, representa el cálculo pesimista entregando un 20.5 ha contaminadas.

En este caso, la superficie probable de contaminación en la zona de estudio es de 17.9 ha, aunque es posible que se encuentre un 4% más producto de la presencia de zonas de incertidumbre en los límites de las zonas potencialmente contaminadas.

Mientras más precisa sea la estimación, menores diferencias habrá entre las superficies pesimistas y optimistas. Además, será más sencilla la decisión que se debiera tomar respecto a estrategia de descontaminación.

5.5 Estimación machine learning

Antes de realizar estimación de los contaminantes en el área de estudio, se debe entrenar los modelos de redes neuronales que se desean implementar. Para esto se utilizó el 80% de los datos en entrenamiento y el restante para evaluar la calidad del modelo.

Es importante mencionar que cada técnica de estimación tiene diferentes métricas que le permiten calcular bondad de ajuste del modelo. Por ejemplo, se tiene el R^2, ME, RSME o el MSPE para casos de kriging. Por lo tanto, distintos modelos tienen indicadores de evaluación de calidad que no son compatibles o comparables. Es por esto, que se utilizaron métricas que permitieran comparar modelos geoestadísticos y machine learning como la exactitud, precisión, exhaustividad y las tablas de clasificación o matriz de confusión.

5.5.1 Estimación variables continuas

Hidrocarburo

A continuación, se presentan los resultados de la fase de entrenamiento de las redes neuronales entregándoles como información de entrada coordenadas Este y Norte, además concentración de elemento contaminante.

Tabla 23: *Entrenamiento redes neuronales HC (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_200	3.084	0.692	0.479	0.692	0.307
NN 1000_200	16.897	0.692	0.479	0.692	0.307
NN 100_2000	27.609	0.692	0.779	0.692	0.780
NN 1000_2000	145.715	0.738	0.814	0.738	0.824

De la tabla 23, se puede observar que el modelo que presentó mejores resultados en la fase de entrenamiento fue el que tenía mayor cantidad de neuronas e iteraciones. Se aprecia que el tiempo de entrenamiento está relacionado directamente con la cantidad de iteraciones que se le piden a la red. Además, el proceso de entrenamiento de la red es bastante rápido, ya que los primeros 3 modelos se entrenan en menos de 30 segundos. Mientras que, el modelo con parámetros más altos obtiene resultados pasado a los 2 minutos.

A continúan, se presentan los resultados de clasificación en matriz de confusión.

Tabla 24: *Matriz de confusión modelo NN 100_200 con datos de entrenamiento HC (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	106	0	106
	1	47	0	47
	Total	153	0	153

Tabla 25: *Matriz de confusión modelo NN 1000_200 con datos de entrenamiento HC (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	106	0	106
	1	47	0	47
	Total	153	0	153

Tabla 26: *Matriz de confusión modelo NN 100_2000 con datos de entrenamiento HC (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	66	40	106
	1	7	40	47
	Total	73	80	153

Tabla 27: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento HC (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	71	35	106
	1	5	42	47
	Total	76	77	153

De las tablas anteriores, se observa que se utilizaron 153 datos para el entrenamiento, donde el 0 significa datos bajo criterio de contaminación y el 1 sobre criterio de contaminación. Dicho esto, inicialmente se tiene 106 datos que corresponden a muestras que no superan criterio de contaminación para HC, mientras que, 47 muestras superan criterio de contaminación. El entrenamiento nos dice que los modelos con 200 iteraciones no son capaces de predecir datos como contaminados. Esto significa que no son modelos óptimos para este caso de estudio, ya que lo interesante es poder predecir datos como contaminados. Por lo tanto, los modelos NN 100_2000 y NN 1000_2000 podrían ser capaces de estimar la contaminación producida en el área de estudio, aunque se equivoquen al predecir datos 0 como 1. Esto es bueno en estas situaciones, ya que les da a los modelos un poco de flexibilidad, permitiendo que sean generales y no sufran sobreajustes. Este último aspecto es sumamente importante, porque tener un modelo con sobreajuste implica que

nuevos datos tengan estimaciones inadecuadas por no tener compatibilidad con el modelo entrenado.

El color rojo de la tabla significa que el modelo ha cometido un error su clasificación. Por ejemplo, dato que debiese ser clasificado como 0 fue clasificado como 1. Mientras que, el color verde significa datos que fueron clasificados correctamente por el modelo.

Para realizar predicción del área de estudio, se entrega al modelo solamente las coordenadas a estimar. Los resultados para caso continuo se observan desde ilustración 53 hasta 56.

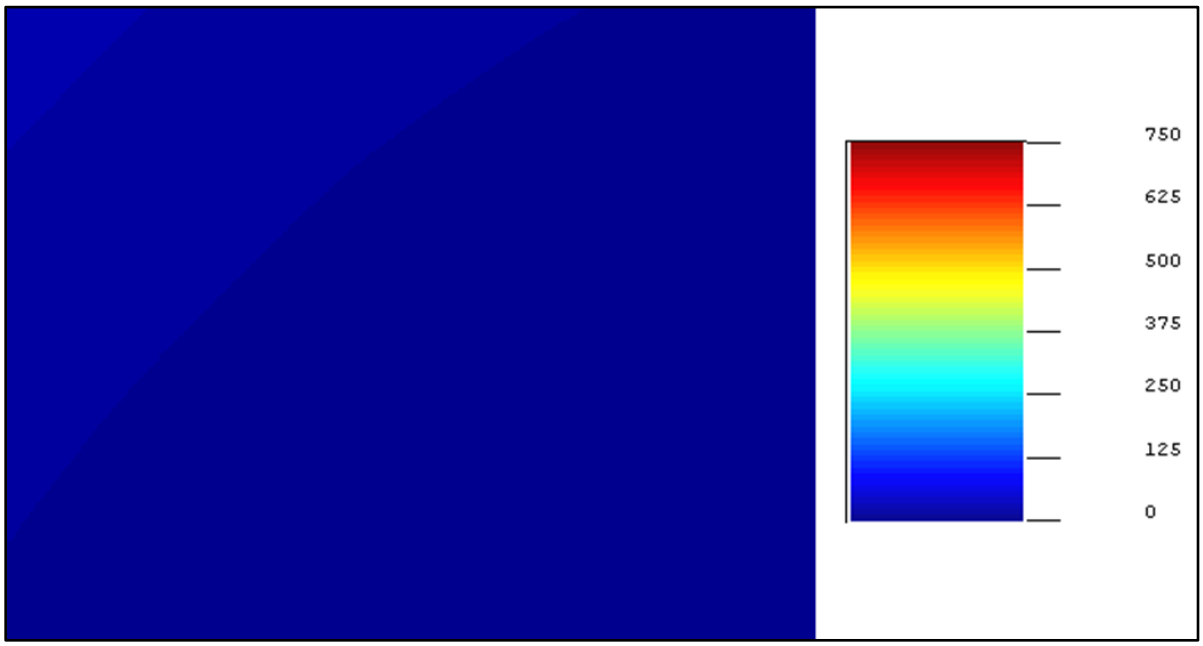

Ilustración 53: *Estimación modelo NN 100_200, variables continuas HC (Elaboración propia).*

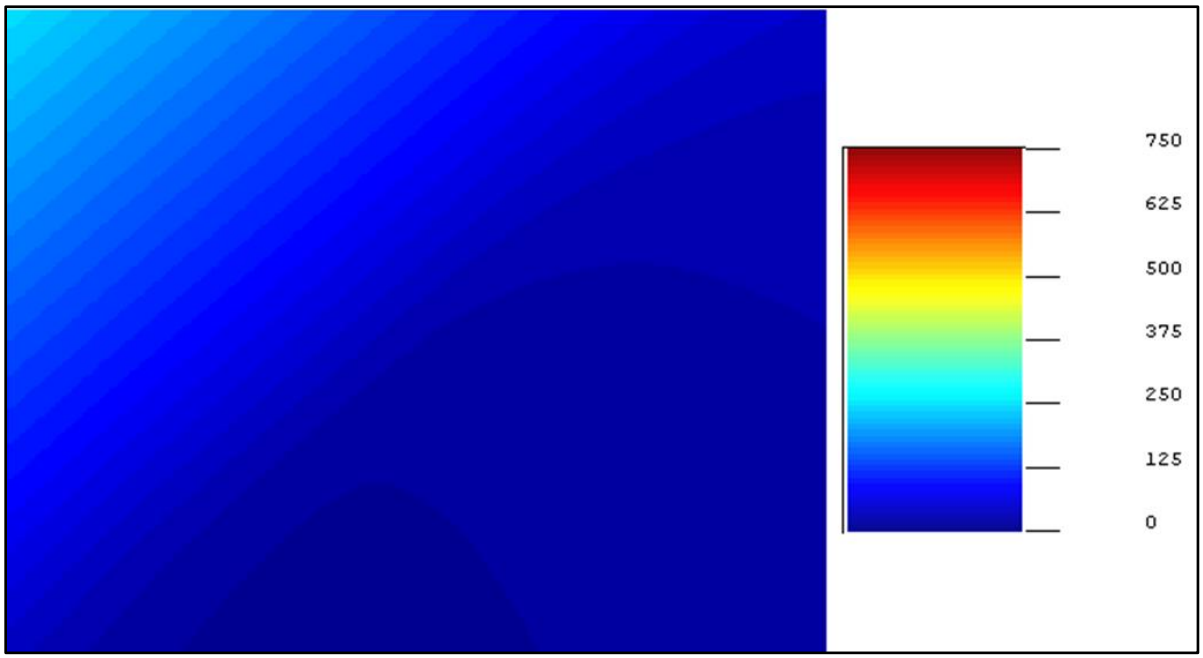

Ilustración 54: *Estimación modelo NN 1000_200, variables continuas HC (Elaboración propia).*

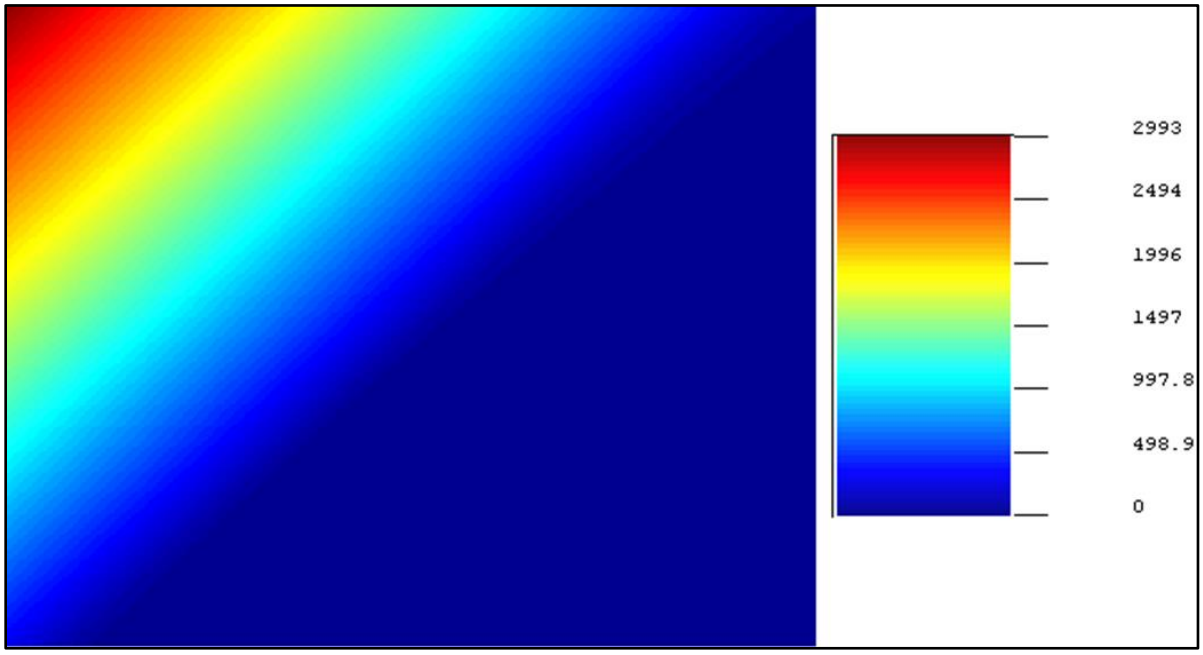

Ilustración 55: *Estimación modelo NN 100_2000, variables continuas HC (Elaboración propia).*

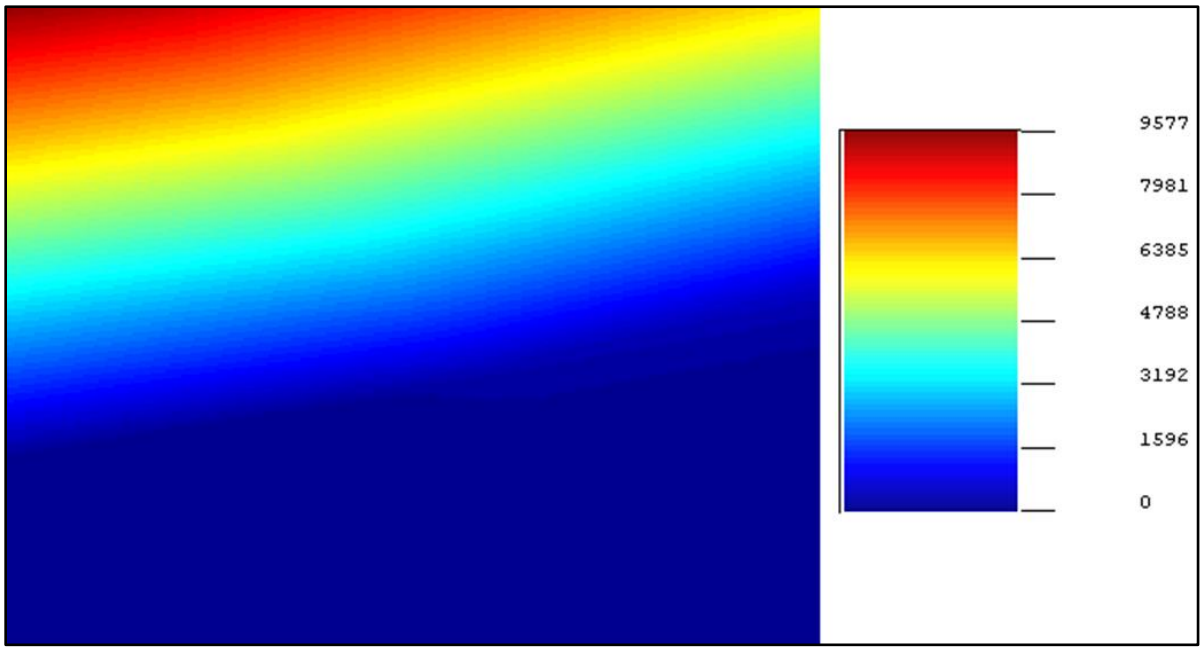

Ilustración 56: *Estimación modelo NN 1000_2000, variables continuas HC (Elaboración propia).*

De los cuatro modelos generados, solamente aquellos que tienen como parámetro 2000 iteraciones son capaces de predecir zonas contaminadas. Aunque, ninguno es capaz de decir con claridad alguna zona en particular, solamente que hacia el sector noroeste hay indicios de contaminación.

Para evaluar su calidad, se compara valor predicho con valor real utilizando el 20% de los datos que no se ocuparon en el entrenamiento.

Tabla 28: *Prueba redes neuronales HC (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_200	15727.249	0.717	0.533	0.743	-
NN 1000_200	15680.842	0.717	0.533	0.743	-
NN 100_2000	15194.586	0.692	0.752	0.692	-
NN 1000_2000	14049.994	0.487	0.692	0.487	-

Tabla 29: *Matriz de confusión modelo NN 100_200 con datos de prueba HC (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	28	0	28
	1	11	0	11
	Total	39	0	39

Tabla 30: *Matriz de confusión modelo NN 1000_200 con datos de prueba HC (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	28	0	28
	1	11	0	11
	Total	39	0	39

Tabla 31: *Matriz de confusión modelo NN 100_2000 con datos de prueba HC (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	19	9	28
	1	3	8	11
	Total	22	17	39

Tabla 32: *Matriz de confusión modelo NN 1000_2000 con datos de prueba HC (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	10	18	28
	1	2	9	11
	Total	12	27	39

La tabla 28 es la prueba del modelo y muestra que hay un aumento significativo del tiempo de prueba. Además, se ve que los modelos que estaban peor evaluados en la fase de entrenamiento tienen mejores resultados. Esto ocurre principalmente, porque estos modelos en su entrenamiento están enfocados solamente en predecir datos como 0, por lo que no hay flexibilidad para predecir datos como contaminados, por lo tanto, tienen mayor porcentaje de acierto ya que existen más datos clasificados como 0. Las ilustraciones de estimaciones y las estadísticas descriptivas demuestran lo anteriormente dicho, donde sus valores máximos no alcanzan a superar los 750 ppm.

Tabla 33: *Estadísticas descriptivas estimación redes neuronales HC (Elaboración propia).*

Modelo	Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
NN 100_200	83200	2.5	25.4	7.1	5.1	23.5	9.2
NN 1000_200	83200	5.1	251.8	58.1	31.8	3186	89.9
NN 100_2000	83200	-0.4	2993.4	547.4	126	519438	968.8
NN 1000_2000	83200	-464.5	9577	2438.4	1443.2	7.7E+06	4728.3

Níquel

Con lo expuesto anteriormente, se comprende que cuando se trabaja con modelos de 200 iteraciones hay problemas para estimar sitios contaminados. Por ende, desde este punto en adelante se trabaja con modelos de 2000 iteraciones. Los resultados de entrenamiento y prueba se observan en apéndice G.

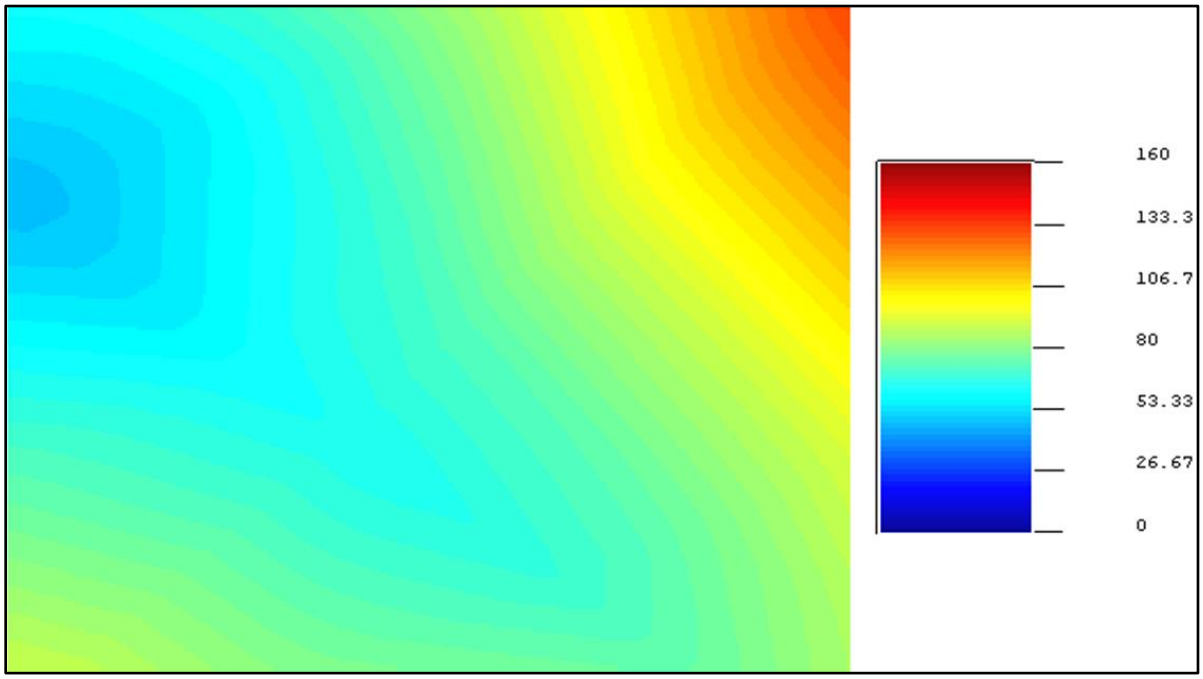

Ilustración 57: *Estimación modelo NN 100_2000, variables continuas Ni (Elaboración propia).*

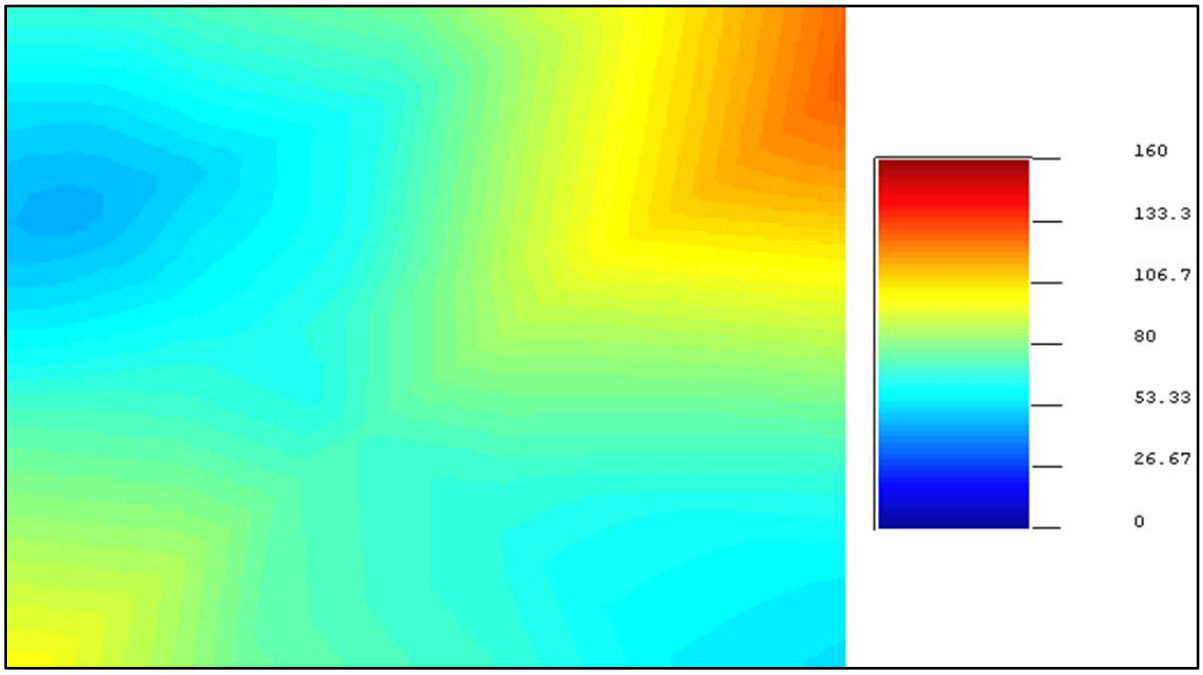

Ilustración 58: *Estimación modelo NN 1000_2000, variables continuas Ni (Elaboración propia).*

De acuerdo con los resultados, en direcciones diagonales pareciera haber contaminación por níquel, pero las estimaciones no son precisas para distinguir alguna fuente de contaminación.

5.5.2 Estimación variables categóricas

Hidrocarburo

Dado que los resultados de modelos de red neuronal para variables continuas no son precisos en cuanto a forma y lugar de contaminación, se estima utilizando variables categóricas para determinar probabilidad de superar criterio de contaminación.

Los resultados de entrenamiento y prueba de modelos con variable categórica se pueden observar en apéndice G. Mientras que los resultados de estimación, se observan a continuación.

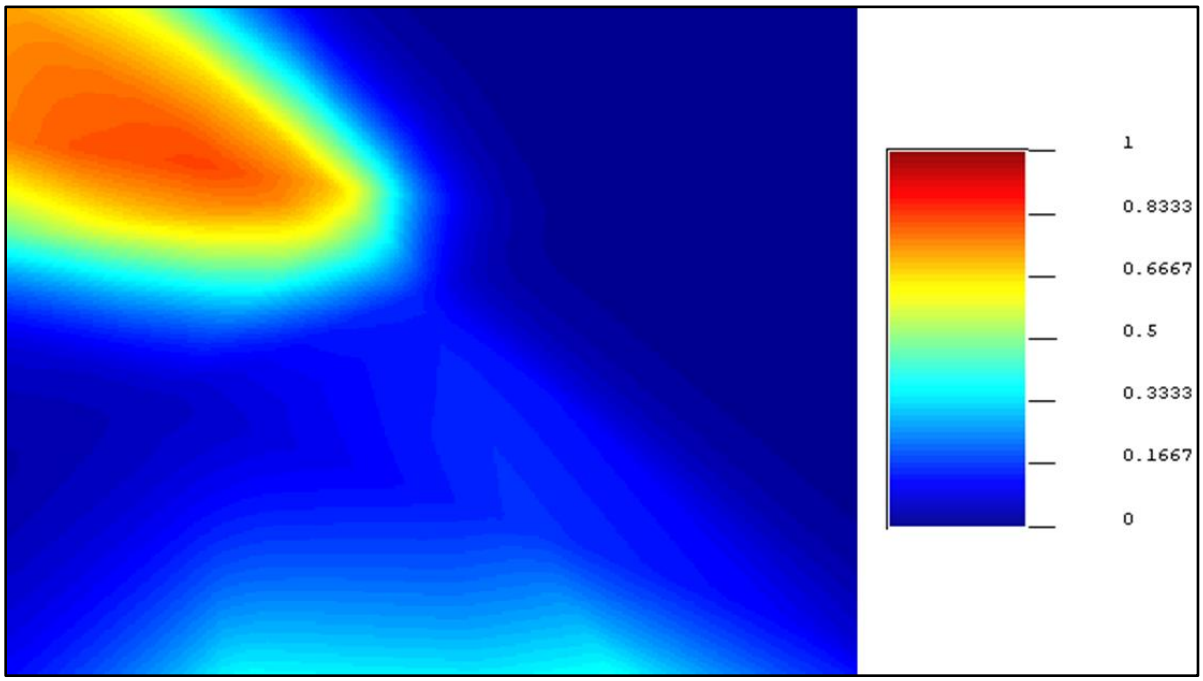

Ilustración 59: *Estimación modelo NN 100_2000, variables categóricas HC (Elaboración propia).*

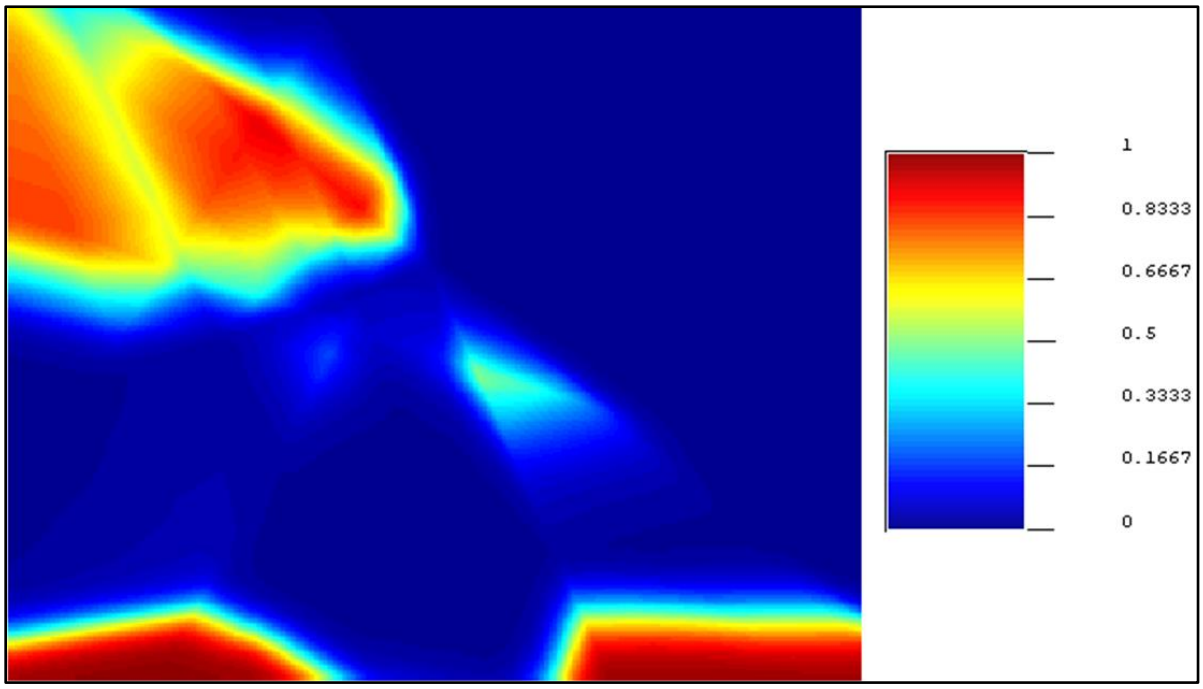

Ilustración 60: *Estimación modelo NN 1000_2000, variables categóricas HC (Elaboración propia).*

Respecto a la predicción de los modelos, se observa que cuando se trabaja con variables categóricas los modelos son capaces de ser mucho más precisos en lugar y forma de contaminación. La diferencia entre modelos neuronales es; modelo NN 100_2000 dice que hay gran acumulación de elemento contaminante en sitio de derrame y que luego se fue escurriendo de Norte a Sur por las vías del ferrocarril hasta terminar por acumularse en la zona Sur del área de estudio. En cambio, el modelo NN 1000_2000 dice que el escurrimiento no es solo por una dirección, sino que hay una bifurcación que produce que aparezcan dos zonas de acumulación de hidrocarburo. Por último, sobre estas dos últimas zonas mencionadas, el modelo NN 1000_2000 está prediciendo con seguridad que ahí existen zonas contaminadas, ya que de acuerdo con la escala dice que hay probabilidad del 80% de superar criterio de contaminación, mientras que el modelo NN 100_2000 intuye que puede haber sitios contaminados en el área sur de estudio, pero no esta tan seguro ya que sus probabilidades bordean el 30 al 50%.

Níquel

Es importante mencionar que, de los 192 datos iniciales solo 17 muestras de níquel superan criterio de contaminación. Por lo tanto, la fase de entrenamiento y prueba trabajan con pocas muestras que figuran como contaminadas. Los resultados de estimación se aprecian a continuación.

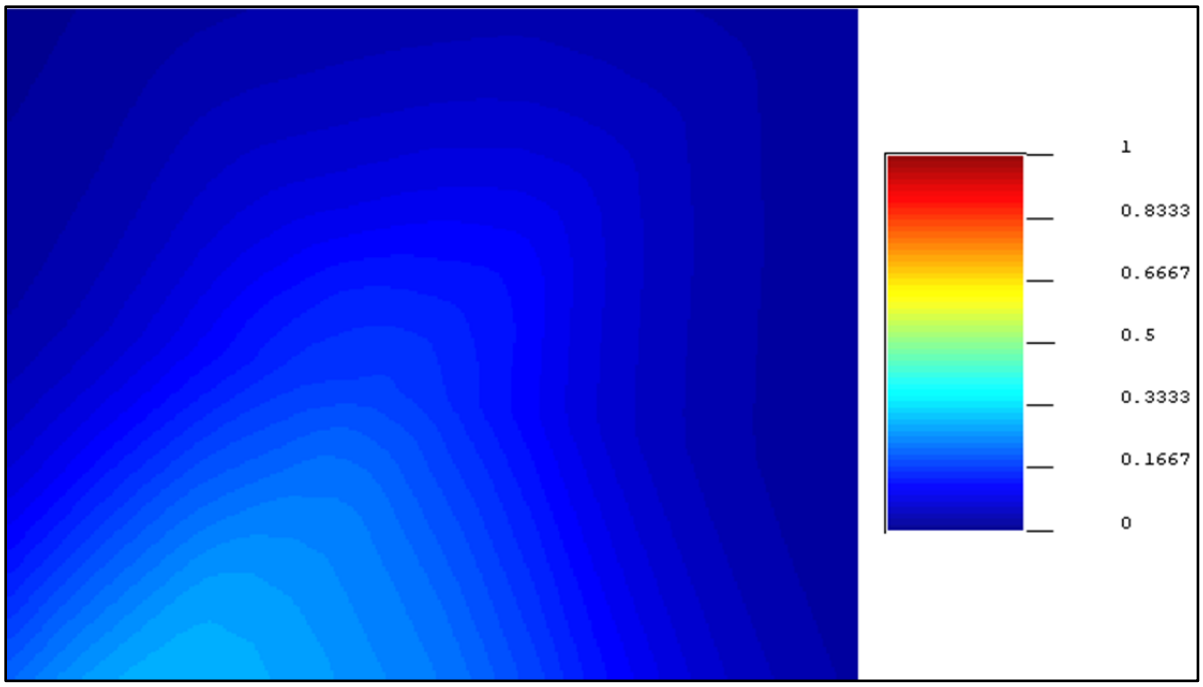

Ilustración 61: *Estimación modelo NN 100_2000, variables categóricas Ni (Elaboración propia).*

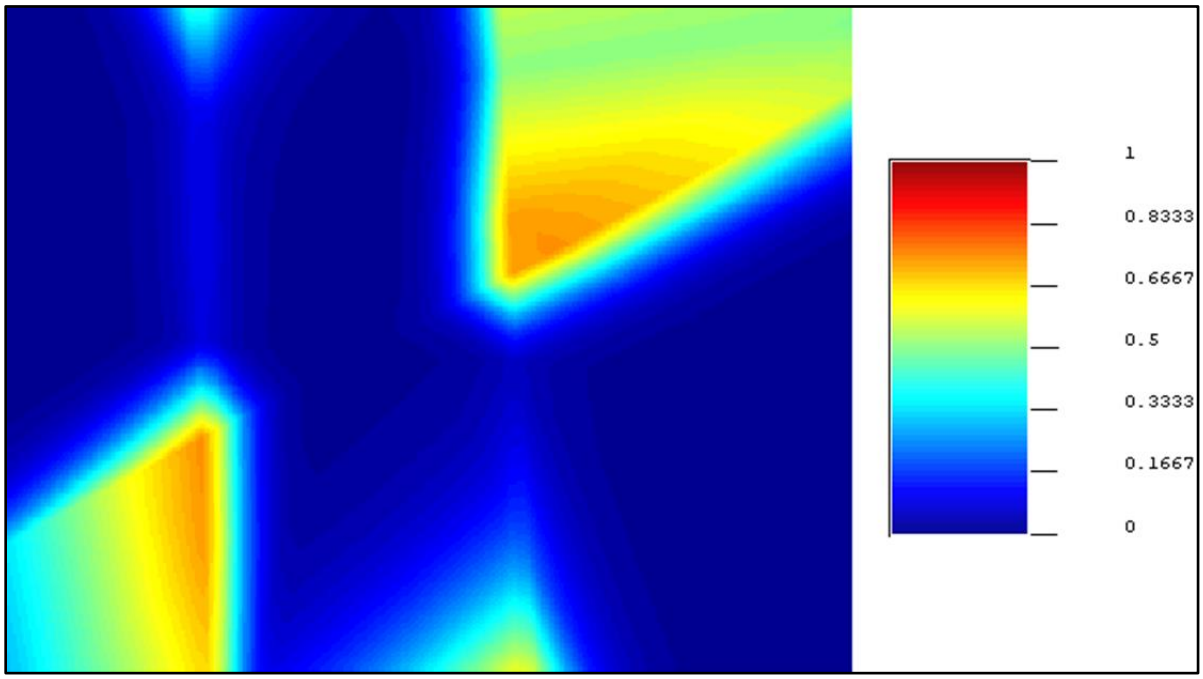

Ilustración 62: *Estimación modelo NN 1000_2000, variables categóricas Ni (Elaboración propia).*

A pesar de que el modelo NN 100_2000 predice que la contaminación se ha ido depositando a través de estratos o capas, en su fase de entrenamiento y prueba no fue capaz de estimar valores como contaminados, a diferencia del modelo NN 1000_2000 que si fue capaz de hacerlo. Esto nos dice, que las redes neuronales requieren de datos balanceados. En caso de no ser balanceados, para

predecir casos del que se tiene poca información, se requieren muchas neuronas, lo que no significa que un modelo con muchas neuronas sea un buen modelo.

Variables categóricas: Caso hidrocarburo con 210 muestras

Dado que los resultados de los métodos anteriores no han sido del todo satisfactorio, es que se busca la manera de ayudar a las redes neuronales a predecir mejor.

Sabemos que la contaminación ha ocurrido en una zona residencial, por lo tanto, el modelo no debiese estimar contaminación en edificios y calles del área de estudio, por lo que es importante que pueda contar con dicha información. Para esto, en las coordenadas de los edificios y calle se agrega un 0, porque corresponden a zonas que no debiese tener contaminación.

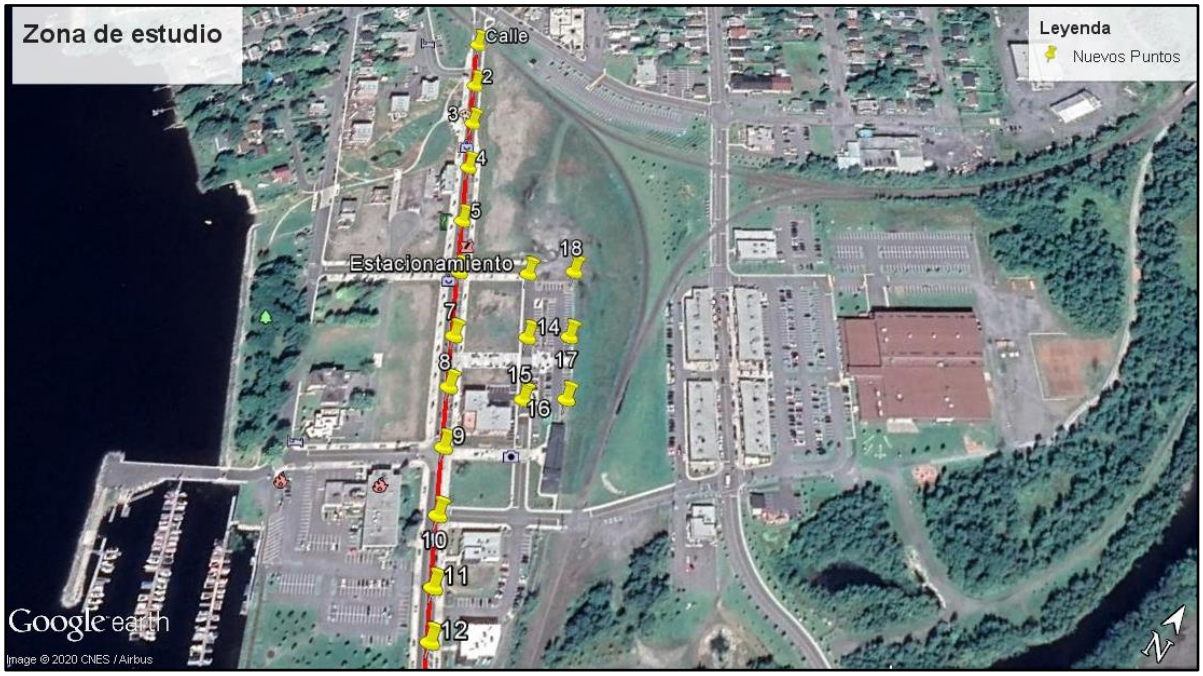

Ilustración 63: *Zona de estudio con nuevos puntos agregados (Google earth, 2020).*

En la ilustración 63, se observa que cada 100 metros se agregó un dato en la avenida principal. También, se agregaron datos en un área de estacionamiento. Con estas 18 nuevas muestras, se esperaría que la estimación por redes neuronales fuese capaz de diferenciar la avenida principal y área de estacionamiento como no contaminadas. Los resultados de estimación se aprecian en las siguientes ilustraciones.

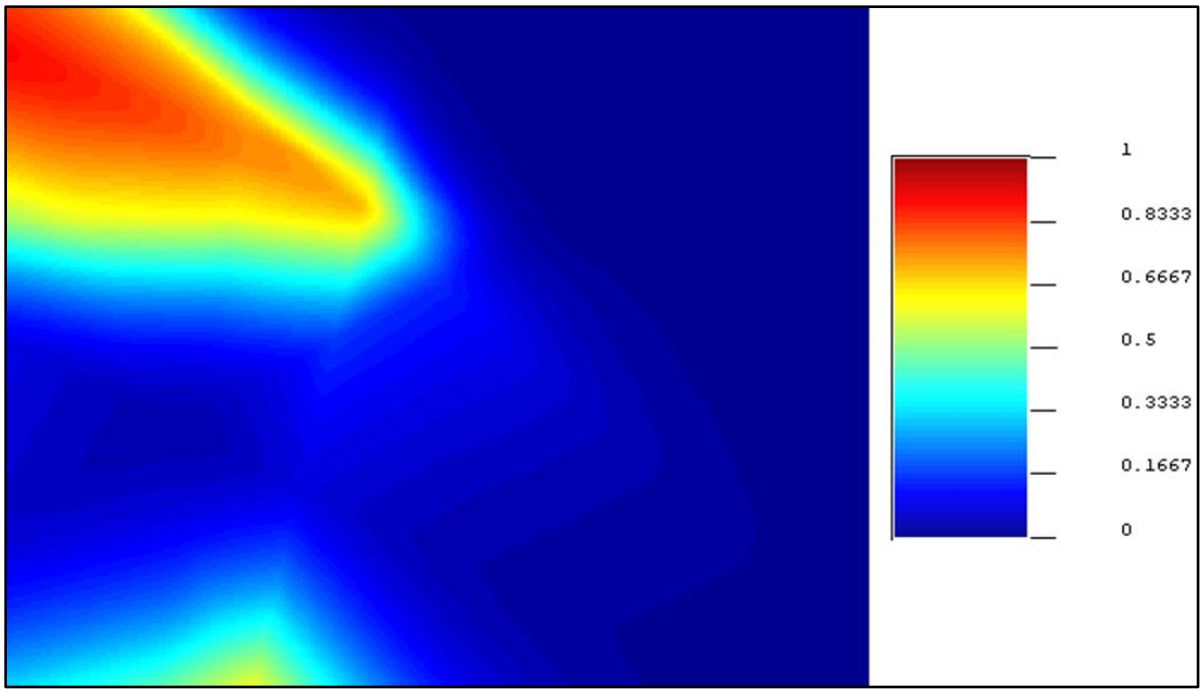

Ilustración 64: Estimación modelo NN 100_2000 HC caso 210 muestras (Elaboración propia).

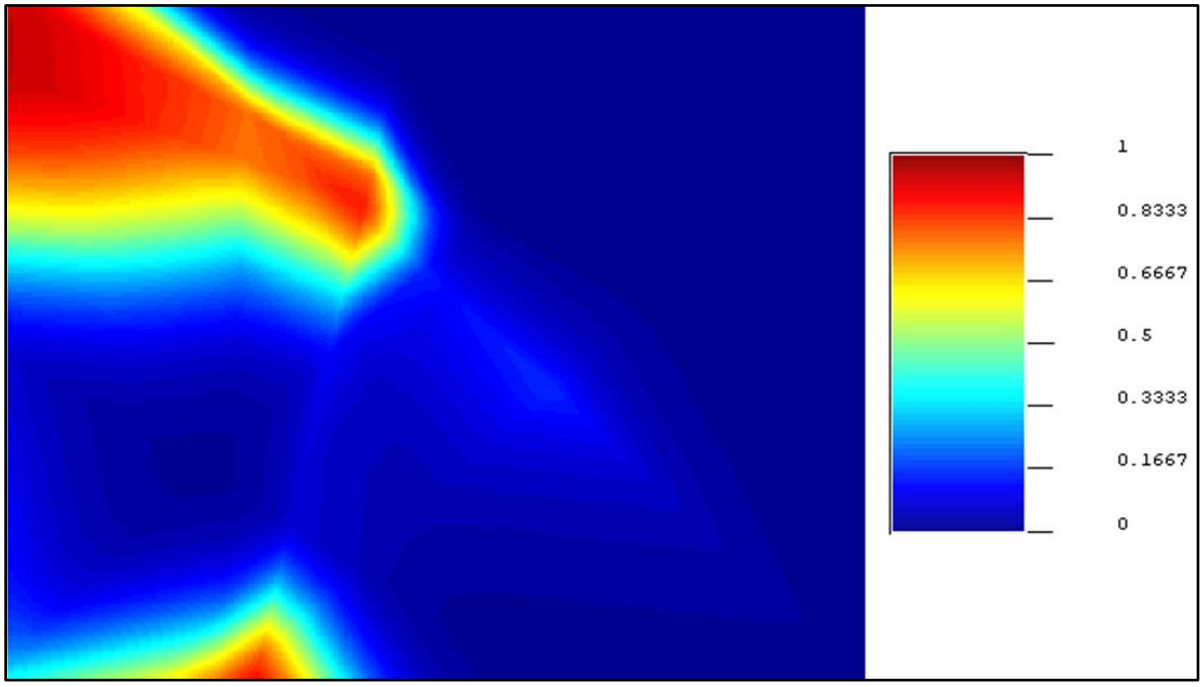

Ilustración 65: *Estimación modelo NN 1000_2000 HC caso 210 muestras (Elaboración propia).*

Los resultados obtenidos muestran que la red neuronal reconoce la calle, por lo que estima baja probabilidad de que haya contaminación en dicho lugar. Incluso, ambos modelos estiman que la contaminación escurrió a través de la avenida hasta acumularse en el área Sur de la zona de estudio y depositándose por capas a las zonas Oeste-Este.

Variables categóricas: Caso hidrocarburo muestreado con grilla

Como se mencionó en un inicio, la toma de muestras fue hecha al azar, es decir, sin que la selección de puntos se haya realizado a distancias uniformes. Debido a esto, es que se vuelve necesario analizar con detalle la distribución de la toma de muestras, con el fin de determinar si existen sesgos hacia el hidrocarburo, puesto que pueden influir en las predicciones del modelo de kriging y de redes neuronales.

