UNIVERSIDAD DE SALAMANCA
DEPARTAMENTO DE ESTADÍSTICA

Modelo de predicción subespacial: Regresión Multivariante Gaussiana Subespacial

Víctor Vicente Palacios

2017

Modelo de predicción subespacial: Regresión Multivariante Gaussiana Subespacial

---

Memoria para optar al Grado de Doctor por el Departamento de Estadística de la Universidad de Salamanca presentada por:

Víctor Vicente Palacios

Salamanca

2017

VNiVERSiDAD
D SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

SANTIAGO VICENTE TAVERA

Profesor del Departamento de Estadística
de la Universidad de Salamanca

CERTIFICA:

Que **Víctor Vicente Palacios**, Ingeniero Industrial, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que, para optar al Grado de Doctor, presenta con el título: "Modelo de predicción subespacial: Regresión Multivariante Gaussiana Subespacial"; y para que conste, firma el presente certificado en Salamanca, en Mayo de 2017.

A mi familia
y a Martina
por su gran apoyo

“Que un gato negro se cruce en tu camino,
significa que el animal va a alguna parte.”
Groucho Marx

Agradecimientos

Son muchas las personas que me han apoyado durante estos tres años y aunque quisiera nombrar a todos, no podría. Todos y cada uno de los que se han interesado, por mínimo que fuera, en mis avances, investigaciones y sobre todo en mis momentos difíciles, merecen una mención. Esta tesis es un poco de cada uno de ellos.

Si hay alguien que ha sabido conjugar tanto apoyo moral como técnico, ha sido mi padre. Admiro y admiraré su espíritu positivo del que yo carezco muchas veces, la decisión con la que afronta los problemas diarios y sobre todo su humildad. Durante todos estos años ha conseguido muchos triunfos personales y aún así ha sabido priorizar aquello que realmente importa, su familia. Gracias.

Quiero agradecer a mi madre por haberme escuchado, haber aguantado charlas larguísimas sobre métodos matemáticos, estadística e informática. Sin sus consejos y su confianza no habría podido sacar adelante nada de lo que he conseguido hasta la fecha.

A mi hermana que ha conseguido contagiarme su coraje en mil ocasiones. Me ha animado en todo momento a seguir adelante y no ha dudado ni un solo momento de mí.

A todos los miembros de mi familia que de una manera u otra, durante toda mi vida, han creído en mí.

A Martina quiero agradecerle haber estado ahí en todo momento, haber se-

guido con mucha atención mis progresos y haberme ayudado a sortear los obstáculos. Ahora ella comienza el mismo camino que yo hace tres años, sólo espero corresponder con el mismo apoyo y amor que ella me ha brindado.

Una especial mención al profesor José Luis Vicente Villardón por sus sabios consejos durante la primera parte de mi tesis.

Al profesor Jhonny R. Demey porque sin él nada de esta tesis habría sido posible, gracias a él descubrí el apasionante mundo de la geoestadística. Descanse en paz.

A Pedro Luis Sánchez y a Adolfo Íñigo por creer en el algoritmo y permitirme aplicarlo a sus investigaciones.

A todos mis compañeros de doctorado por su comprensión y apoyo. Por su implicación y participación activa en DataLab, que sin ellos no sería la realidad que es ahora.

Al Departamento de Estadística por haberme brindado la oportunidad de realizar esta tesis y en especial a todos aquellos que se han interesado en el desarrollo de la misma.

A todos los chicos de la Oficina Verde, por haberme ayudado a sonreír en cada descanso. Voy a extrañar mucho los cafés de las 11 en el Mandala.

A los chicos de Obture, porque sin ellos el primer año habría sido menos llevadero. Gracias por vuestros consejos informáticos y por alegraros tanto de mis éxitos como yo de los vuestros.

A todo Medialab, por haberme aceptado tan bien en su familia. En menos de un año puedo llamarlos amigos y es un honor. Sin ellos DataLab no existiría.

A la Fundación del Centro de Supercomputación de Castilla y León (FCSC)

por habernos prestado sus clústers de computación para desarrollar nuestro algoritmo.

Y por último, y no por ello menos importante quiero agradecer a todos mis amigos por cada uno de los momentos que hemos compartido en estos años. A todos mis amigos de Salamanca, porque con ellos soy un poco más feliz. A los chicos de la resi, que son como mis hermanos. A los amigos de Francia, porque comprendieron y apoyaron mi decisión de volver a casa.