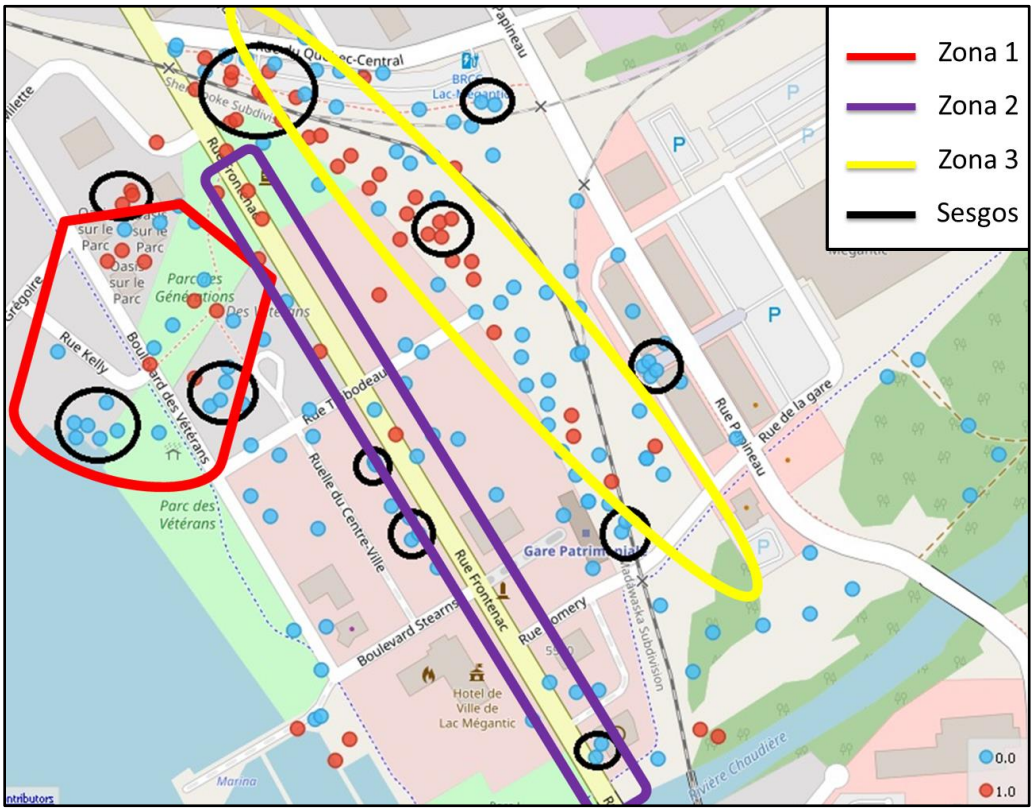

Ilustración 66: *Sesgos hacia las muestras de HC (Elaboración propia).*

Ciertamente, existe una serie de puntos que están siendo sobre representados, por lo que podrían estar influyendo en las estimaciones. Dentro de este marco, se identificaron 3 zonas, donde se infiere que una cuadrilla se dedicó a muestrear a través de la línea férrea, otra siguiendo los charcos de hidrocarburo que escurrían por la avenida principal. Por último, una cuadrilla dedicada a muestrear los parques. Por lo tanto, esto puede dar motivo para eliminar puntos en las zonas donde hay mayor densidad de muestreo, con el fin de evitar sesgos. Sin embargo, la pregunta que viene a continuación ¿Cómo debiese hacerse el muestreo?

De acuerdo con la guía metodológica para la gestión de suelos con potencial presencia de contaminantes elaborada por el MMA y CORFO, se estipula una serie de recomendaciones para

determinar presencia de contaminantes dentro un emplazamiento, siendo clave el diseño muestral del suelo, donde independientemente de la homogeneidad de la distribución de la contaminación se recomienda el uso de una grilla. Por lo tanto, siguiendo estas recomendaciones es que se repite el proceso de cálculo, donde las muestras cercanas o que intersectan la grilla se utilizan para la fase de entrenamiento de la red neuronal. En cambio, las muestras no seleccionadas se usan para la fase de prueba. A continuación, se presenta la grilla con muestras de entrenamiento y prueba.

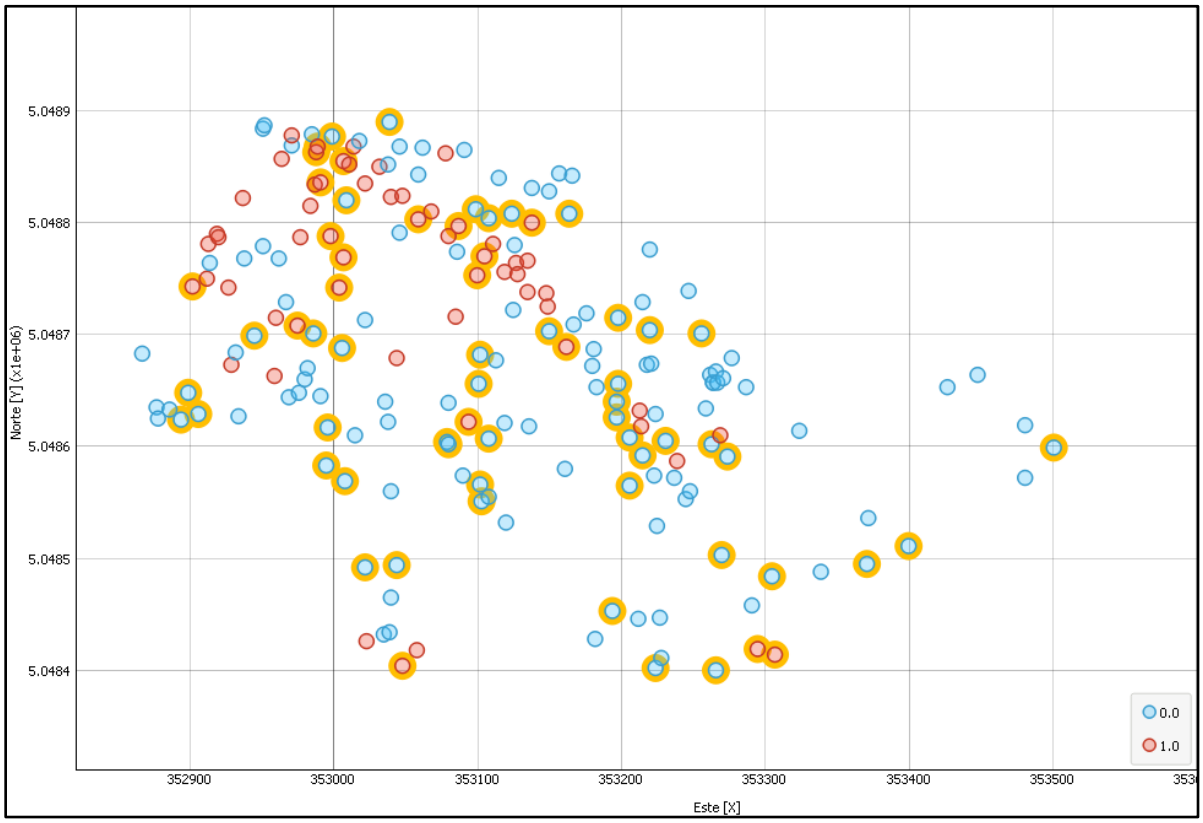

Ilustración 67: *Muestras de entrenamiento con grilla (Elaboración propia).*

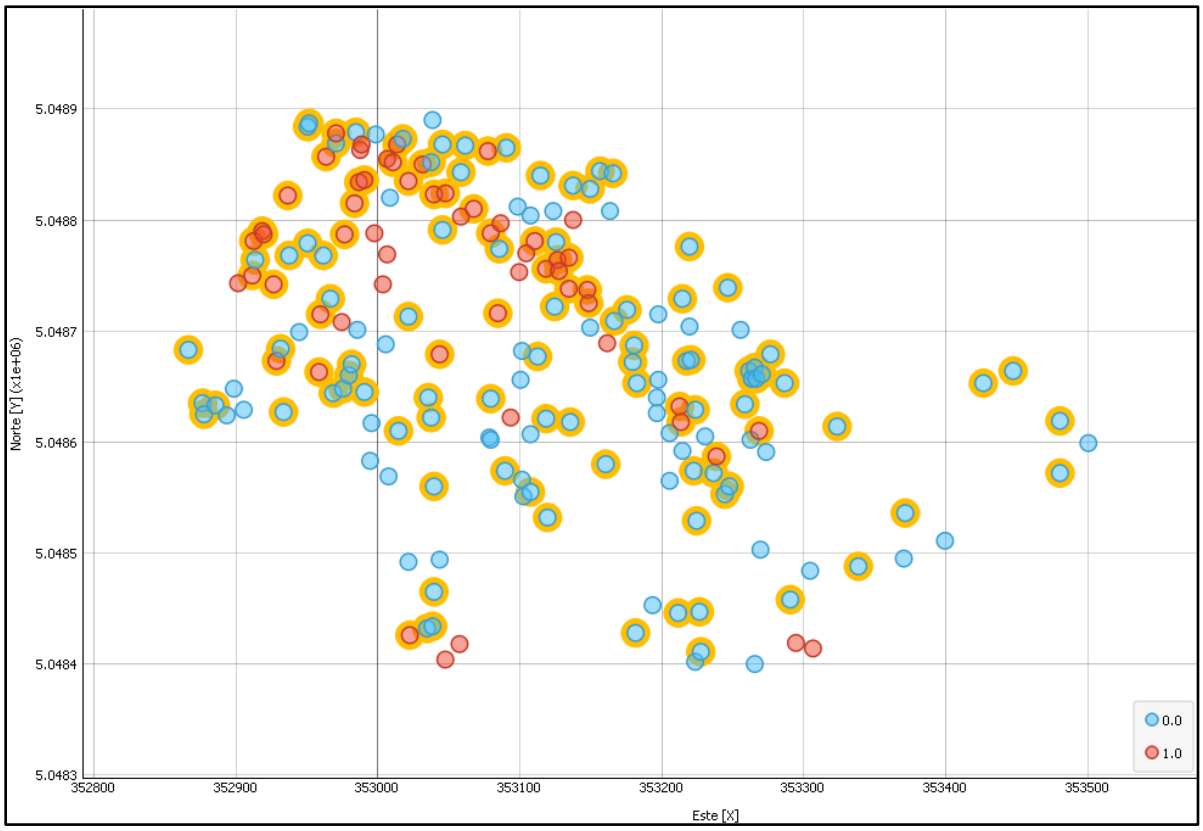

Ilustración 68: *Muestras de prueba con grilla (Elaboración propia).*

Utilizando el criterio anterior, la red neuronal fue entrenada con 65 muestras, mientras que, para la fase de prueba se utilizaron 127 muestras. Los resultados de fase de entrenamiento y prueba se pueden observar en apéndice G.

Con este criterio es posible eliminar zonas con sesgos, por lo tanto, ya no existen áreas sobre muestreadas. Los resultados de distribución espacial utilizando como base una grilla se observan en ilustración 69. En cambio, los resultados de estimación de redes neuronales cuando se emplea una grilla como base, se observan en ilustración 70 y 71.

Ilustración 69: *Distribución espacial muestras con grilla (Elaboración propia).*

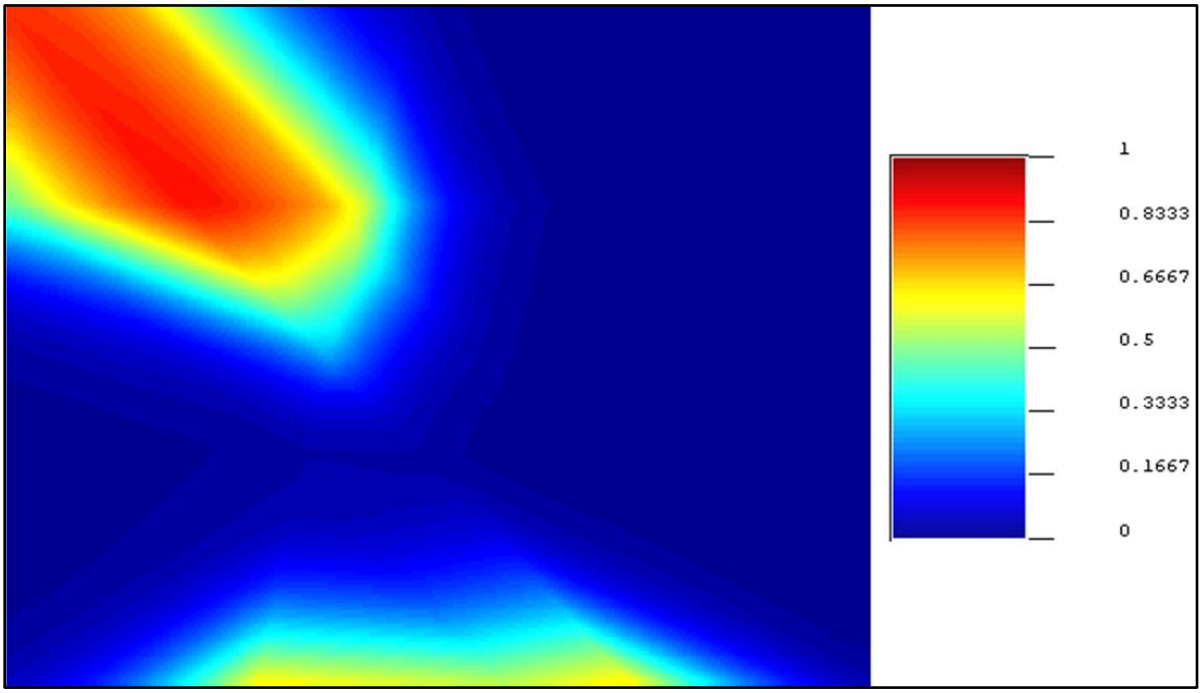

Ilustración 70: *Estimación modelo NN 100_2000, caso HC muestreado con grilla (Elaboración propia).*

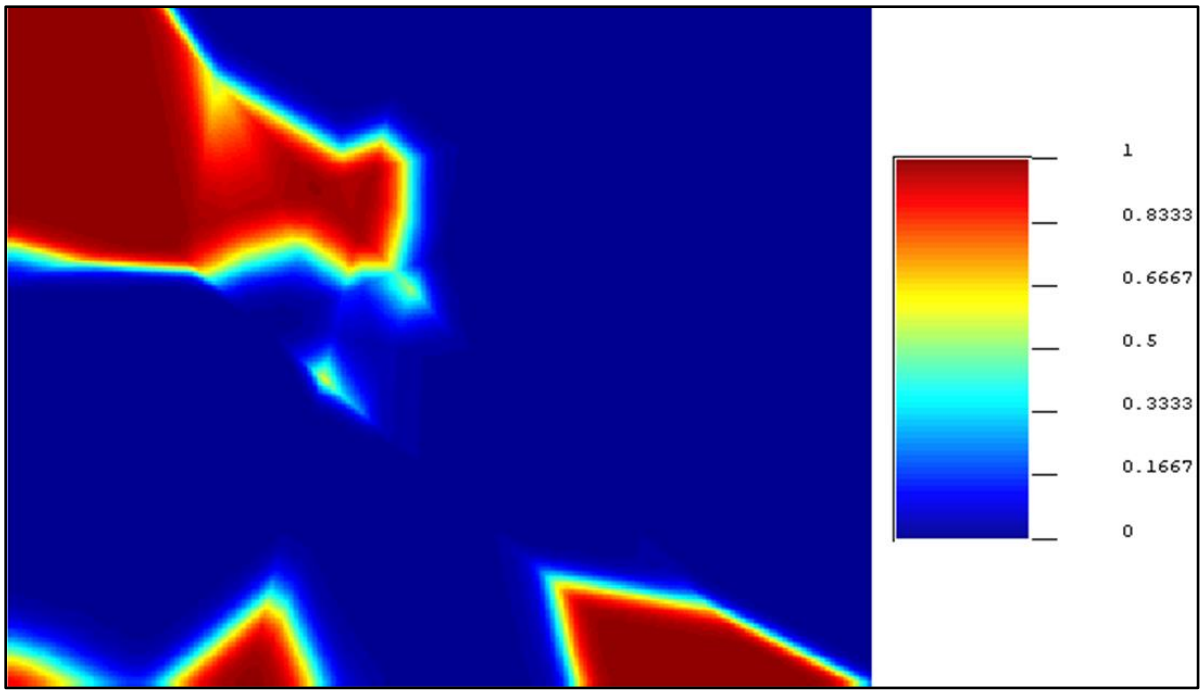

Ilustración 71: *Estimación modelo NN 1000_2000, caso HC muestreado con grilla (Elaboración propia).*

Los resultados obtenidos demuestran que la clave para obtener una adecuada estimación está relacionada con la calidad de las muestras. Esto queda en evidencia cuando ambos modelos de red neuronal para este caso, estiman prácticamente las mismas zonas que los modelos que fueron entrenados con la totalidad de los datos. En consecuencia, se tienen algunas ventajas como: estudiar menos datos y gastar menos recursos. Esto es clave, porque permite optimizar tiempos y disminuir costos. En síntesis, una gran cantidad de muestras o datos no implica que las estimaciones tengan mejores resultados, sino que, está relacionado intrínsicamente con la calidad de las muestras.

Finalmente, ¿Qué sucedería si la red neuronal tuviera como información de entrada adicional la distancia que hay entre las muestras? ¿Se obtendrían mejores estimaciones?

Caso estimación con información de entrada: Distancia entre muestras

Recapitulando, los casos anteriormente mencionados entrenaban la red neuronal conociendo las coordenadas y la concentración de HC o Ni que había en la muestra. En este caso, se agrega un dato adicional que es la distancia que hay entre las muestras. Su razón, es para ayudar a la red neuronal a tener estimaciones más precisas, ya que cuando quiera estimar alguna coordenada conocerá las distancias que hay desde ese punto hacia todas las otras muestras, de manera que al realizar la predicción la red neuronal tomará como referencia las muestras que están más cercanas a su alrededor, como si fuera una especie del método K vecinos más cercanos. Es así como, que se

calculó la distancia euclidiana que había entre cada punto, dado que se trata de un análisis bidimensional. Este procedimiento se empleó para modelo NN 1000_2000, ya que presentó mejores resultados. Para ejemplificar lo mencionado, se muestran en este apartado resultados para variable continua y variable categórica.

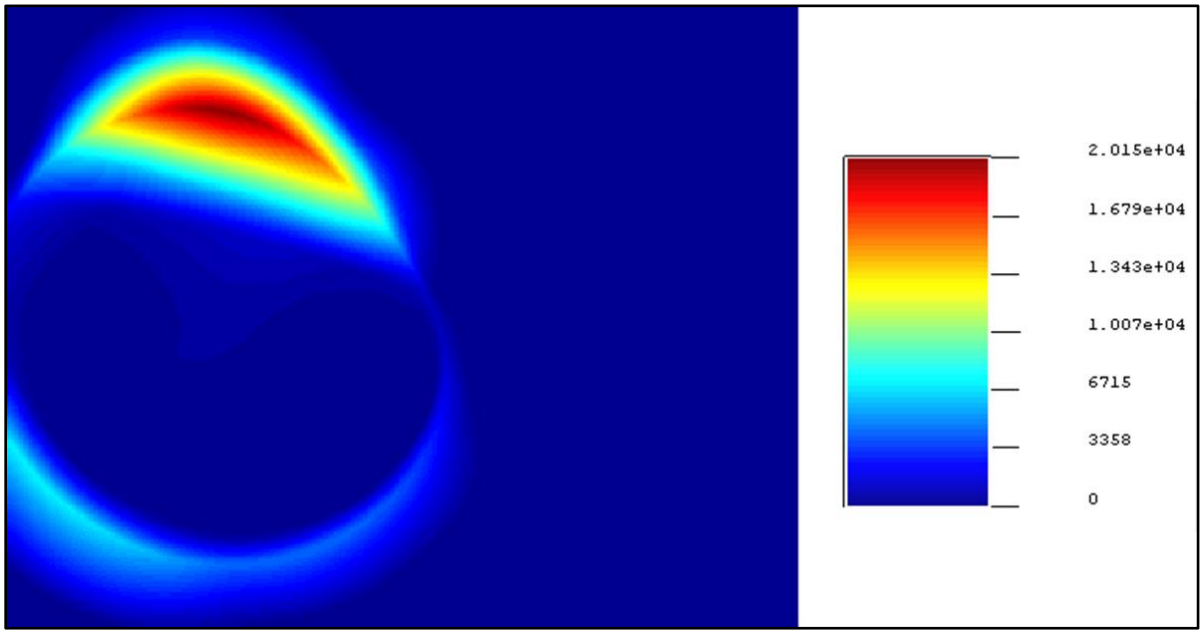

Ilustración 72: *Estimación variable continua modelo NN 1000_2000 HC con distancia entre puntos (Elaboración propia).*

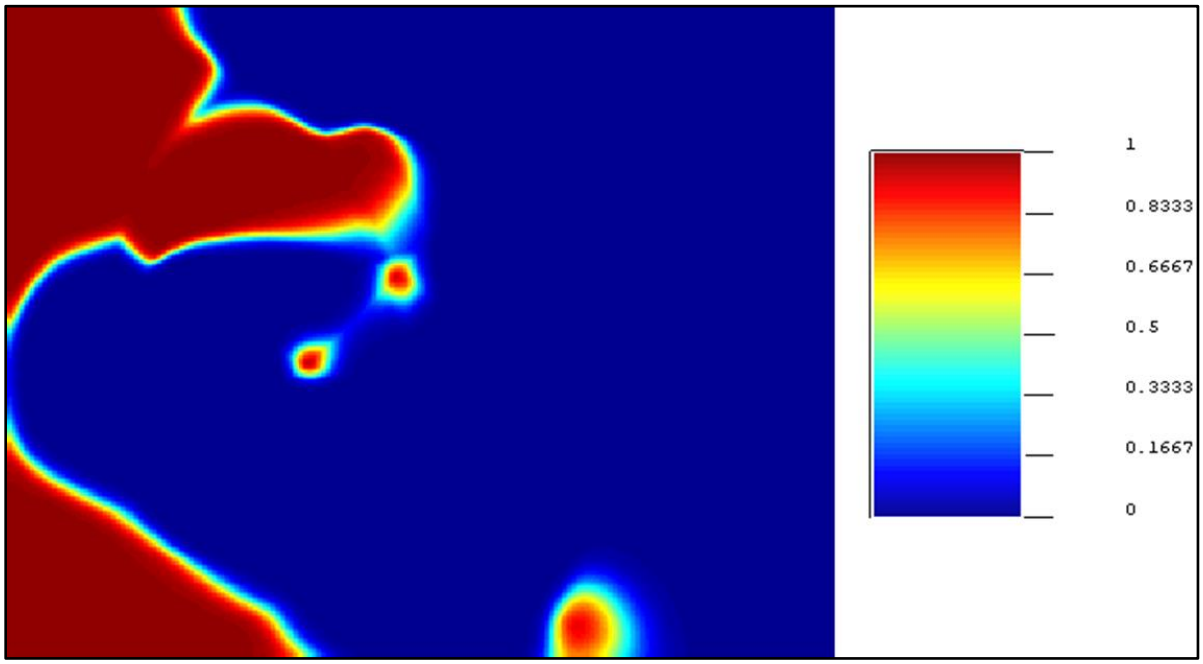

Ilustración 73: *Estimación variable categórica modelo NN 1000_2000 HC muestreado con grilla y distancia entre puntos (Elaboración propia).*

Comparando los resultados de caso continuo sin distancia entre puntos *(ver ilustración 56)* versus caso continuo con distancia entre puntos, se observa que cuando el modelo conoce dicha información, es capaz de estimar con precisión la fuente de contaminación. En cambio, los modelos estimados para variables categóricas no sufren grandes cambios en cuanto a forma y fuente de contaminación, pero si cambia respecto a seguridad con que estima la contaminación, pasando de valores de 60% a cercanos al 95%.

5.6 Zonas contaminadas método machine learning

A continuación, se presentan las zonas con riesgo de contaminación del modelo NN 1000_2000 con los distintos casos mencionados en el apartado anterior.

Variables continuas

La ilustración 74 y 75 exponen las zonas que superan criterio de contaminación establecido para elementos contaminantes: hidrocarburo y níquel.

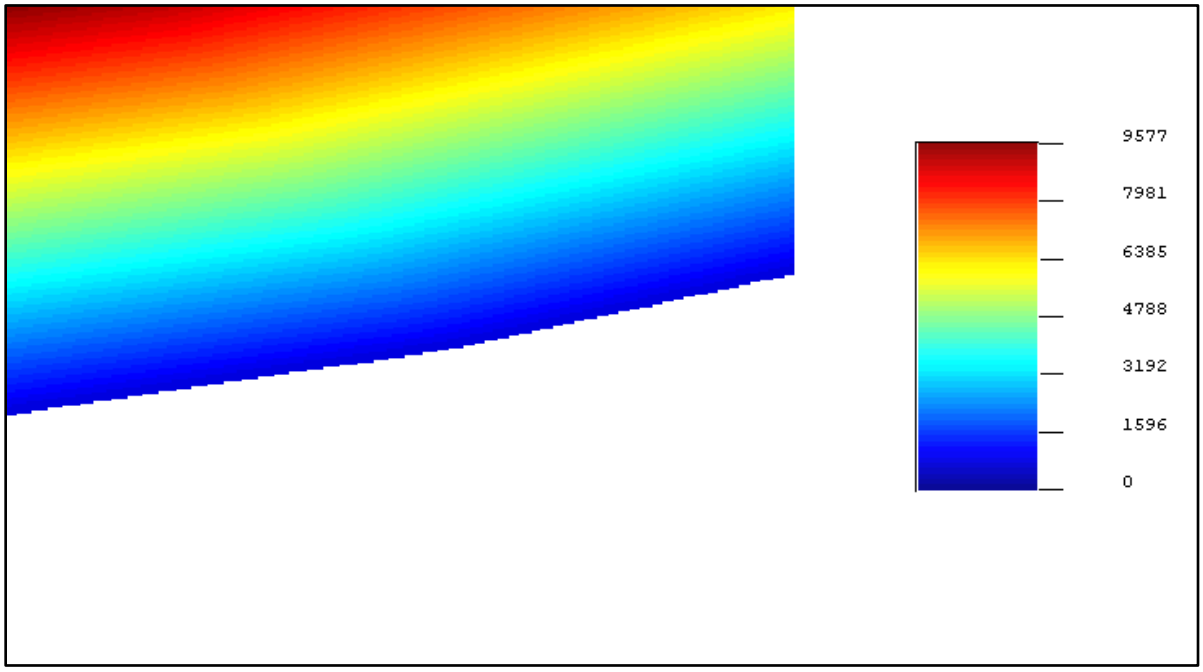

Ilustración 74: *Zonas que superan criterio de contaminación establecido modelo NN 1000_2000 HC (Elaboración propia).*

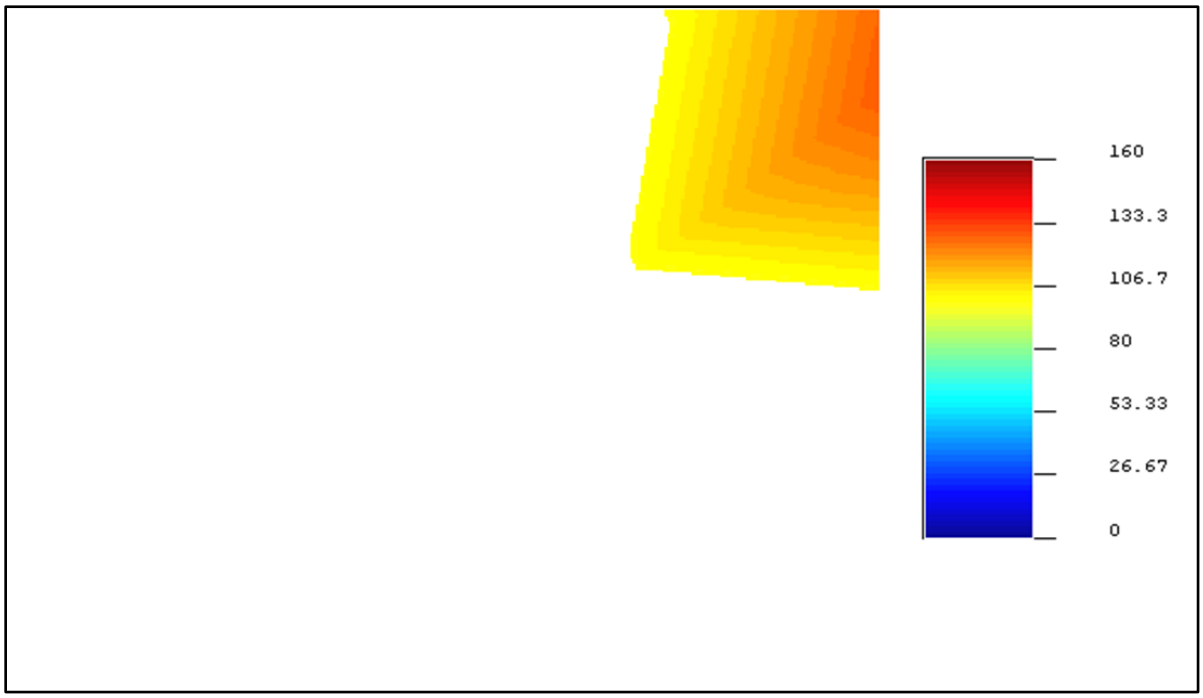

Ilustración 75: *Zonas que superan criterio de contaminación establecido modelo NN 1000_2000 Ni (Elaboración propia)*

La ilustración 76 corresponde a las zonas que superan criterio de contaminación de HC cuando modelo red neuronal conoce la distancia entre puntos

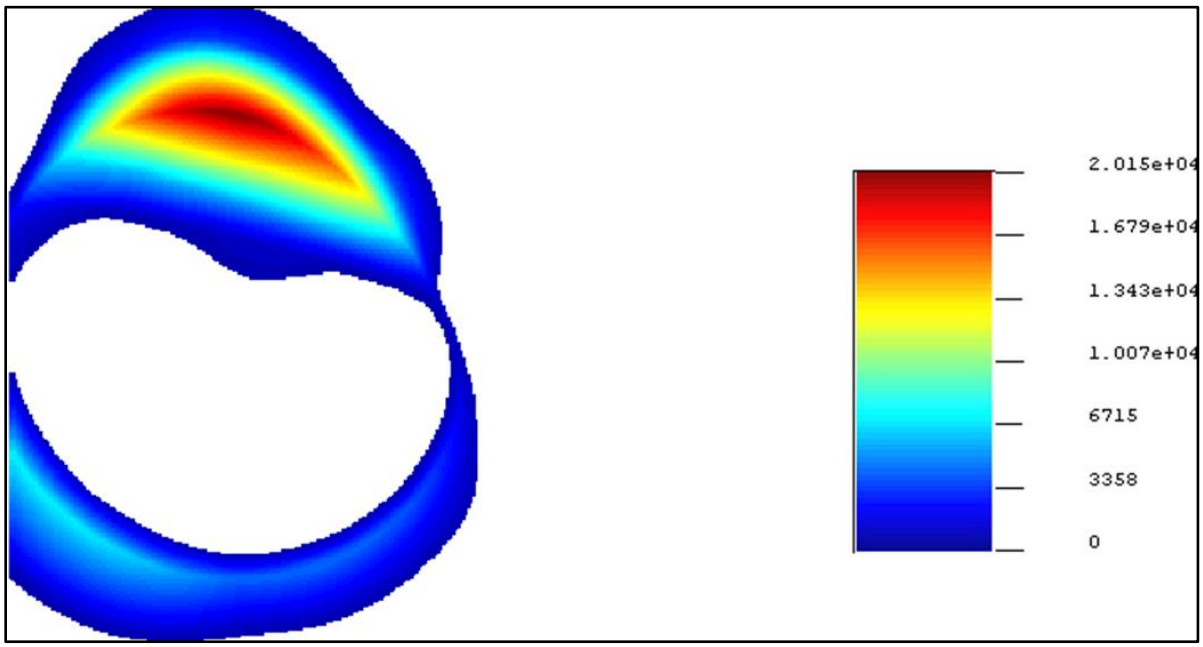

Ilustración 76: *Zonas que superan criterio de contaminación establecido modelo NN 1000_2000 HC con distancia entre puntos (Elaboración propia)*

Variables categóricas

Los resultados de contaminación por hidrocarburo cuando se emplean 192 muestras se presentan desde la ilustración 77 hasta la 79. En cambio, las zonas por contaminación de níquel se pueden observar en apéndice I.

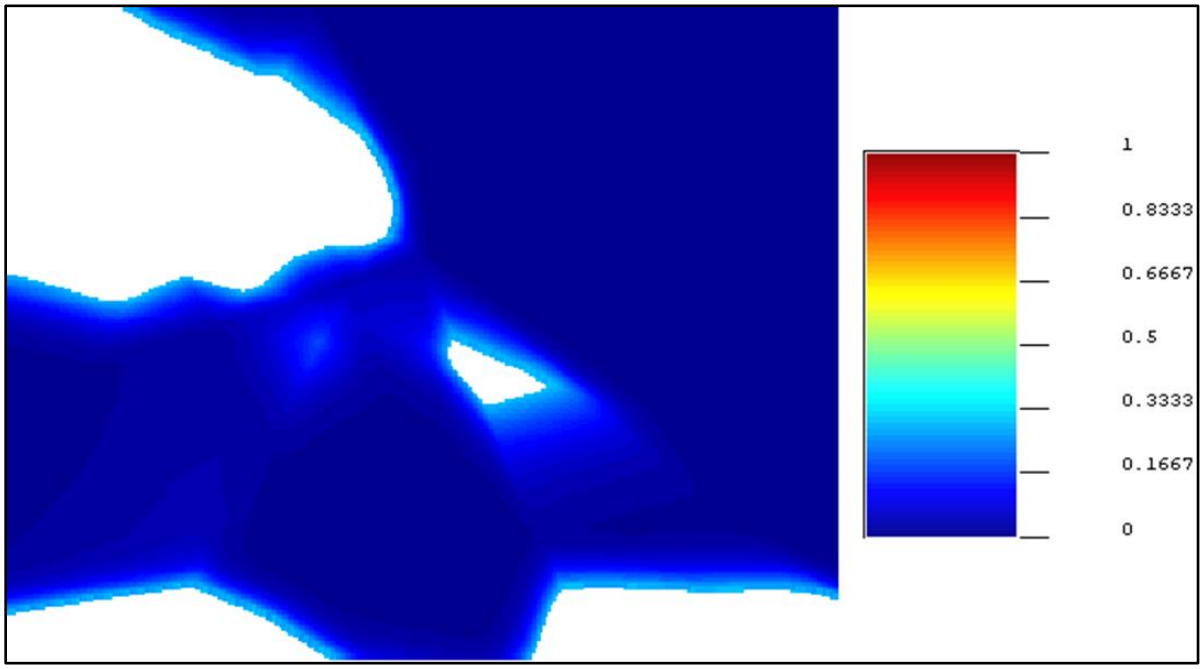

Ilustración 77: *Zonas con bajo riesgo de contaminación modelo NN 1000_2000 HC (Elaboración propia).*

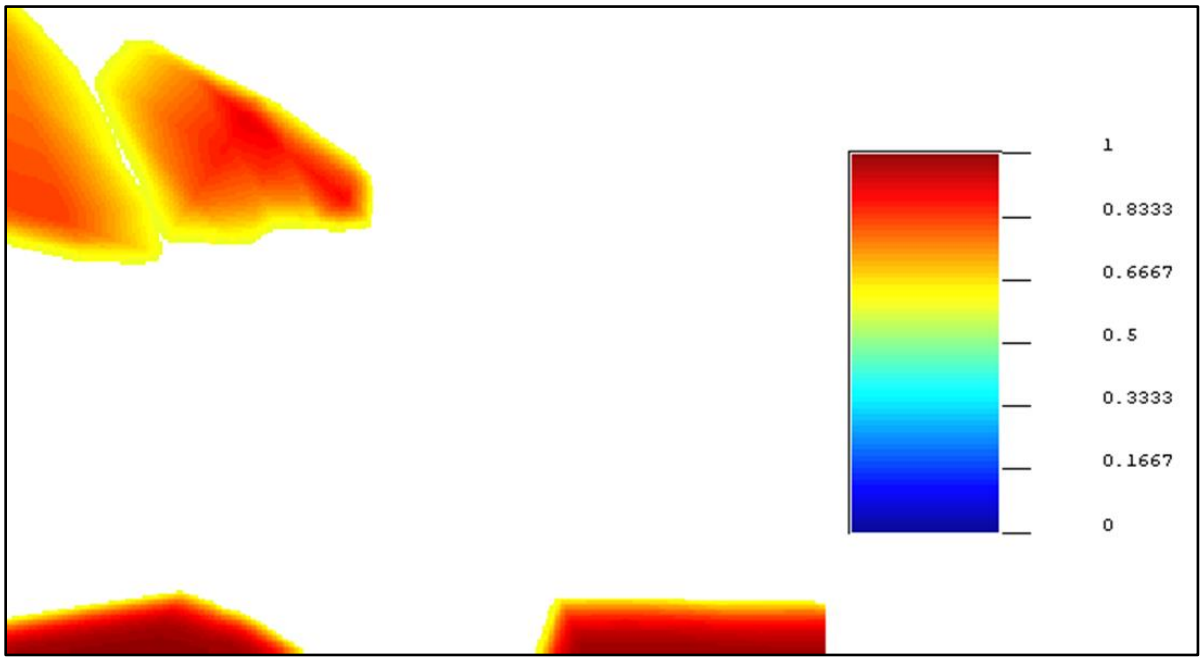

Ilustración 78: *Zonas con alto riesgo de contaminación modelo NN 1000_2000 HC (Elaboración propia).*

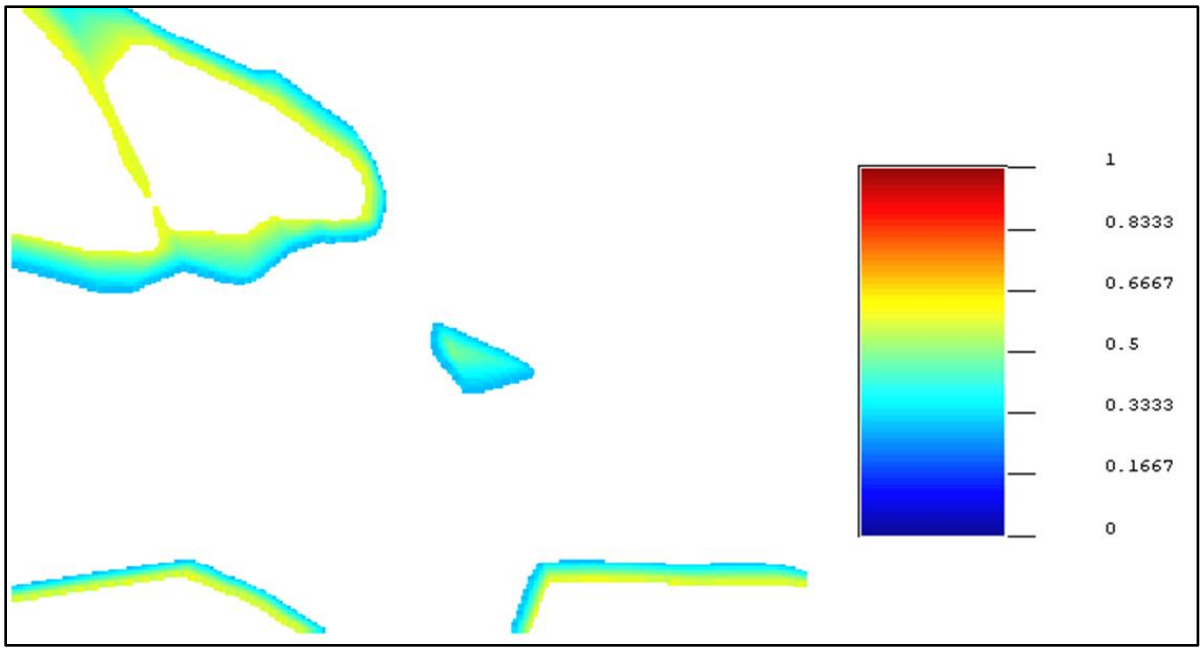

Ilustración 79: *Zonas de incertidumbre modelo NN 1000_2000 HC (Elaboración propia).*

Caso hidrocarburo con 210 muestras

Las siguientes tres ilustraciones corresponden a las zonas con riesgo de contaminación cuando el modelo estima con 210 muestras.

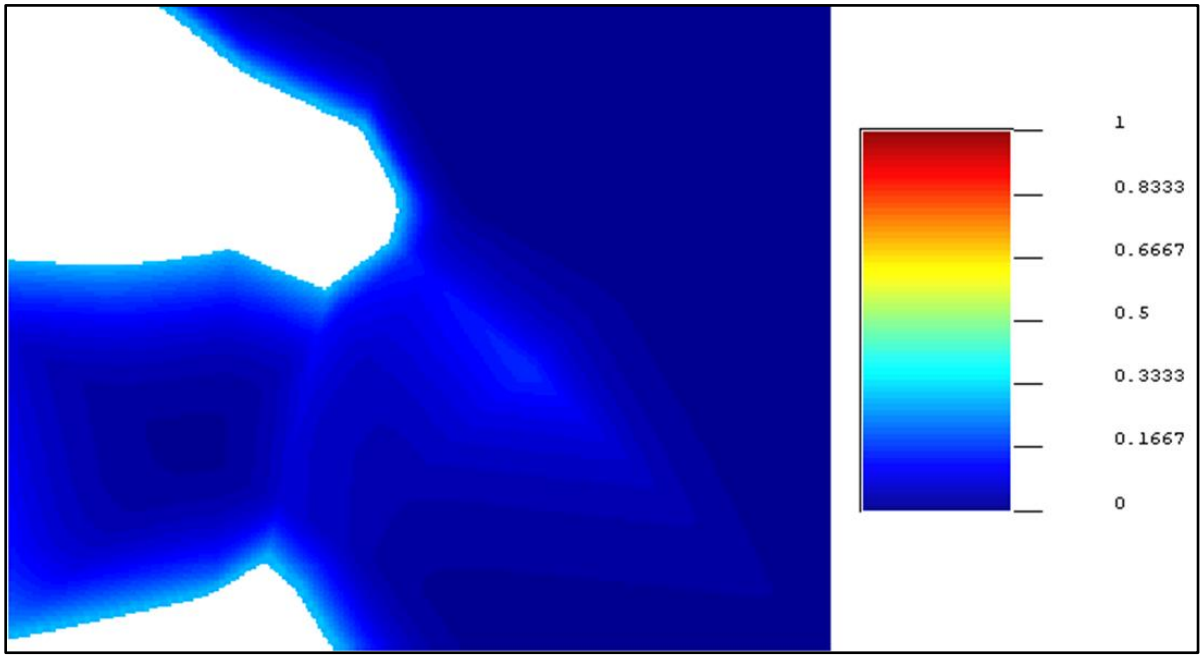

Ilustración 80: *Zonas con bajo riesgo de contaminación modelo NN 1000_2000 HC caso 210 muestras (Elaboración propia).*

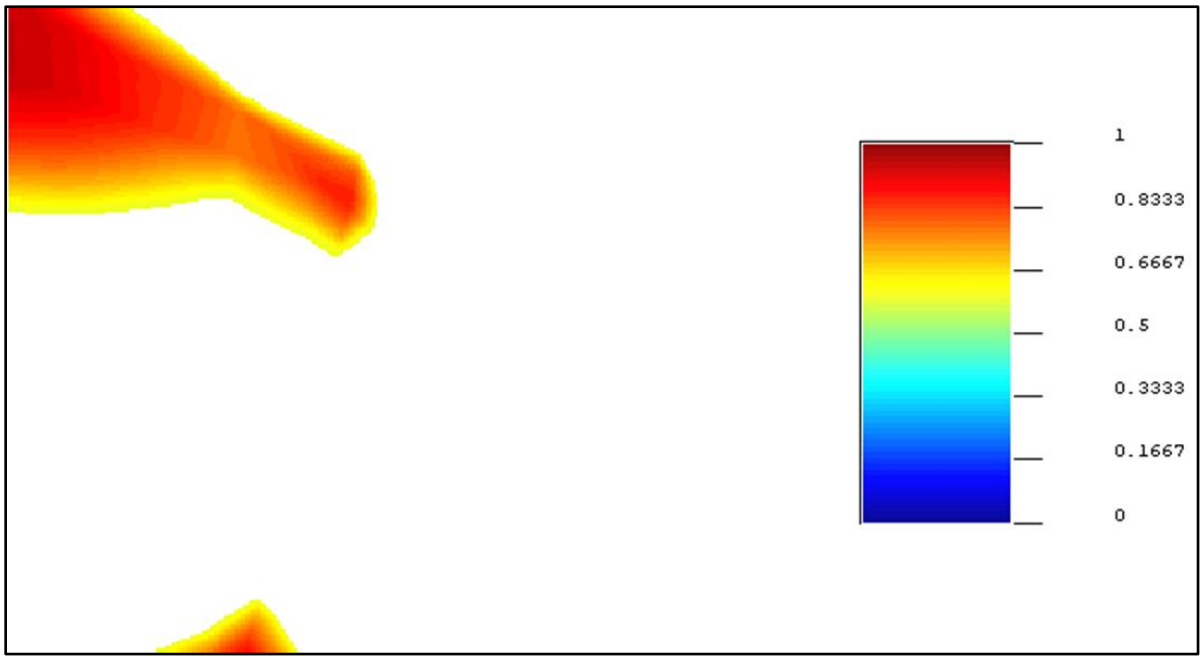

Ilustración 81: *Zonas con alto riesgo de contaminación modelo NN 1000_2000 HC caso 210 muestras (Elaboración propia).*

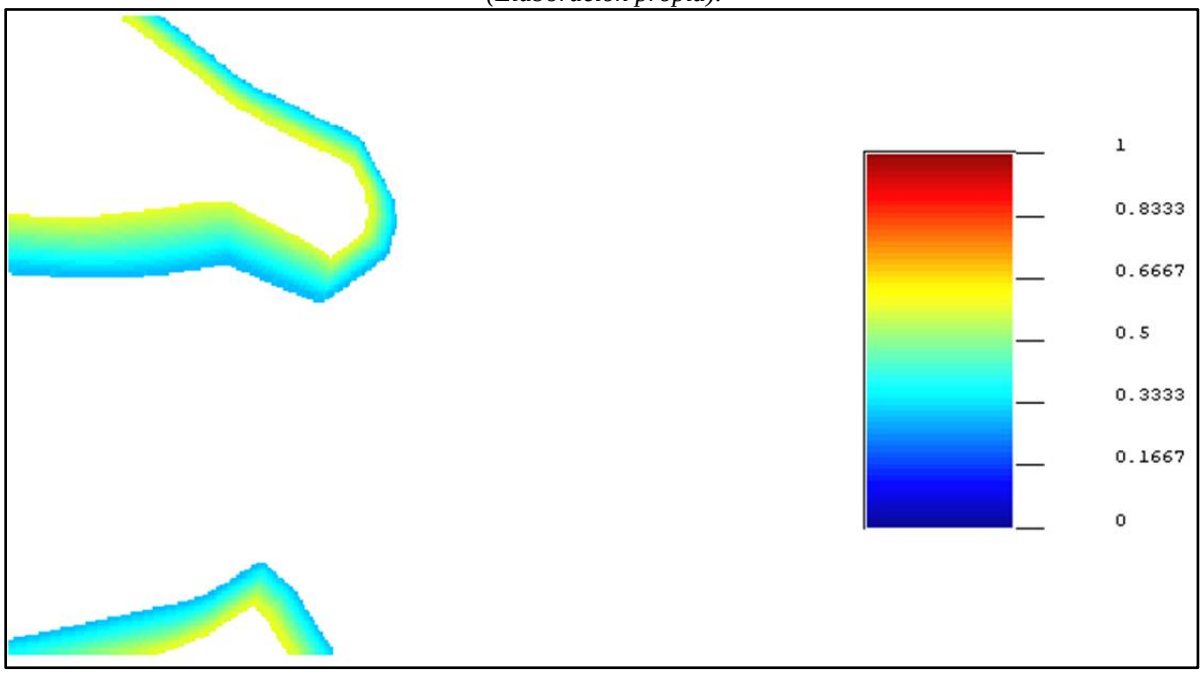

Ilustración 82: *Zonas de incertidumbre modelo NN 1000_2000 HC caso 210 muestras (Elaboración propia).*

Caso hidrocarburo muestreado con grilla

Las siguientes ilustraciones corresponden a las zonas con riesgo de contaminación cuando el modelo utiliza grilla en la toma de muestras del hidrocarburo.

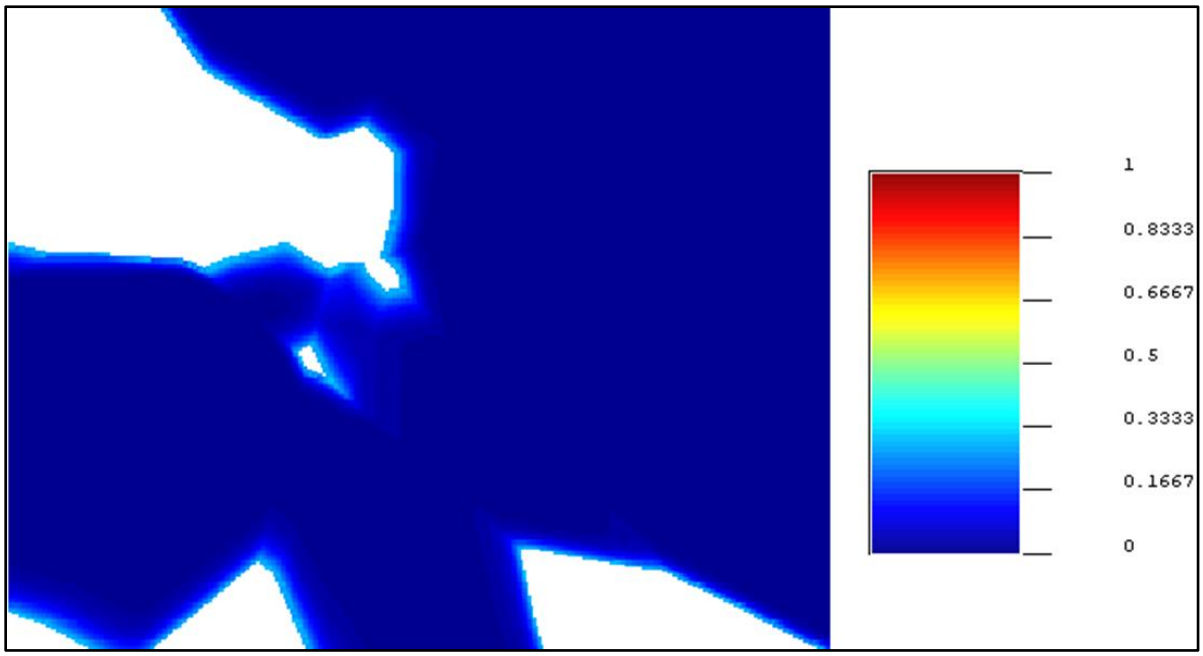

Ilustración 83: *Zonas con bajo riesgo de contaminación modelo NN 1000_2000, caso HC muestreado con grilla (Elaboración propia).*

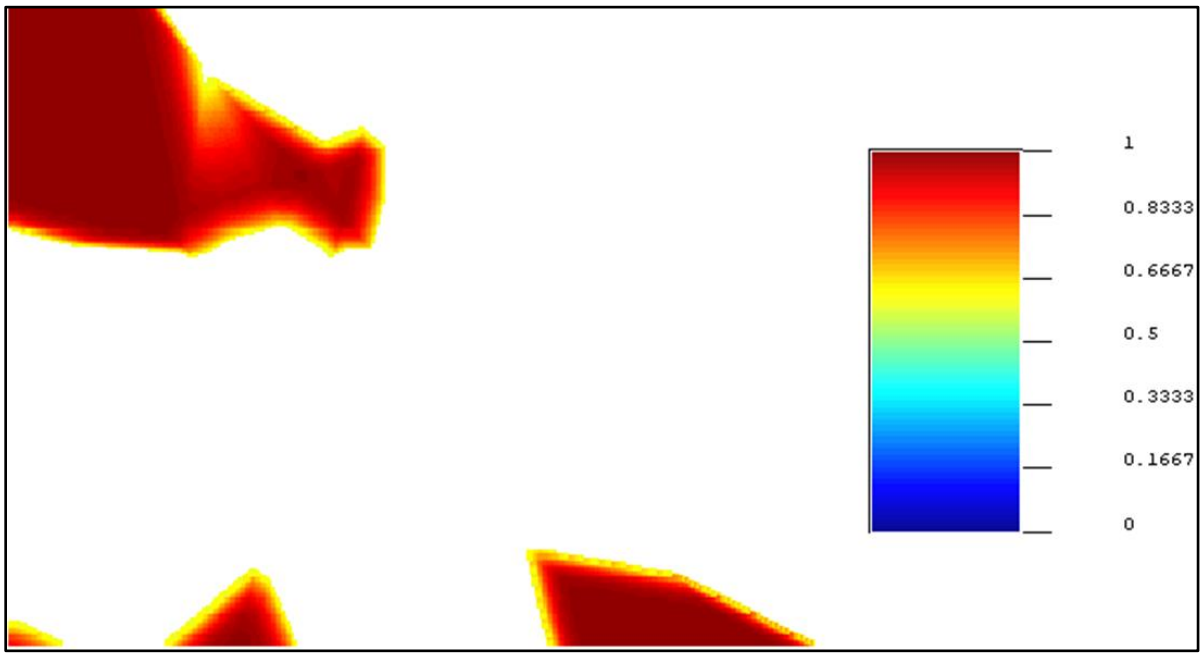

Ilustración 84: *Zonas con alto riesgo de contaminación modelo NN 1000_2000, caso HC muestreado con grilla (Elaboración propia).*

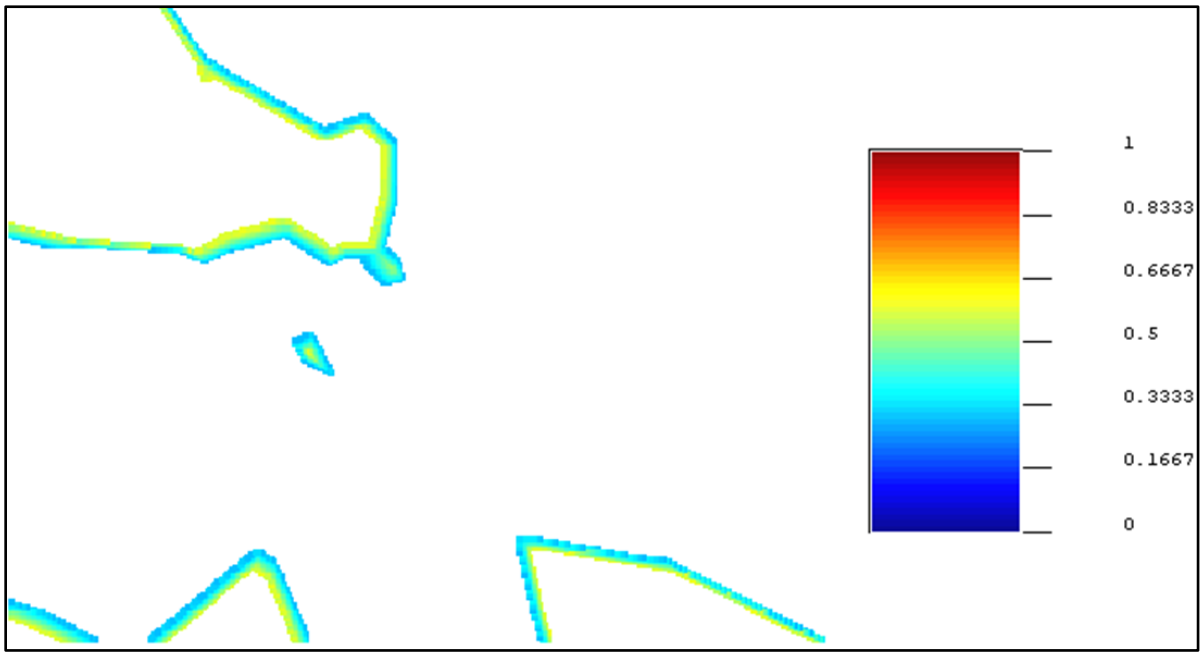

Ilustración 85: *Zonas de incertidumbre modelo NN 1000_2000, caso HC muestreado con grilla (Elaboración propia).*

Caso hidrocarburo muestreado con grilla con información de entrada: Distancia entre puntos

A continuación, se presentan las zonas con riesgo de contaminación cuando el modelo tiene como información la distancia que existe entre los puntos o muestras.

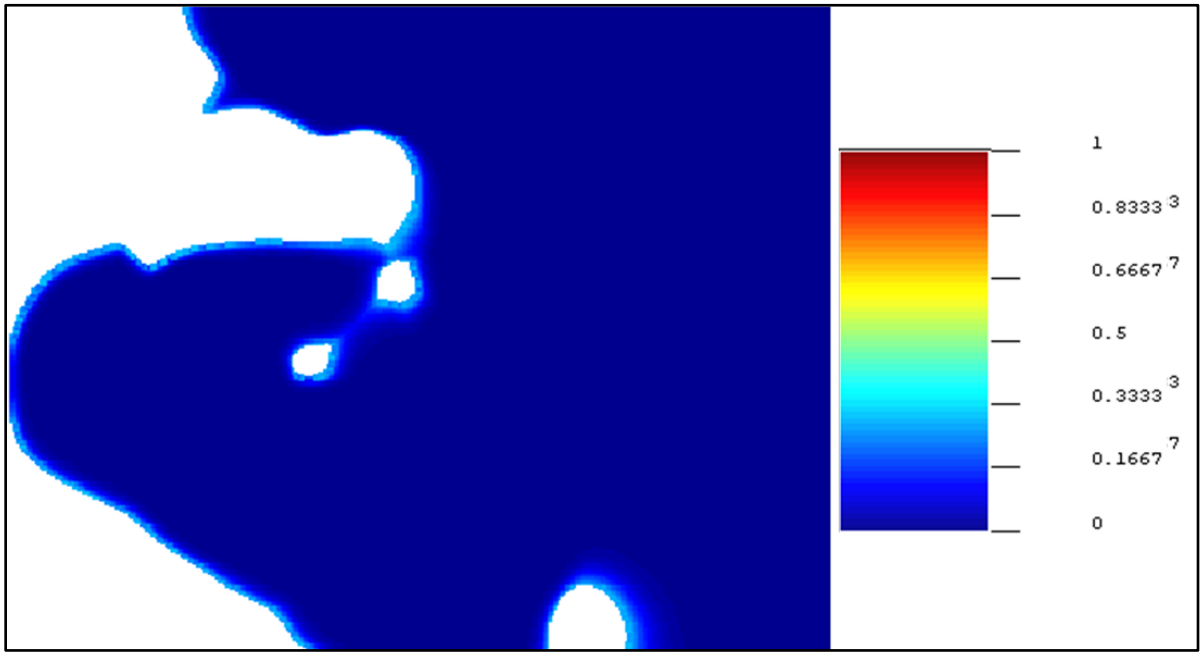

Ilustración 86: *Zonas con bajo riesgo de contaminación modelo NN 1000_2000 HC caso muestreado con grilla con distancia entre puntos (Elaboración propia).*

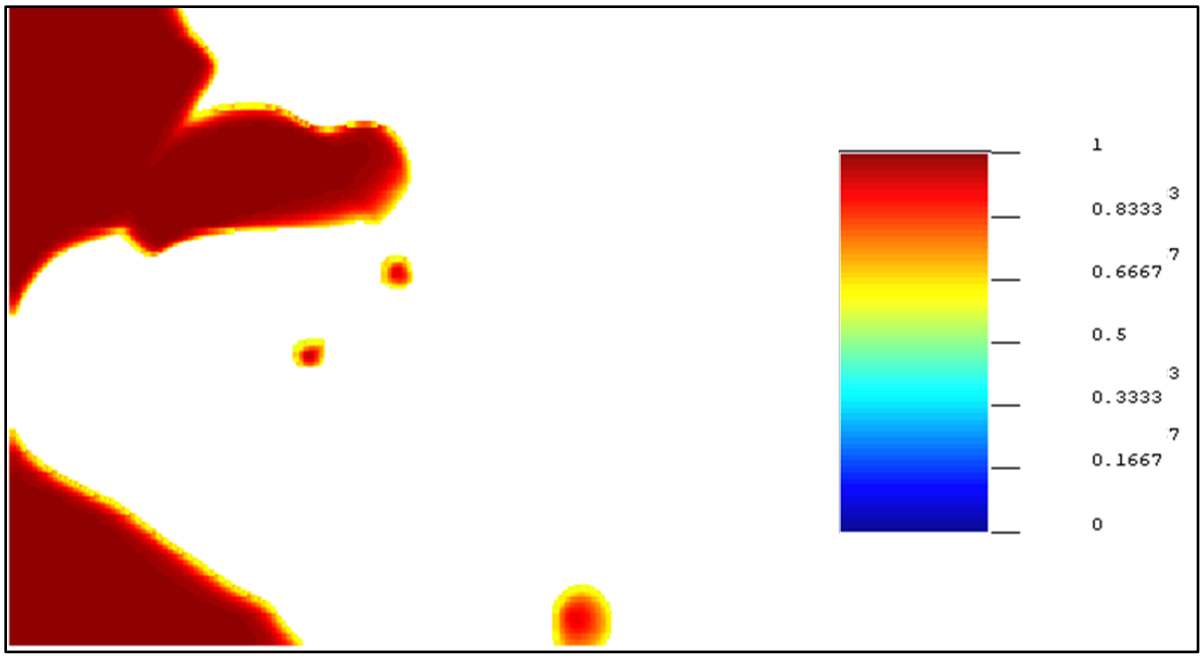

Ilustración 87: *Zonas con alto riesgo de contaminación modelo NN 1000_2000 HC caso muestreado con grilla con distancia entre puntos (Elaboración propia).*

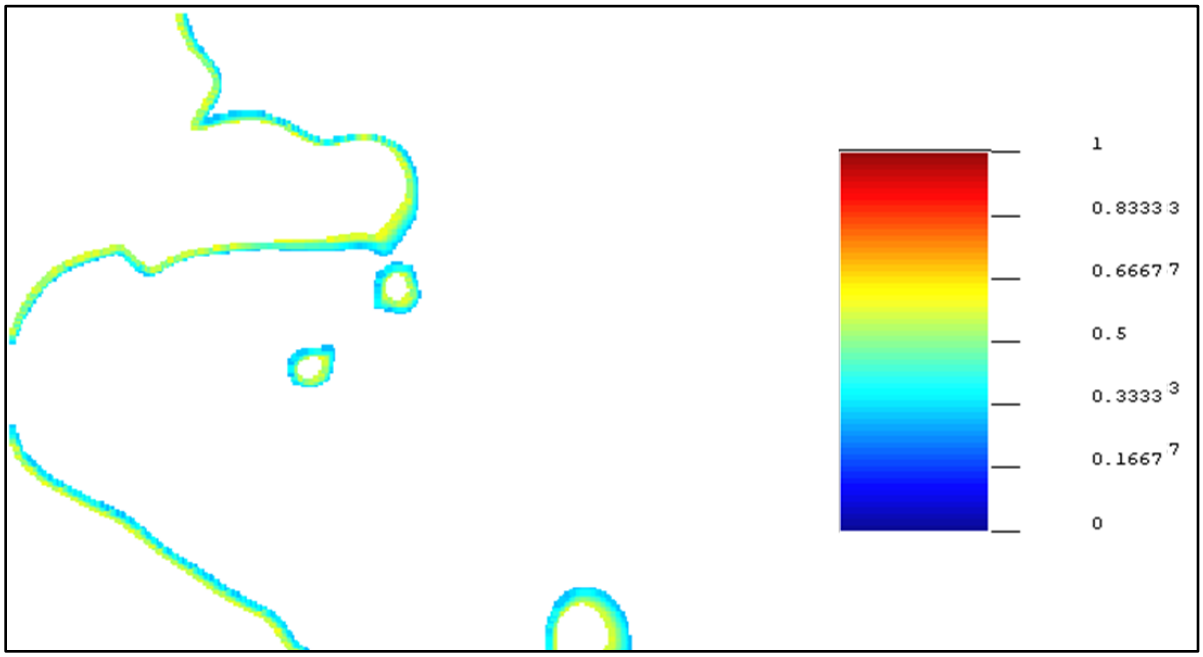

Ilustración 88: *Zonas de incertidumbre modelo NN 1000_2000 HC caso muestreado con grilla con distancia entre puntos (Elaboración propia).*

Sin un análisis exhaustivo, se aprecia que los resultados por el modelo NN 1000_2000 entrega pocas zonas de incertidumbre y se ubican preferentemente alrededor de la fuente de contaminación.

5.6.1 Superficies de contaminación

De acuerdo con los resultados de entrenamiento, prueba y estimación, el modelo que presenta mejores resultados en estos tres aspectos fue el NN 1000_2000, pero se agrega también el NN 100_2000 para demostrar la diferencia que hay cuando se tienen menos neuronas al estimar.

Variables continuas

A continuación, se presentan los bloques que superan criterio de contaminación cuando se estima mediante redes neuronales variables continuas.

Tabla 34: *Superficie de contaminación de variables continuas, modelos redes neuronales (Elaboración propia).*

Caso	Variable	Modelo	Bloques que superan criterio de contaminación	Superficie de contaminación total (ha)	Superficie de contaminación total (%)
Sin distancia entre puntos	**HC**	**NN 100_2000**	25516	10.21	30.67
		NN 1000_2000	45980	18.39	55.26
	Ni	**NN 100_2000**	8300	3.32	9.98
		NN 1000_2000	8663	3.47	10.41
Con distancia entre puntos	**HC**	**NN 100_2000**	33052	13.22	39.73
		NN 1000_2000	21270	8.51	25.56
	Ni	**NN 100_2000**	14204	5.68	17.07
		NN 1000_2000	8205	3.28	9.86

Como se vio en el apartado “estimación mediante machine learning, los modelos que trabajan con variables continuas tienen problemas al estimar cuando no tienen como información de entrada la distancia que hay entre puntos; no son precisos en forma y fuente de contaminación, más bien intuyen de donde proviene, por lo que existe una gran diferencia entre el mismo modelo NN 1000_2000 cuando estima con o sin distancia entre puntos la variable hidrocarburo. En cambio, el níquel no presenta una gran variación entre casos, siendo aproximadamente la superficie de contaminación de un 10%.

Respecto al volumen de contaminación, se utilizó el mismo espesor que fue calculado previamente por kriging.

Tabla 35: *Volumen de contaminación variables continuas, modelos redes neuronales (Elaboración propia).*

Caso	Variable	Modelo	Volumen total de contaminación (m^3)
Sin distancia entre puntos	**HC**	**NN 100_2000**	82439.03
		NN 1000_2000	135076.13
	Ni	**NN 100_2000**	20197.3
		NN 1000_2000	21317.5
Con distancia entre puntos	**HC**	**NN 100_2000**	105207.6
		NN 1000_2000	68966.7
	Ni	**NN 100_2000**	38022.1
		NN 1000_2000	23930.1

Variables categóricas

A continuación, se presentan las superficies probables de contaminación cuando se estima mediante redes neuronales variables categóricas **sin distancia entre puntos.**

Caso estimación con datos de inicio

Tabla 36: *Superficie de contaminación probable, redes neuronales caso datos de inicio sin distancia entre puntos (Elaboración propia).*

Variable	Modelo	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
HC	NN 100_2000	0-30	Bajo	66597	26.64	80.04
		30-60	No se sabe	7999	3.20	9.61
		60-100	Alto	8604	3.44	10.34
	NN 1000_2000	0-30	Bajo	63589	25.44	76.43
		30-60	No se sabe	6322	2.53	7.60
		60-100	Alto	13289	5.32	15.97
Ni	NN 100_2000	0-30	Bajo	83200	100	83200
		30-60	No se sabe	0	0.00	0.00
		60-100	Alto	0	0	0.00
	NN 1000_2000	0-30	Bajo	61263	24.5	73.63
		30-60	No se sabe	14844	6	17.85
		60-100	Alto	7093	2.84	8.52

Caso estimación 210 muestras

Tabla 37: *Superficie de contaminación probable, redes neuronales caso 210 muestras sin distancia entre puntos (Elaboración propia).*

Variable	Modelo	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
HC	NN 100_2000	0-30	Bajo	67309	26.92	80.90
		30-60	No se sabe	7894	3.16	9.49
		60-100	Alto	7997	3.20	9.61
	NN 1000_2000	0-30	Bajo	67983	27.19	81.71
		30-60	No se sabe	6190	2.48	7.44
		60-100	Alto	9027	3.61	10.85

Caso estimación con grilla

Tabla 38: *Superficie de contaminación probable, redes neuronales caso grilla sin distancia entre puntos (Elaboración propia).*

Variable	Modelo	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
HC	NN 100_2000	0-30	Bajo	64649	25.86	77.70
		30-60	No se sabe	9300	3.72	11.18
		60-100	Alto	9251	3.70	11.12
	NN 1000_2000	0-30	Bajo	66548	26.62	79.99
		30-60	No se sabe	2631	1.05	3.16
		60-100	Alto	14021	5.61	16.85

Los resultados anteriores demuestran que modelos con mayor cantidad de neuronas reducen las zonas de incertidumbre. Sobre el hidrocarburo, se presentan porcentajes similares respecto zonas con alto riesgo de contaminación, siendo un mínimo de 10% y pudiendo alcanzar como máximo el 17% de la zona de estudio. Por otra parte, el único modelo capaz de estimar zonas contaminadas para el níquel es modelo NN 1000_2000 con un 9% de probabilidad de presentar zonas con riesgo de contaminación. Por último, si comparamos resultados de caso de estimación con datos de inicio versus estimación con grilla, se puede apreciar que estiman prácticamente las mismas superficies contaminadas, lo que significa que se pueden lograr estimaciones muy precisas a pesar de que existan pocas muestras, por lo tanto, si se quieren obtener modelos con alta calidad la etapa del muestreo es la clave en modelos de contaminación de suelos.

Respecto a superficies probables de contaminación cuando modelo tiene como **información de entrada distancia entre puntos**, se tienen los siguientes resultados.

Caso estimación con datos de inicio

Tabla 39: *Superficie de contaminación probable, redes neuronales caso datos de inicio con distancia entre puntos (Elaboración propia).*

Variable	Modelo	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
HC	NN 100_2000	0-30	Bajo	65223	26.09	78.39
		30-60	No se sabe	3895	1.56	4.68
		60-100	Alto	14082	5.63	16.93
	NN 1000_2000	0-30	Bajo	61719	24.69	74.18
		30-60	No se sabe	4643	1.86	5.58
		60-100	Alto	16838	6.74	20.24
Ni	NN 100_2000	0-30	Bajo	76439	30.58	91.87
		30-60	No se sabe	2804	1.12	3.37
		60-100	Alto	3957	1.58	4.76
	NN 1000_2000	0-30	Bajo	76664	30.67	92.14
		30-60	No se sabe	2280	0.91	2.74
		60-100	Alto	4256	1.70	5.12

Caso estimación 210 muestras

Tabla 40: *Superficie de contaminación probable, redes neuronales caso 210 muestras con distancia entre puntos (Elaboración propia).*

Variable	Modelo	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
HC	NN 100_2000	0-30	Bajo	67222	26.89	80.80
		30-60	No se sabe	3509	1.40	4.22
		60-100	Alto	12469	4.99	14.99
	NN 1000_2000	0-30	Bajo	67532	27.01	81.17
		30-60	No se sabe	3615	1.45	4.34
		60-100	Alto	12053	4.82	14.49

Caso estimación con grilla

Tabla 41: *Superficie de contaminación probable, redes neuronales caso grilla con distancia entre puntos (Elaboración propia).*

Variable	Modelo	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
HC	NN 100_2000	0-30	Bajo	67136	26.85	80.69
		30-60	No se sabe	2214	0.89	2.66
		60-100	Alto	13850	5.54	16.65
	NN 1000_2000	0-30	Bajo	64893	25.96	78.00
		30-60	No se sabe	1774	0.71	2.13
		60-100	Alto	16533	6.61	19.87

Los resultados demuestran que cuando el modelo tiene como información de entrada la distancia que hay entre las muestras, es capaz de reducir al mínimo las zonas de incertidumbre y que se limitan a las zonas de contacto. Esto tiene como consecuencia un aumento hacia las zonas con alto riesgo de probabilidad de contaminación con valores entre un 15 y 21%.

5.7 Comparación modelos

Para realizar comparación entre modelos geoestadísticos y machine learning se debe cumplir un requisito. Ambos modelos deben estimar con los mismos datos. De esta manera están compitiendo en igualdad de condiciones. Sin embargo, como modelos de redes neuronales estiman con la porción que fue destinada a fase de entrenamiento, se realizó una nueva estimación geoestadística usando los mismos datos de entrenamiento de la red neuronal y que corresponden al 80% de los datos originales. Posteriormente, se compara sus resultados con modelo NN 1000_2000 de red neuronal mediante el 20% de datos destinado a prueba.

A continuación, se presenta comparación de modelo geoestadístico vs machine learning cuando **estiman variables continuas.**

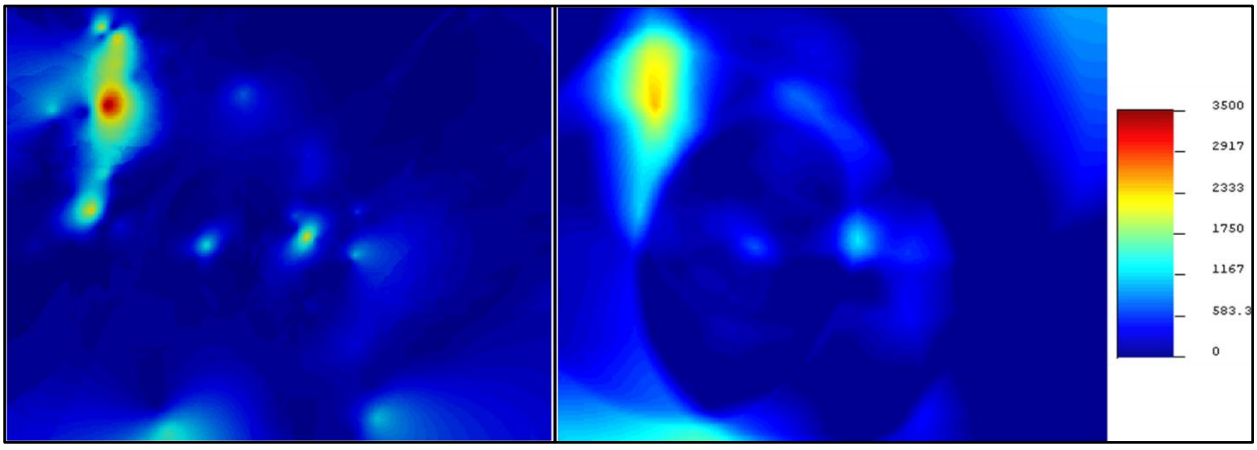

Ilustración 89: *Estimación geoestadística (izquierda) vs estimación machine learning con distancia entre puntos (derecha) usando variables continuas (Elaboración propia).*

La ilustración 89 corresponde a modelos que estiman variables continuas sin datos atípicos, debido a que modelo geoestadístico no es capaz de estimar cuando existe una alta variabilidad entre los datos. Este aspecto es importante de mencionar, ya que marca la primera diferencia entre modelos, dado que las redes neuronales son capaces de realizar predicción, aunque haya una alta varianza de los datos, lo que implica que pueda predecir con el 100% de las muestras de hidrocarburo. Agregar que, las redes neuronales mejoran su estimación en casos continuos cuando tienen como

información de entrada la distancia que existe entre las muestras. Por otra parte, el dato original con mayor concentración tiene un valor de 3500 ppm. Sobre esto, el kriging ordinario empleado predice un valor máximo de concentración mayor que modelo de red neuronal, siendo este de 3435 ppm versus un 2429 ppm. Sin embargo, cuando se trata de estimar valores mínimos, las redes neuronales pueden predecir valores bastantes lejanos al mínimo original, siendo de -753 cuando el original es de 100 ppm. Para más detalle de comparación acerca de estadísticas descriptivas, se puede observar la siguiente tabla.

Tabla 42: *Estadísticas descriptivas de estimación, modelo geoestadístico vs machine learning (Elaboración propia).*

Modelo	Mínimo	Máximo	Media	Mediana	Varianza	Q3
Geoestadística	-196.7	3435	256.3	130	113938	305
Machine Learning	-753.9	2429	206.5	115.8	196469	357.5

Acerca de la distribución de valores estimados, se pueden apreciar en histograma 90 y 91.

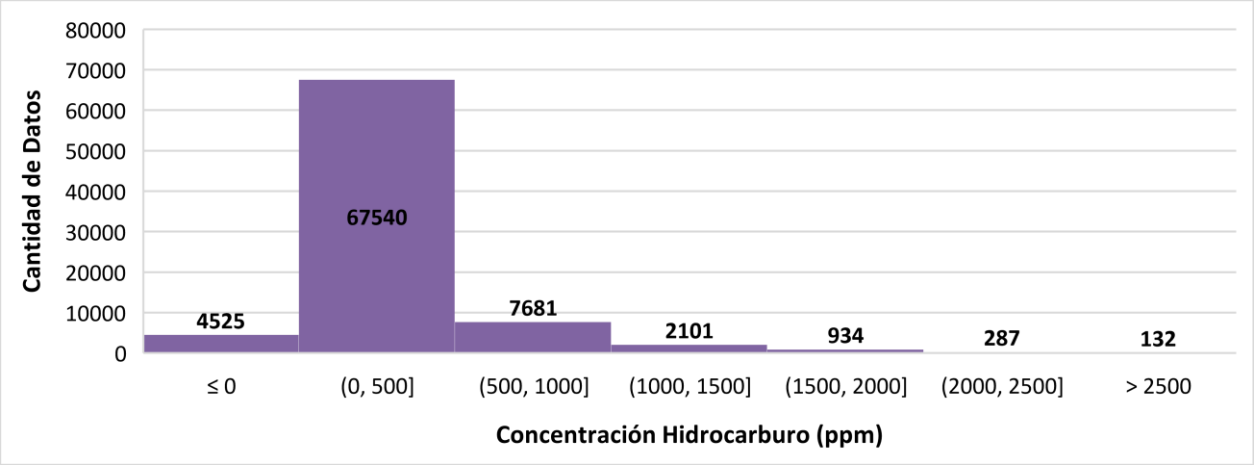

Ilustración 90: *Histograma de estimación geoestadística con 80% de datos, caso sin datos atípicos HC (Elaboración propia).*

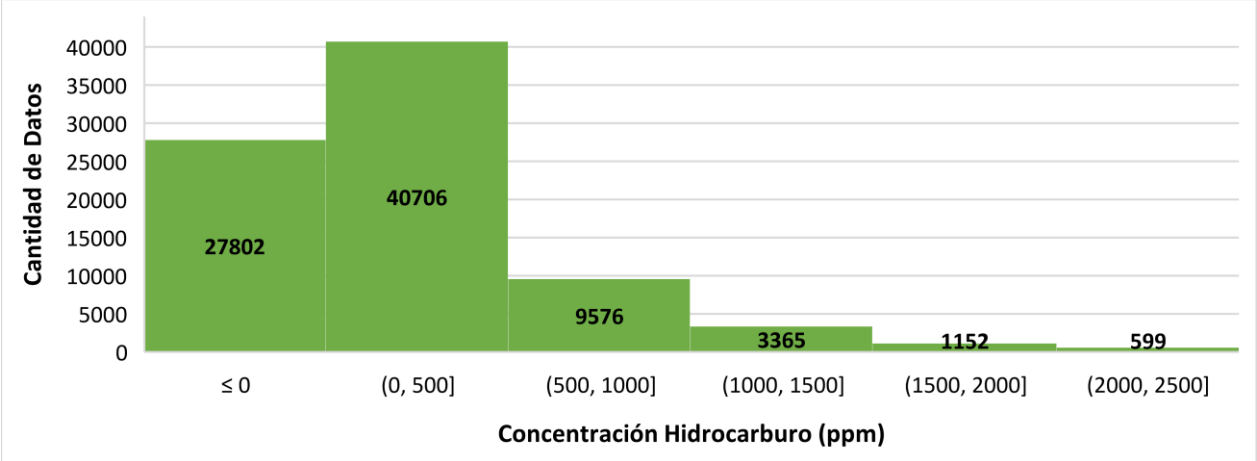

Ilustración 91: *Histograma de estimación machine learning, caso sin datos atípicos HC (Elaboración propia).*

Observando los histogramas, se aprecia que aproximadamente el 80% de los datos estimados mediante kriging ordinario se concentran en un intervalo de concentración, dejando muy pocos datos en los extremos. Esto es típico por el efecto del suavizado del kriging. En cambio, redes neuronales en el mismo intervalo concentran aproximadamente el 48% de los datos.

Respecto a bloques que superan criterio de contaminación, se tiene siguiente comparación:

Tabla 43: *Bloques que superan criterio de contaminación, modelo geoestadístico vs machine learning (Elaboración propia).*

Modelo	Bloques que superan criterio de contaminación	Superficie de contaminación total (ha)	Superficie de contaminación total (%)
Geoestadística	6036	2.41	7.3
Machine Learning	8652	3.46	10.4

Finalmente, para evaluar modelos se utiliza el 20% de datos que no fueron utilizados en estimación, comparando valor real con valor estimado. Para esto, se utiliza matriz de clasificación porque permite comparar modelos que tienen diferentes indicadores de evaluación. Los resultados se aprecian en las siguientes tablas.

Tabla 44: *Evaluación modelo geoestadístico vs machine learning con variables continuas (Elaboración propia).*

Modelo	Exactitud	Precisión	Exhaustividad
Geoestadística	0.866	0.866	0.866
Machine Learning	0.833	0.909	0.833

Tabla 45: *Matriz de confusión modelo geoestadístico vs machine learning con variables continuas (Elaboración propia).*

		Predicción				
		Geoestadística		**Machine Learning**		
Real		**0**	**1**	**0**	**1**	**Total**
	0	26	2	24	4	28
	1	2	0	1	1	2
	Total	28	2	25	5	30

Los resultados demuestran que existen mínimas diferencias entre kriging y redes neuronales cuando trabajan con variables continuas. Sin embargo, las redes neuronales predicen puntos de alta contaminación mucho mejor, clasificando un 16% de los datos como contaminados, mientras que el kriging solo un 6% de los datos.

A continuación, se presenta comparación de modelos cuando trabajan con **variables categóricas** y redes neuronales tienen como **información de entrada distancia entre puntos.**

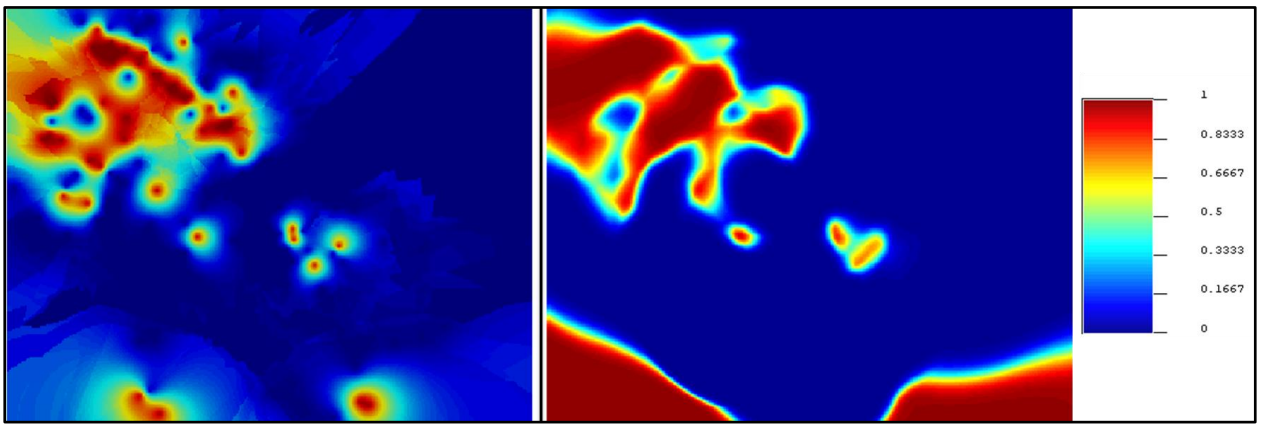

Ilustración 92: *Estimación geoestadística (izquierda) vs estimación machine learning con distancia entre puntos (derecha) usando variables categóricas (Elaboración propia).*

Observando la ilustración 92, se observa que predicciones de redes neuronales se mueve en pocos intervalos de probabilidad. Por ejemplo, de los 61719 bloques con bajo riesgo de contaminación, 94% de dichos bloques tiene una probabilidad del 10% de estar contaminado, mientras que solo el 3% de los bloques entrega una probabilidad de estar contaminado entre el 20 y 30%. Esto mismo ocurre con las zonas de alto riesgo de contaminación, donde 11060 bloques tienen probabilidades de estar contaminados del 90 a 100%. Esto tiene como consecuencia, que la estimación por redes neuronales entregue pocas zonas de incertidumbre. En cambio, la distribución de las estimaciones del kriging es más pareja. Esto se aprecia en las ilustraciones 93 y 94.

Acerca de los resultados probables de superficies contaminadas, se aprecian en la tabla 46.

Tabla 46: *Comparación de superficies probables de contaminación mediante modelo geoestadístico y machine learning (Elaboración propia).*

Modelo	Probabilidad (%)	Riesgo de contaminación	N° de bloques	Superficie total (ha)	Superficie total (%)
Geoestadística	0-30	Bajo	66075	26.43	79.4
	30-60	No se sabe	8540	3.41	10.3
	60-100	Alto	8585	3.43	10.3
Machine Learning	0-30	Bajo	61719	24.69	74.18
	30-60	No se sabe	4643	1.86	5.58
	60-100	Alto	16838	6.74	20.24

De la tabla anterior, se concluye que las redes neuronales reducen las zonas de incertidumbre a la mitad de lo que estima el kriging. Además, de estimar el doble de superficies con alto riesgo de contaminación, alcanzando un 21% del área de estudio de estar contaminada. La diferencia ocurre

en la zona sur del área de estudio, donde el kriging estima algunas zonas puntuales con alta probabilidad de contaminación, mientras que las redes neuronales estiman que la contaminación por hidrocarburo se acumuló en los costados del área Sur, dejando libre de contaminación las vías del ferrocarril.

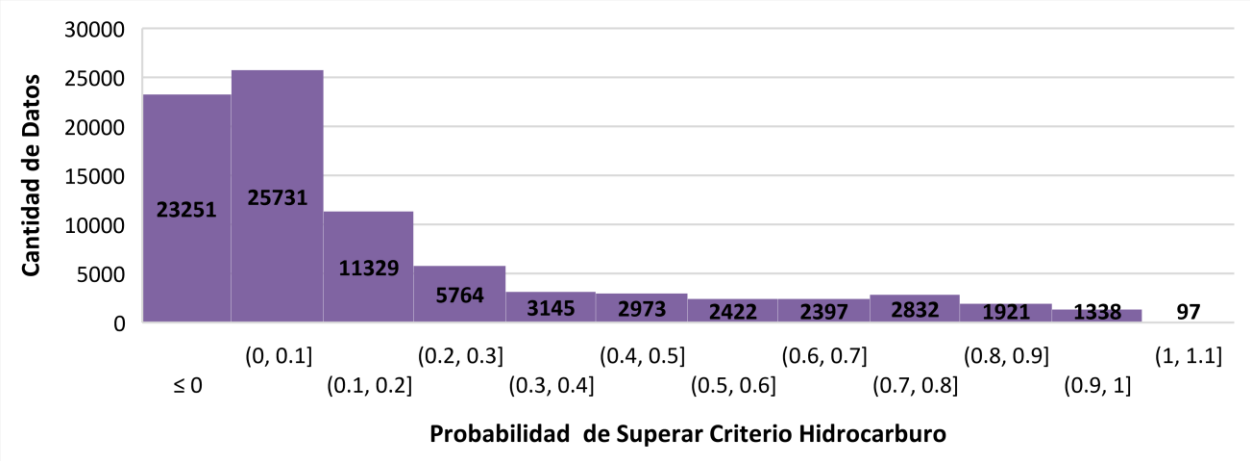

Ilustración 93: *Histograma de estimación geoestadístico, probabilidad de superar criterio HC (Elaboración propia).*

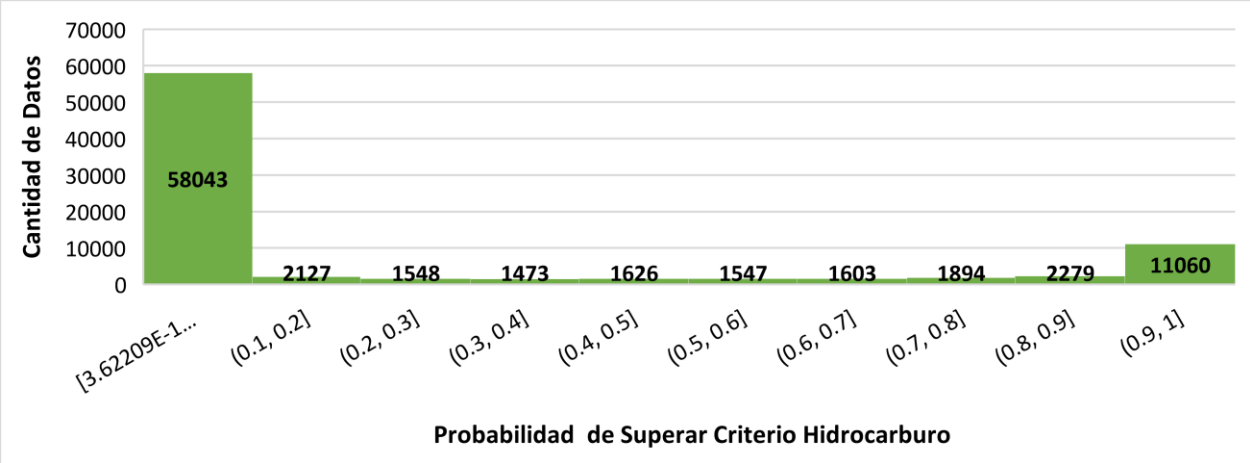

Ilustración 94: *Histograma de estimación machine learning, probabilidad de superar criterio HC (Elaboración propia).*

Por último, se presentan resultados de evaluación de modelos mediante matriz de clasificación en las siguientes tablas:

Tabla 47: *Evaluación modelo geoestadístico vs machine learning con variables categóricas (Elaboración propia).*

Modelo	Exactitud	Precisión	Exhaustividad
Geoestadística	0.717	0.711	0.717
Machine Learning	0.817	0.774	0.735

Tabla 48: *Matriz de confusión modelo geoestadístico vs machine learning con variables categóricas (Elaboración propia).*

		Predicción				
		Geoestadística		**Machine Learning**		
Real		**0**	**1**	**0**	**1**	**Total**
	0	22	5	24	3	27
	1	6	6	4	8	12
	Total	28	11	28	11	39

Para evaluar el modelo, se utilizaron 39 muestras, donde 27 de estas corresponden a datos clasificados como no contaminados, mientras que, 12 datos clasificados como contaminados.

Los resultados de evaluación demuestran que ambos modelos estiman el 72% de los datos como no contaminados, mientras que el 28% restante como contaminados. Sin embargo, en estimación de datos que se conoce su valor, machine learning acierta en más casos que modelo geoestadístico, siendo sus porcentajes de acierto de 82% y 72% respectivamente. En otras palabras, para este tipo de casos las redes neuronales tienen mejores resultados que estimación geoestadística.

Conclusión

En primer lugar, con la realización de esta memoria, se logró identificar que en Chile existe un vacío legal en materia medio ambiental, específicamente en normativas o leyes que protejan, regulen y caractericen suelos chilenos. Esto es sumamente importante, dado que actualmente, no hay una base que permita definir cuando se está en presencia de elementos que no forman parte de las condiciones normales de un suelo, ni parámetros que determinen cuando se puede hablar de contaminación, pese a que se tienen las herramientas técnicas para poder crearlas. En consecuencia, sin una política clara, no se tienen noción acerca que es lo correcto y que se debe hacer en los suelos, más aún en aquellos que se produce la pérdida natural de su calidad debido a la presencia de sustancias tóxicas, provocando efectos negativos en la naturaleza y el hombre. Por ello, debe ser prioridad regular y fiscalizar sitios con potencial presencia de contaminantes (SPPC) por medio de una ley que tenga una mirada global, considerando aspectos como erosión de suelos, desertificación que enfrentan algunas zonas por cambio climático y contaminación causada por distintos motivos. Con esto, se podrían establecer límites máximos permisibles de concentración de elementos químicos como: arsénico, boro, mercurio, plomo y otros elementos en función del riesgo. También, agregar un plan geoquímico de suelos a nivel nacional para establecer como varían las concentraciones de ciertos elementos en distintas zonas, analizar que aporta la naturaleza y que es por actividad humana, para así obtener una imagen de la composición actual del suelo para desarrollar un plan de monitoreo sistematizado, pero hasta ahora, se tiene escasa información sobre procedimientos a aplicar en SPPC. De hecho, se cuenta con la Guía Metodológica para la Gestión de Suelos con Potencial Presencia de Contaminantes del Ministerio del Medio Ambiente, que expone procedimientos para levantar información y evaluar sitios con SPPC en base al nivel de riesgo. Sin embargo, no existe una metodología de base que permita definir zonas contaminadas. Por lo tanto, el trabajo hecho en esta memoria se podría aplicar a futuros trabajos de identificación y remediación de suelos contaminados. Aunque, volviendo a recalcar lo anterior, sin una ley general de suelos es difícil orientarse en cómo actuar y cuáles son los objetivos que se deben cumplir. Por lo tanto, evidentemente se requiere una norma, que cuente con la participación de universidades, organizaciones civiles, empresas privadas y especialistas, basándose no solo en las referencias internacionales, sino que, con información de la realidad chilena en vista que se tiene una diversidad de suelos a lo largo del país.