A todos, muchas gracias.

Índice general

1. Introducción **1**

2. Marco Teórico **5**
2.1. Reducción dimensional 6
2.1.1. Análisis de Componentes Principales 7
2.1.2. Análisis de Correspondencias 8
2.1.3. Análisis Factorial 8
2.1.4. Sistema Gifi 9
2.1.5. Métodos Biplot 10
2.2. Procesos Gaussianos 13
2.2.1. Geoestadística 15
2.2.2. Cokriging 29

3. Regresión Multivariante Gaussiana Subespacial **35**
3.1. Coordenadas Subespaciales 36
3.1.1. Variogramas Cruzados Subespaciales 37
3.2. Modelo Lineal Corregionalizado 39
3.2.1. Matriz Varianza-Covarianza Subespacial 40
3.2.2. Algoritmo LMC Subespacial 41
3.2.3. Distribución Cuadrática 42
3.3. Mallado 43
3.4. Cokriging Simple 44
3.5. Validación Cruzada 45
3.6. Predicciones 46
3.7. Software 47

4. Aplicaciones **49**
4.1. El modelo MGSR aplicado a la predicción de los efectos del stent (Postoperatorio y Follow-Up) en enfermos de infarto de miocardio 50
4.1.1. Introducción 50
4.1.2. Resultados 54
4.1.3. Validez y estabilidad del modelo 60
4.2. Predicción de los efectos de la cristalización de fosfatos en el envejecimiento de conglomerados silíceos 72
4.2.1. Introducción 72
4.2.2. Resultados 78

5. Conclusiones **91**

A. Tutorial Software MGSR **99**
A.1. Coordenadas Subespaciales 103
A.2. Variograma Cruzado 105
A.3. Modelo Lineal de Corregionalización (LMC) 107
A.4. Mallado 108
A.5. Cokriging 109
A.6. Validación Cruzada 110
A.7. Predicciones 110

B. Publicación del autor **113**

Capítulo 1

Introducción

El objetivo de esta tesis doctoral es contribuir al desarrollo de nuevos métodos que permitan mejorar los modelos clásicos de regresión. Los modelos de regresión clásicos buscan la manera de hallar una variable respuesta en función de un conjunto de variables independientes. Estas técnicas persiguen un objetivo en concreto, predecir la variable dependiente. Esta situación provoca que las variables independientes con las que construimos nuestro modelo cobren una gran importancia. El modelo construido necesita estas variables de entrada para hallar la variable o variables (regresión lineal multivariante) de salida. La relación entre ambos conjuntos de variables tiene una única dirección. En el caso de querer cambiar variables de salida por variables de entrada o viceversa necesitamos construir nuevos modelos.

Estas limitaciones provocan que los procedimientos sean rígidos. Su adaptabilidad a diferentes situaciones es susceptible a cambios en su estructura original, lo que provoca nuevos cálculos computacionales que hacen más complejo el desarrollo. Una de las motivaciones principales de esta tesis es la búsqueda de un modelo que rompa con esta rigidez permitiendo a las variables tener diferentes roles sin por ello perder poder predictivo.

Para poder lograr nuestro objetivo hemos asociado dos campos estadísticos de diversa índole, técnicas de reducción dimensional y procesos gaussianos. Las técnicas de reducción dimensional nos permiten, entre otras cosas, conocer mejor la estructura de nuestros datos, simplificar métodos complejos o reducir la multicolinealidad. Por otro lado, los procesos gaussianos multivariantes son capaces de calcular un conjunto de variables correlacionadas por un dominio continuo (espacio o tiempo principalmente).

Las técnicas de reducción dimensional son en su mayoría métodos exploratorios que permiten describir de forma intrínseca datos multivariantes. Estos procedimientos reproducen sobre planos factoriales hipotéticos nuestros datos en función de las variables que los representan. Aunque estas técnicas están muy presentes en el análisis multivariante, su poder predictivo es bajo.

Los procesos gaussianos son modelos estadísticos en los que las observaciones suceden en un dominio continuo como espacio o tiempo. El caso espacial es descrito por las técnicas de *krigeaje*. Estos métodos de interpolación basan su poder de predicción en la denominada covarianza espacial y/o temporal y la distribución normal de sus variables. La idea básica de estas técnicas es predecir los valores en un punto desconocido del espacio calculando un promedio ponderado de los valores cercanos conocidos.