Con respecto a las técnicas de estimación aplicadas en esta memoria, no solo las redes neuronales son desconocidas en temas ambientales, la geoestadística recientemente comienza a relacionarse con el área ambiental; gestión de plagas y aplicación de suelos contaminados. De este último, se demostró que existen una serie de técnicas que se pueden implementar de acuerdo con el caso que se desea resolver. Por ejemplo, al estimar mediante kriging ordinario variables continuas, el modelo geoestadístico en primera instancia no obtiene resultados adecuados producto de la alta variabilidad que presentan los datos. En cambio, las redes neuronales cuando no tiene como información de entrada las distancias entre las muestras, es capaz de intuir hacía que sector fluye la contaminación y luego mejora notablemente cuando conoce la distancia que hay entre las muestras. Por otra parte, el uso de bloques pequeños como soporte de estimación, permite tener una mejor visibilidad de la variabilidad de la contaminación. Sin embargo, al pasar a trabajos de descontaminación, se sugiere escoger soportes adaptados a la maquinaria que se pretende utilizar.

Otro aspecto importante a destacar, es que la técnica del kriging ordinario provoca un suavizado de las concentraciones, haciendo desaparecer valores extremos y llevándolos hacia la media de la distribución. También, por medio de la varianza del kriging, permite identificar las zonas donde la interpolación es menos precisa, en función del muestreo realizado y del modelado de variograma, pero entrega solamente un resultado de estimación y no proporciona una noción sobre la incertidumbre de la estimación, aunque esto cambia si se trabaja con variables categóricas. En cambio, la simulación genera más predicciones factibles, varianzas y se puede conocer la incertidumbre asociada a la estimación.

Los resultados de estimación por kriging cuando trabaja con variables continuas, mostraron una superficie contaminada de 2.4 hectáreas equivalente al 7% del área de estudio. En el caso de las redes neuronales, se obtuvo una superficie de contaminación de 8.5 hectáreas, lo que equivale al 26% del emplazamiento en estudio. Aunque, es importante mencionar que para estimar por kriging, previamente se debió eliminar las muestras consideradas atípicas, por lo tanto, dicha estimación está siendo subestimada, pero cuando se comparan ambos modelos estimando sin las muestras atípicas y con el 80% de los datos, las diferencias no son tan grandes en cuanto a superficie de contaminación, teniendo resultados de 7.3% para kriging y 10.4% para redes neuronales.

Respecto a estimaciones que se realizan cuando se trabajan con variables categóricas, se pueden obtener distintos niveles de contaminación según el riesgo que se esté dispuesto asumir. Además,

de tener certeza de cuáles son las zonas de incertidumbre, para posteriormente poder realizar futuros trabajos de muestreo o incorporar a suelos contaminados según el criterio del especialista. Los resultados por kriging de indicadores mostraron un 10% de obtener zonas con incertidumbre, simulaciones un 62% y redes neuronales entre un 3 y 7% dependiendo de los distintos casos expuestos. La diferencia de las zonas de incertidumbre de las simulaciones con los otros métodos, es porque dicha técnica aumenta la varianza debido a que puede estimar valores fuera de los rangos iniciales.

En el caso de la evaluación de los modelos, las redes neuronales fueron capaces de estimar mucho mejor los datos reales, teniendo una exactitud del 81% y el kriging del 71%, lo que demuestra que las redes neuronales si son capaces de estimar suelos contaminados y pueden seguir mejorando si tuvieran más información.

Otro aspecto importante de mencionar, es que independiente del método que se elija, la toma de datos debe ser de calidad para obtener resultados precisos. Esto se confirma con el caso expuesto "estimación con grilla" , donde se estimó utilizando 65 muestras y los resultados mostraron que se obtenían estimaciones muy parecidas cuando se trabaja con la totalidad de los datos. En consecuencia, la toma de muestras es una etapa crucial para este tipo de proyectos, porque son la base de los modelos geoestadísticos o machine learning de contaminación de suelos. Por lo tanto, esta es una etapa que podría denominarse primera fuente de incertidumbre de un proyecto, por lo que se sugiere que a la hora del muestreo no se permita a las cuadrillas salirse de la planificación previa.

Sobre la decisión de implementar una técnica, dependerá del conocimiento del campo de aplicación y de las ventajas que tenga una técnica frente a la otra. Por ejemplo, los métodos deterministas son de ejecución rápida, porque son fáciles de aplicar. Por ende, se necesita poca inversión, ya que cualquier software geoespacial los incluye y sin importar la persona que realice el estudio, se presentan los mismos resultados. Sin embargo, sus desventajas son que solo admite tratamiento 2D, no se puede verificar la calidad de la estimación realizada, ya que no considera la estructura del fenómeno en estudio, es decir, no cuantifica el error o la incertidumbre asociada al proceso y por lo general, presentan sesgo condicional, el cual se traduce en sobreestimación o subestimación de los valores.

En el caso de la geoestadística, se pueden realizar estimaciones tridimensionales, integra variables auxiliares para realizar estudios multivariantes, permite analizar errores asociados al cálculo y determinar zonas de incertidumbre, pero para su aplicación, se requiere de ciertas nociones técnicas para la compresión del método y de softwares específicos que hay que saber manejar, por lo tanto, su tiempo de ejecución es mucho mayor. Además, en casos donde la contaminación presenta comportamiento errático, una aplicación directa del método puede resultar compleja.

Finalmente, las ventajas que poseen las redes neuronales respecto al método geoestadístico es que no requiere de un modelo variográfico para entender el comportamiento espacial de las variables. Si bien es cierto, que el procedimiento de estimación por kriging puede ser más rápido, dado que las redes neuronales o cualquier técnica machine learning necesita la construcción de modelos iterativos para encontrar el modelo adecuado, es que, si se desea agregar más atributos o información a la red, el procedimiento es simple, dado que significa otra información de entrada al modelo. En cambio, el kriging puede requerir múltiples modelos variográfico.

Para finalizar, una de las desventajas de las redes neuronales es que, a mayor cantidad de atributos o neuronas, los tiempos de entrenamiento irán aumentando, por ende, se requiere de mayor tiempo para obtener el resultado esperado.

Referencias bibliográficas

Aduvire, O. (2006). *Drenaje ácido de mina : generación y tratamiento.* Madrid: Instituto Geológico y Minero de España.

Alcaino Concha, G. I. (2012). *Análisis y comparación de tecnologías de remediación para suelos contaminados con metales.* Santiago: Universidad de Chile.

Alfaro Sironvalle, M. A. (2007). *Estimación de recursos mineros.* Santiago, Chile: Universidad de Chile.

Alonso, M. J. (2019). *Introducción al machine learning.* Fundación Telefónica: BigMl.

Alpaydin, E. (2014). *Introduction to machine learning, second edition.* Massachusetts: Massachusetts Institute of Technology.

Alva Santos, A. (S/F). *Análisis de los datos e interpretación de los resultados.*

Barberena, M., & Gnoza, N. (2018). *Estudio de factibilidad del uso de machine learning con múltiples fuentes de datos en el pronóstico del tiempo.* Universidad ORT Uruguay.

Belmonte Quispe, S., & Tito Cruz, E. L. (2012). *Determinación de los contaminantes de origen minero y alternativa de tratamiento del río Sayaquira municipio Ichoca provincia Inquisivi La Paz.* La Paz: Universidad Mayor de San Andrés.

Betrie, G., Sadiq, R., Morin, K., & Tesfamariam, S. (2014). *Uncertainty quantification and integration of machine learning techniques for predicting acid rock drainage chemistry: A probability bounds approach.* Science of the Total Environment.

Canadian Council of Ministers of the Environment,. (2007). *Canadian Soil Quality Guidelines for the Protection of Environmental and Human Health.*

Carmona Suárez, E. (2013). *Tutorial sobre Máquinas de Vectores de Soporte (SVM).* Madrid: Universidad Nacional de Educación a Distancia.

Castaño Agudelo, A. F., & Vergara Elorza, F. (2004). *Simulación geoestadística aplicada al modelamiento de yacimientos de petróleo.* Medellín: Univerisdad Nacional de Colombia.

Castro Altamirano, F. J. (2014). *Opciones reales aplicadas a un problema de secuenciamiento minero.* Santiago: Universidad de Chile.

Cely Pulido, J. W., Siabato Vaca, W. L., Sánchez Ipia, A. H., & Rangel Sotter, A. P. (2002). *Geoestadística aplicada a estudios de contaminación ambiental.* Universidad Distrital Francisco José de Caldas.

Chaparro Leal, L. T. (2015). Drenajes Ácidos de Mina: Formación y Manejo. *Revista ESAICA*, 53-57.

Cimarra Muñoz, D. (2018). *Experimentos de predicción con gradient boosting y random forest.* Madrid: Universidad Politécnica de Madrid.

Comesaña García, Y., Dago-Morales, Á., Talavera Bustamante, I., Núñez Cuadra, O., & Hernández González, N. (2010). Modelo de regresión de máquinas de vectores de soporte de mínimos cuadrados para la predicción de la cristalinidad de catalizadores de craqueo por espectroscopia infraroja. *Revista CENIC Ciencias Biológicas*, 43-48.

Consejo de Minería de la Columbia Británica. (s/f). *Drenaje Ácido de la Minería : Minería y contaminación de agua en la Columbia Británica, Canadá.*

Diaz, N. C. (2006). *Técnicas de muestreo. Sesgos más frecuentes.*

Dupouy Berrios, C. (2014). *Aplicación de árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa de interés en Chile.* Santiago: Universidad de Chile.

Earthworks . (s/f). Drenaje Ácido Generado por la Minería. *Niparajá*, 1-2.

Emery, X. (2007). *Apunte de geoestadística* . Santiago, Chile: Universidad de Chile.

Emery, X. (2008). *Simulación estocástica y geoestadística no lineal.* Santiago, Chile: Universidad de Chile.

Emery, X. (S/F). *Simulación Geoestadística.* Santiago, Chile: Universidad de Chile.

Frez Ríos, T. (2014). *Kriging y simulación secuencial de indicadores con proporciones localmente variables.* Santiago: Universidad de Chile.

García Cárdenas, S. A. (2013). *Modelación del potencial de drenaje ácido de botaderos: calibración a partir de celdas húmedas y granulometría.* Santiago: Universidad de Chile.

Giraldo Henao, R. (S/F). *Introducción a la geoestadística.* Bogotá: Universidad Nacional de Colombia.

Gobierno de México. (2012). *NORMA Oficial Mexicana NOM-138-SEMARNAT/SSA1-2012,.* Diario Oficial de la Federación.

Godoy, S. (2019). *Curso de geoestadística aplicada a la contaminación de suelos.* Ingeoexpert.

Guerrero, M. (2016). *Contaminación del suelo en la zona minera de Rasgatá Bajo.* Ciencia e ingeniería neogranadina.

(2009). *Guía Global de Drenaje Ácido de Roca (GARD).* (INAP), Red Internacional para la Prevencion de Ácido.

Haro Bustamante, V. F. (2007). *Legislación de suelos y su protección ambiental.* Punta Arenas: Universidad de Chile.

Jepson, W. (2002). Un Caso de Estudio: Mina Zortman-Landusky. *Helena Independt Record*, págs. 1-2.

Journel, A., & Huijbregts, J. (1978). *Mining Geostatistics.* New York: Academics Press.

Life-Etad. (2012). *Proyecto Life-Etad : Ecological Treatment of Acid Drainage.*

Llona, M. (25 de Marzo de 2019). Expertos buscan reactivar discusión para una norma de suelo, la mayor deuda en materia de contaminación en Chile. *País Circular*.

Mallea Álvarez, M. I. (s/f). *Remediación de suelos contaminados y análisis de un proyecto piloto en Chile, en el marco del sistema de evaluación de impacto ambiental.* Santiago: Unidad de Medio Ambiente Consejo de Defensa del Estado.

Matich, D. J. (2001). *Redes neuronales: conceptos básicos y aplicaciones.* Universidad Tecnológica Nacional.

Meier, M., De Souza, E., Rocha Francelino, M., Fernandes Filho, E. I., & Goncalves Reynaud Schaefer, C. E. (2018). Digital soil mapping using machine learning algorithms in a tropical mountainous area. *Revista Brasileria de Ciência Do Solo*, 1-22.

Ministerio de Salud Chile. (2004). *Reglamento sanitario sobre manejo de residuos peligrosos.* Gobierno de Chile.

Moral García, F. J. (2004). Aplicación de la geoestadística en las ciencias ambientales. *Revista científica y técnica de ecología y medio ambiente*, 78-86.

Okwuashi, O., & Ndehedehe, C. (2014). Digital terrain model height estimation using support vector machine regression. 1-5.

Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO). (2015). *La contaminación del suelo: una realidad oculta.*

Pedraza Camelo, J. C. (S/F). *Propotio de un modelo de machine learning para la predicción de partículas de contaminación atmósferica finas, en la localidad de Kennedy en Bogotá*. Universidad Distrital Francisco José De Caldas.

Pereira, T., Robaina, A., Peiter, M., Torres, R., & Bruning, J. (2018). The use of artificial intelligence for estimating soil resistance to penetration. *Engenharia Agrícola, Jaboticabal*, 142-148.

Pérez, N., Schwarz, A., & Urrutia, H. (2017). Tratamiento del drenaje ácido de minas: estudio de reducción de sulfato en mezclas orgánicas. *Tecnología y Ciencias del Agua, vol. VIII*, 53-64.

Pérez, R. (2008). *Tratamientos de drenaje ácido de minas División el Teniente - CODELCO Chile.* Universidad Católica de Valparaiso.

Petitgas, P. (1996). *Geostatistics and their applications to fisheries survey data.* Londres: Computers and fisheries.

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahab, H., . . . Tien Bui, D. (2019). Predicting uncertainty of machine learning models for modelling nitatre pollution of groundwater using quantile regression and UNEEC methods. *Science of the Total Environment*, 855-866.

Rath, C. (1999). *Estimación de la superficie y número de árboles en bosques de pinus radiata D. Don en fotografías aéreas, mediante el uso de redes neuronales artificiales.* Santiago: Universidad de Chile.

Rodríguez Eugenio, N., McLaughlin, M., & Pennock, D. (2019). *La contaminación del suelo: una realidad oculta.* Roma: FAO.

Rodríguez-Sahagún Alesanco, P. (2018). *Aplicación de redes neuronales convulucionales y recurrentes al diagnóstico de autismo a partir de resonancias magnéticas funcionales.* Madrid: Universidad Politécnica de Madrid.

Sánchez, J. (2014). *Análisis de técnicas machine learning para la estimación de medidas corporales.* Universitat Autónoma de Barcelona.

SEA. (4 de Marzo de 2020). *Servicio de Evaluación Ambiental*. Obtenido de Servicio de Evaluación Ambiental: www.sea.gob.cl/documentacion/permisos-autorizaciones-ambientales/normativa-ambiental-aplicable

Seijas Fossi, C., & Caralli D'Ambrosio, A. (2004). Uso de las máquinas de vectores de soporte para la estimación del potencial de acción celular. *Revista ingeniería UC*, 56-61.

SERNAGEOMIN. (2012). *Guía metodológica para la gestión de suelos con potencial presencia de contaminantes.* Fundación Chile.

SERNAGEOMIN. (2015). *Guía Metodológica para la Estabilidad Química de Faenas e Instalaciones Mineras.* Fundación Chile.

Solutions, M. (2018). *Machine learning, una pieza clave en la transformación de los modelos de negocio.* Management Solutions .

Sommer, I., Fernández, P., Rivas, H., & Gutiérrez, M. (200). *La geoestadística como herramienta en estudios de contaminación de suelos. Análisis de caso: Afectación por arsénico, plomo y cadmio contenidos en jales mineros.* México: Instituto de Geografía, UNAM.

Suárez, A., Jiménez, A., Castro-Franco, M., & Cruz-Roa, A. (2017). *Classification and automatic mapping of land covers in satellite images using convulutional neural networks.* Universidad de los Llanos.

Urquidi, M. (s/f). Drenaje ácido generado por la minería. *Niparajá*, 1-2.

Valenzuela Dupre, A. D. (2012). *Construcción de modelo geoestadístico para la generación y complementación de información hidrogeológica.* Santiago, Chile: Universidad de Chile.

Vergara Bustos, D. R. (2013). *Estimación multivariable y sesgo condicional.* Santiago, Chile: Universidad de Chile.

Villalba Mata, D. (2000). *Construcción y utilización de un modelo estocástico para la simulación de estrategias de manejo invernal en rebaños de vacas nodrizas.* LLeida: Universitat de Lleida.

Villarroel, L., Miller, J., Lechler, P., Germanoski, D., & Puch, E. (2007). Contaminación por metales pesados del sistema de drenaje Río Chilco - Río Tupiza, Sur de Bolivia. *Ecología en Bolivia*, 48-71.

Weeberb, Requia, Coull, B., & Koutrakis, P. (2019). Evaluation of predictive capabilities of ordinary geostatistical interpolation, hybrid interpolation, and machine learning methods for estimating PM 2.5 constituents over space. *Environmental Research*, 421-433.

Zamora, D. (2013). *Métodos machine learning aplicados para estimar la concentración de los contaminantes de la DQO y de los SST en hidrosistemas de saneamiento urbano a partir de espectrometría UV-Visible.* Bogotá: Pontificía Universidad Javeriana.

Zevallos Santivañez, J. F. (2016). *Estabilización del drenaje ácido de mina (DAM) de la empresa Paraíso Pérdido Apata.* Huancayo: Universidad Nacional del Centro del Perú.

Apéndice A: Análisis exploratorio de datos inicial

En este apéndice, se presentan los resultados del análisis exploratorio de la variable espesor con la base de datos inicial. Además, resultados de base de datos transformada al aplicar método de kriging de indicadores y simulaciones condicionales gaussianas.

Caso base de datos inicial

La tabla A.1 y la ilustración A.1 corresponde al EDA aplicado para la variable espesor.

Tabla A. 1: *Estadísticas descriptivas espesor (Elaboración propia).*

Espesor (m)	
Cantidad de datos	192
Mínimo	0.05
Máximo	1.8
Media	0.66
Mediana	0.6
Varianza	0.2
Error estándar	0.03
Desviación estándar	0.4
Q1	0.3
Q3	1
Skewness	0.7
Curtosis	-0.2
Coeficiente de variación	59.8

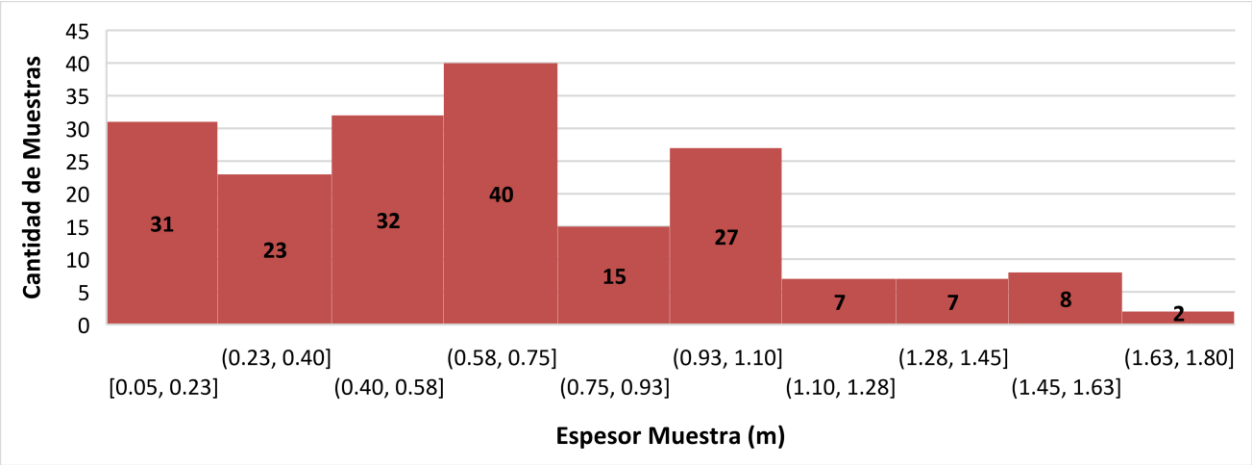

Ilustración A. 1: *Histograma espesor de la muestra (Elaboración propia).*

Caso kriging de indicadores

La tabla A.2 y las ilustraciones A.2, A.3 corresponden al EDA aplicado para la base de datos transformada por kriging de indicadores. Se observa que al aplicar este método disminuye significativamente la variabilidad de los datos, dado que son sustituidos por 0 y 1.

Tabla A. 2: *Estadísticas descriptivas elementos contaminantes caso kriging de indicadores (Elaboración propia).*

	Hidrocarburos (ppm)	Níquel (ppm)
Cantidad de datos	192	192
Mínimo	0	0
Máximo	1	1
Media	0.3	0.09
Mediana	0	0
Varianza	0.21	0.08
Error estándar	0.03	0.02
Desviación estándar	0.5	0.3
Q1	0	0
Q3	1	0
Skewness	0.9	2.9
Curtosis	-1.3	6.6
Coeficiente de variación	152.4	321.7

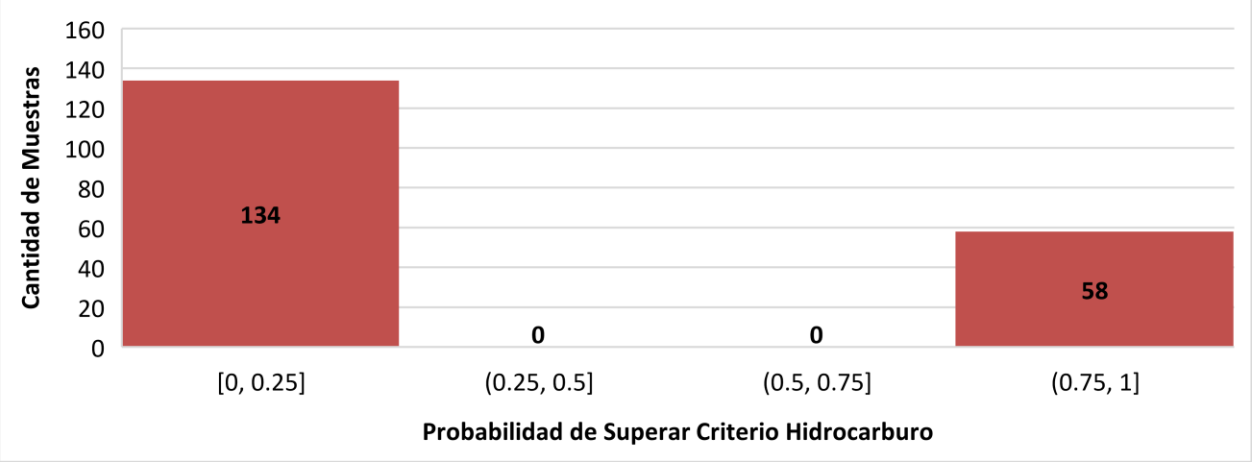

Ilustración A. 2: *Histograma HC caso kriging de indicadores (Elaboración propia).*

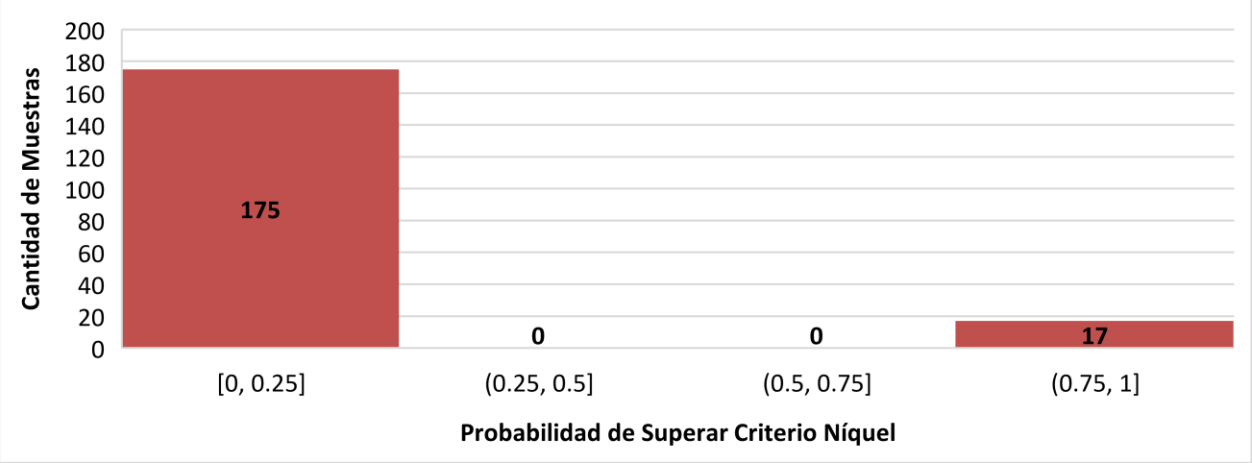

Ilustración A. 3: *Histograma Ni caso kriging de indicadores (Elaboración propia).*

Caso simulaciones condicionales gaussianas

La tabla A.3 y la ilustración A.4 corresponde al EDA aplicado para la base de datos transformada por simulaciones condicionales gaussianas. Aquí disminuyó la varianza y los datos se concentran en intervalos medios de la distribución.

Tabla A. 3: *Estadísticas descriptivas HC caso simulación condicional gaussiana (Elaboración propia).*

	Hidrocarburos (ppm)
Cantidad de datos	192
Mínimo	-4160.7
Máximo	11845.4
Media	3695.6
Mediana	3761.2
Varianza	1.01E+07
Error estándar	229.7
Desviación estándar	3182.5
Q1	1150.4
Q3	59774.1
Skewness	-0.006
Curtosis	-0.4
Coeficiente de variación	86.1

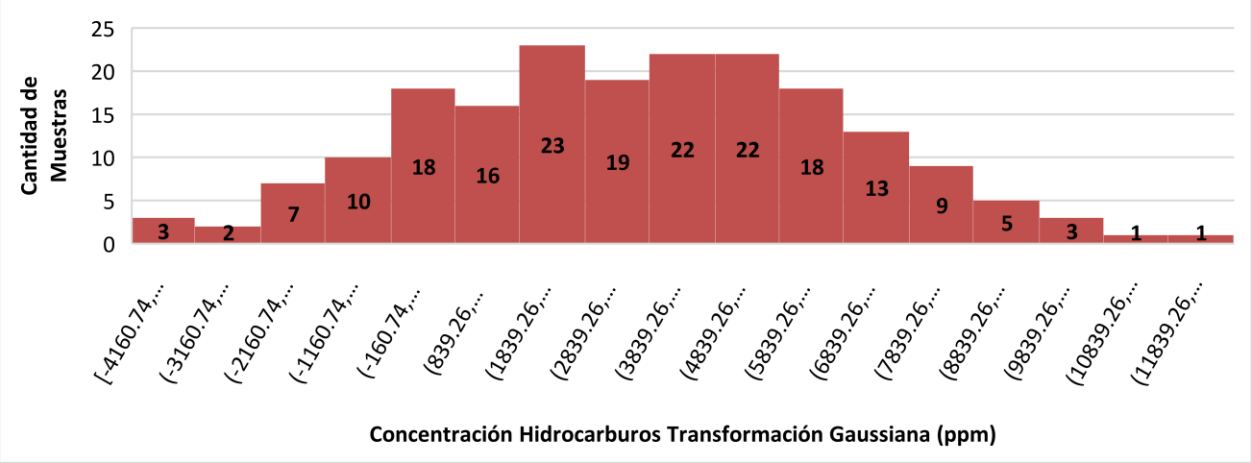

Ilustración A. 4: *Histograma HC caso simulaciones condicionales gaussianas (Elaboración propia).*

Caso sin datos atípicos

La tabla A.4 y la ilustración A.5 corresponde al EDA aplicado para la base de datos transformada caso sin datos atípicos, donde se tiene un total de 152 muestras con las que se realiza estimación.

Tabla A. 4: *Estadísticas descriptivas HC caso sin datos atípicos (Elaboración propia).*

	Hidrocarburos (ppm)
Cantidad de datos	152
Mínimo	100
Máximo	3500
Media	328.8
Mediana	100
Varianza	302378.7
Error estándar	44.6
Desviación estándar	549.9
Q1	100
Q3	210
Skewness	3.2
Curtosis	11.1
Coeficiente de variación	167.2

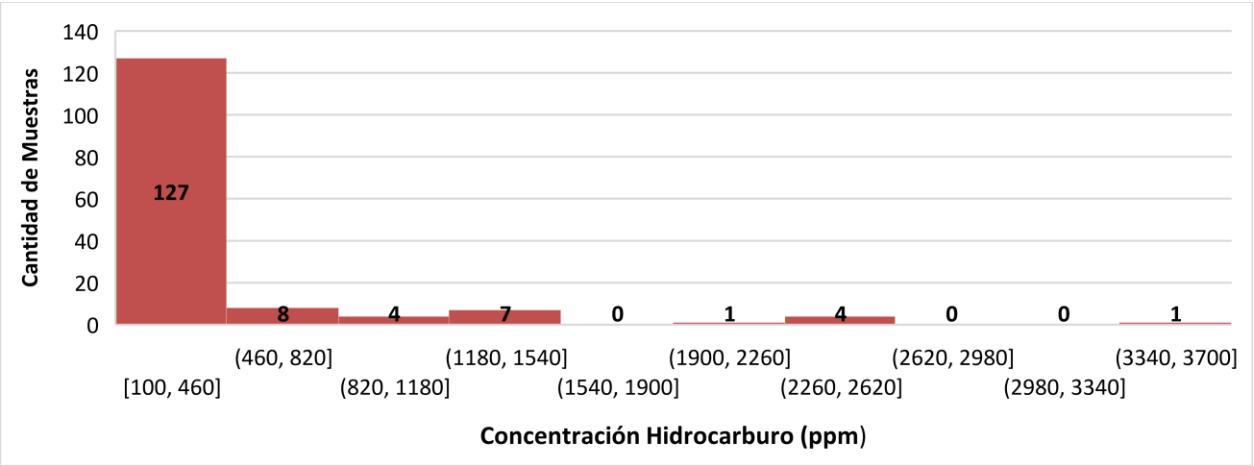

Ilustración A. 5: *Histograma HC caso sin datos atípicos (Elaboración propia).*

Caso hidrocarburo muestreado con grilla

La tabla A.5 y la ilustración A.6 corresponde al EDA aplicado para la base de datos cuando se trabaja con grilla, donde se utilizaron 65 muestras para realizar estimación.

Tabla A. 5: *Estadísticas descriptivas HC caso muestreado con grilla (Elaboración propia).*

	Hidrocarburos (ppm)
Cantidad de datos	65
Mínimo	0
Máximo	1
Media	0.3
Mediana	0
Varianza	0.2
Error estándar	0.1
Desviación estándar	0.5
Q1	0
Q3	1
Skewness	0.9
Curtosis	-1.2
Coeficiente de variación	156.8

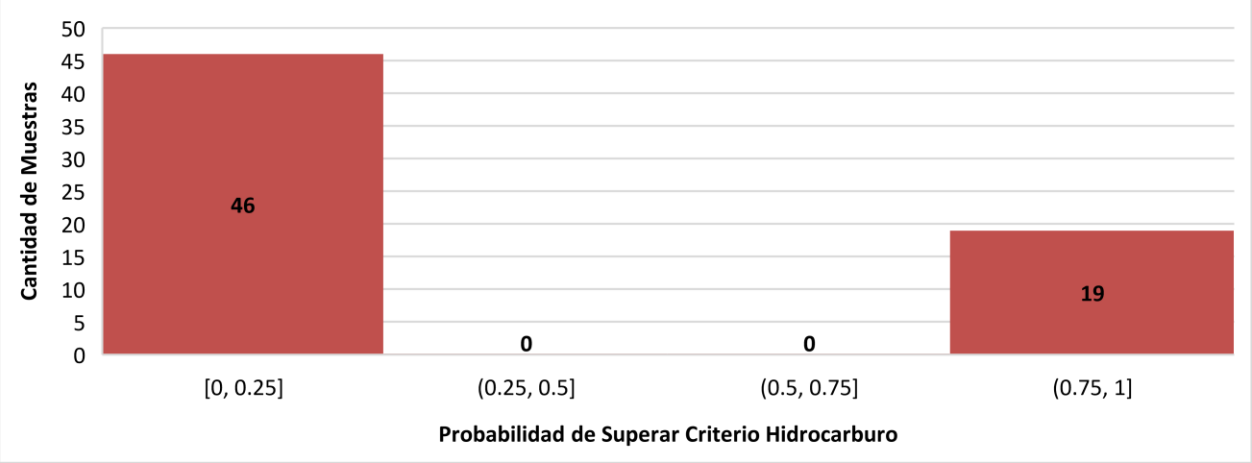

Ilustración A. 6: *Histograma HC caso sin datos atípicos (Elaboración propia).*

Apéndice B: Variogramas

En este apéndice, se aprecian los variogramas experimentales y modelados para los casos de kriging ordinario , indicadores y simulaciones condicionales gaussianas que no fueron expuesto en el cuerpo principal de la memoria.

Caso kriging ordinario

Los variogramas experimentales y modelados de las variables níquel y espesor, se observan desde la ilustración B.1 hasta la B.6

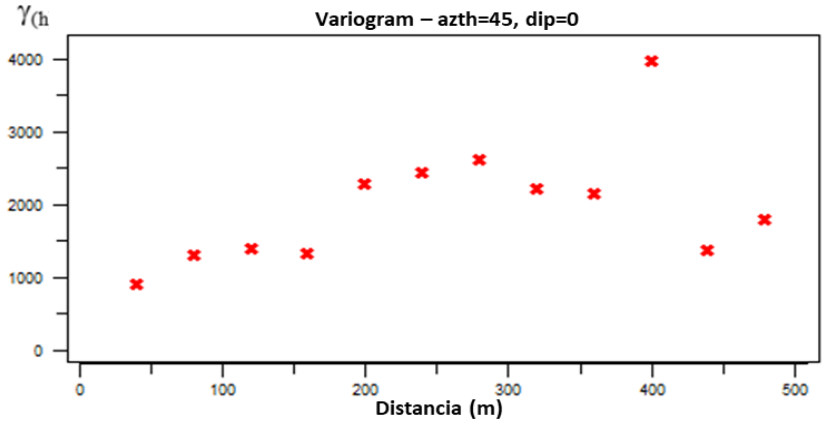

Ilustración B. 1: *Variograma experimental 45° Ni (Elaboración propia).*

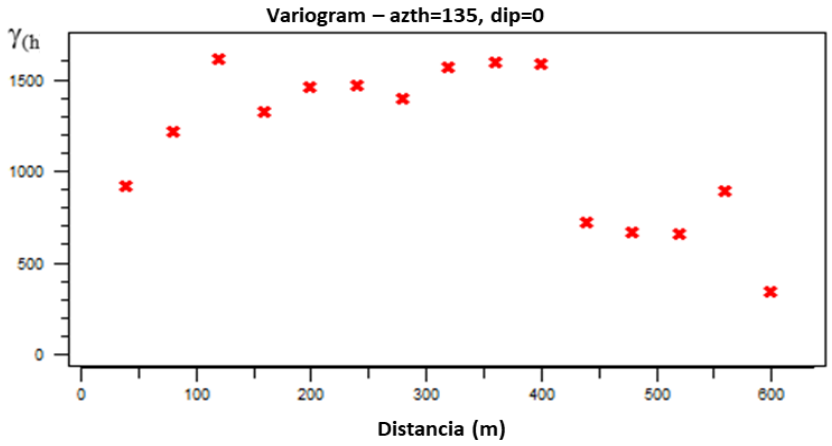

Ilustración B. 2: *Variograma experimental 135° Ni (Elaboración propia).*

El variograma experimental del níquel fue ajustado a un modelo exponencial con una meseta aproximadamente a la varianza de los datos del níquel, asignándole un efecto pepita de 0.1. Ambas direcciones presentan alcance a los 280 metros, es decir, la distancia hasta donde una muestra tiene influencia sobre otra muestra.

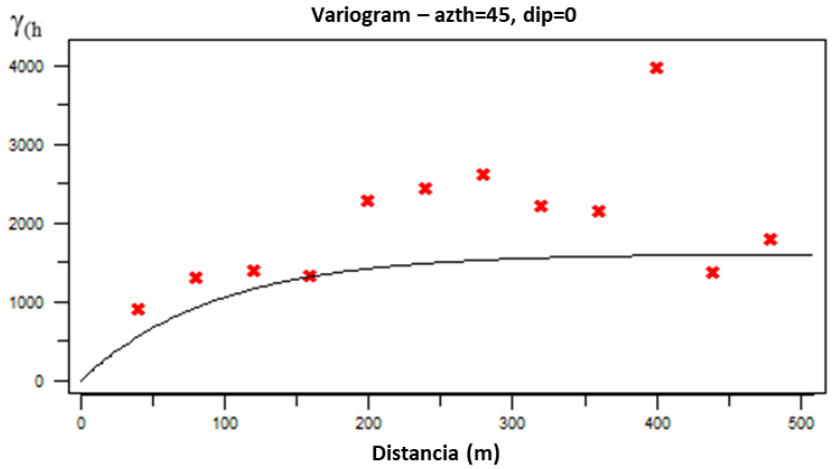

Ilustración B. 3: *Variograma modelado 45° Ni (Elaboración propia).*

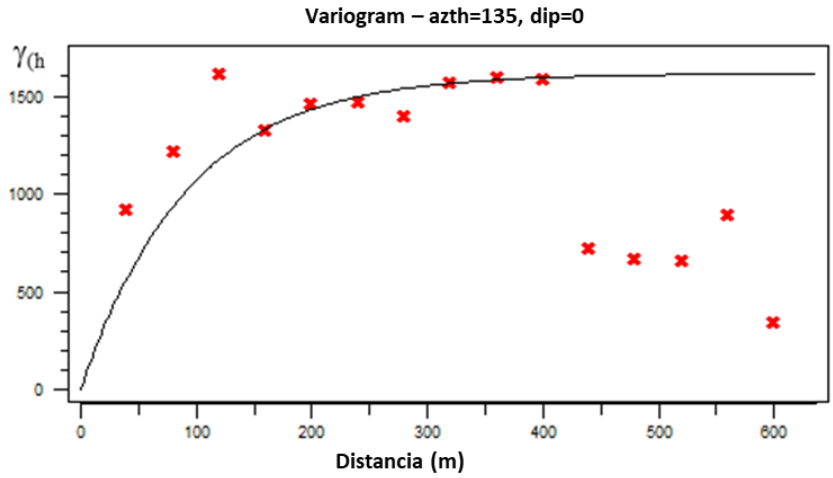

Ilustración B. 4: *Variograma modelado 135° Ni (Elaboración propia).*

Para el espesor se utiliza el variograma omnidireccional, ya que no se necesita detectar direcciones de anisotropía.

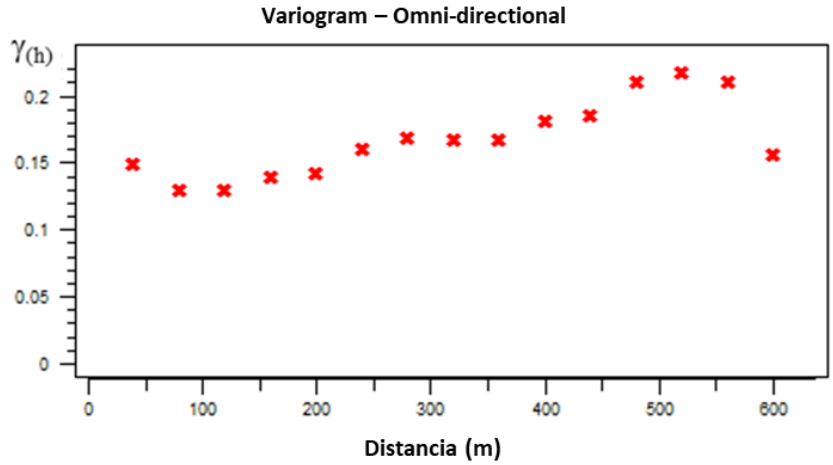

Ilustración B. 5: *Variograma experimental omnidireccional espesor (Elaboración propia).*

El variograma experimental para el caso espesor, fue ajustado a un modelo exponencial con una meseta aproximadamente la varianza de los datos del espesor, asignándole un efecto pepita de 0.02, observándose alcance a los 294 metros.

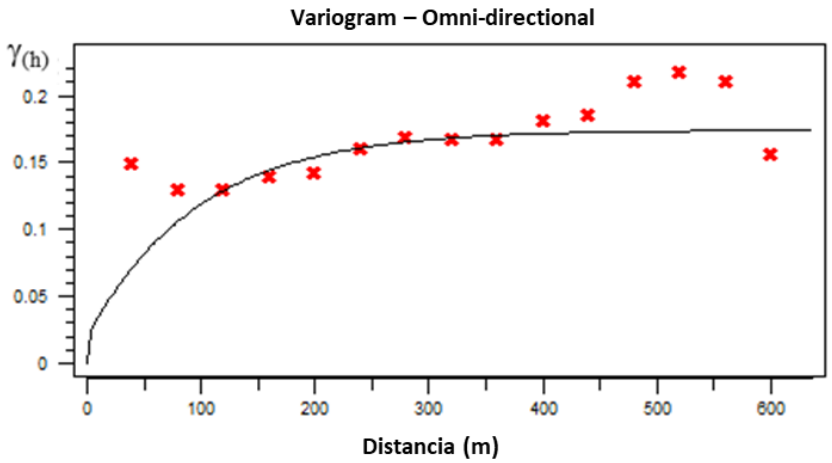

Ilustración B. 6: *Variograma modelado omnidireccional espesor (Elaboración propia).*

Caso kriging de indicadores

Los variogramas experimentales y modelados para este caso, se aprecian desde la ilustración B.7 hasta la B.14, donde ambas variables presentan como eje mayor 45° azimuth y 135° azimuth como eje menor.

Al aplicar este método, se aprecian variogramas menos sensibles a los valores extremos de la distribución.

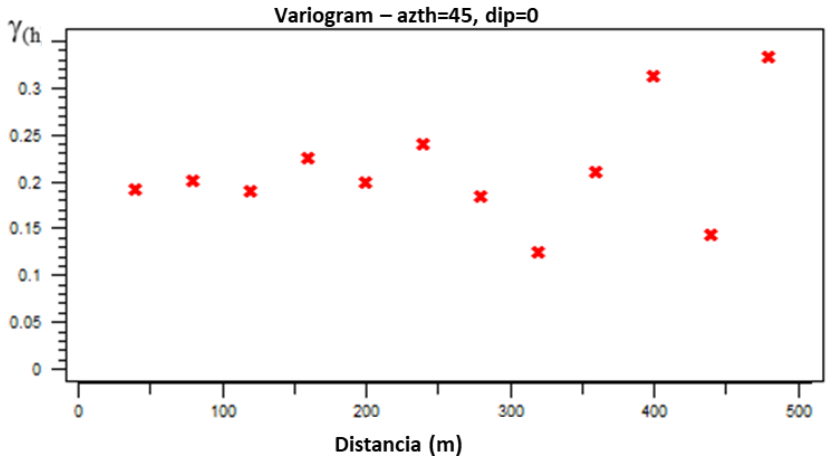

Ilustración B. 7: *Variograma experimental 45° HC caso kriging de indicadores (Elaboración propia).*

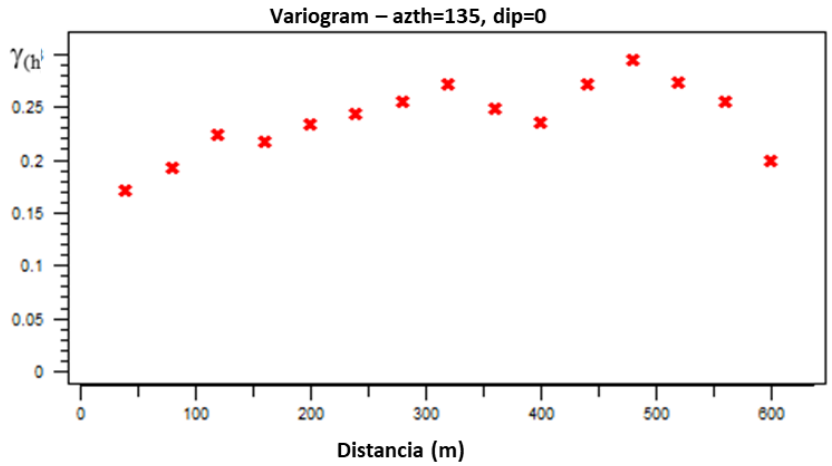

Ilustración B. 8: *Variograma experimental 135° HC caso kriging de indicadores (Elaboración propia).*

Los variogramas experimentales para la variable hidrocarburo, fueron ajustados a un modelo exponencial con una meseta aproximadamente la varianza de los datos del hidrocarburo, asignándole un efecto pepita de 0.01 y un alcance que llega a los 108 metros en ambas direcciones.