El modelo que desarrollamos "Regresión Multivariante Gaussiana Subespacial" (MGSR) conjuga ambas corrientes estadísticas. Partiendo de unas coordenadas subespaciales generadas por una técnica cualquiera de reducción dimensional y asociando a éstas sus valores reales, podemos construir una nueva matriz sobre la cual aplicar un proceso gaussiano como el cokriging (kriging multivariante). Este proceso nos permite adivinar múltiples combinaciones entre las variables analizadas y a partir de ellas construir nuestros modelos predictivos.

La presente memoria se estructura en tres apartados. La primera parte desarrolla los fundamentos teóricos en los que se basa el algoritmo propuesto (MGSR). En primer lugar se describen las principales técnicas de reducción dimensional. A continuación se expone las bases teóricas de los procesos gaussianos y más en concreto las técnicas geoestadísticas.

El segundo apartado detalla los sucesivos pasos que se han de llevar a cabo para desarrollar el algoritmo MGSR. La primera parte del mismo especifica el vínculo de unión entre las técnicas factoriales y el cokriging (las coordenadas subespaciales). Posteriormente se desarrolla el algoritmo en sí.

Por último, se llevaron a cabo dos aplicaciones a datos reales:

La primera de ellas fue aplicada a una base de datos de pacientes intervenidos de angioplastia con implantación de stent. El objetivo del estudio era predecir la evolución del paciente en etapas posteriores a la intervención y hallar valores ausentes de las distintas variables de estudio.

La segunda aplicación tuvo como objetivo la predicción de cambios de color en conglomerados silíceos. Este tipo de conglomerados está presente en monumentos históricos, los cuales están expuestos a condiciones ambientales que provocan cambios en su coloración y estructura. Para simular estas condiciones, muestras de cantera fueron expuestas a tratamientos de envejecimiento artificial en laboratorio. El objetivo principal del análisis fue predecir este comportamiento más allá de lo recogido en los experimentos evitando ensayos largos y costosos.

El sistema de citas sigue la norma IEEE (Institute of Electrical and Electronics Engineers).

Capítulo 2

Marco Teórico

Hasta la fecha los modelos de regresión intentan calcular una variable respuesta en función de variables independientes. Los modelos, más o menos complejos, tienen una estructura funcional. A través de un conjunto explicativo de variables calculamos el valor de una variable dependiente.

Las técnicas multivariantes han permitido que se puedan predecir conjuntos de variables dependientes en función de otro conjunto de variables independientes. Es el caso de la regresión multivariante. En este tipo de análisis aunque nos permiten predecir más de una variable a la vez, sigue existiendo una división entre los conjuntos dependientes e independientes.

Una de las principales limitaciones de las regresiones clásicas es su direccionalidad. Si queremos, por ejemplo, cambiar una variable independiente en respuesta, tenemos que cambiar el modelo por completo.

A continuación introducimos dos amplios campos estadísticos en los que vamos a basar nuestro algoritmo.

2.1. Reducción dimensional

La reducción de la dimensionalidad tiene como objetivo representar de forma fiel datos multivariantes en espacios de menor dimensión. La reducción dimensional puede utilizarse para fines de aminoración o comprensión de datos, pero la componente descriptiva sigue siendo su aplicación principal en una gran variedad de disciplinas.

La interpretación de datos multivariantes es una tarea ardua para el cerebro humano. Principalmente porque nuestra visión no está acostumbrada a procesar espacios de dimensiones mayores a tres. Los procesos de reducción dimensional pretenden representar esta multidimensionalidad de una forma interpretable. Además de la representación, otras de las aplicaciones de las técnicas de

reducción de dimensionalidad son el preprocesado o filtrado de datos como paso previo a otras técnicas estadísticas.

Obviamente, los procesos de reducción dimensional deben proporcionar una representación de baja dimensión que es significativa en algún modo. Independientemente del modelo, la idea general es representar elementos de datos similares en términos de cercanía, manteniendo distancias mayores entre aquellos que son diferentes.
En la práctica, el objetivo es preservar las propiedades principales de los datos mostrando a su vez las similaridades o diferencias entre ellos.

Entre las múltiples técnicas de reducción dimensional, introducimos a continuación una serie de ellas. Lo cual no implica que sean las únicas pero son las que consideramos más utilizadas.