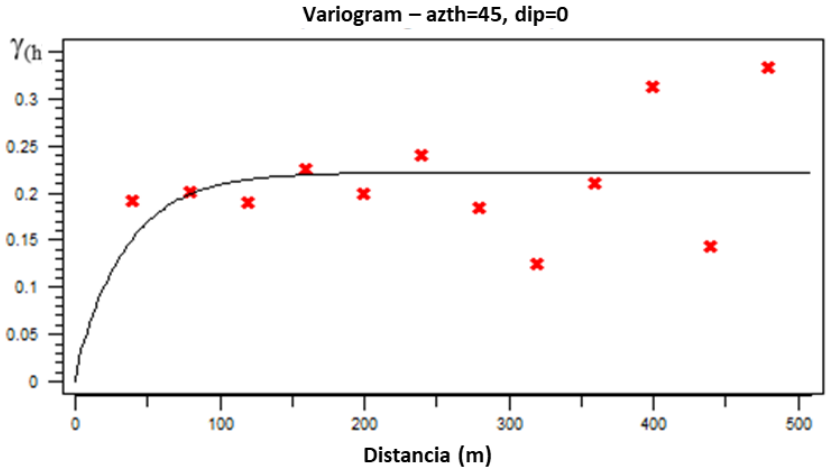

Ilustración B. 9: *Variograma modelado 45° HC caso kriging de indicadores (Elaboración propia).*

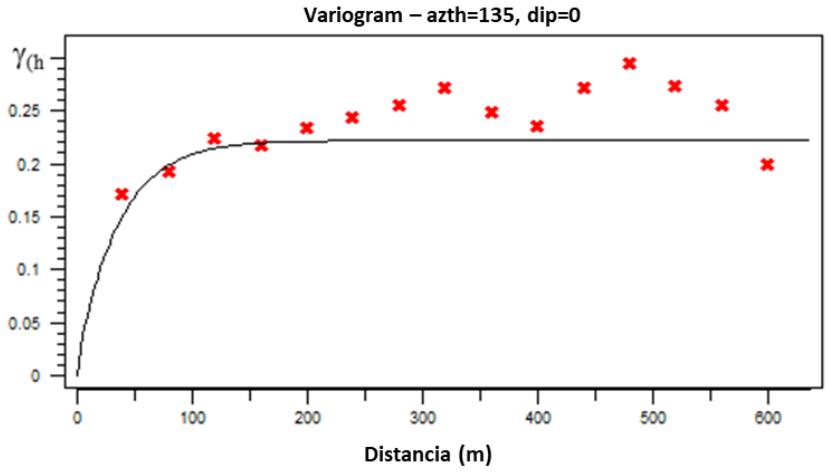

Ilustración B. 10: *Variograma modelado 135° HC caso kriging de indicadores (Elaboración propia).*

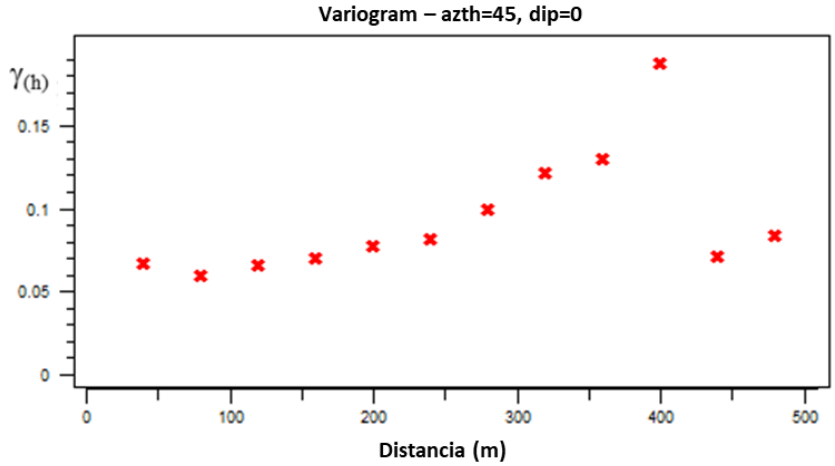

Ilustración B. 11: *Variograma experimental 45° Ni caso kriging de indicadores (Elaboración propia).*

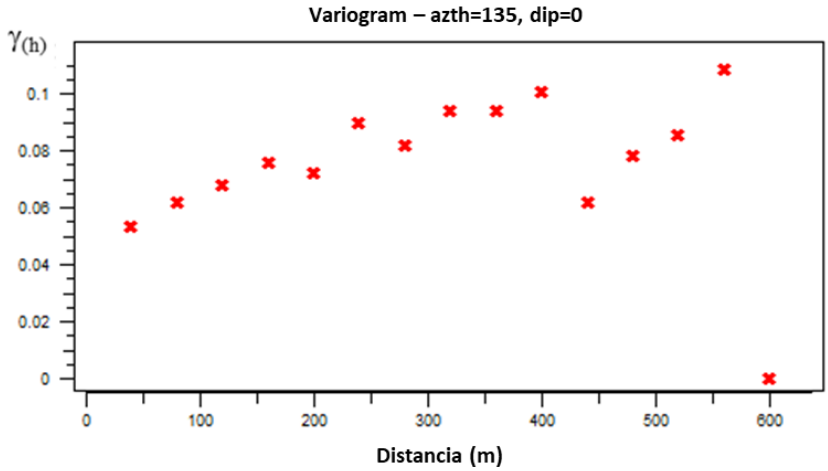

Ilustración B. 12: *Variograma experimental 135° Ni caso kriging de indicadores (Elaboración propia).*

Los variogramas experimentales para la variable níquel, fueron ajustados a un modelo exponencial con una meseta aproximadamente la varianza de los datos del níquel, asignándole un efecto pepita de 0.01 y un alcance que llega a los 198 metros en ambas direcciones.

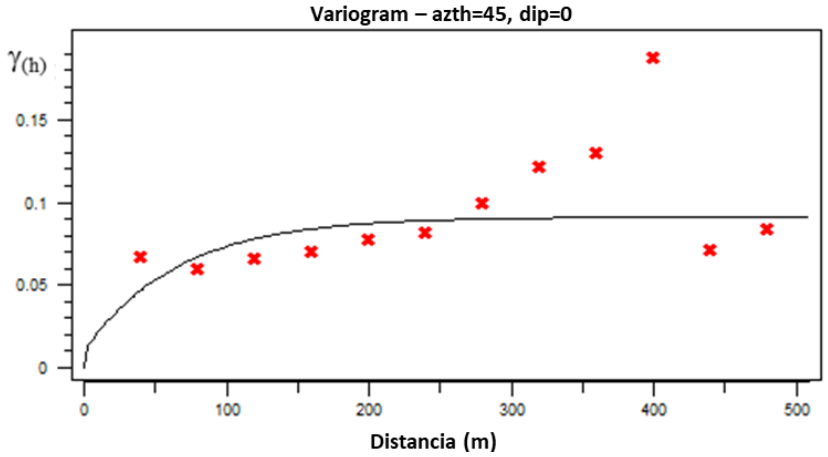

***Ilustración B. 13**: Variograma modelado 45° Ni caso kriging de indicadores (Elaboración propia).*

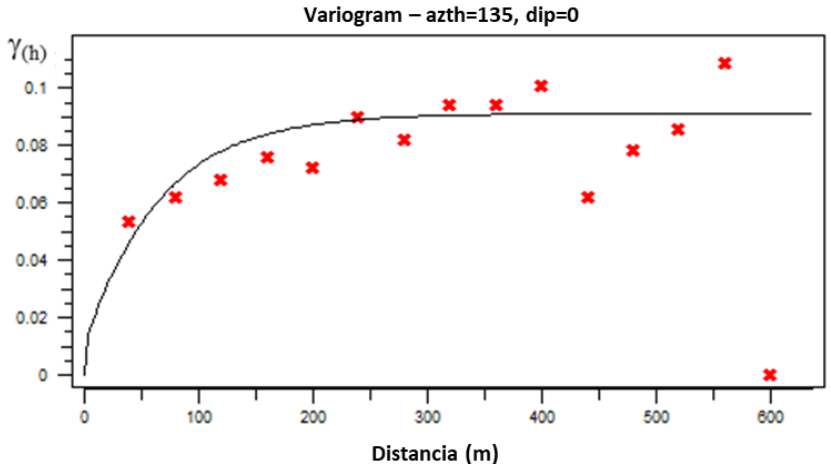

***Ilustración B. 14:** Variograma modelado 135° Ni caso kriging de indicadores (Elaboración propia).*

Caso simulaciones condicionales gaussianas

Los variogramas experimentales y modelados para este caso, se aprecian desde la ilustración B.7 hasta la B.14, donde el eje mayor corresponde a 45° azimuth, mientras que el eje menor 135° azimuth para variable hidrocarburo.

La dirección 45° azimuth presenta un comportamiento poco estructurado. Mientras que, dirección 135° azimuth se logra apreciar un comportamiento lineal en el origen de los datos.

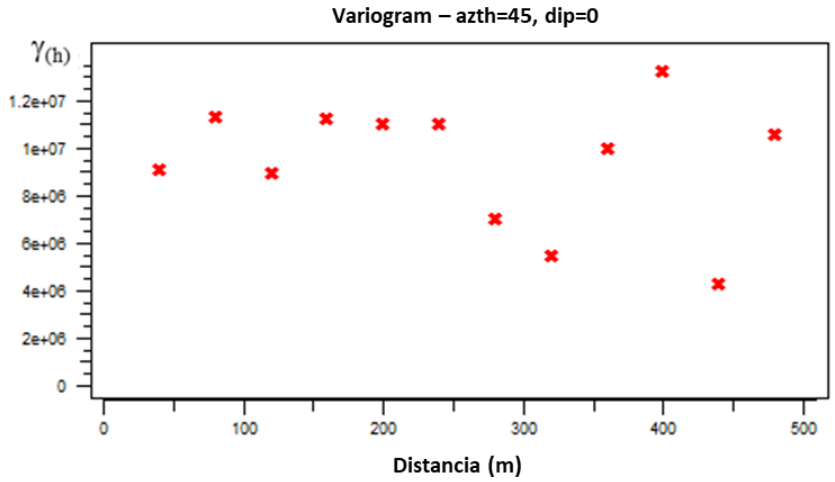

Ilustración B. 15: *Variograma experimental 45° HC caso simulaciones condicionales gaussianas (Elaboración propia).*

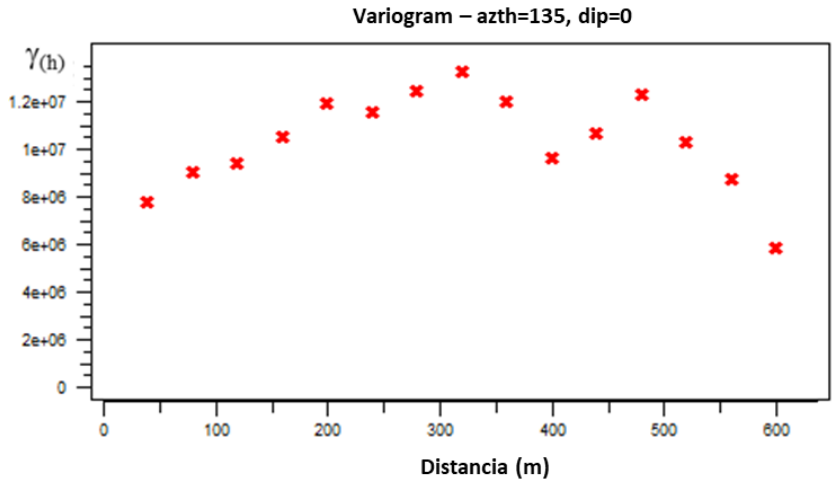

Ilustración B. 16: *Variograma experimental 135° HC caso simulaciones condicionales gaussianas (Elaboración propia).*

Los variogramas experimentales de la variable hidrocarburo con distribución gaussiana, fueron ajustados a un modelo exponencial con una meseta aproximadamente a la varianza de los datos del HC sin efecto pepa. Presentan un alcance a los 144 metros en ambas direcciones, es decir, la distancia hasta donde una muestra tiene influencia sobre otra muestra.

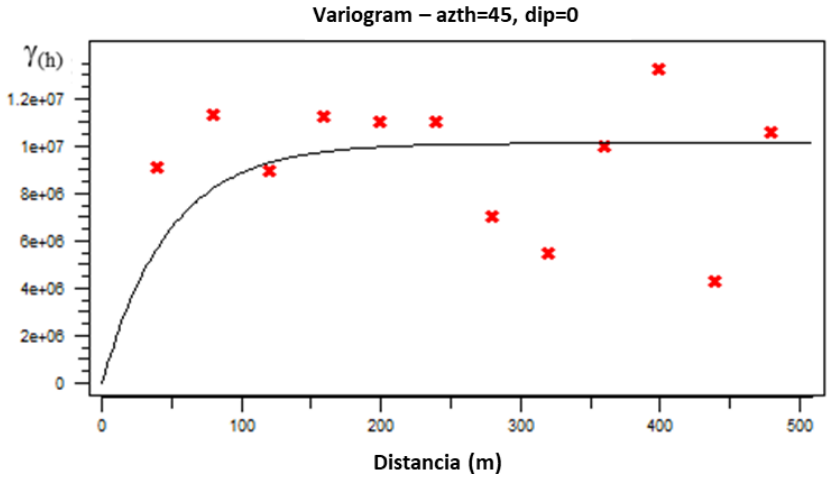

Ilustración B. 17: *Variograma modelado 45° HC caso simulaciones condicionales gaussianas (Elaboración propia).*

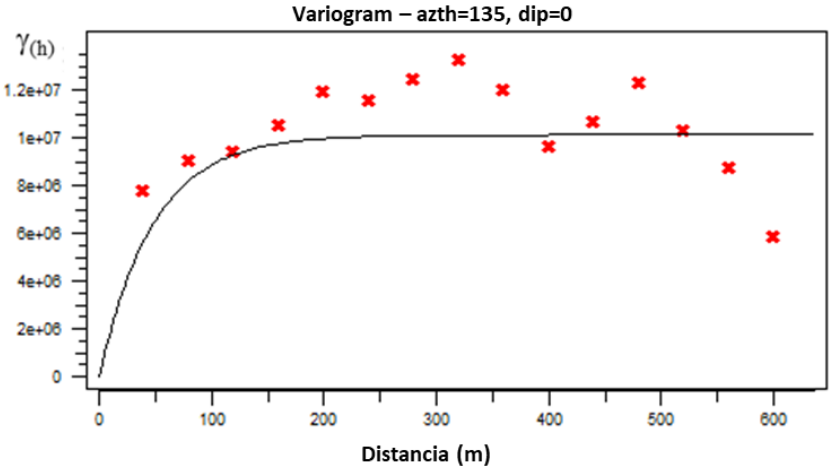

Ilustración B. 18: *Variograma modelado 135° HC caso simulaciones condicionales gaussianas (Elaboración propia).*

Caso sin datos atípicos

Los variogramas experimentales y modelados para este caso, se aprecian desde la ilustración B.19 hasta la B.22.

Al igual que en los casos anteriores, eje mayor corresponde a 45° azimuth, mientras que eje menor a 135° azimuth.

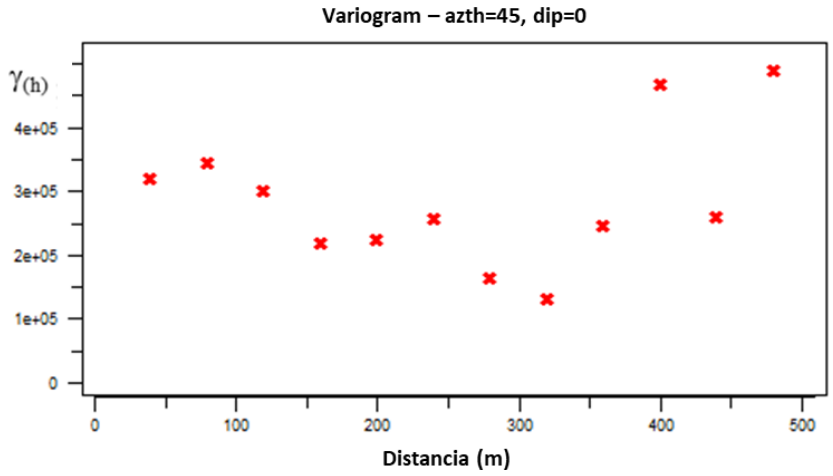

Ilustración B. 19: *Variograma experimental 45° HC caso sin datos atípicos (Elaboración propia).*

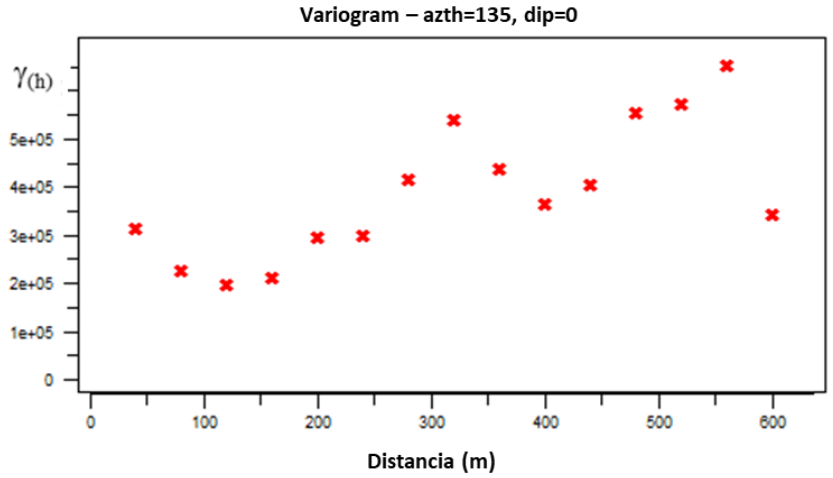

Ilustración B. 20: *Variograma experimental 135° HC caso sin datos atípicos (Elaboración propia).*

Los variogramas fueron ajustado a una curva esférica con efecto pepita de 0.01 y meseta igual a la varianza de los datos del hidrocarburo. En dirección 45° azimuth presenta un rango de 294 metros y en dirección 135° azimuth de 240 metros.

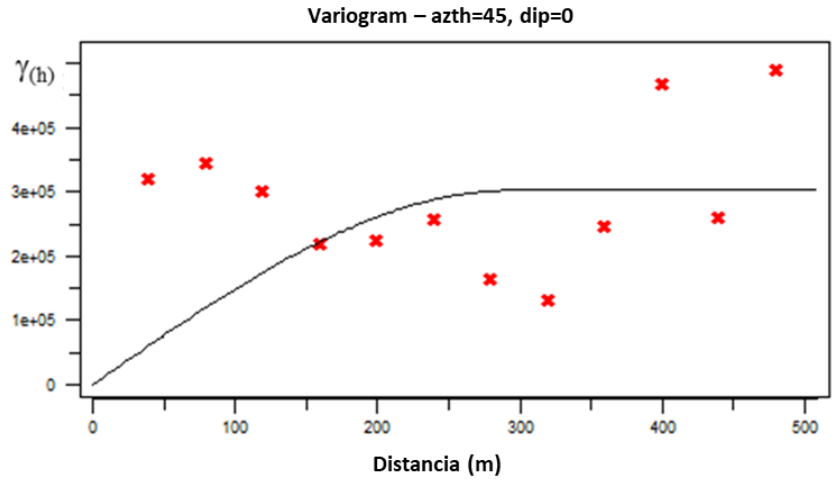

Ilustración B. 21: *Variograma modelado 45° HC caso sin datos atípicos (Elaboración propia).*

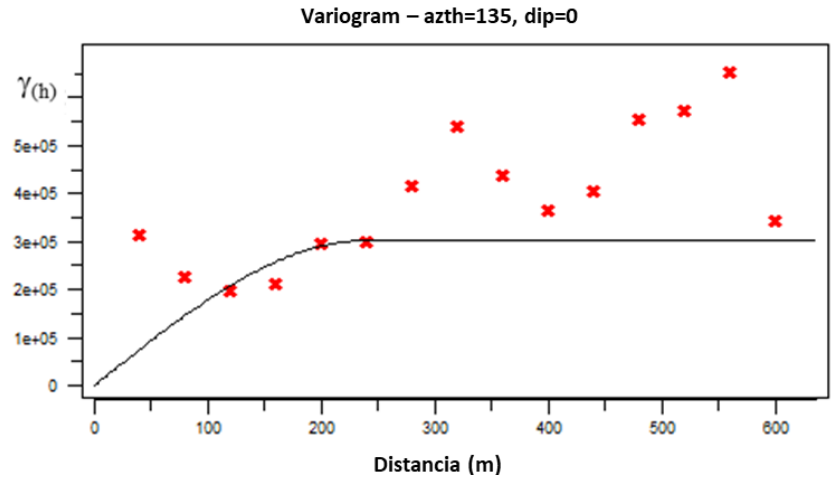

Ilustración B. 22: *Variograma modelado 135° HC caso sin datos atípicos (Elaboración propia).*

Apéndice C: Análisis exploratorio de estimaciones geoestadísticas

Kriging ordinario

Los resultados de las estadísticas descriptivas de estimación, se pueden observar a partir desde la tabla C.1 hasta la C.3 y de la ilustración C.1 hasta la C.5

Tabla C. 1: *Estadísticas descriptivas estimación kriging ordinario HC caso sin datos atípicos bloques 10x10 (Elaboración propia).*

Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
3380	-196.01	2927.6	239.3	112.1	103703	276.8

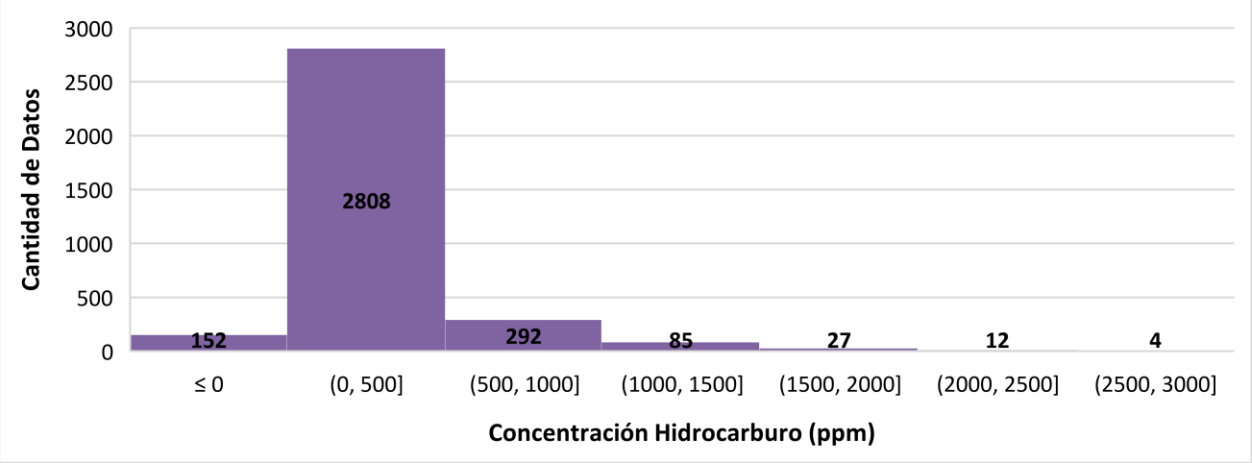

***Ilustración C. 1**: Histograma de estimación kriging ordinario HC caso sin datos atípicos bloques 10x10 (Elaboración propia).*

***Tabla C. 2:** Estadísticas descriptivas estimación kriging ordinario Ni (Elaboración propia).*

Bloques	Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
2x2	83200	26.3	340	72.5	63.01	962.9	86.4
10x10	3380	28.9	295.3	72.4	63	950	86.2

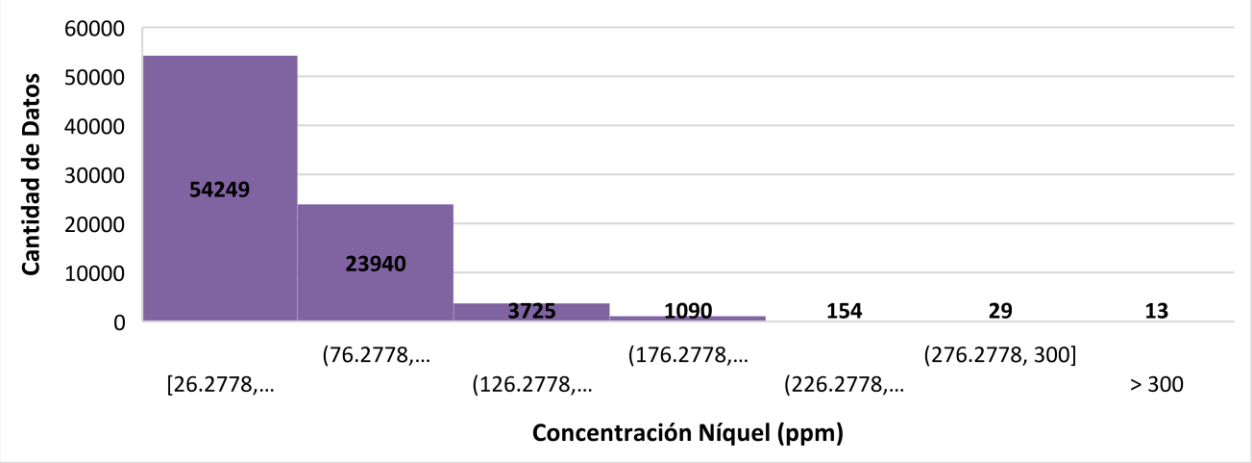

***Ilustración C. 2:** Histograma de estimación kriging ordinario Ni bloques 2x2 (Elaboración propia).*

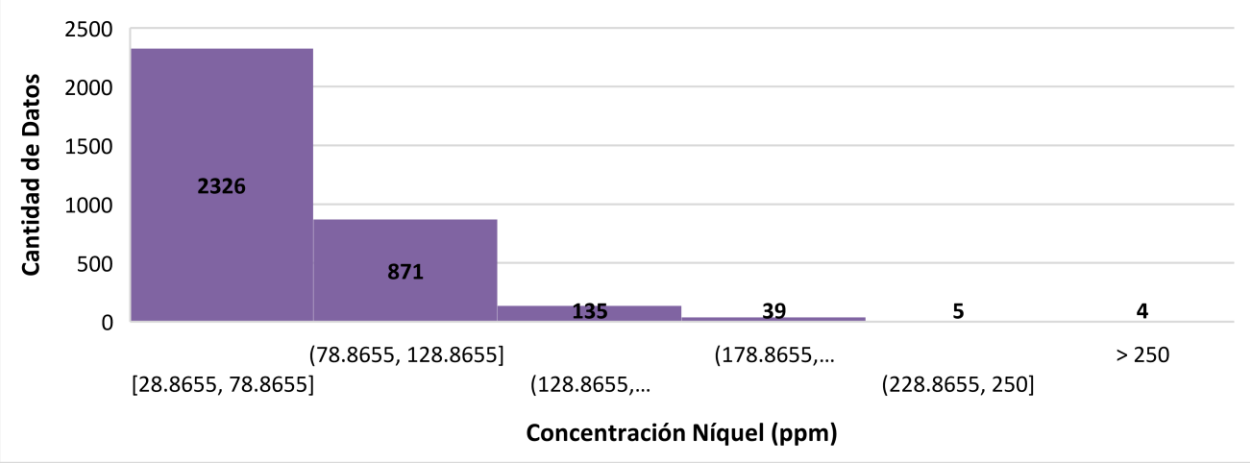

Ilustración C. 3: *Histograma de estimación kriging ordinario Ni bloques 10x10 (Elaboración propia).*

Tabla C. 3: *Estadísticas descriptivas estimación kriging ordinario espesor (Elaboración propia).*

Bloques	Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
2x2	83200	0.05	1.7	0.7	0.7	0.05	0.8
10x10	3380	0.1	1.5	0.7	0.7	0.05	0.8

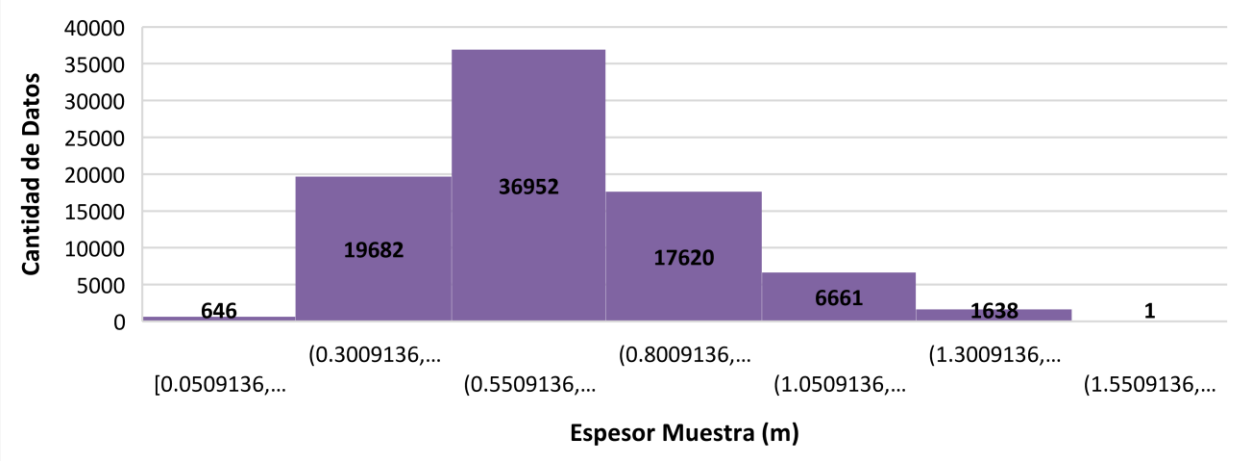

Ilustración C. 4: *Histograma de estimación kriging ordinario espesor bloques 2x2 (Elaboración propia).*

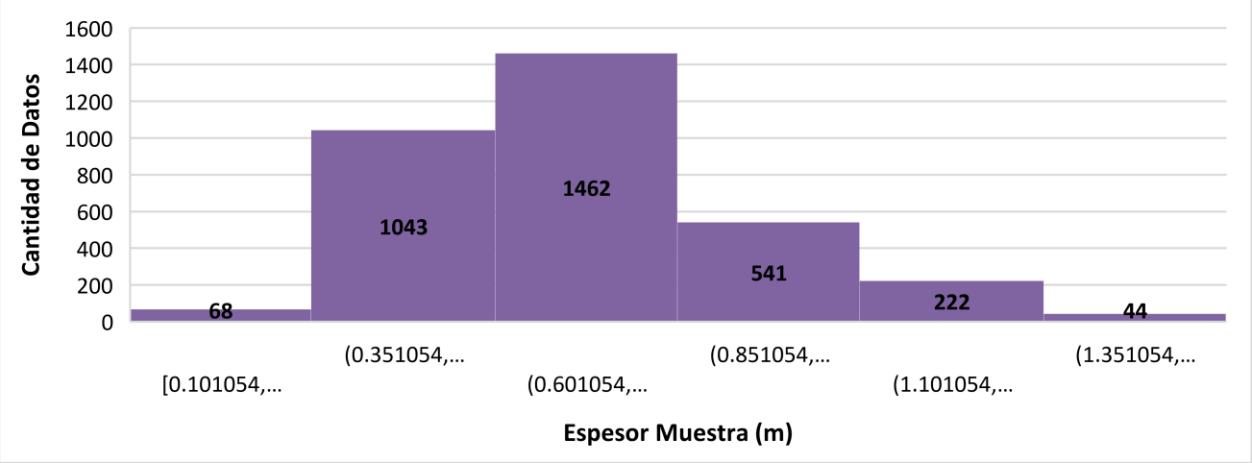

Ilustración C. 5: *Histograma de estimación kriging ordinario espesor bloques 10x10 (Elaboración propia).*

Kriging de indicadores

A continuación, se presenta el histograma de estimación por kriging de indicadores.

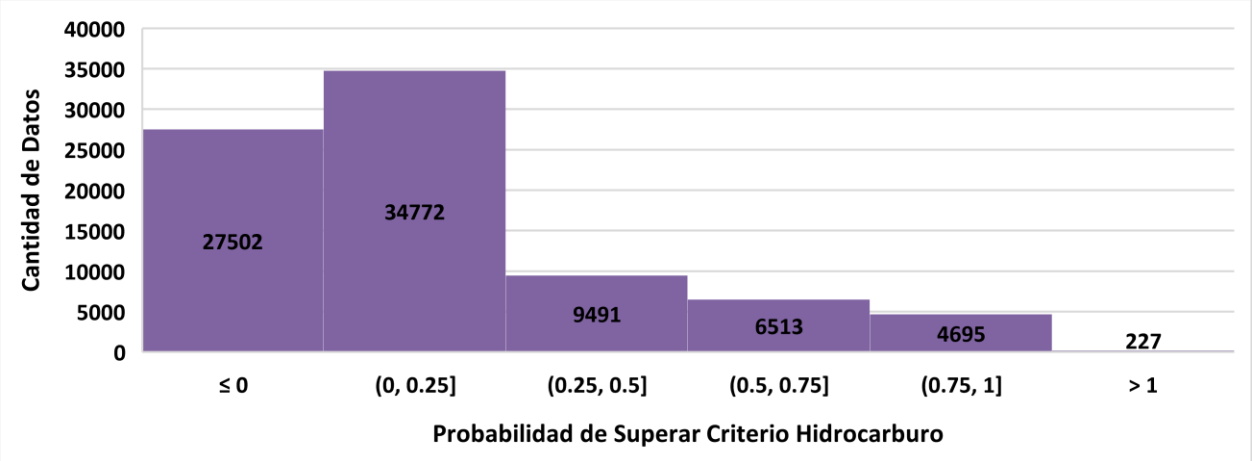

Ilustración C. 6: *Histograma de estimación kriging de indicadores HC bloques 2x2 (Elaboración propia).*

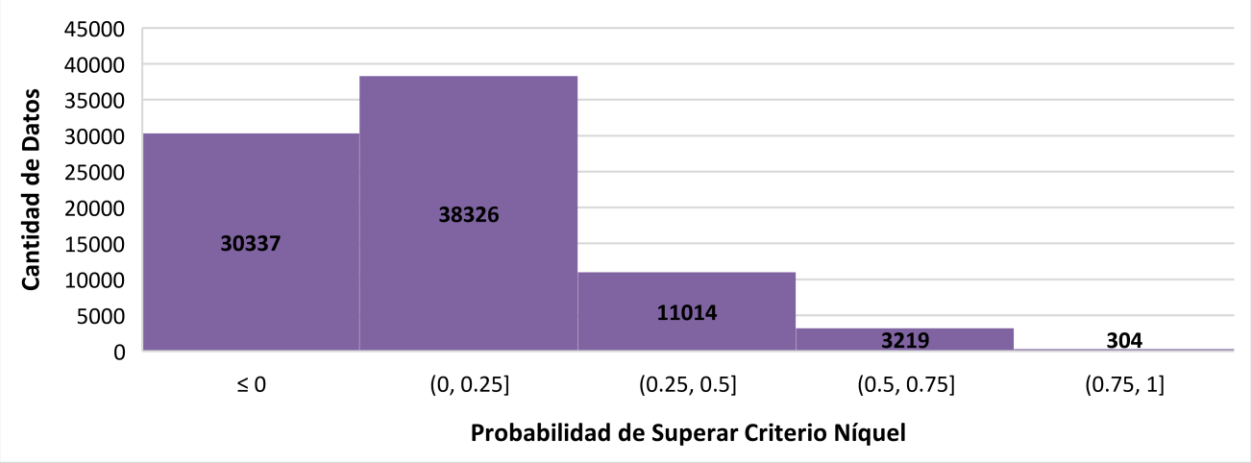

Ilustración C. 7: *Histograma de estimación kriging de indicadores Ni bloques 2x2 (Elaboración propia).*

Simulaciones condicionales gaussianas

Las estadísticas descriptivas de estimación e histogramas del post-tratamiento se aprecian a continuación.

Tabla C. 4: *Estadísticas descriptivas estimación probabilidad de superar criterio de contaminación HC (Elaboración propia).*

Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
83200	0	1	0.5	0.5	0.05	0.64

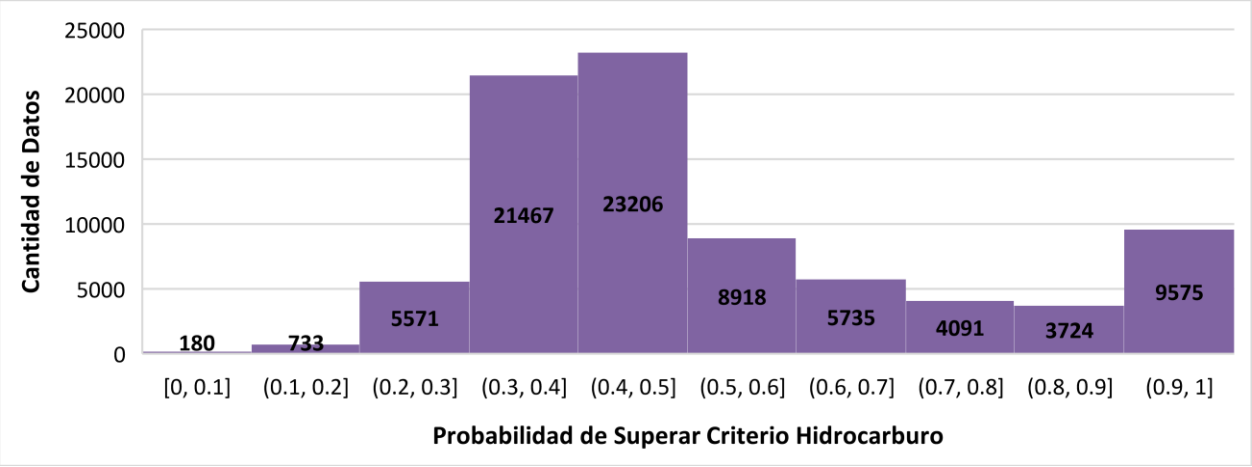

Ilustración C. 8: *Histograma de probabilidad de superar criterio HC bloques 2x2 (Elaboración propia).*

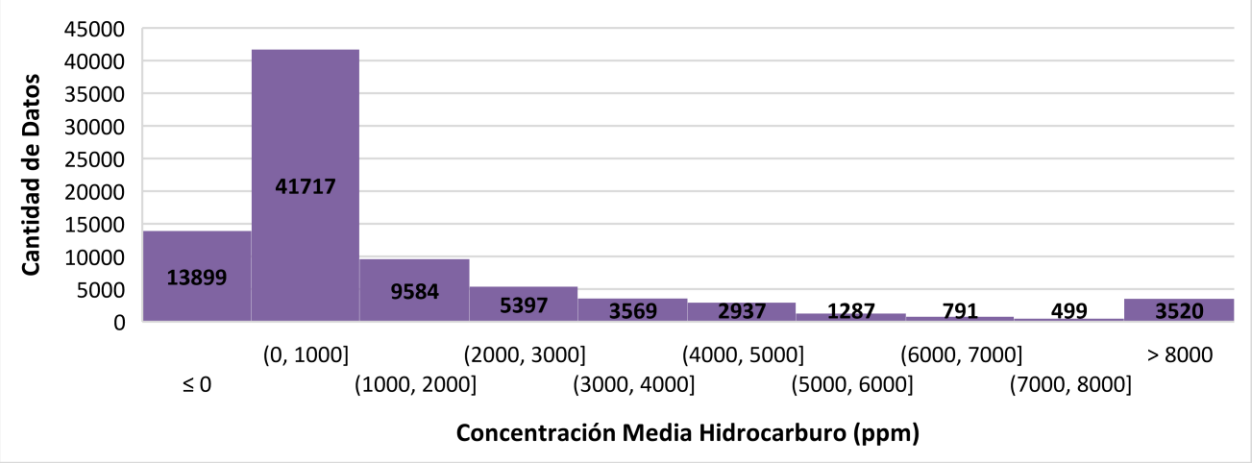

Ilustración C. 9: *Histograma de estimación concentración media de simulaciones realizadas HC bloques 2x2 (Elaboración propia).*

Apéndice D: Análisis exploratorio de estimaciones machine learning

Variables continuas: Hidrocarburo

A continuación, se muestran estadísticas descriptivas de estimación e histogramas de modelos generados por redes neuronales que no son parte del cuerpo principal de la memoria.

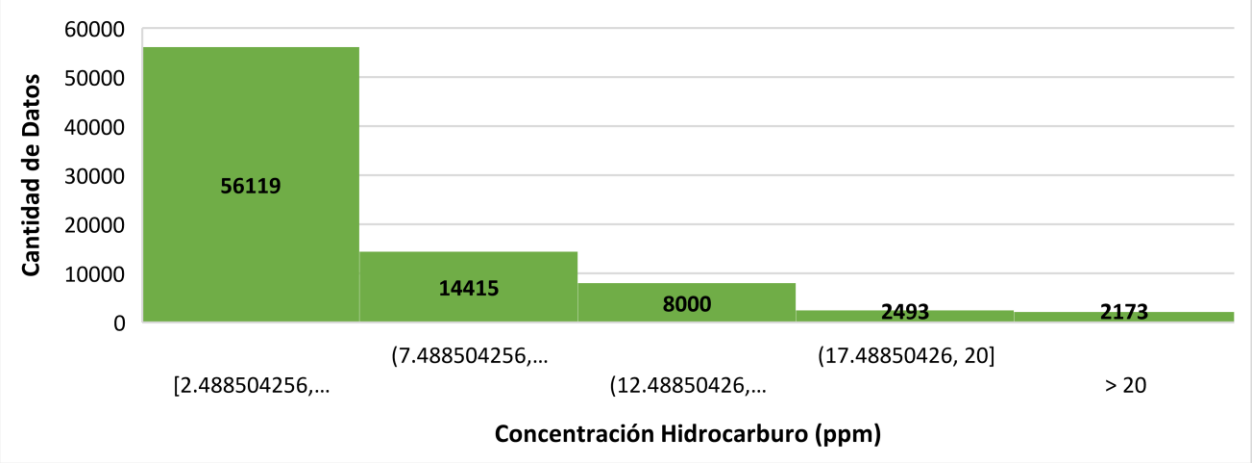

Ilustración D. 1: *Histograma de estimación modelo NN 100_200 HC variable continua (Elaboración propia).*

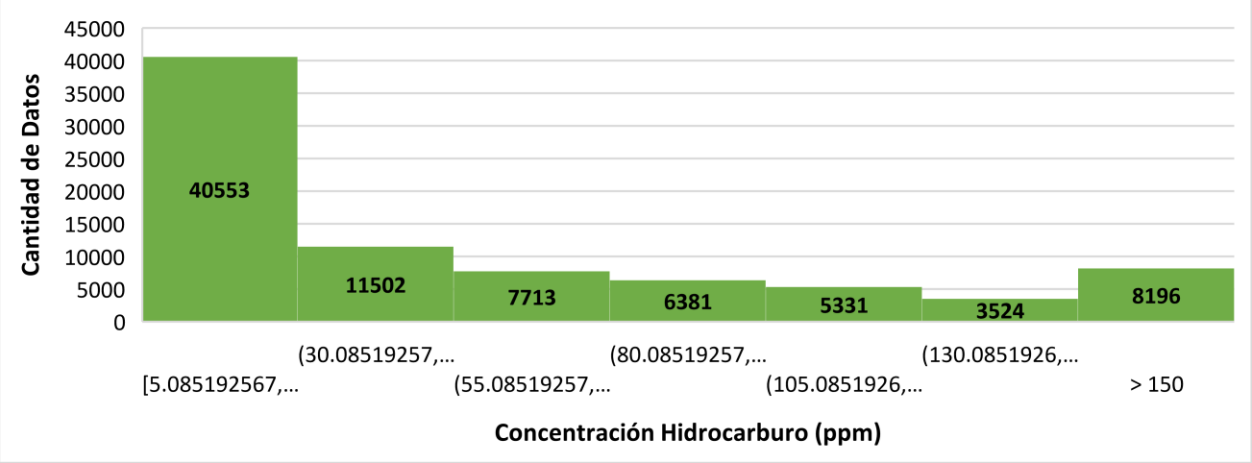

Ilustración D. 2: *Histograma de estimación Modelo NN 1000_200 HC variable continua (Elaboración propia).*

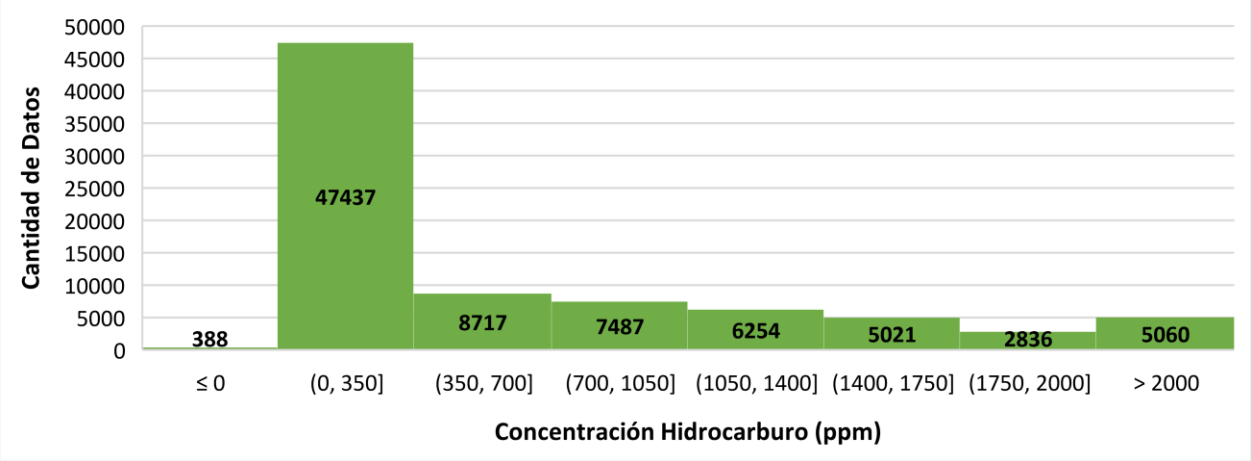

Ilustración D. 3: *Histograma de estimación Modelo NN 100_2000 HC variable continua (Elaboración propia).*

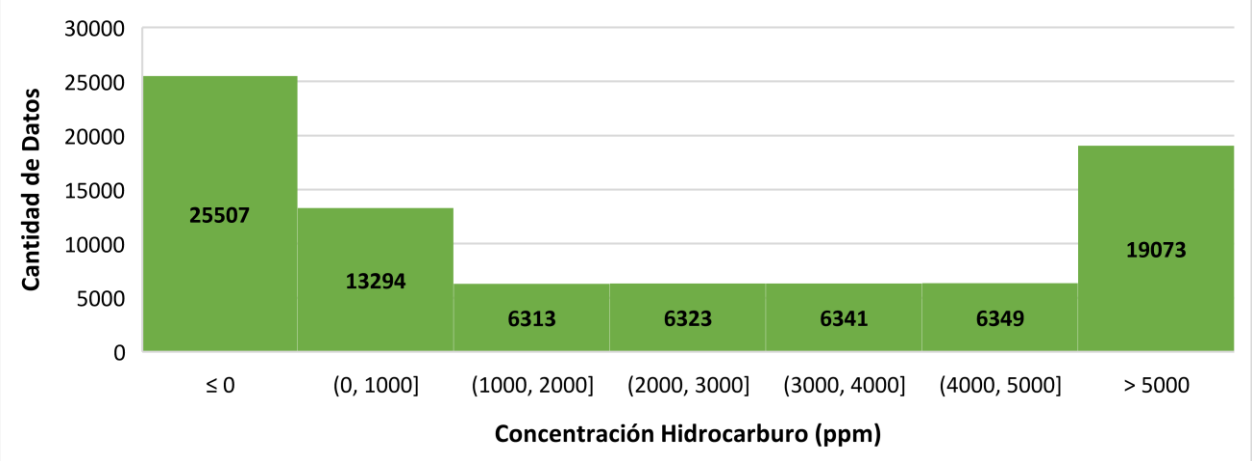

Ilustración D. 4: *Histograma de estimación Modelo NN 100_2000 HC variable continua (Elaboración propia).*

Variables continuas: Níquel

A continuación, se muestran estadísticas descriptivas de estimación e histogramas de modelos generados por redes neuronales que no son parte del cuerpo principal de la memoria.