2.1.1. Análisis de Componentes Principales

Es la principal técnica de reducción dimensional, y la más antigua. El análisis de componentes principales (ACP) fue inventado por Karl Pearson[1] hace más de un siglo.

Técnicamente, el ACP busca la mejor representación de datos a través de mínimos cuadrados, convirtiendo este conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales.

En la práctica, se construye la matriz de covarianza de los datos y se calculan los vectores propios de la misma a través de la Descomposición en Valores Singulares (DVS) de la matriz original. Aquellos vectores propios que corresponden a los valores propios más grandes son usados para reproducir los datos originales lo más fielmente posible.

El ACP, como muchas de las técnicas de reducción dimensional, se utiliza

principalmente como una herramienta de exploración de datos y en muchos casos sirve como paso previo a la construcción de modelos predictivos.

Para el caso de variables cualitativas, un análisis estadístico similar al ACP es el Análisis de Coordenadas Principales[2] (AcoP) que permite obtener una representación parecida a la del ACP basada en matrices de distancias (similaridades).

2.1.2. Análisis de Correspondencias

El análisis de correspondencias[3] (AC) tiene como punto de partida una tabla de contingencia y como resultado un gráfico en pocas dimensiones que resume el comportamiento de perfiles filas y columna. En este análisis se proyectan los perfiles fila en el subespacio correspondiente a las columnas y viceversa, siendo el subespacio resultante común a ambas. Esto posibilita representar en un mismo subespacio ambos perfiles. El resultado final nos permite jerarquizar y determinar las dependencias existentes entre ellos.

Por otra parte, el análisis de correspondencias múltiples[4] (ACM) es una extensión del análisis de correspondencias aplicado a un conjunto de matrices cuyas filas representan individuos y cuyas columnas representan variables caracterizadoras de éstos pero que no tienen porqué ser iguales en cada matriz. El resultado es una representación gráfica de estos individuos semejante al ACP pero aplicado a variables categóricas, pero en este caso el resultado es más completo al permitir la representación conjunta de individuos y variables.

2.1.3. Análisis Factorial

En muchos casos, la interpretación del ACP no es muy clara. Una variable de nuestra muestra puede contribuir significativamente en más de una componente de nuestros datos. La situación ideal la encontramos sólo cuando existe una vinculación única entre variable y componente, pero esta situación muchas veces no sucede. Uno de los objetivos del análisis factorial es tratar de interpre-

tar los datos a través de factores.
La principal motivación del análisis factorial tiene que ver con que los datos observados son función de una cantidad menor de variables no observadas denominadas factores, las cuales no pueden ser medidas previamente.

Charles Spearman[5] plantea por primera vez el análisis factorial en 1904 planteando una teoría sobre la inteligencia basada en la existencia de un factor común. Las aplicaciones desde entonces han sido numerosas en diversos campos como la psicología, la sociología o la medicina entre otras.

En los años 80 Escofier y Pagès[6] proponen una variación con respecto al clásico análisis factorial, el análisis factorial múltiple (AFM). Este método factorial permite el estudio de individuos por un conjunto de variables (cualitativas o cuantitativas) estructuradas en grupos.

El núcleo del AFM es un análisis factorial en el que las variables son ponderadas a través del ACP para los casos de datos cuantitativos y del AC para el caso de las variables cualitativas. Estas ponderaciones son idénticas para las variables pertenecientes al mismo grupo, variando de un grupo a otro. En otras palabras, aplicando el ACP o AC, según convenga, el AFM asigna a cada variable del grupo correspondiente un peso igual a la inversa del primer valor propio del ACP o AC, haciendo que la inercia axial máxima de cada grupo sea igual a 1.

Introduciendo diferentes grupos de variables implícitamente en el análisis factorial asumimos un equilibrio entre ellos. Este equilibrio debe tener en cuenta que un grupo multidimensional influye en más ejes que el caso univariante, siendo las ponderaciones las que cumplen este rol.

2.1.4. Sistema Gifi

Los sistemas Gifi son un conjunto de técnicas multivariantes no lineales. La denominación de estos métodos estadísticos como "Sistema Gifi"[7] proviene del seudónimo introducido por la escuela holandesa de la Universidad de Lei-

den a principio de los años 80.