Tabla D. 1: *Estadísticas descriptivas estimación redes neuronales Ni (Elaboración propia).*

Modelo	Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
NN 100_2000	83200	49	130.5	76.2	73	257	84.4
NN 1000_2000	83200	46	127.3	75.2	71.1	288.4	85

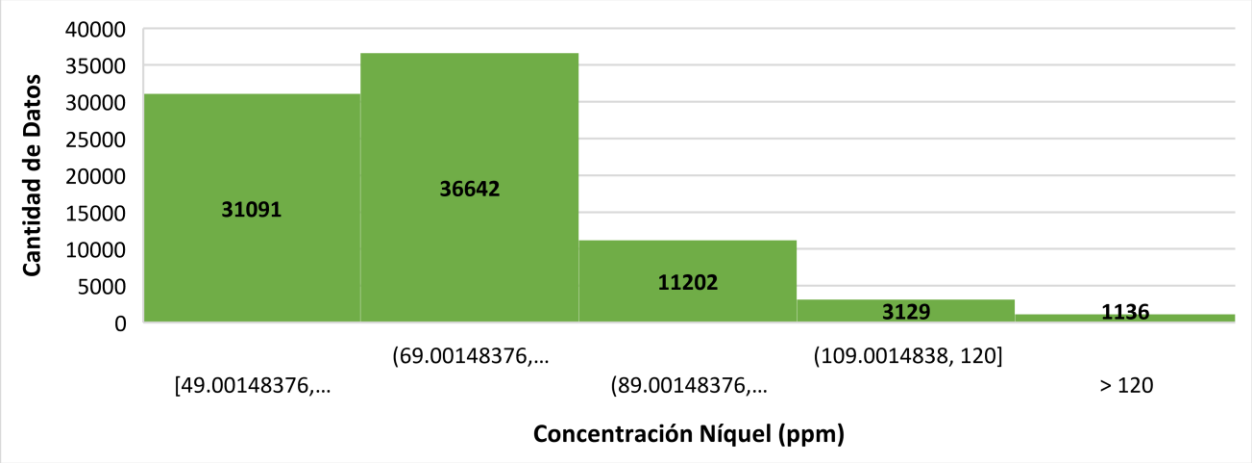

Ilustración D. 5: *Histograma de estimación Modelo NN 100_2000 Ni variable continua (Elaboración propia).*

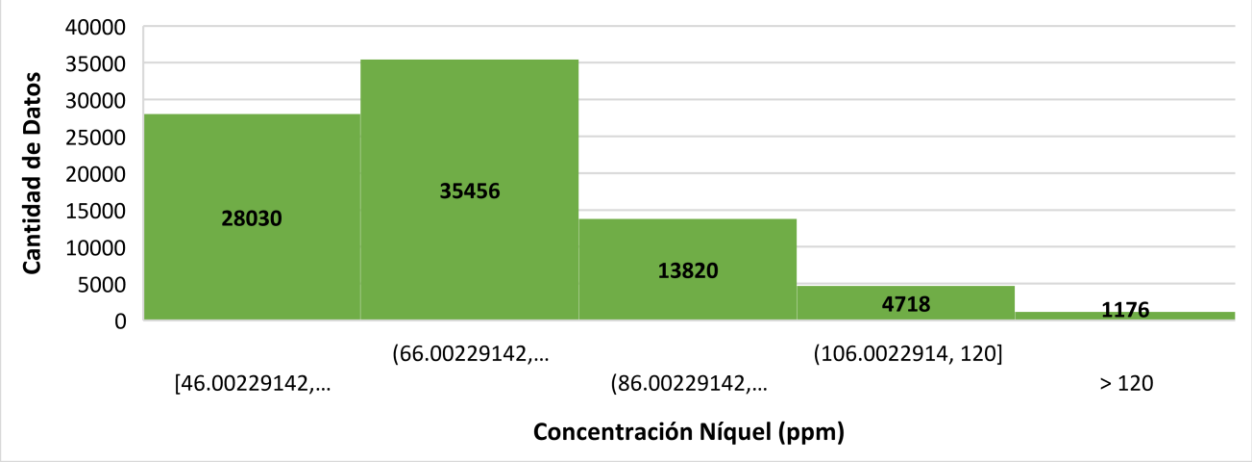

Ilustración D. 6: *Histograma de estimación Modelo NN 1000_2000 Ni variable continua (Elaboración propia).*

Variables categóricas

A continuación, se muestran estadísticas descriptivas de estimación e histogramas de modelos generados por redes neuronales que no son parte del cuerpo principal de la memoria.

Tabla D. 2: *Estadísticas descriptivas estimación redes neuronales variables categóricas (Elaboración propia).*

Variable	Modelo	Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
HC	**NN 100_2000**	83200	1.7E-07	0.8	0.2	0.1	0.05	0.2
	NN 1000_2000	83200	2.7E-25	0.9	0.2	0.02	0.08	0.2
Ni	**NN 100_2000**	83200	0.007	0.3	0.1	0.1	0.01	0.1
	NN 1000_2000	83200	3.2E-05	0.7	0.2	0.05	0.05	0.3

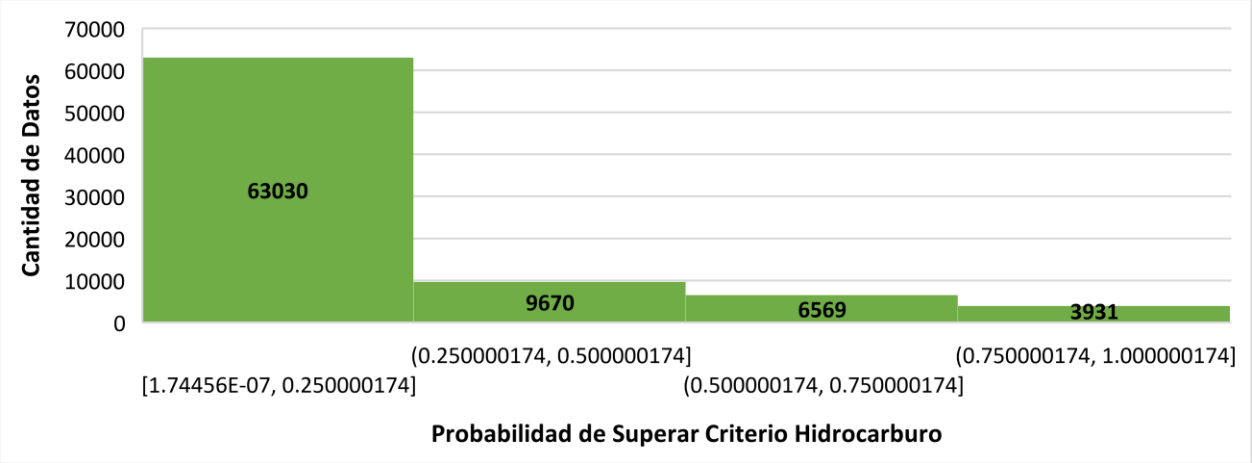

Ilustración D. 7: *Histograma de estimación Modelo NN 100_2000 HC variable categórica (Elaboración propia).*

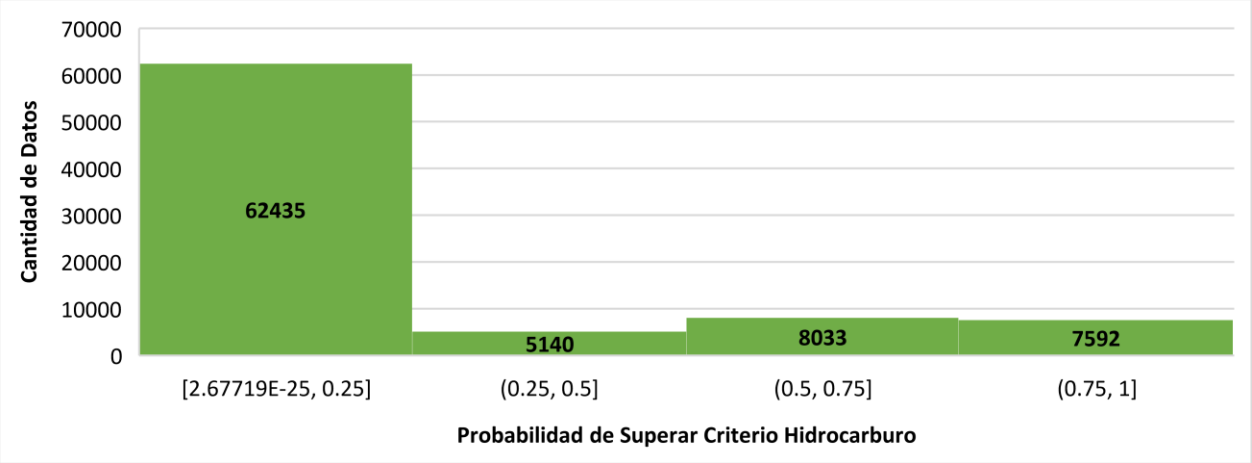

Ilustración D. 8: *Histograma de estimación Modelo NN 1000_2000 HC variable categórica (Elaboración propia).*

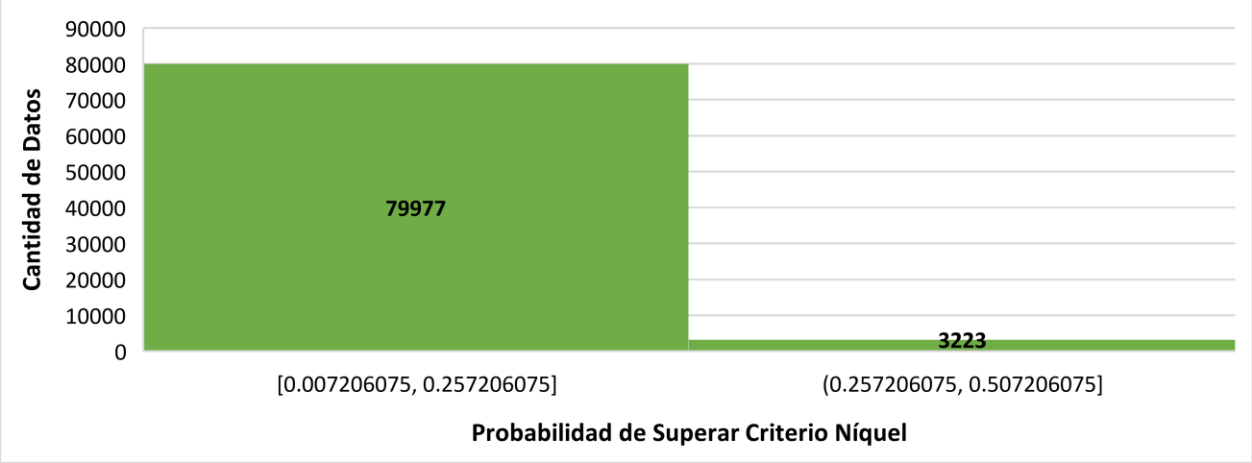

Ilustración D. 9: *Histograma de estimación Modelo NN 100_2000 Ni variable categórica (Elaboración propia).*

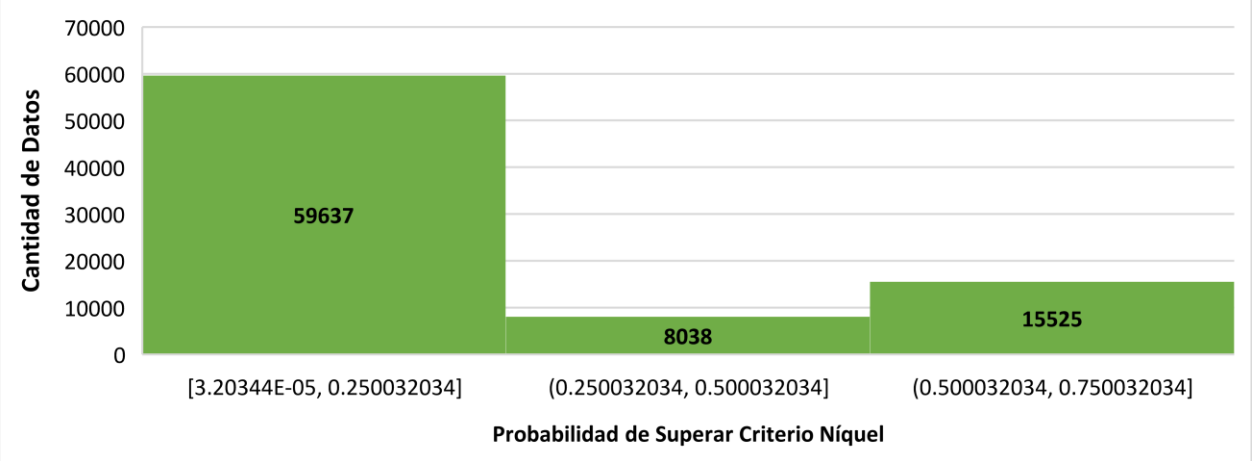

Ilustración D. 10: *Histograma de estimación Modelo NN 1000_2000 Ni variable categórica (Elaboración propia).*

Variables categóricas: Caso hidrocarburo con 210 muestras

A continuación, se muestran estadísticas descriptivas de estimación e histogramas de modelos redes neuronales para caso de 210 muestras de hidrocarburo.

Tabla D. 3: *Estadísticas descriptivas estimación redes neuronales variables categóricas, caso HC 210 muestras (Elaboración propia).*

Modelo	Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
NN 100_2000	83200	4.9E-09	0.9	0.2	0.05	0.05	0.2
NN 1000_2000	83200	2.5E-14	0.9	0.2	0.04	0.06	0.1

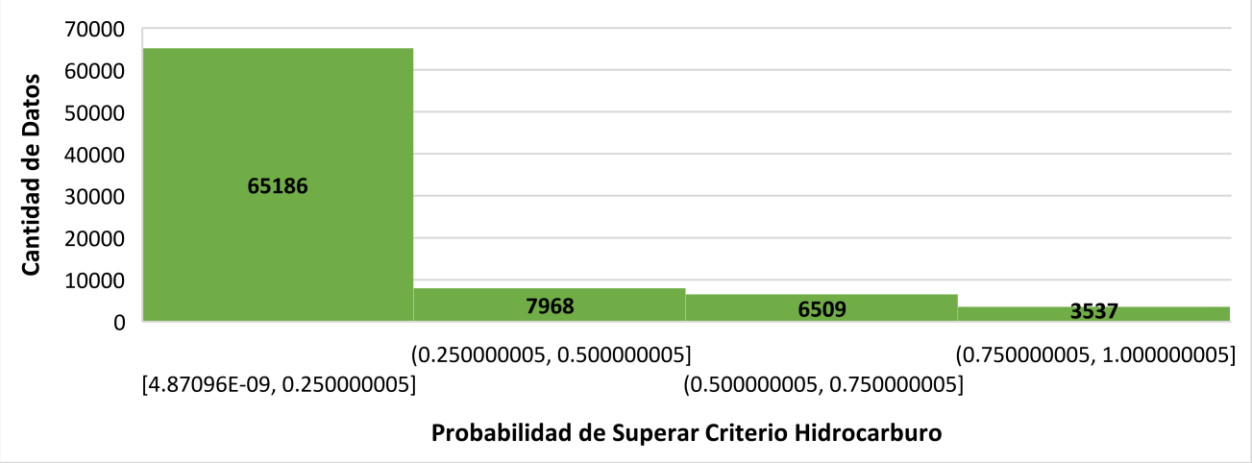

Ilustración D. 11: *Histograma de estimación Modelo NN 100_2000 caso HC 210 muestras (Elaboración propia).*

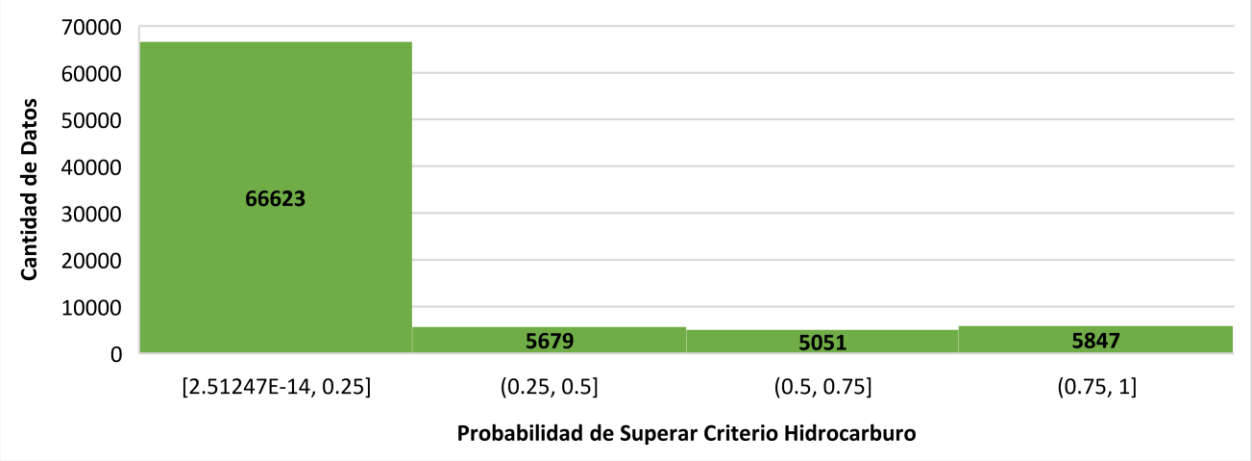

Ilustración D. 12: *Histograma de estimación Modelo NN 1000_2000 caso HC 210 muestras (Elaboración propia).*

Variables categóricas: Caso hidrocarburo muestreado con grilla

A continuación, se muestran estadísticas descriptivas de estimación e histogramas de modelos redes neuronales para caso HC muestreado con grilla.

Tabla D. 4: *Estadísticas descriptivas estimación redes neuronales variables categóricas, caso HC muestreado con grilla (Elaboración propia).*

Modelo	Cantidad de datos	Mínimo	Máximo	Media	Mediana	Varianza	Q3
NN 100_2000	83200	3.7E-07	0.9	0.2	0.03	0.06	0.2
NN 1000_2000	83200	1.5E-07	1	0.2	0.0004	0.1	0.1

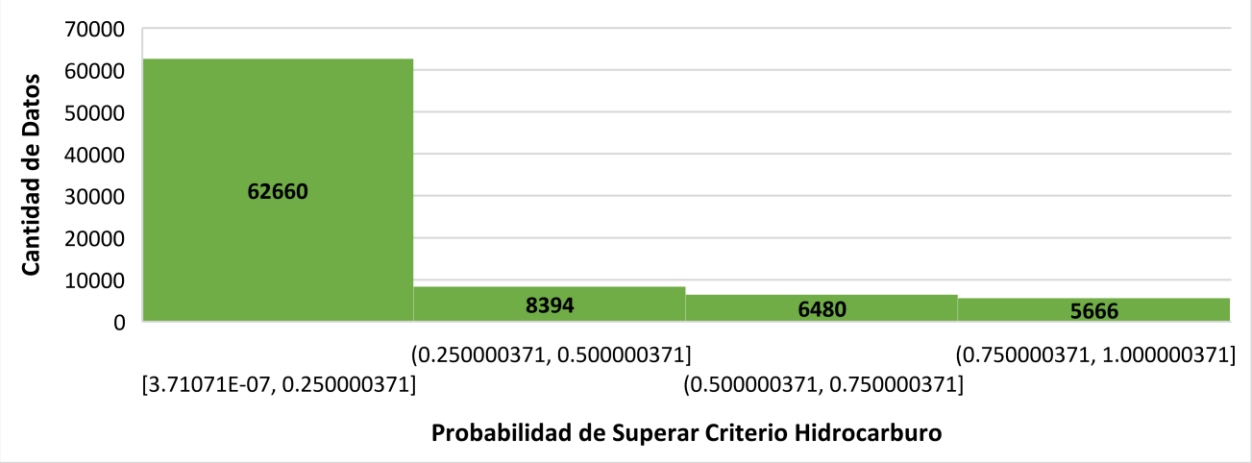

Ilustración D. 13: *Histograma de estimación Modelo NN 100_2000 caso HC muestreado con grilla (Elaboración propia).*

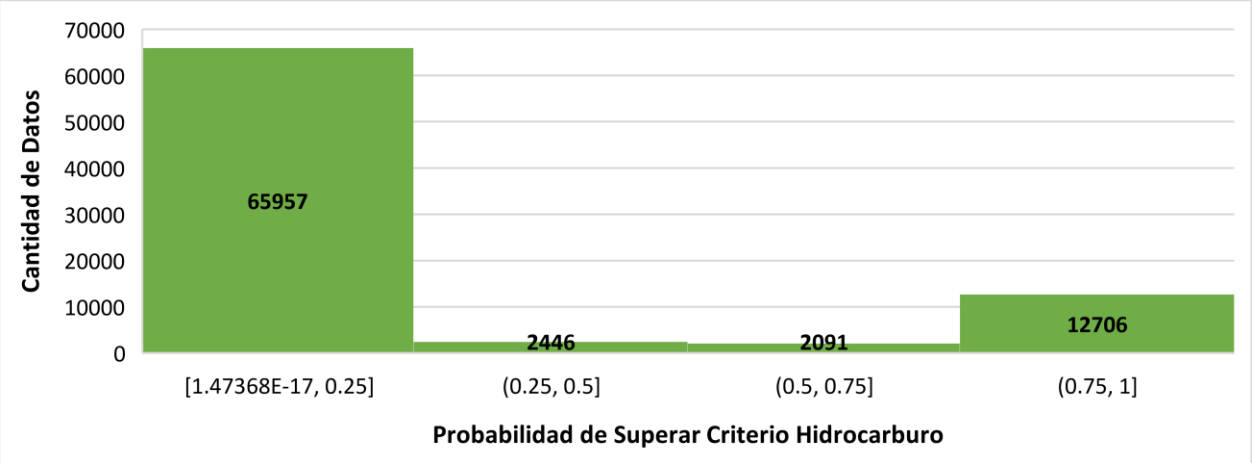

Ilustración D. 14: *Histograma de estimación Modelo NN 1000_2000 caso HC muestreado con grilla (Elaboración propia).*

Caso distancia entre muestras como información de entrada

A continuación, se muestran estadísticas descriptivas de estimación e histogramas de modelos redes neuronales para casos que estimaron sabiendo distancia entre muestras.

Tabla D. 5: *Estadísticas descriptivas estimación redes neuronales, caso distancia entre muestras (Elaboración propia).*

Modelo	Caso	Variable	Mínimo	Máximo	Media	Mediana	Varianza	Q3
NN 1000_2000	Continuo	HC	-10180	20146	240.2	-208,8	1.7E+07	811.4
		Ni	46	127.3	75	71	288.4	85
	Categóricas	HC	3.6E-13	1	0.2	0.0004	0.1	0.4
		Ni	6E-07	1	0.1	3.26E-06	0.04	0.002
	210 datos	HC	3E-06	1	0.2	0.01	0.1	0.2
	Grilla	HC	4E-15	1	0.2	7.4E-07	0.2	0.1

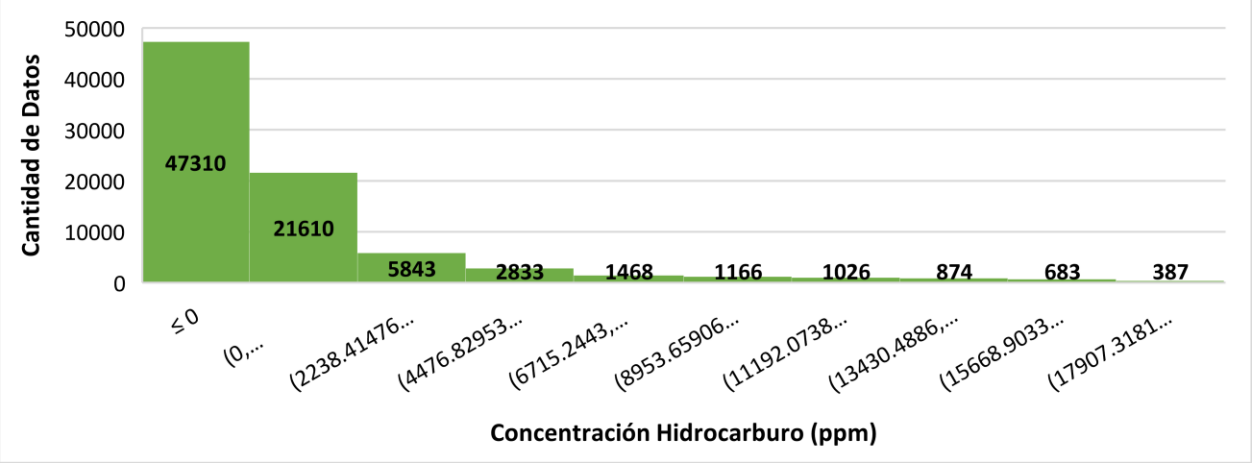

Ilustración D. 15: *Histograma de estimación Modelo NN 1000_2000 variable continua, caso HC distancia entre muestras (Elaboración propia).*

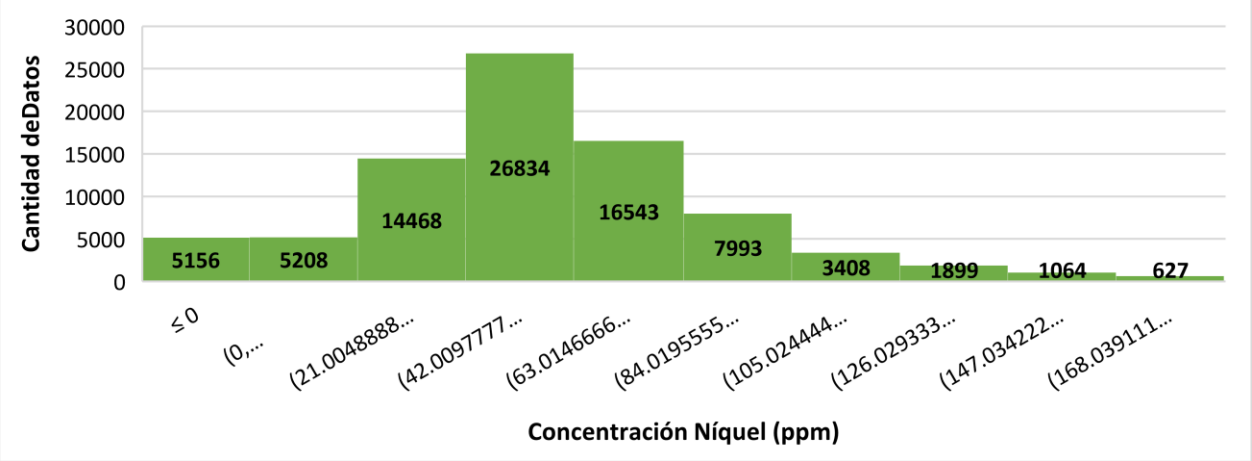

Ilustración D. 16: *Histograma de estimación Modelo NN 1000_2000 variable continua, caso Ni distancia entre muestras (Elaboración propia).*

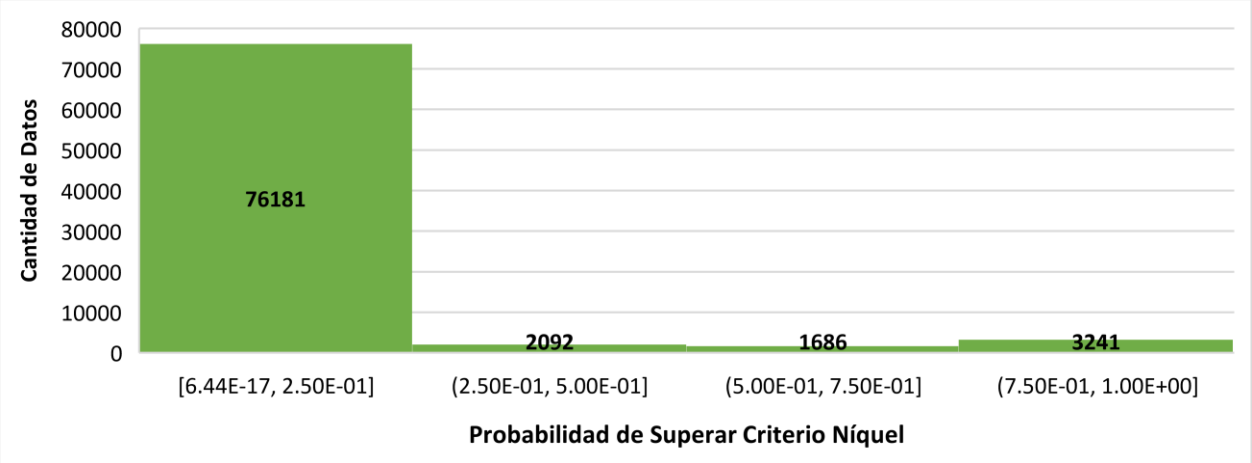

Ilustración D. 17*: Histograma de estimación Modelo NN 1000_2000 variable categórica, caso Ni distancia entre muestras (Elaboración propia).*

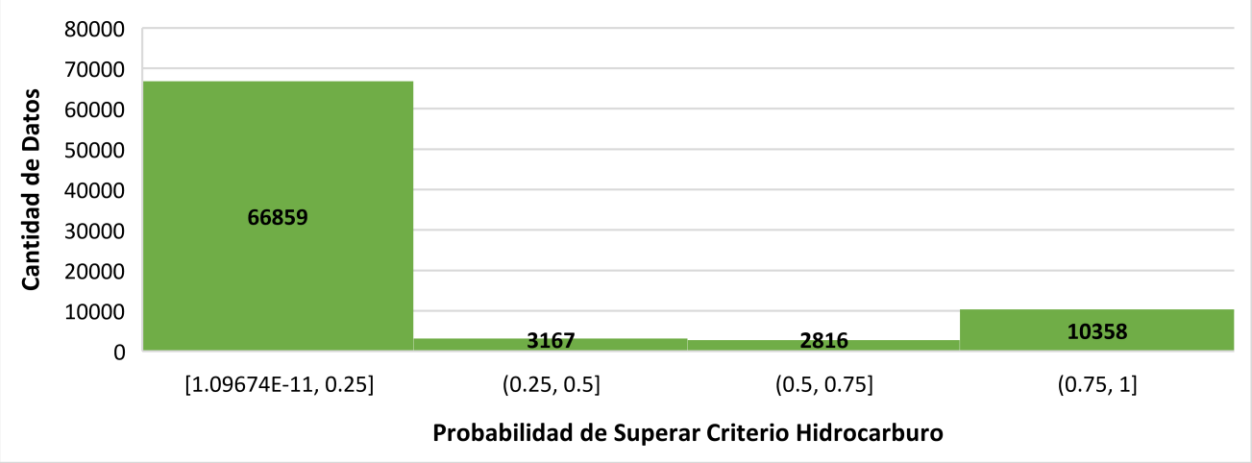

Ilustración D. 18: *Histograma de estimación Modelo NN 1000_2000 variable categórica, caso HC 210 muestras y distancia entre muestras (Elaboración propia).*

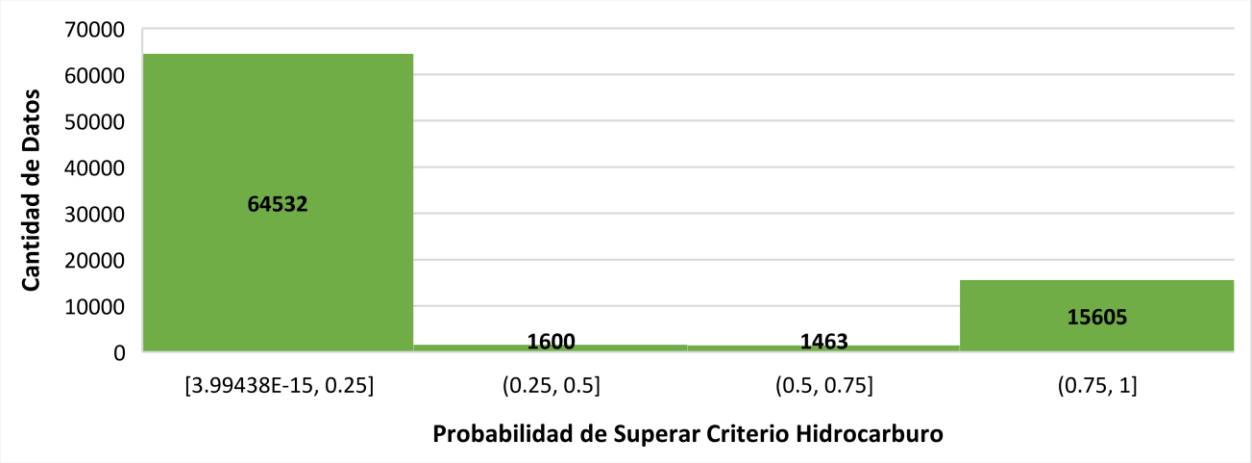

Ilustración D. 19: *Histograma de estimación Modelo NN 1000_2000 variable categórica, caso HC muestreado con grilla y distancia entre muestras (Elaboración propia).*

Apéndice E: Estimación geoestadística

Kriging ordinario

Desde la ilustración E.1 hasta E.6, se observan los resultados de las variables níquel y espesor que no aparecen en el cuerpo principal de la memoria.

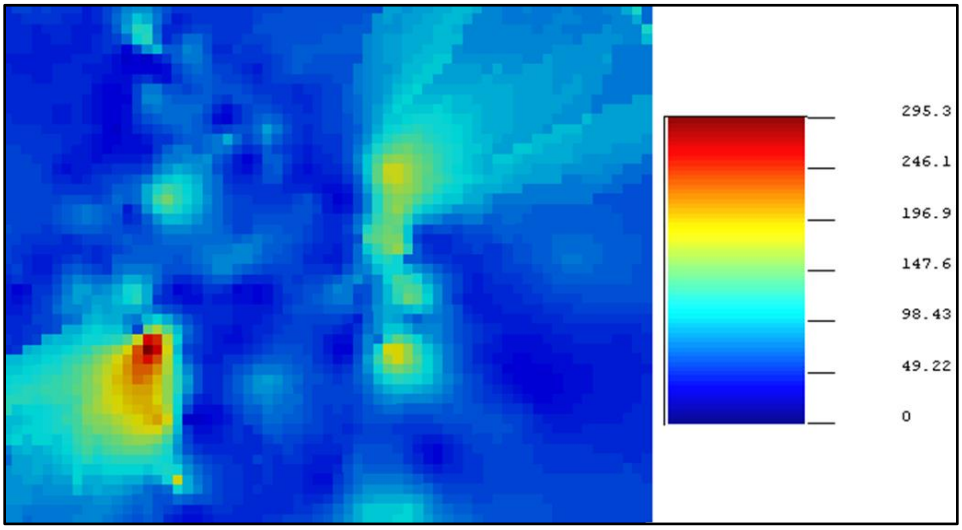

Ilustración E. 1: *Estimación kriging ordinario Ni bloques 10x10 (Elaboración propia).*

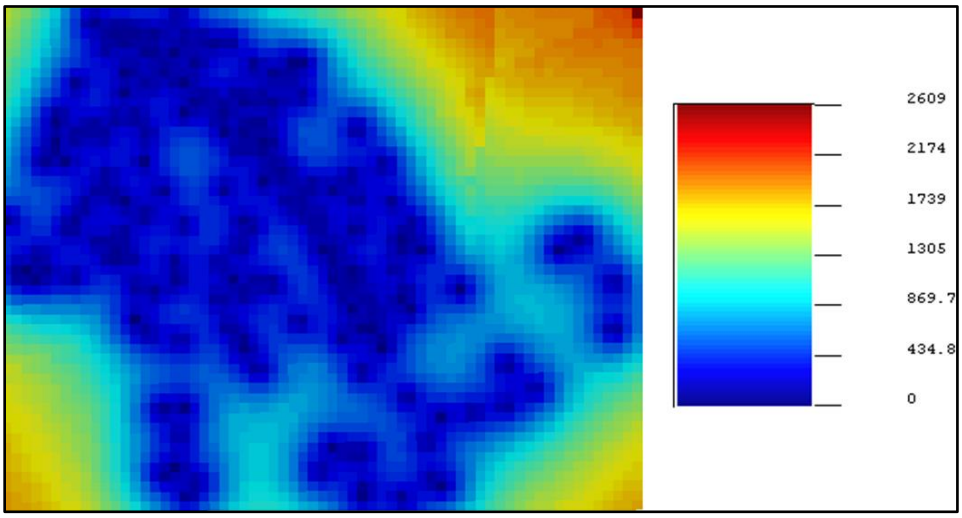

Ilustración E. 2: *Varianza kriging ordinario Ni bloques 10x10 (Elaboración propia).*

El resultado de la estimación del espesor, nos dice que en la zona donde se produjo el descarrilamiento el tamaño de la muestra es bastante menor. En cambio, en las zonas alejadas del accidente se predice tamaños de la capa de contaminación mayores a 1.2 metros.

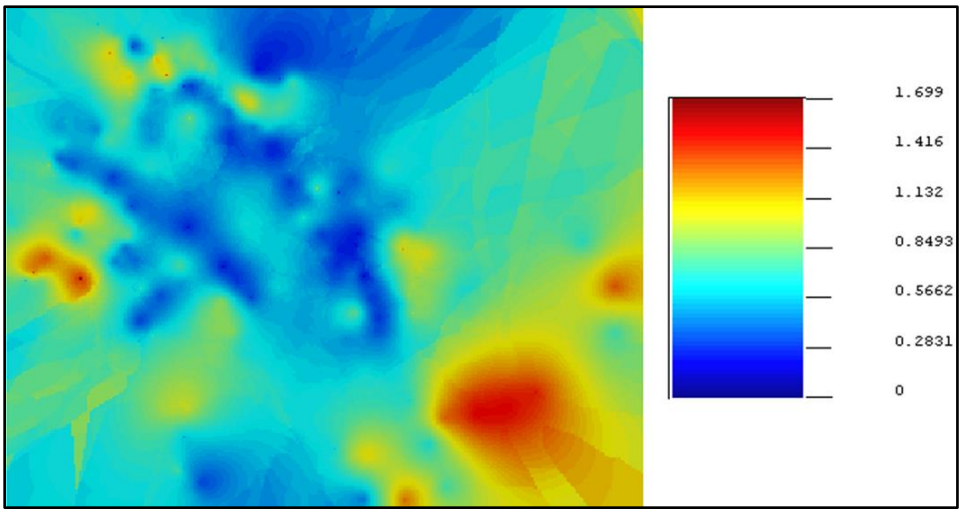

Ilustración E. 3: *Estimación kriging ordinario espesor bloques 2x2 (Elaboración propia).*

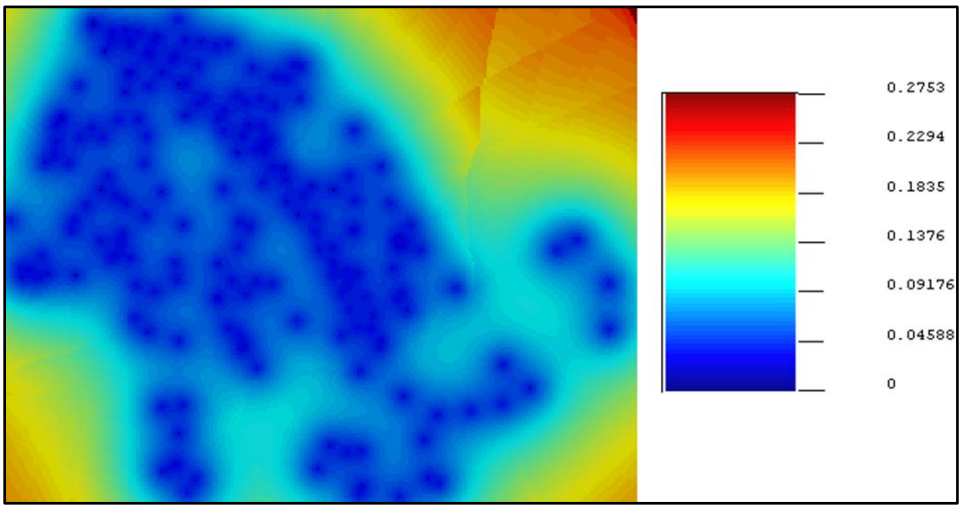

Ilustración E. 4: *Varianza kriging ordinario espesor bloques 2x2 (Elaboración propia).*

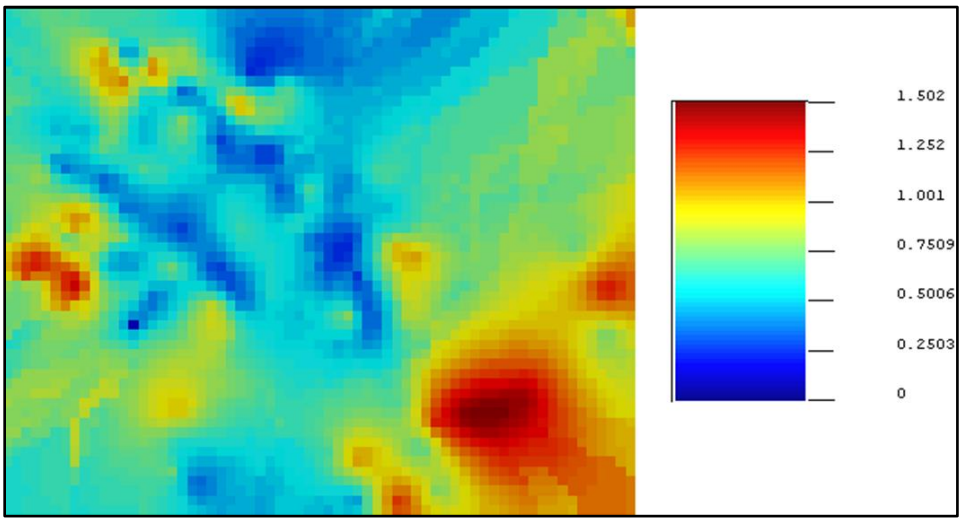

Ilustración E. 5: *Estimación kriging ordinario espesor bloques 10x10 (Elaboración propia).*

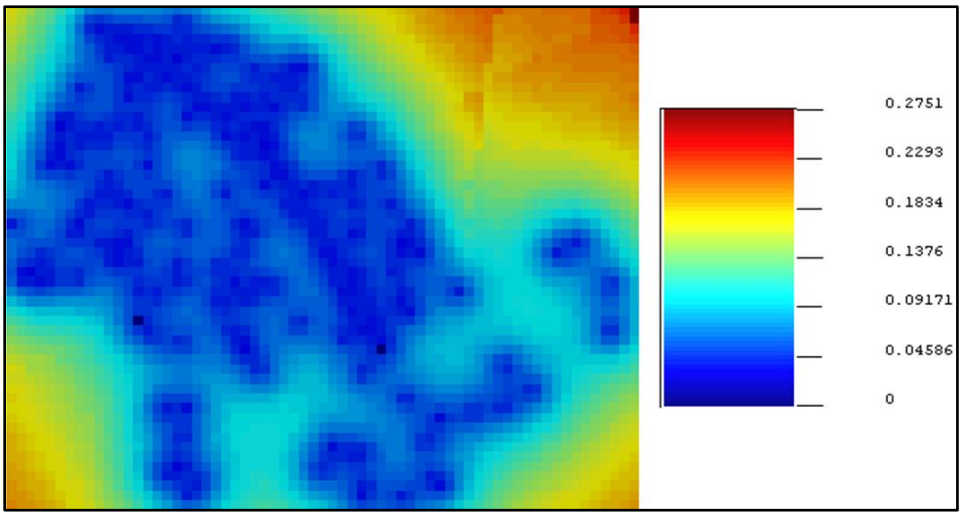

Ilustración E. 6: *Varianza kriging ordinario espesor bloques 10x10 (Elaboración propia).*

Simulaciones condicionales gaussianas

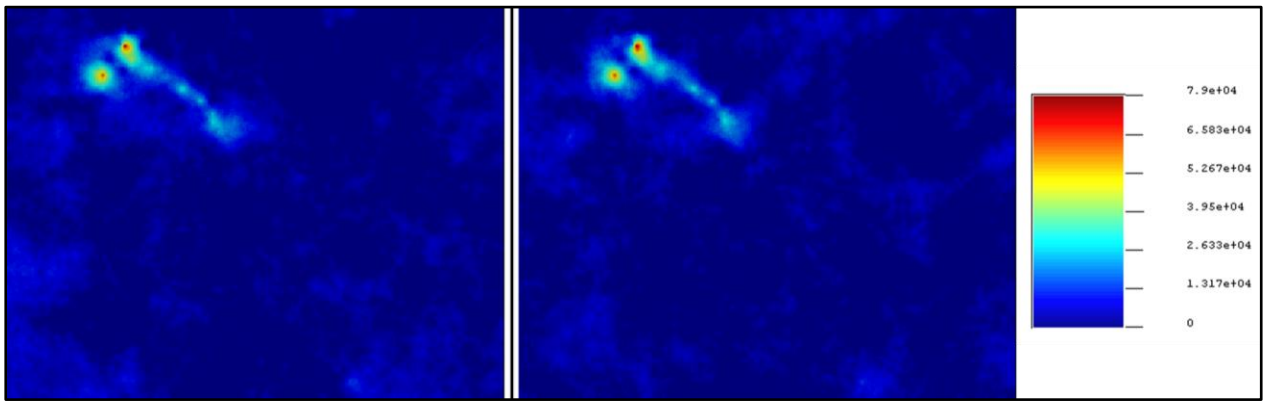

Ilustración E. 7: *Simulación condicional 1 (izquierda) y simulación condicional 2 (derecha) (Elaboración propia).*

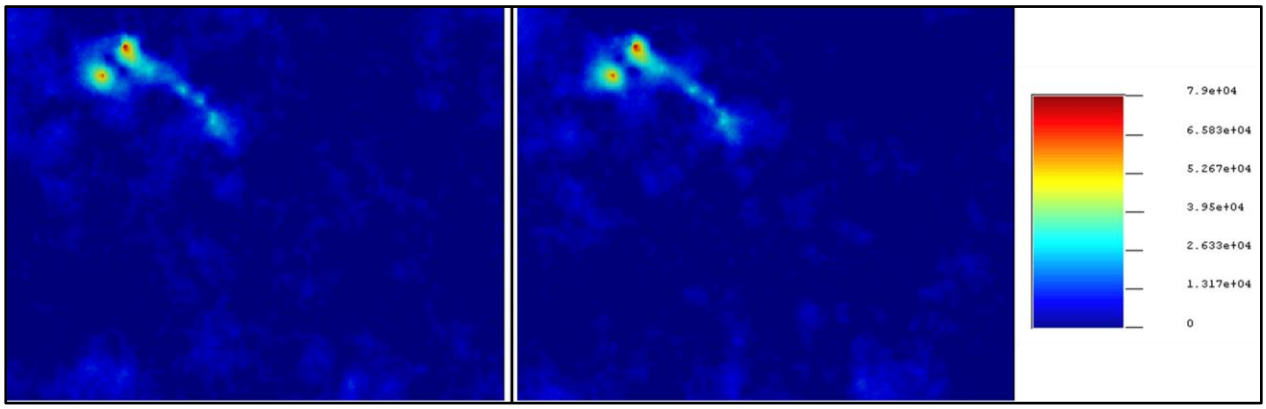

Ilustración E. 8: *Simulación condicional 3 (izquierda) y simulación condicional 4 (derecha) (Elaboración propia).*

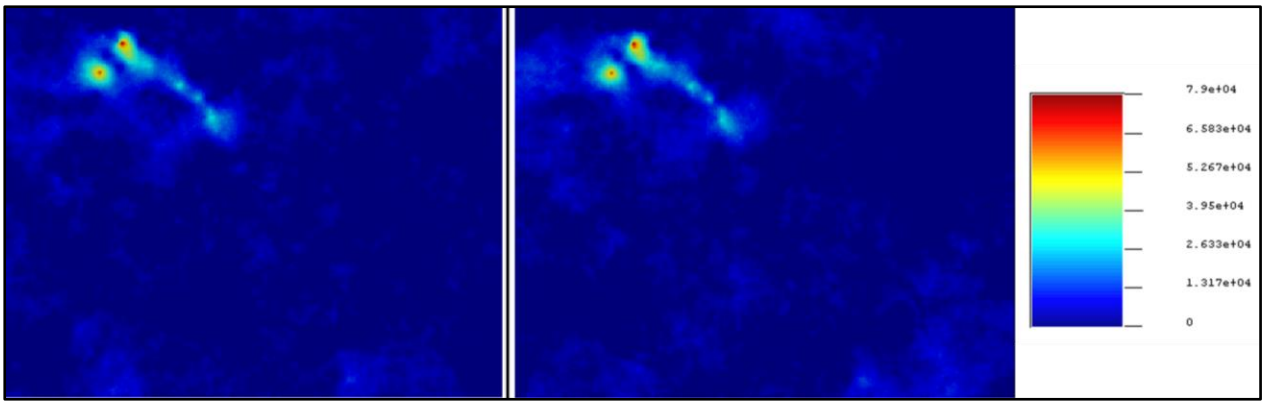

Ilustración E. 9: *Simulación condicional 5 (izquierda) y simulación condicional 6 (derecha) (Elaboración propia).*

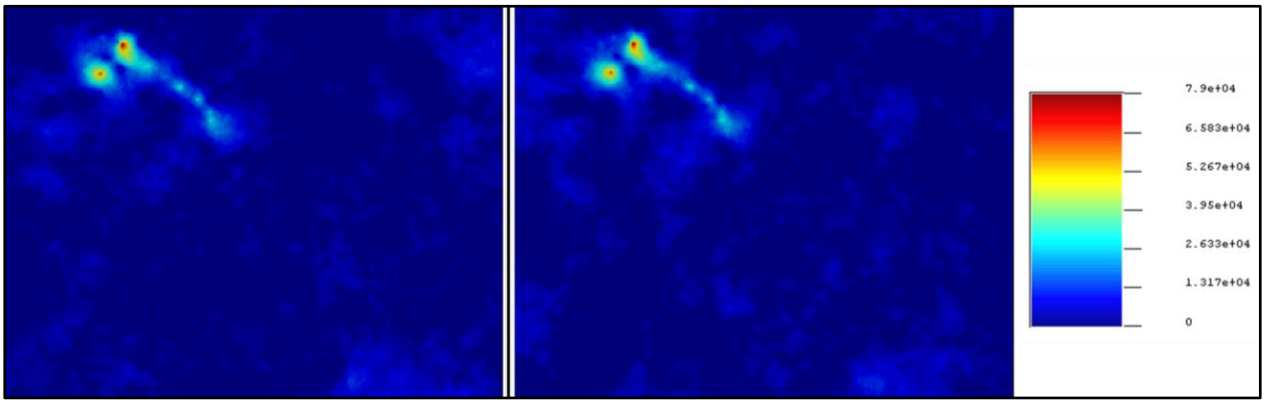

Ilustración E. 10: *Simulación condicional 7 (izquierda) y simulación condicional 8 (derecha) (Elaboración propia).*

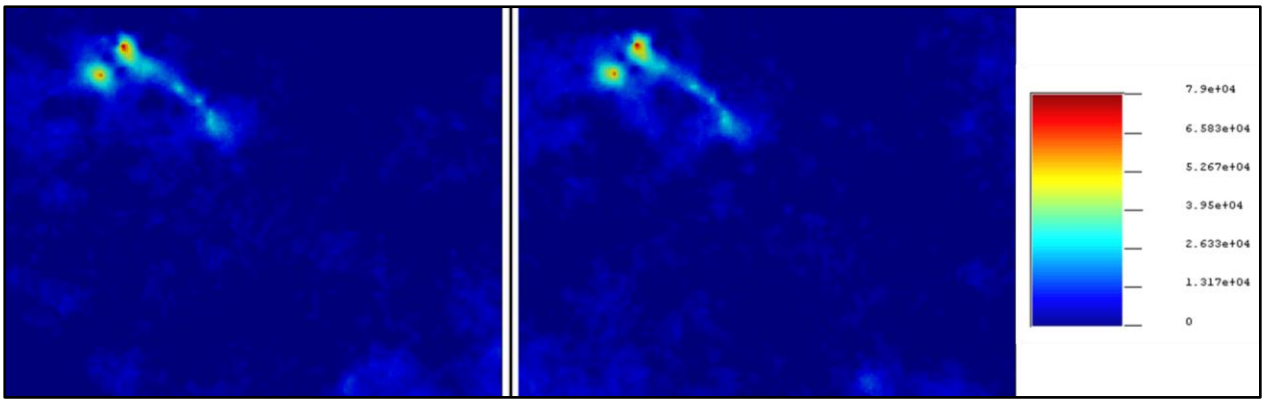

Ilustración E. 11: *Simulación condicional 9 (izquierda) y simulación condicional 10 (derecha) (Elaboración propia).*

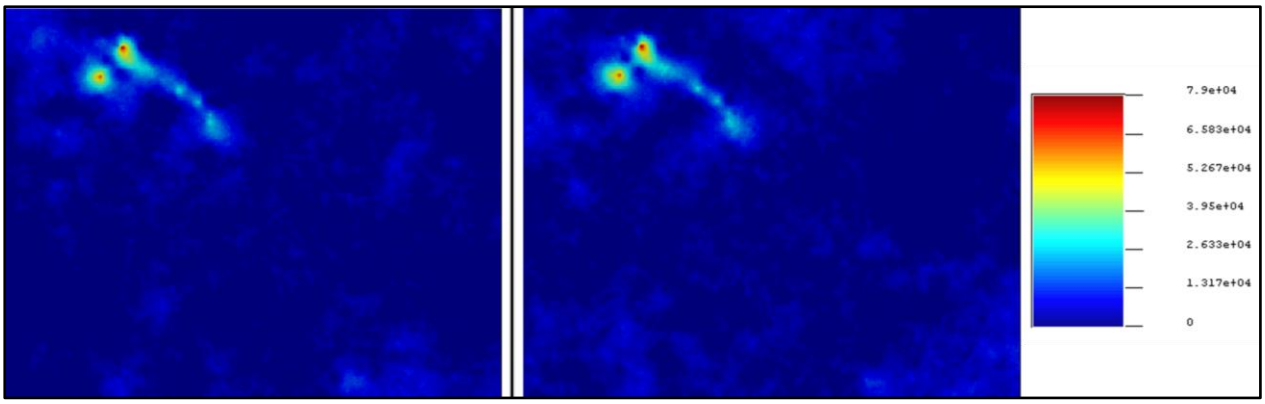

Ilustración E. 12: *Simulación condicional 11 (izquierda) y simulación condicional 12 (derecha) (Elaboración propia)*

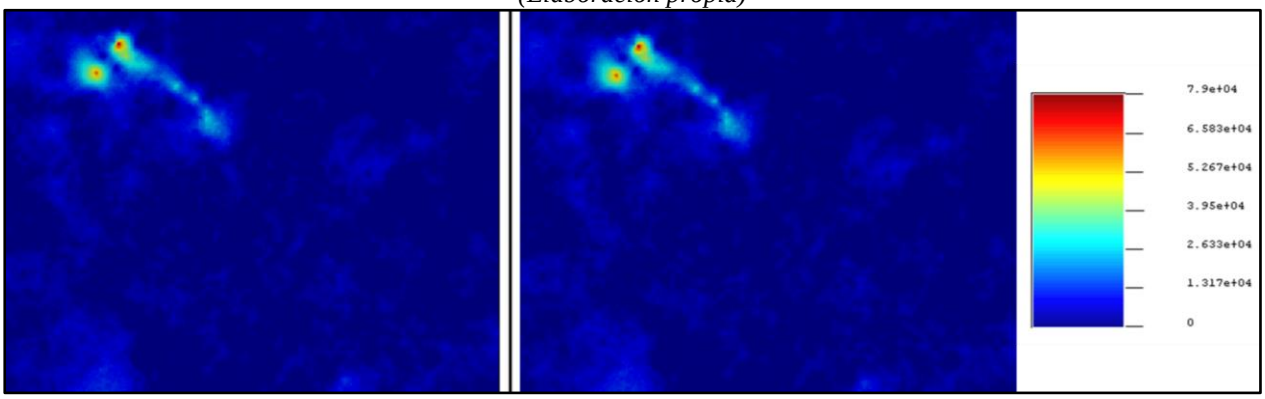

Ilustración E. 13: *Simulación condicional 13 (izquierda) y simulación condicional 14 (derecha) (Elaboración propia).*

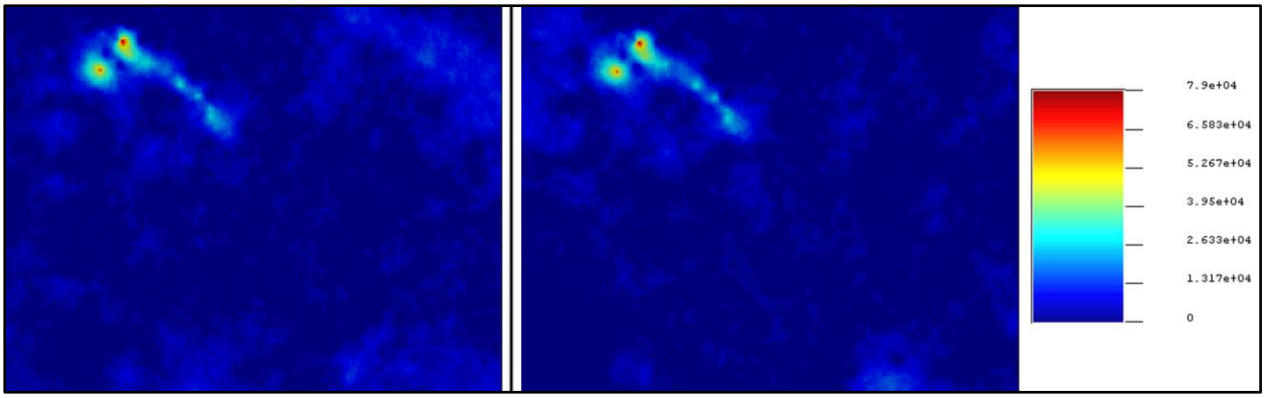

Ilustración E. 14: *Simulación condicional 15 (izquierda) y simulación condicional 16 (derecha) (Elaboración propia).*

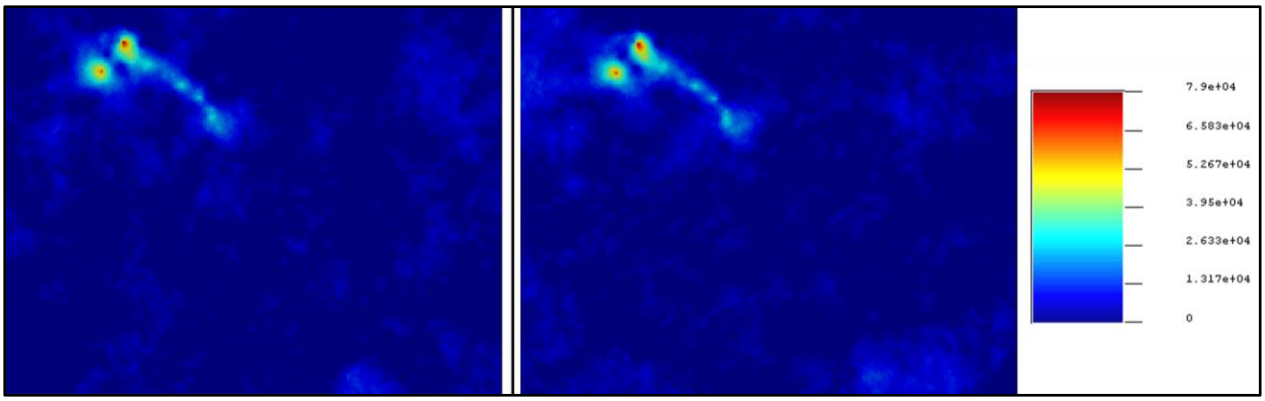

Ilustración E. 15: *Simulación condicional 17 (izquierda) y simulación condicional 18 (derecha) (Elaboración propia).*

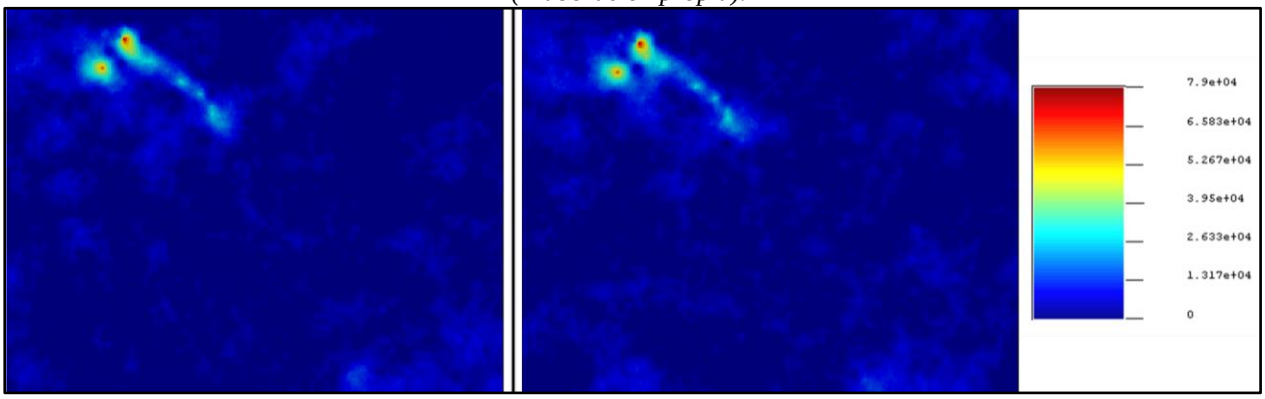

Ilustración E. 16: *Simulación condicional 19 (izquierda) y simulación condicional 20 (derecha) (Elaboración propia).*

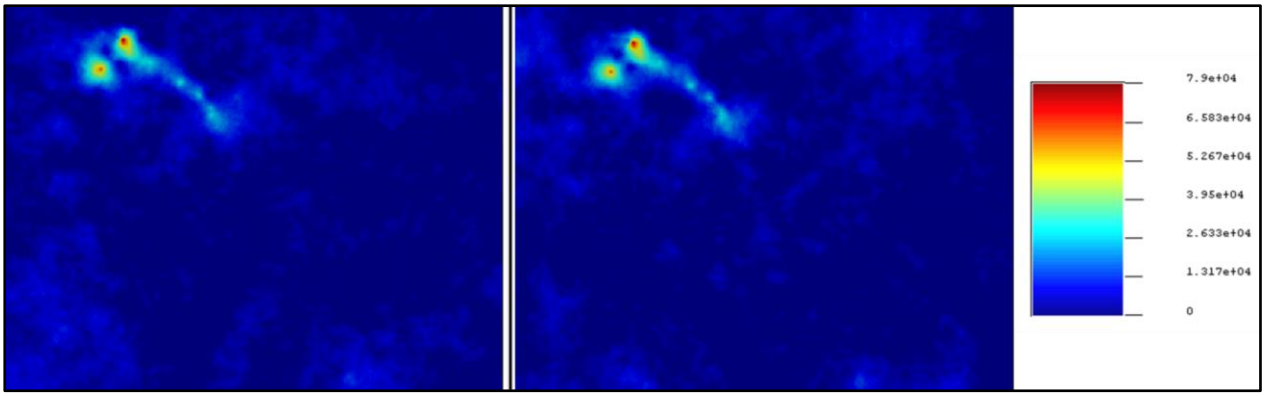

Ilustración E. 17: *Simulación condicional 21 (izquierda) y simulación condicional 22 (derecha) (Elaboración propia).*

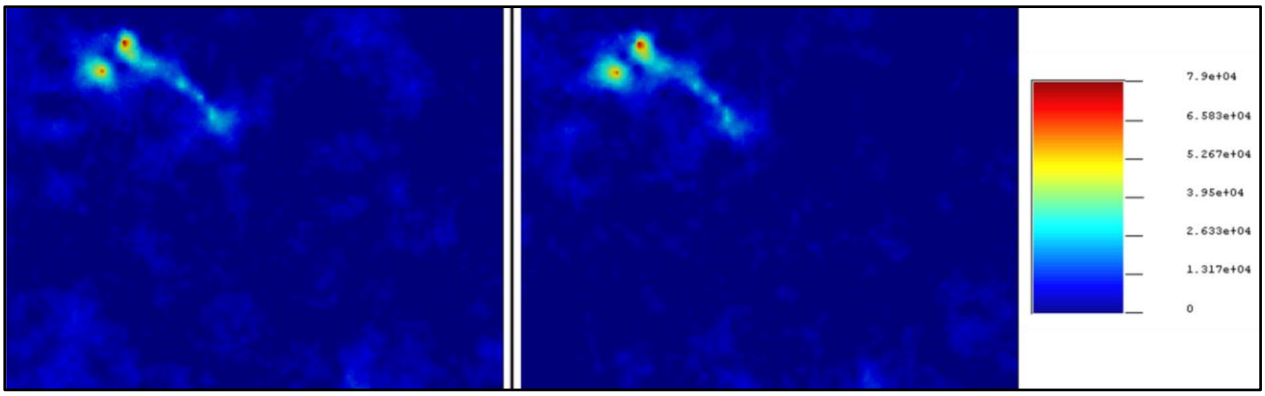

Ilustración E. 18: *Simulación condicional 23 (izquierda) y simulación condicional 24 (derecha) (Elaboración propia).*

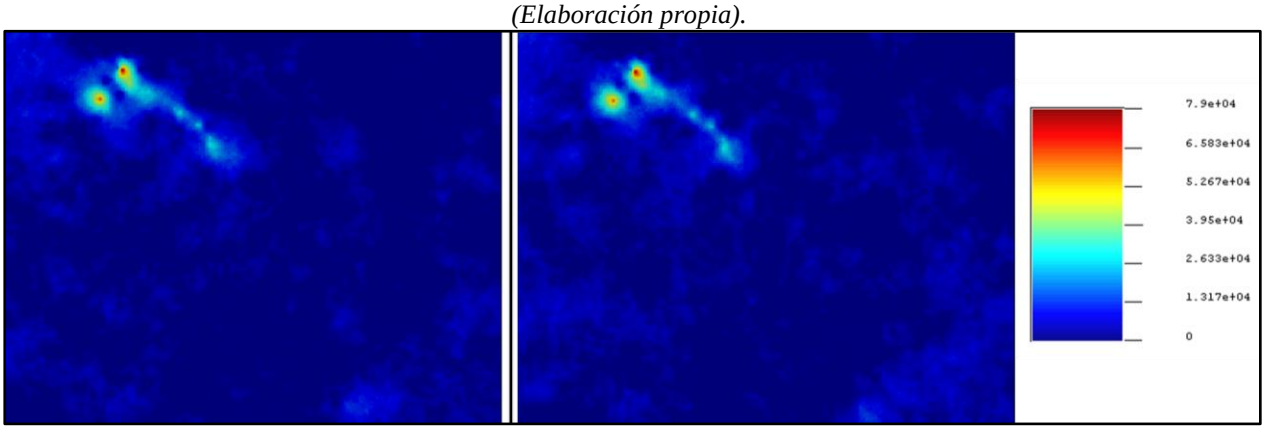

Ilustración E. 19: *Simulación condicional 25 (izquierda) y simulación condicional 26 (derecha) (Elaboración propia).*

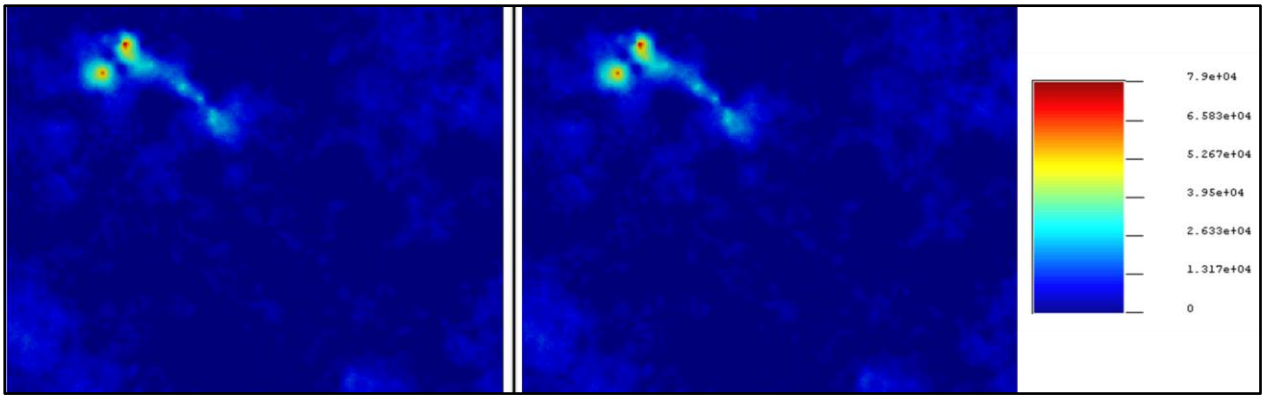

Ilustración E. 20: *Simulación condicional 27 (izquierda) y simulación condicional 28 (derecha) (Elaboración propia).*

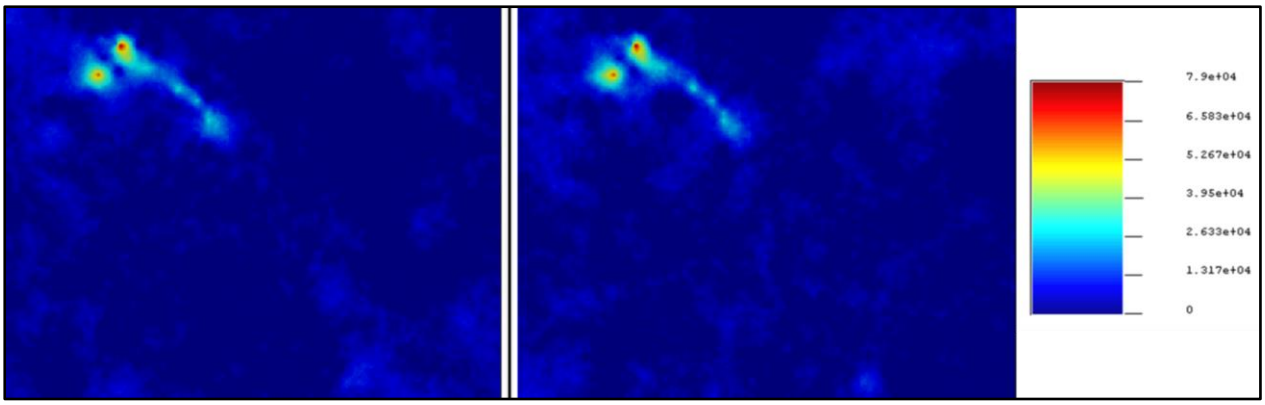

Ilustración E. 21: *Simulación condicional 29 (izquierda) y simulación condicional 30 (derecha) (Elaboración propia).*

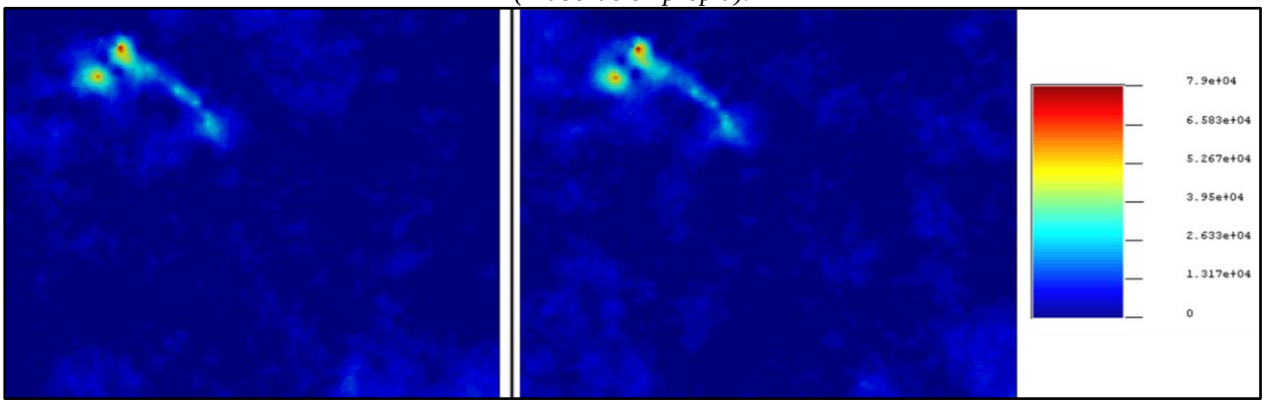

Ilustración E. 22: *Simulación condicional 31 (izquierda) y simulación condicional 32 (derecha) (Elaboración propia).*

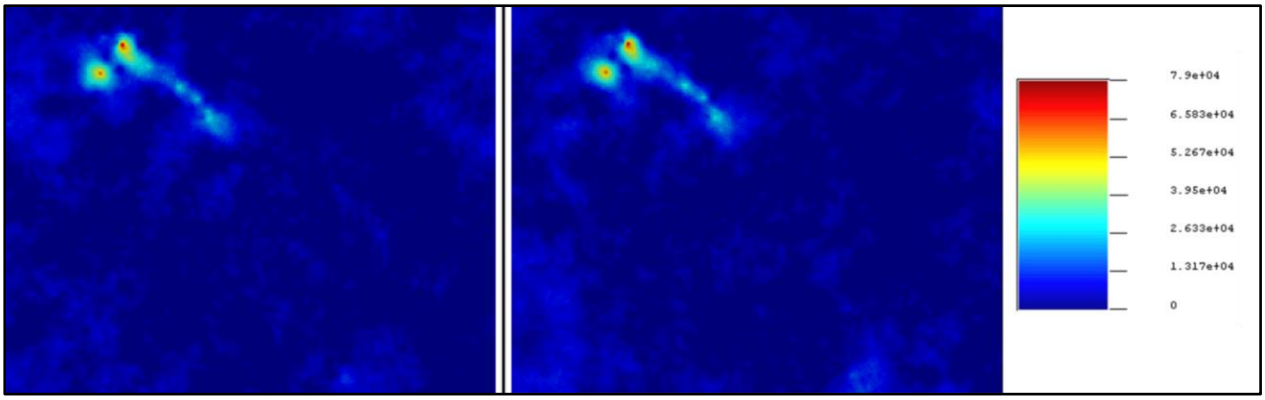

Ilustración E. 23: *Simulación condicional 33 (izquierda) y simulación condicional 34 (derecha) (Elaboración propia).*

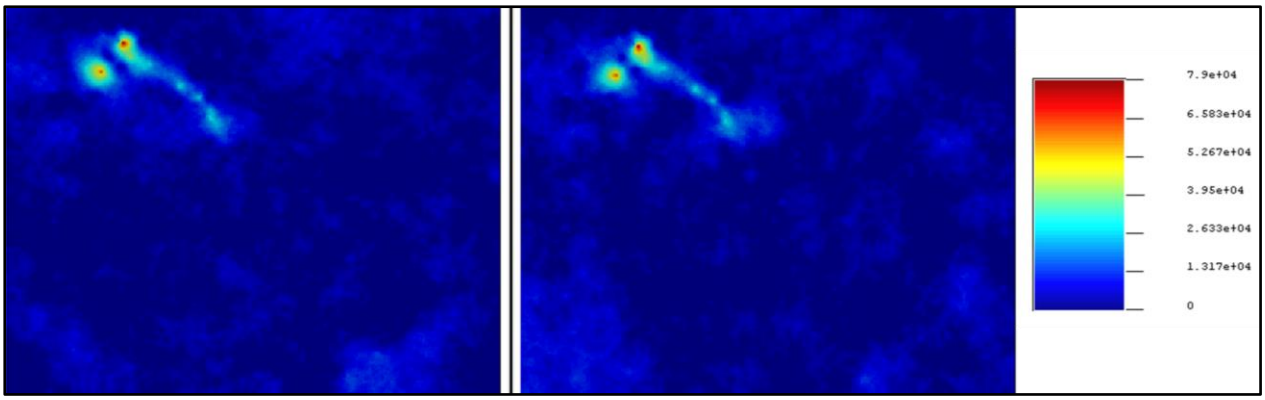

Ilustración E. 24: *Simulación condicional 35 (izquierda) y simulación condicional 36 (derecha) (Elaboración propia).*

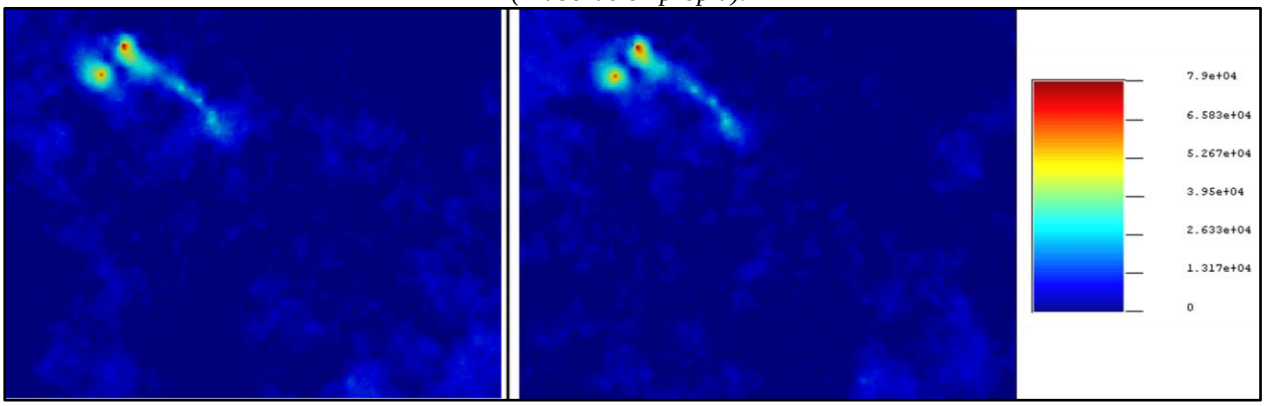

Ilustración E. 25: *Simulación condicional 37 (izquierda) y simulación condicional 38 (derecha) (Elaboración propia).*

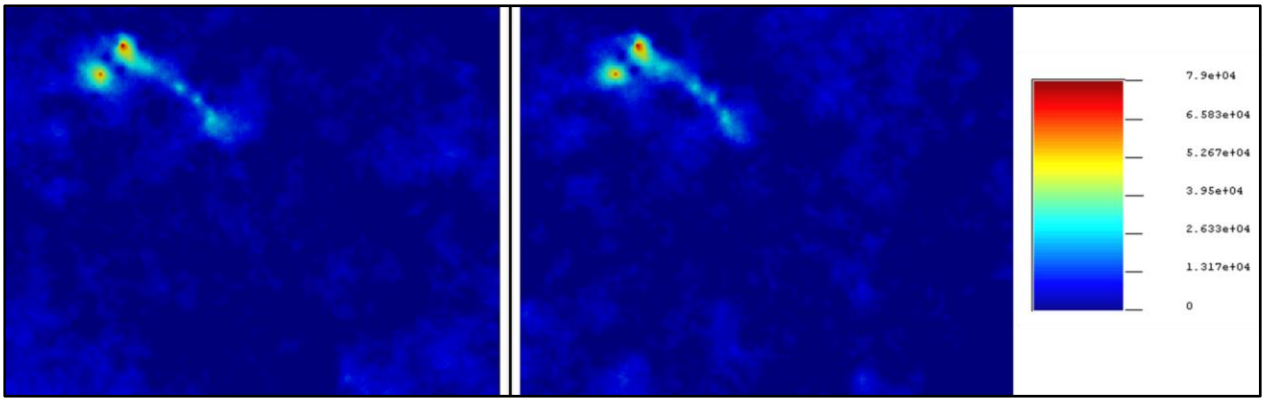

Ilustración E. 26: *Simulación condicional 39 (izquierda) y simulación condicional 40 (derecha) (Elaboración propia).*

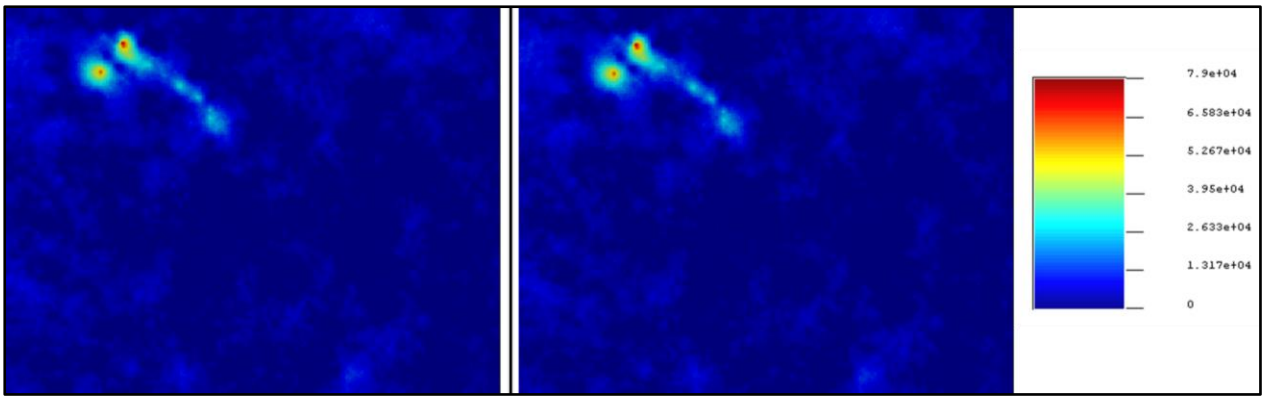

Ilustración E. 27: *Simulación condicional 41 (izquierda) y simulación condicional 42 (derecha) (Elaboración propia).*

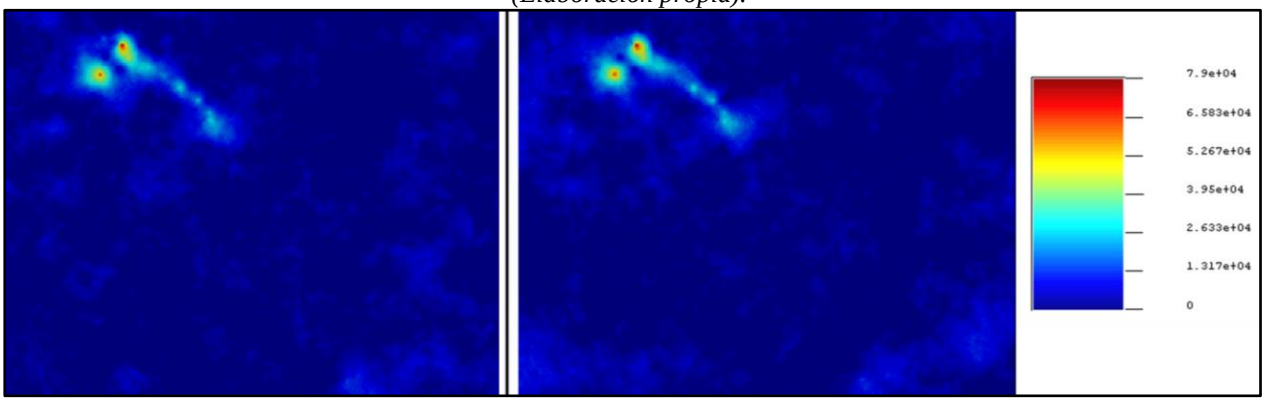

Ilustración E. 28: *Simulación condicional 43 (izquierda) y simulación condicional 44 (derecha) (Elaboración propia).*

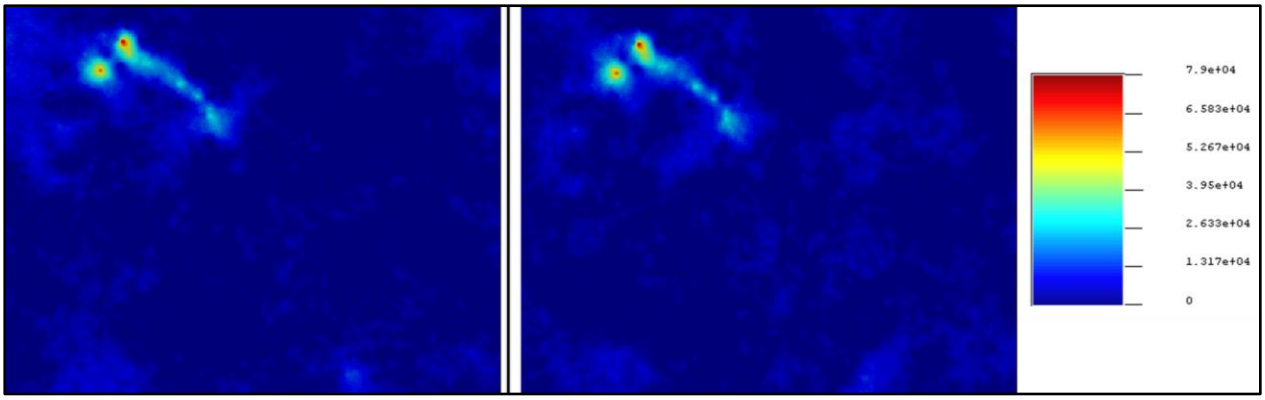

Ilustración E. 29: *Simulación condicional 45 (izquierda) y simulación condicional 46 (derecha) (Elaboración propia).*

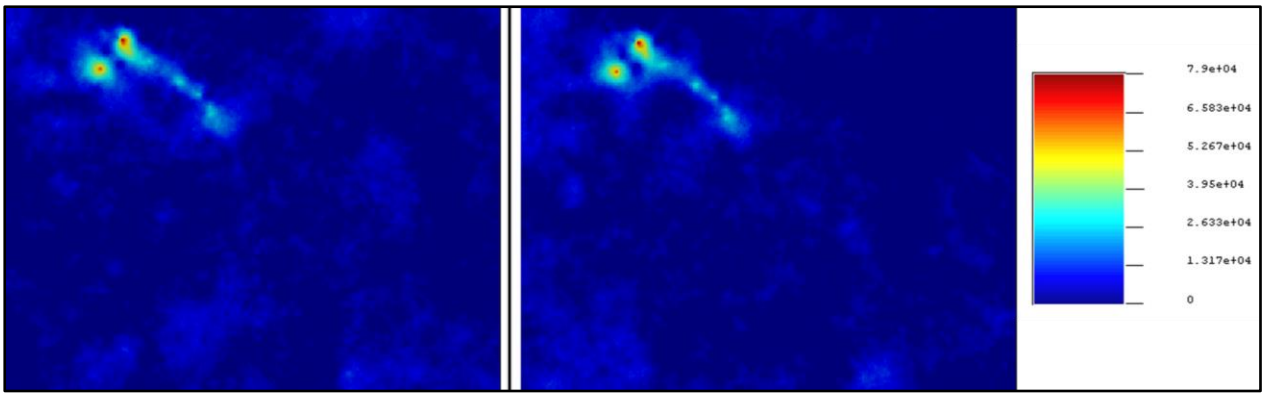

Ilustración E. 30*: Simulación condicional 47 (izquierda) y simulación condicional 48 (derecha) (Elaboración propia).*

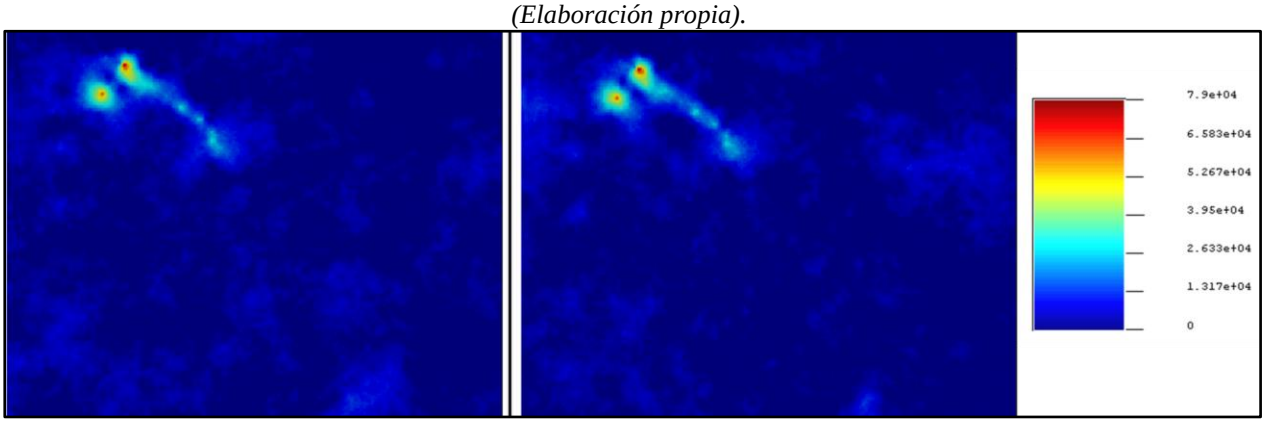

Ilustración E. 31: *Simulación condicional 49 (izquierda) y simulación condicional 50 (derecha) (Elaboración propia).*

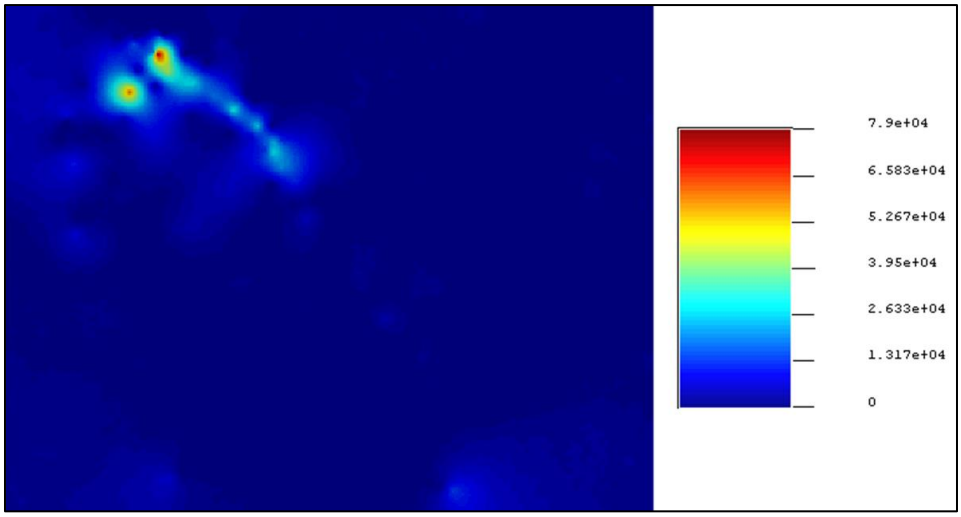

Ilustración E. 32: *Estimación concentración media HC bloques 2x2 (Elaboración propia).*

Apéndice F: Estimación machine learning

En este apéndice, se muestran resultados de estimación para casos que utilizaron como información de entrada distancia entre muestras y no aparecen en el cuerpo principal de la memoria.