La característica común en todas las técnicas Gifi es la minimización de la función de pérdida (*loss function*) a través de Mínimos Cuadrados Alternados (ALS) y la transformación de las variables para cuantificar las distintas categorías presentes en nuestro análisis.

Al igual que las técnicas introducidas anteriormente, los métodos Gifi tienen como principal propósito la exploración y modelado de la relación entre dos o más conjuntos de variables. Aunque existen diferentes técnicas, las más conocidas son: OVERALS y HOMALS.

El método estadístico OVERALS[8] es una técnica de análisis de correlación canónica aplicada a dos o más conjuntos de variables. Esta técnica se puede aplicar a cualquier tipo de tabla de tres vías, independientemente del tipo de datos, ya que es capaz de manejar datos de tipo numérico, ordinal o nominal.
Este método permite el análisis de distintos conjuntos de datos de manera conjunta buscando el subespacio que describa mejor un grupo de variables medidas en un mismo subespacio de referencia.

El método HOMALS[9] es una extensión del ACM salvo que en vez de usar DVS utilizamos ALS para minimizar la homogeneidad de nuestros datos de partida.

2.1.5. Métodos Biplot

Un Biplot[10] es una representación gráfica de datos multivariantes que permite representar variables e individuos simultáneamente. Se trata de una generalización de un clásico gráfico de dispersión de dos variables, salvo que estos métodos son capaces de representar múltiples variables.
En este tipo de representaciones los individuos de una matriz (filas) se representan como puntos y las variables (columnas) como vectores.

Sea $\boldsymbol{X}_{N\times P}$ la matriz de partida compuesta por P variables cuantitativas y N individuos. Un Biplot es una representación gráfica de la matriz $\boldsymbol{X}$ por marcadores fila $\boldsymbol{r_1}, ..., \boldsymbol{r_N}$ y marcadores columna $\boldsymbol{c_1}, ..., \boldsymbol{c_P}$, con el propósito de obtener $x_{ij} \approx r_i^\intercal c_j$. En forma matricial, $\boldsymbol{X} \approx \boldsymbol{R}\boldsymbol{C}^\intercal$.

Para obtener esta aproximación utilizamos una DVS. Si $T = rank(\boldsymbol{X})$ entonces la factorización de la matriz $\boldsymbol{X}$ se obtiene tal que,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\intercal = \sum_{t=1}^{T} \lambda_t r_t c_t \tag{2.1}$$

donde $\boldsymbol{U}$ es una $N \times T$ matriz unitaria, $\boldsymbol{\Lambda}$ es una $T \times T$ matriz diagonal no negativa, y $\boldsymbol{V}$ es una $P \times T$ matriz unitaria.

Existe un rango T que hace que $\boldsymbol{X}$ sea tal que,

$$\boldsymbol{X} \cong \boldsymbol{U}_{(T)}\boldsymbol{\Lambda}_{(T)}\boldsymbol{V}_{(T)}^\intercal = \sum_{t=1}^{T} \lambda_t r_t c_t \tag{2.2}$$

Siendo λ_t los valores propios, r_t y c_t los vectores propios de filas y columnas respectivamente.
Por tanto, $\boldsymbol{R}$ y $\boldsymbol{C}$ son fácilmente obtenibles tal que,

$$\boldsymbol{R} = \boldsymbol{U}_{(T)}\boldsymbol{\Lambda}_{(T)}^{\psi}, \boldsymbol{C} = \boldsymbol{V}_{(T)}\boldsymbol{\Lambda}_{(T)}^{1-\psi} \tag{2.3}$$

con $0 \leq \psi \leq 1$. Con $\psi = 1$, $\boldsymbol{R}$ representa las coordenadas correspondientes al ACP y $\boldsymbol{C}$ los vectores principales de la matriz de covarianza. Este tipo de biplots se denomina JK-Biplot. El caso contrario, $\psi = 0$ se denomina GH-Biplot, en el que prepondera la representación de las variables.

La interpretación del Biplot es muy sencilla. En el plano resultante podemos visualizar tanto los elementos fila, representados como puntos en el subespacio, y los elementos columna, representados como vectores. Las proyecciones ortogonales de los marcadores fila sobre los columna aproximan el orden de

los individuos y sus similaridades con respecto a las variables analizadas. El producto escalar entre marcadores columna aproxima la covarianza entre variables. La longitud de los elementos columna aproxima la desviación estándar de las variables. Por último, el ángulo entre dos variables aproxima su correlación correspondiente.