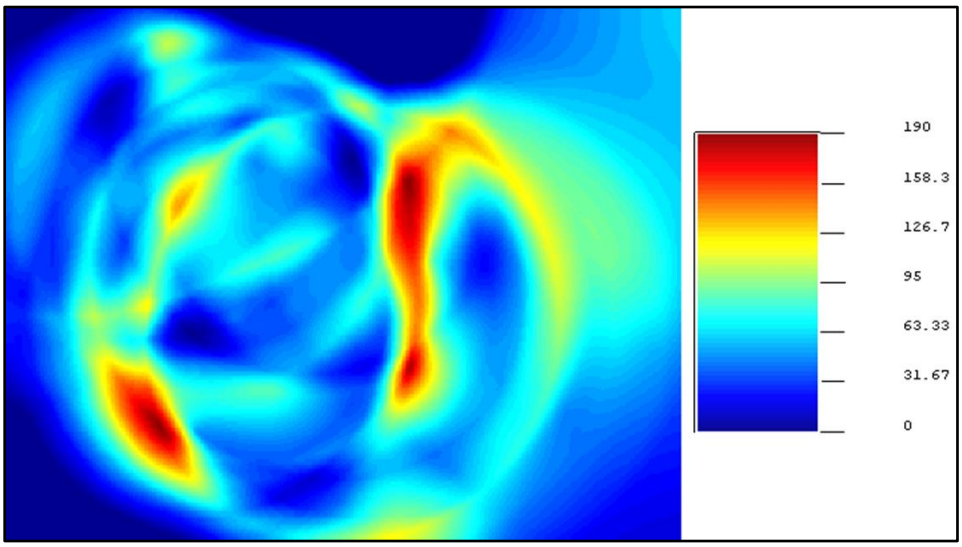

Ilustración F. 1: *Estimación modelo NN 1000_2000 variable continua Ni, caso distancia entre puntos como entrada (Elaboración propia)*

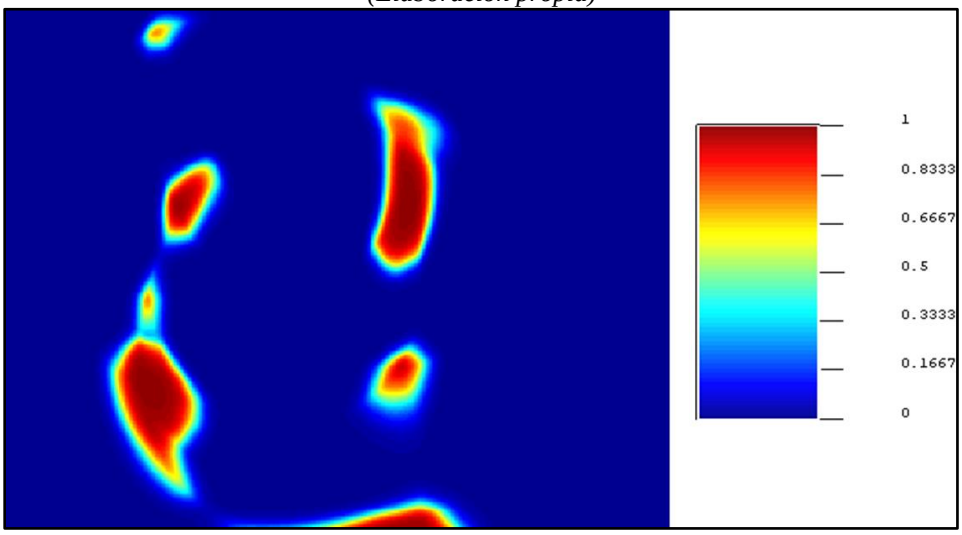

Ilustración F. 2*: Estimación modelo NN 1000_2000 variable categórica Ni, caso distancia entre puntos como entrada (Elaboración propia)*

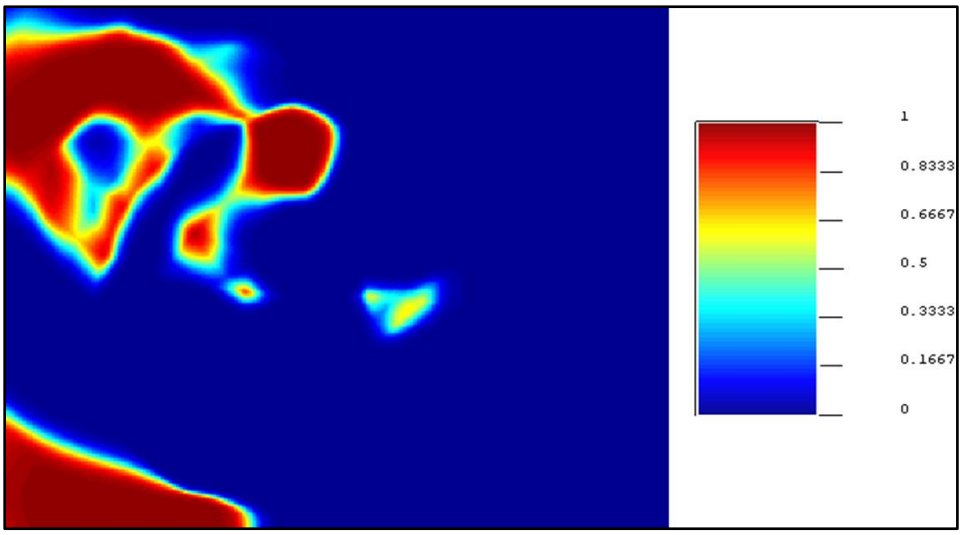

Ilustración F. 3: *Estimación modelo NN 1000_2000 variable categórica HC, caso 210 muestras con distancia entre puntos como entrada (Elaboración propia)*

Apéndice G: Entrenamiento y prueba modelos machine learning

En este apéndice se muestran los resultados de entrenamiento y prueba de los casos que se mencionan, pero que no aparecen en el cuerpo principal de la memoria.

Variables continuas: Níquel

Desde tabla G.1 hasta G.6 se muestran los resultados de fase de entrenamiento y prueba del níquel cuando trabaja con variables continuas.

Tabla G. 1: *Entrenamiento redes neuronales Ni variables continuas (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	19.225	0.901	0.819	0.901	0.090
NN 1000_2000	121.850	0.908	0.825	0.908	0.091

Tabla G. 2: *Matriz de confusión modelo NN 100_2000 con datos de entrenamiento Ni variables continuas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	138	1	139
	1	14	0	14
	Total	152	1	153

Tabla G. 3: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento Ni variables continuas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	139	0	139
	1	14	0	14
	Total	153	0	153

Tabla G. 4: *Prueba redes neuronales Ni variables continuas (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	54.391	0.923	0.852	0.923	-
NN 1000_2000	54.019	0.923	0.852	0.923	-

Tabla G. 5: *Matriz de confusión modelo NN 100_2000 con datos de prueba Ni variables continuas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	36	0	36
	1	3	0	3
	Total	39	0	39

Tabla G. 6: *Matriz de confusión modelo NN 100_2000 con datos de prueba Ni variables continuas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	36	0	36
	1	3	0	3
	Total	39	0	39

Variables categóricas: Hidrocarburo

Desde tabla G.7 hasta G.12 se muestran los resultados de fase de entrenamiento y prueba del hidrocarburo cuando trabaja con variables categóricas.

Tabla G. 7: *Entrenamiento redes neuronales HC variables categóricas (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	13.711	0.810	0.804	0.810	0.695
NN 1000_2000	116.534	0.758	0.766	0.758	0.697

Tabla G. 8: *Matriz de confusión modelo NN 100_2000 con datos de entrenamiento HC variables categóricas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	96	11	107
	1	18	28	46
	Total	114	39	153

Tabla G. 9: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento HC variables categóricas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	86	21	107
	1	16	30	46
	Total	102	51	153

Tabla G. 10: *Prueba redes neuronales HC variables categóricas (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	0.769	0.790	0.463	0.759	-
NN 1000_2000	0.795	0.807	0.435	0.770	-

Tabla G. 11: *Matriz de confusión modelo NN 100_2000 con datos de prueba HC variables categóricas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	21	6	27
	1	4	8	12
	Total	25	14	39

Tabla G. 12: *Matriz de confusión modelo NN 1000_2000 con datos de prueba HC variables categóricas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	23	4	27
	1	4	8	12
	Total	27	12	39

Variables categóricas: Níquel

Desde tabla G.13 hasta G.18 se muestran los resultados de fase de entrenamiento y prueba del níquel cuando trabaja con variables categóricas.

Tabla G. 13: *Entrenamiento redes neuronales Ni variables categóricas (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	3.025	0.908	0.825	0.908	0.091
NN 1000_2000	56.279	0.882	0.843	0.882	0.153

Tabla G. 14: *Matriz de confusión modelo NN 100_2000 con datos de entrenamiento Ni variables categóricas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	139	0	139
	1	14	0	14
	Total	153	0	153

Tabla G. 15: *Matriz de confusión modelo NN 1000_200 con datos de entrenamiento Ni variables categóricas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	134	5	139
	1	13	1	14
	Total	147	6	153

Tabla G. 16: *Prueba redes neuronales Ni variables categóricas (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	0.923	0.852	0.265	0.077	-
NN 1000_2000	0.897	0.850	0.324	0.075	-

Tabla G. 17: *Matriz de confusión modelo NN 100_2000 con datos de prueba Ni variables categóricas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	36	0	36
	1	3	0	3
	Total	39	0	39

Tabla G. 18: *Matriz de confusión modelo NN 1000_2000 con datos de prueba Ni variables categóricas (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	35	1	36
	1	3	0	3
	Total	38	1	39

Variables categóricas: Caso hidrocarburo con 210 muestras

Desde tabla G.19 hasta G.24 se muestran los resultados de fase de entrenamiento y prueba del hidrocarburo cuando trabaja con 210 muestras.

Tabla G. 19: *Entrenamiento redes neuronales caso HC 210 muestras (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	8.375	0.815	0.807	0.815	0.659
NN 1000_2000	50.680	0.815	0.812	0.815	0.700

Tabla G. 20: *Matriz de confusión modelo NN 100_2000 con datos de entrenamiento, caso HC 210 muestras (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	110	12	122
	1	22	24	46
	Total	132	36	168

Tabla G. 21: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento, caso HC 210 muestras (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	108	14	122
	1	15	31	46
	Total	123	45	168

Tabla G. 22: *Prueba redes neuronales, caso HC 210 muestras (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	0.690	0.715	0.666	0.626	-
NN 1000_2000	0.714	0.747	0.758	0.686	-

Tabla G. 23: *Matriz de confusión modelo NN 100_2000 con datos de prueba, caso HC 210 muestras (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	22	8	30
	1	5	7	12
	Total	27	15	42

Tabla G. 24: *Matriz de confusión modelo NN 1000_2000 con datos de prueba, caso HC 210 muestras (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	22	8	30
	1	4	8	12
	Total	26	16	42

Variables categóricas: Caso hidrocarburo muestreado con grilla

Desde tabla G.25 hasta G.30 se muestran los resultados de fase de entrenamiento y prueba del hidrocarburo muestreado con grilla.

Tabla G. 25: *Entrenamiento redes neuronales caso HC muestreado con grilla (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	7.662	0.692	0.702	0.692	0.594
NN 1000_2000	51.043	0.738	0.762	0.738	0.706

Tabla G. 26: *Matriz de confusión modelo NN 100_2000 con datos de entrenamiento, caso HC muestreado con grilla (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	35	11	46
	1	9	10	19
	Total	44	21	65

Tabla G. 27: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento, caso HC muestreado con grilla (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	35	11	46
	1	6	13	19
	Total	41	24	65

Tabla G. 28: *Prueba redes neuronales, caso HC muestreado con grilla (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 100_2000	0.780	0.783	0.539	0.717	-
NN 1000_2000	0.780	0.791	1.145	0.745	-

Tabla G. 29: *Matriz de confusión modelo NN 100_2000 con datos de prueba, caso HC muestreado con grilla (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	73	15	88
	1	13	26	39
	Total	86	41	127

Tabla G. 30: *Matriz de confusión modelo NN 1000_2000 con datos de prueba, caso HC muestreado con grilla (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	71	17	88
	1	11	28	39
	Total	82	45	127

Caso distancia entre muestras como información de entrada

Desde tabla G.25 hasta G.30 se muestran los resultados de fase de entrenamiento y prueba de modelos que estimaron utilizando como información de entrada distancia entre muestras.

Tabla G. 31: *Entrenamiento redes neuronales variable continua HC con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	627.790	0.692	0.765	0.692	0.757

Tabla G. 32: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento, variable continua HC con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	68	38	106
	1	9	38	47
	Total	77	76	153

Tabla G. 33: *Prueba redes neuronales variable continua HC con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	13052.754	0.538	0.681	0.538	-

Tabla G. 34: *Matriz de confusión modelo NN 1000_2000 con datos de prueba, variable continua HC con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	13	15	28
	1	3	8	11
	Total	16	23	39

Tabla G. 35: *Entrenamiento redes neuronales variable continua Ni con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	433.622	0.843	0.861	0.843	0.341

Tabla G. 36: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento, variable continua Ni con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	125	14	139
	1	10	4	14
	Total	**135**	**18**	**153**

Tabla G. 37: *Prueba redes neuronales variable continua Ni con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	59.275	0.820	0.879	0.820	-

Tabla G. 38: *Matriz de confusión modelo NN 1000_2000 con datos de prueba, variable continua Ni con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	31	5	36
	1	2	1	3
	Total	**33**	**6**	**39**

Tabla G. 39: *Entrenamiento redes neuronales variable categórica HC con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	188.696	0.718	0.720	0.718	0.618

Tabla G. 40: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento, variable categórica HC con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	85	22	107
	1	21	25	46
	Total	106	47	153

Tabla G. 41: *Prueba redes neuronales variable categórica HC con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	0.821	0.817	0.774	0.735	-

Tabla G. 42: *Matriz de confusión modelo NN 1000_2000 con datos de prueba, variable categórica HC con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	24	3	27
	1	4	8	12
	Total	28	11	39

Tabla G. 43: *Entrenamiento redes neuronales variable categórica Ni con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	108.726	0.849	0.833	0.849	0.149

Tabla G. 44: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento, variable categórica Ni con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	129	10	139
	1	13	1	14
	Total	142	11	153

Tabla G. 45: *Prueba redes neuronales, variable categórica Ni con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	0.923	0.912	0.923	0.418	-

Tabla G. 46: *Matriz de confusión modelo NN 1000_2000 con datos de prueba, variable categórica Ni con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	35	1	36
	1	2	1	3
	Total	37	2	39

Tabla G. 47: *Entrenamiento redes neuronales variable categórica caso HC 210 muestras con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	163.300	0.809	0.804	0.809	0.684

Tabla G. 48: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento, variable categórica caso HC 210 muestras con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	108	14	122
	1	18	28	46
	Total	126	42	168

Tabla G. 49: *Prueba redes neuronales, variable categórica caso HC 210 muestras con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo prueba (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	0.714	0.720	2.121	0.636	-

Tabla G. 50: *Matriz de confusión modelo NN 1000_2000 con datos de prueba, variable categórica caso HC 210 muestras con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	23	7	30
	1	5	7	12
	Total	28	14	42

Tabla G. 51: *Entrenamiento redes neuronales variable categórica caso HC muestreado con grilla y con distancia entre puntos (Elaboración propia).*

Modelo	Tiempo entrenamiento (s)	Exactitud	Precisión	Exhaustividad	Especificidad
NN 1000_2000	30.385	0.723	0.742	0.723	0.669

Tabla G. 52: *Matriz de confusión modelo NN 1000_2000 con datos de entrenamiento, variable categórica caso HC muestreado con grilla y con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	35	11	46
	1	7	12	19
	Total	42	23	65

Tabla G. 53: *Prueba redes neuronales, variable categórica caso HC muestreado con grilla y con distancia entre puntos (Elaboración propia).*

Modelo	**Tiempo prueba (s)**	**Exactitud**	**Precisión**	**Exhaustividad**	**Especificidad**
NN 1000_2000	0.772	0.773	1.915	0.699	-

Tabla G. 54: *Matriz de confusión modelo NN 1000_2000 con datos de prueba, variable categórica caso HC muestreado con grilla y con distancia entre puntos (Elaboración propia).*

		Predicción		
		0	1	**Total**
Real	0	73	15	88
	1	14	25	39
	Total	87	40	127

Apéndice H: Zonas contaminadas métodos geoestadísticos

En este apéndice, se presentan las zonas que superan criterio de contaminación por métodos geoestadísticos.

Kriging ordinario

Las ilustraciones G.1 y G.2 muestran zonas que superan criterio de contaminación cuando se utiliza bloques 10x10.

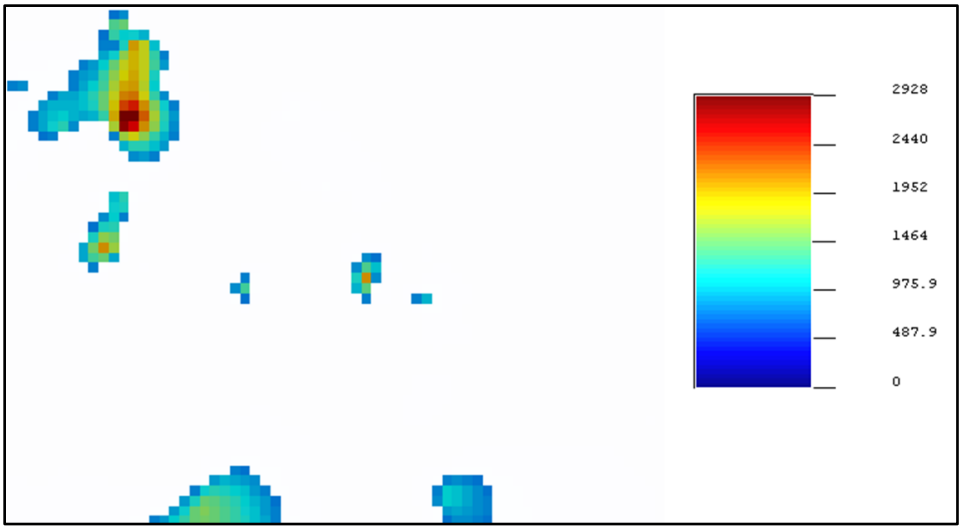

Ilustración G. 1: *Zonas que superan criterio de contaminación establecido, caso sin datos atípicos HC bloques 10x10 (Elaboración propia).*

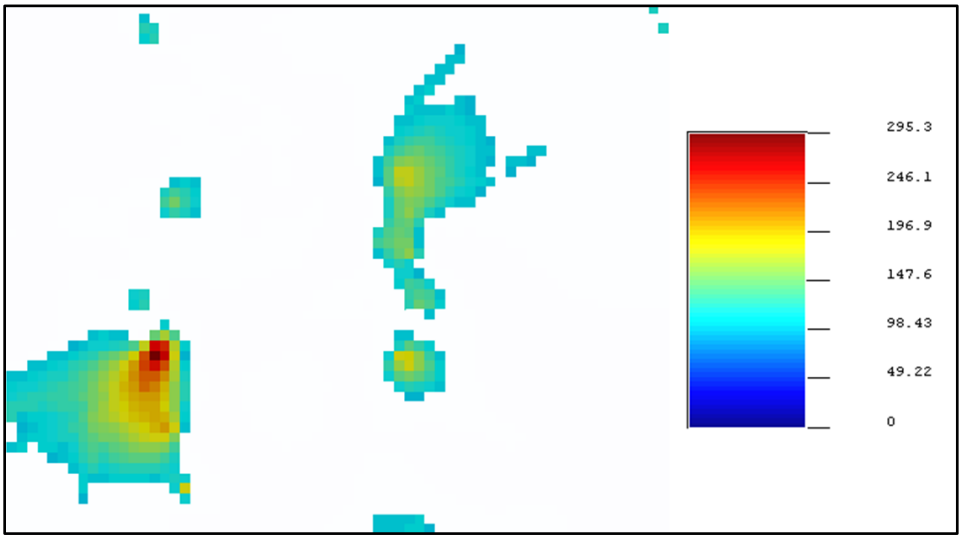

Ilustración G. 2: *Zonas que superan criterio de contaminación establecido Ni bloques 10x10 (Elaboración propia).*

Kriging de indicadores

Las siguientes ilustraciones muestran las zonas con probabilidad de superar criterio de contaminación.

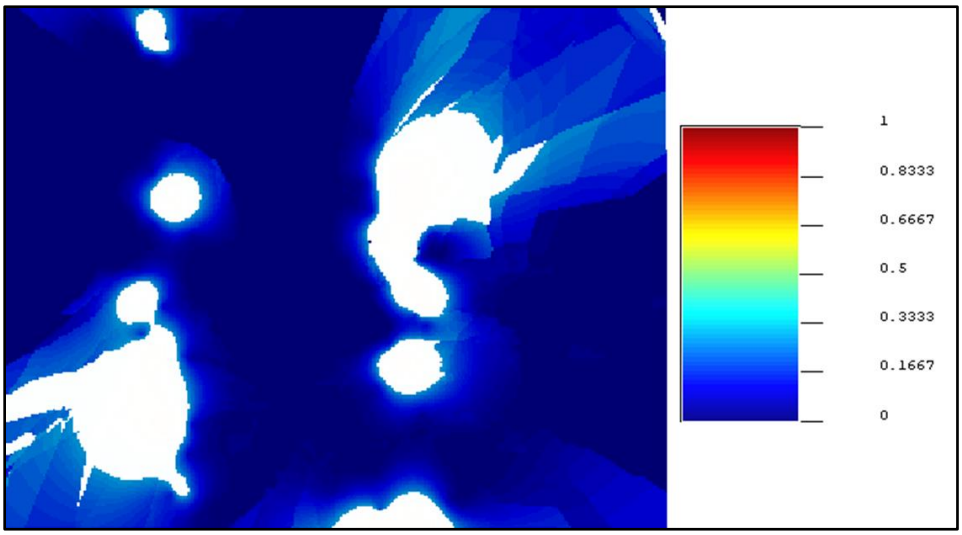

Ilustración G. 3: *Zonas con bajo riesgo de contaminación, kriging de indicadores Ni (Elaboración propia).*

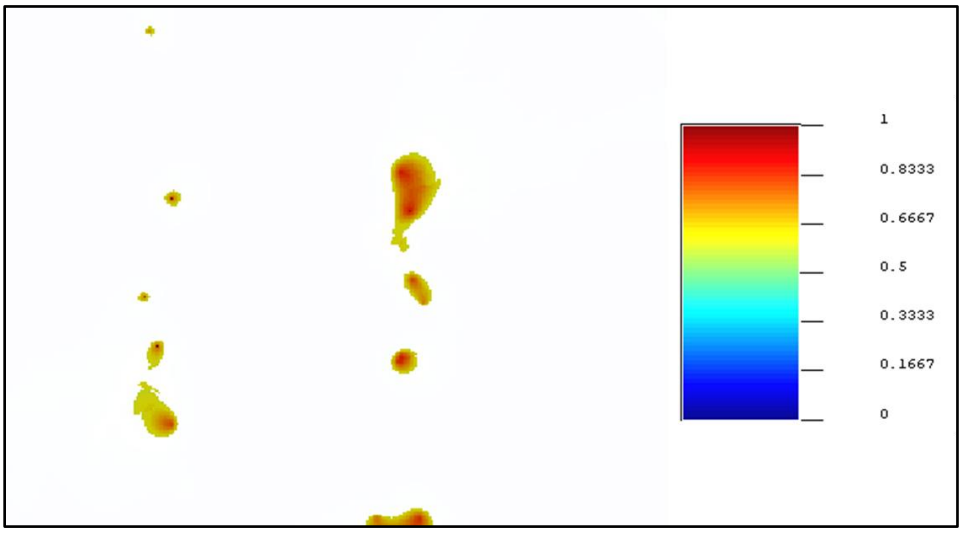

Ilustración G. 4: *Zonas con alto riesgo de contaminación, kriging de indicadores Ni (Elaboración propia).*

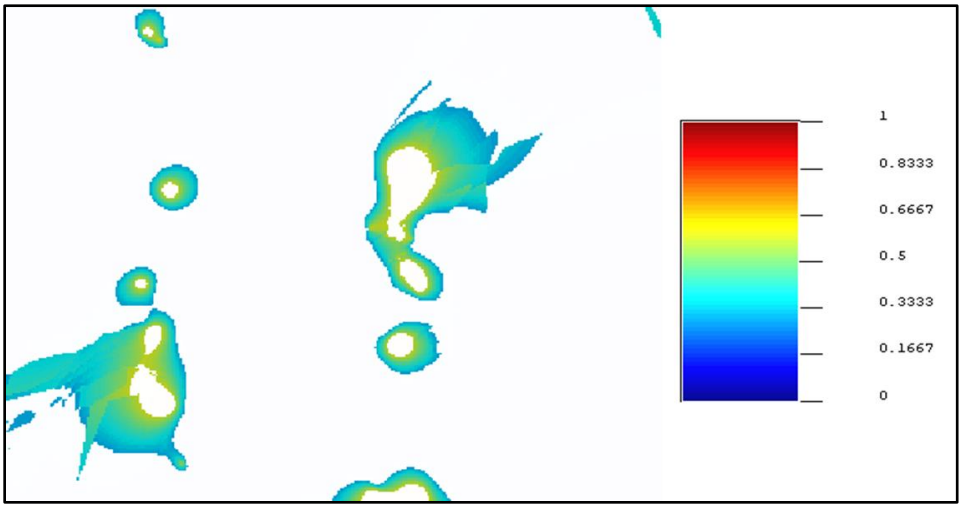

Ilustración G. 5: *Zonas de incertidumbre, kriging de indicadores Ni (Elaboración propia).*

Apéndice I: Zonas contaminadas método machine learning

En este apéndice, se presentan las zonas que superan criterio de contaminación por método machine learning.

Desde la ilustración I.1 hasta I.3, corresponde a las zonas con probabilidad de superar criterio de contaminación cuando modelo NN 1000_2000 estima variables categóricas **sin conocer la distancia entre muestras.** En cambio, desde la ilustración I.4 hasta I.13 las estimaciones fueron realizadas conociendo la **distancia entre muestras.**

Variables categóricas Ni sin distancia entre muestras

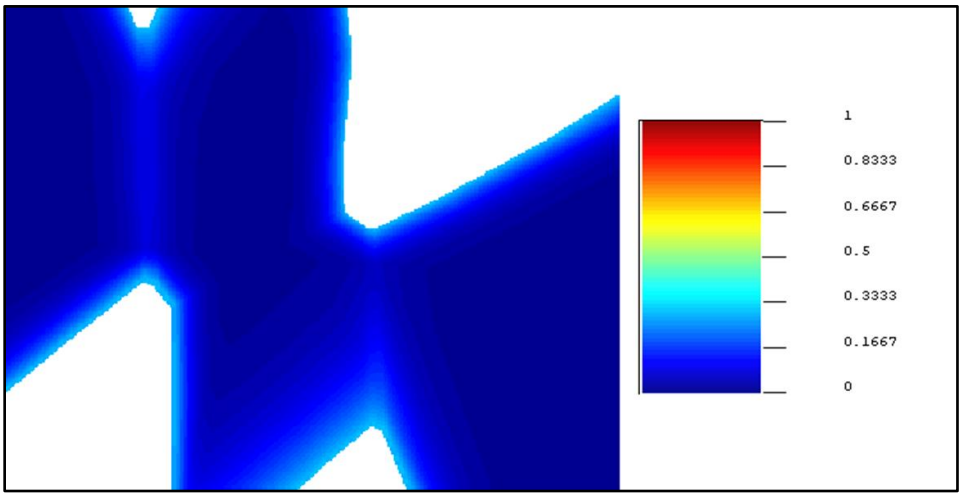

Ilustración I. 1: *Zonas con bajo riesgo de contaminación, modelo NN 1000_2000 variables categóricas Ni (Elaboración propia).*

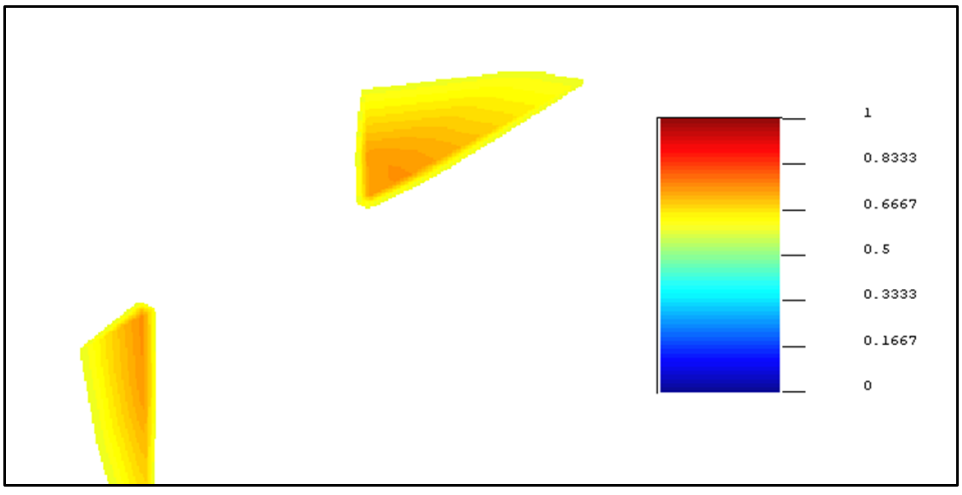

Ilustración I. 2: *Zonas con alto riesgo de contaminación, modelo NN 1000_2000 variables categóricas Ni (Elaboración propia).*

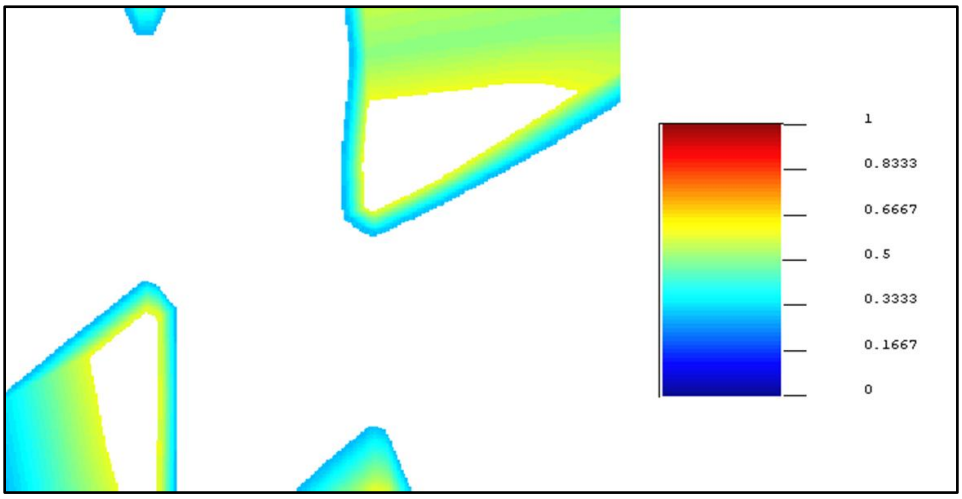

Ilustración I. 3: *Zonas de incertidumbre, modelo NN 1000_2000 variables categóricas Ni (Elaboración propia).*

Variables categóricas HC con distancias entre muestras como información de entrada

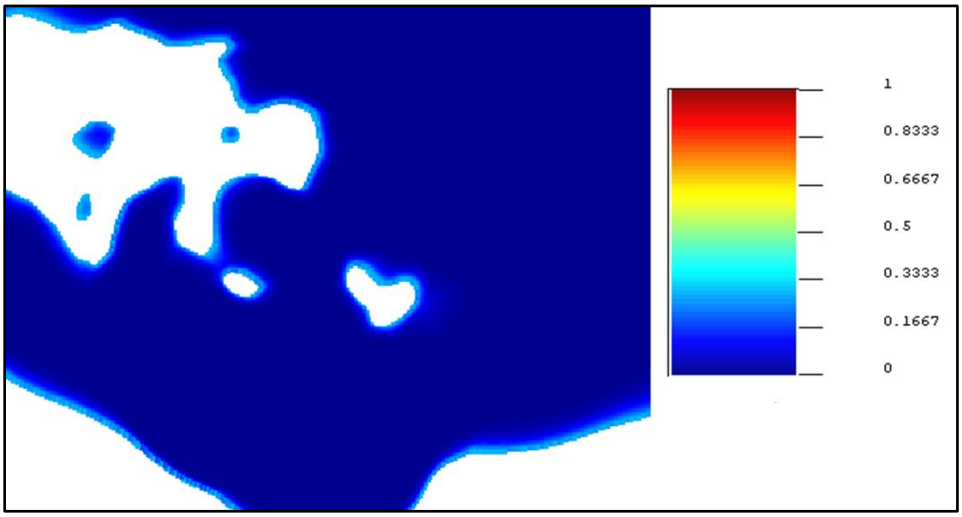

Ilustración I. 4: *Zonas con bajo riesgo de contaminación, modelo NN 1000_2000 variables categóricas HC con distancia entre muestras (Elaboración propia).*

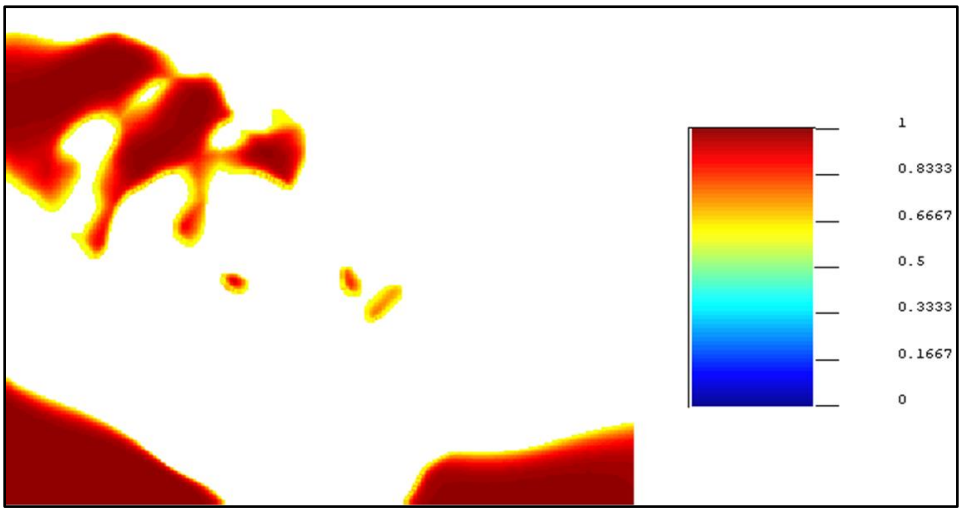

Ilustración I. 5: *Zonas con alto riesgo de contaminación, modelo NN 1000_2000 variables categóricas HC con distancia entre muestras (Elaboración propia).*

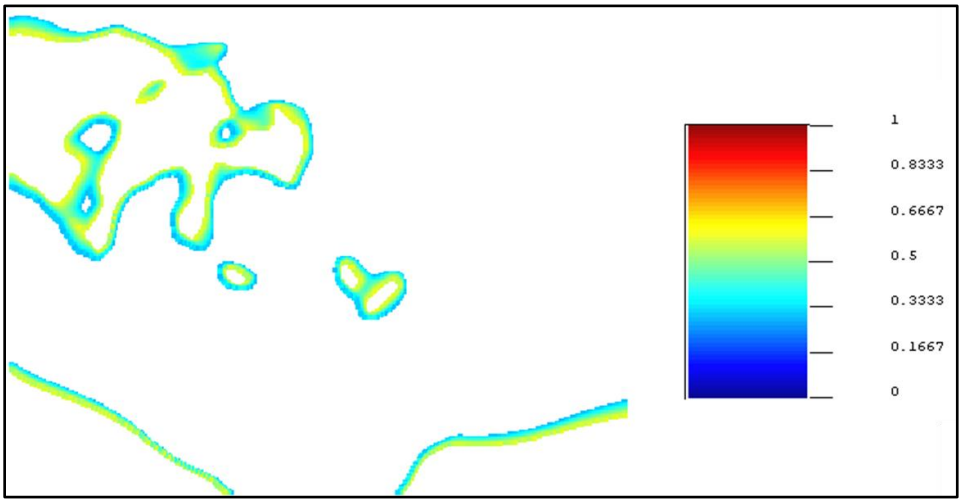

Ilustración I. 6: *Zonas de incertidumbre, modelo NN 1000_2000 variables categóricas HC con distancia entre muestras (Elaboración propia).*

Variables categóricas Ni con distancias entre muestras como información de entrada

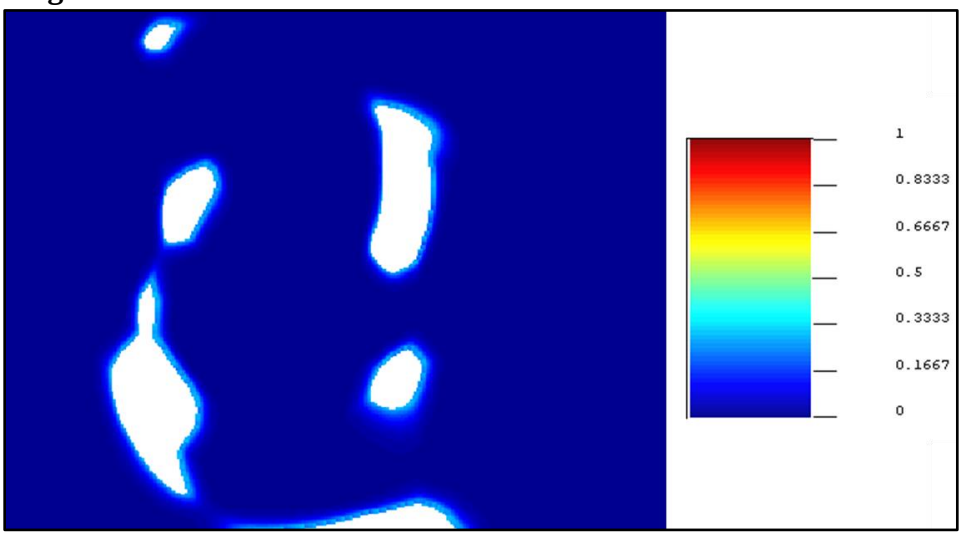

Ilustración I. 7: *Zonas con bajo riesgo de contaminación, modelo NN 1000_2000 variables categóricas Ni con distancia entre muestras (Elaboración propia).*

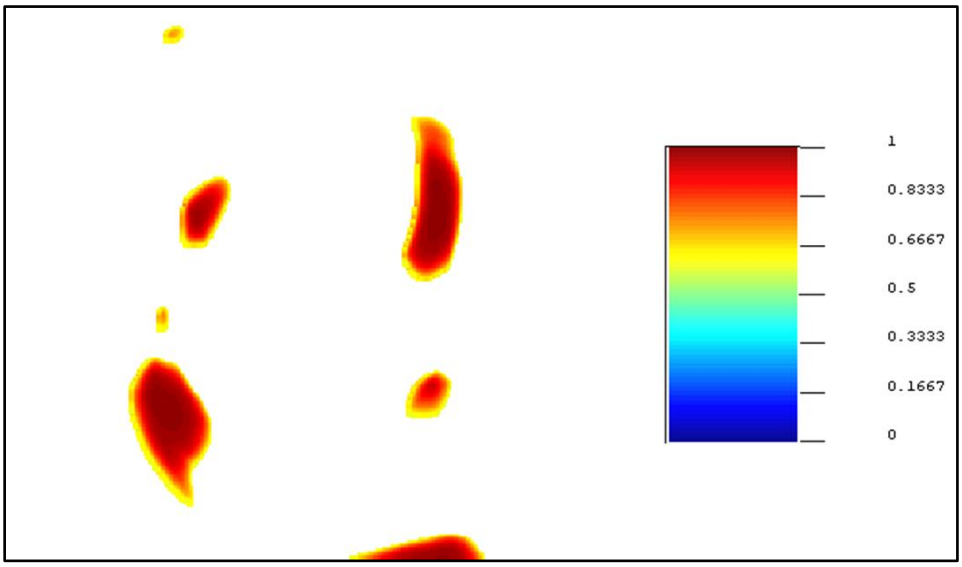

Ilustración I. 8: *Zonas con alto riesgo de contaminación, modelo NN 1000_2000 variables categóricas Ni con distancia entre muestras (Elaboración propia).*

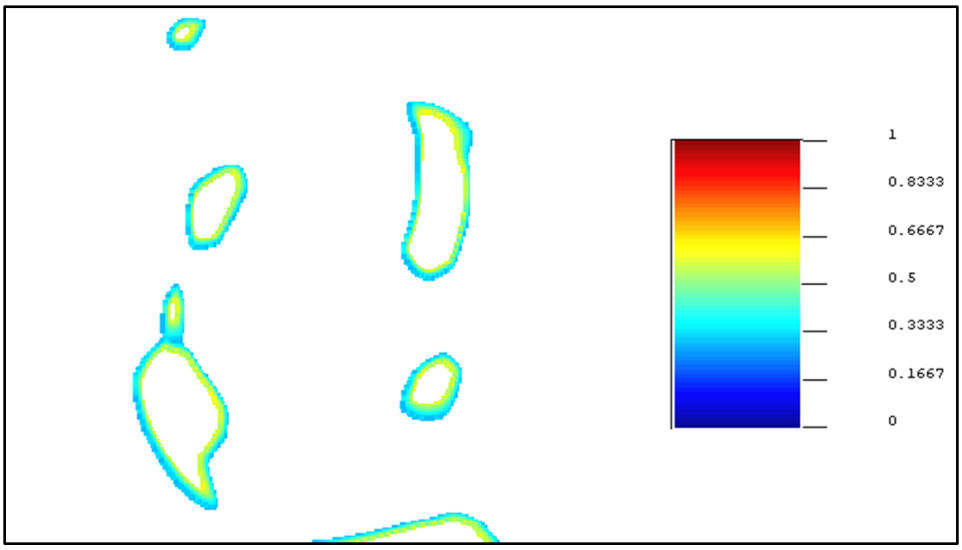

Ilustración I. 9: *Zonas de incertidumbre, modelo NN 1000_2000 variables categóricas Ni con distancia entre muestras (Elaboración propia).*

Caso HC 210 muestras con distancia entre muestras como información de entrada

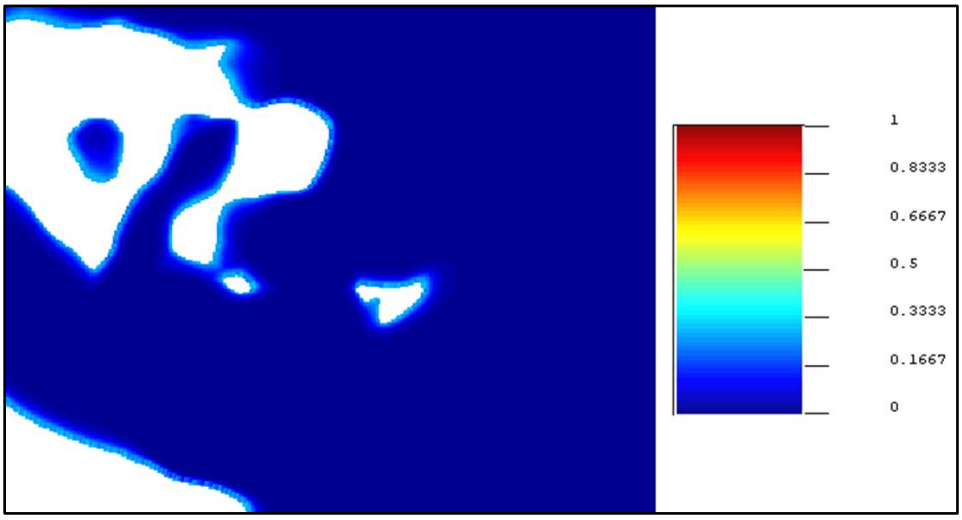

Ilustración I. 10: *Zonas con bajo riesgo de contaminación, modelo NN 1000_2000 variables categóricas caso HC 210 muestras conociendo distancia entre muestras (Elaboración propia).*

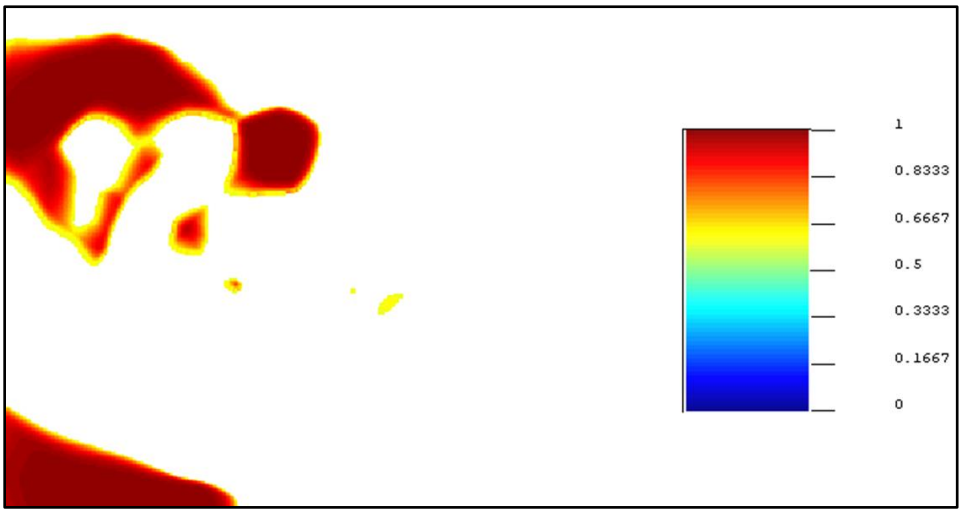

Ilustración I. 11: *Zonas con alto riesgo de contaminación, modelo NN 1000_2000 variables categóricas caso HC 210 muestras conociendo distancia entre muestras (Elaboración propia).*

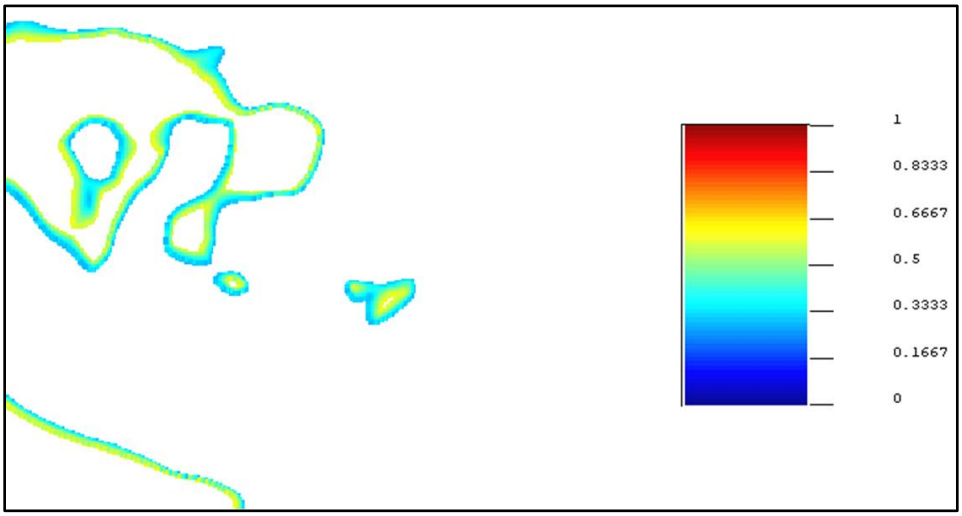

Ilustración I. 12: Zonas de incertidumbre, modelo NN 1000_2000 variables categóricas caso HC 210 muestras conociendo distancia entre muestras (Elaboración propia).

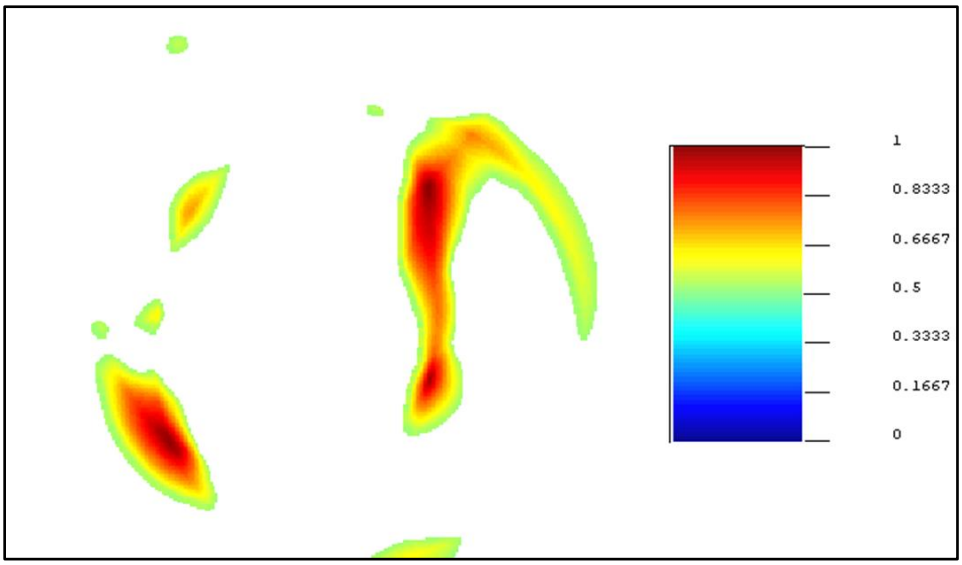

Ilustración I. 13: *Zonas que superan criterio de contaminación establecido, variable continua Ni con distancia entre muestras (Elaboración propia).*