En un Biplot es necesario considerar la calidad global de representación ya que el rango de la matriz $\boldsymbol{X}$ es superior a dos, lo que contribuye a la imposibilidad de representar perfectamente los elementos en un plano. Como indicador de la calidad global definimos $CA = \frac{\lambda_1+\lambda_2}{sum_{t=1}^{T}\lambda_t}$ y el indicador de calidad particular de filas y columnas como $CR = \frac{\boldsymbol{r_{jk}^2}}{\sum_{k=1}^{N}\boldsymbol{r_{jk}^2}}$ y $CC = \frac{\boldsymbol{c_{jk}^2}}{\sum_{k=1}^{p}\boldsymbol{c_{jk}^2}}$

La calidad de representación puede equilibrarse como demuestra Galindo[11] con el HJ-Biplot que, con una interpretación similar al análisis de correspondencias consigue representar filas y columnas con igual calidad.

Aunque en un principio las técnicas Biplot son, en general, técnicas descriptivas o de diagnosis de modelos, se ha contribuido mucho en múltiples aspectos en las últimas décadas ampliando las posibilidades que brindan estas técnicas. Vicente-Villardón[12] propone un Biplot generalizado que permite considerar la importancia de los diferentes individuos y variables. Introduce métricas definidas positivas Ω y Φ en el espacio de filas y columnas respectivamente tal que $\boldsymbol{R^T\Omega R = I}$ y $\boldsymbol{C^T\Phi C = I}$, siendo necesaria la aproximación de $\boldsymbol{X}$ vía DVS Generalizada[4]. Con lo cual podemos obtener Biplots Clásicos con un enfoque diferente y abriendo un espectro de posibilidades muy alentador.

Con un enfoque diferente Gower[13] propone Biplots diferentes basándose en la obtención de marcadores columna a través de una regresión multivariante. A su vez propone Biplots no Lineales que representan las variables a través de trayectorias no lineales que luego proyectan sobre representaciones obtenidas a través de coordenadas principales.
Además, Gower define los Biplots de interpolación y predicción. Los primeros

permiten superponer nuevos individuos proyectándolos sobre el subespacio de representación. Los segundos se puede inferir valores de las variables originales dado un punto sobre la representación subespacial.

Asimismo, existen técnicas multivariantes para descripción de tablas múltiples (de 3 modos) que previa transformación a tablas de 2 vías permiten la aproximación Biplot a través de la DVS. Destacan tres corrientes en este ámbito, configuración consenso (Escofier[14]), comparación de matrices para el análisis de estructuras (Gower[15],Krzanowski[16]) y determinación de componentes latentes (Tucker[17], Tuckals[18], Kroonenberg[19])

Gabriel[20] formula el MANOVA Biplot en el año 1972 aunque no es desarrollado completamente hasta mucho tiempo después.

En los últimos 15 años se ha ampliado mucho la gama de técnicas asociadas al Biplot. Ejemplo de ello son los Biplots para detectar multicolinealidad[21], los Biplot para minería de datos[22] , el Meta-Biplot[23], el Biplot Canónico[24], el Biplot Logístico[25], el Biplot Nominal[26] y el más reciente Co-Tucker3[27].

Si se quiere profundizar sobre las diferentes técnicas asociadas al Biplot se puede consultar el artículo de Cárdenas[28] que hace un recorrido extenso por muchas de las técnicas derivadas del Biplot.

2.2. Procesos Gaussianos

Los procesos gaussianos (PG)[29] son modelos estadísticos en los que las observaciones ocurren en un dominio continuo (tiempo o espacio) y en los que cada punto está asociado a una distribución normal.

Los PG son una extensión de las distribuciones multivariantes gaussianas. Más concretamente, un PG genera datos a través de un dominio continuo de tal forma que cualquier subconjunto finito de este conjunto sigue una distribución

gaussiana multivariante. Además, los PG parten de la asunción de que aquellos puntos que están cercanos son más parecidos que aquellos que no lo están.

Para ilustrar como funcionan los PG supongamos una muestra de n elementos $y = \{y_1, ..., y_n\}$ que tiene una distribución normal con media cero. La función que relaciona los distintos puntos de nuestra muestra es la función de covarianza $k(x, x')$ asociada a la componente continua. Consideremos que esta función k se ajusta a una función exponencial cuadrada (existen otras opciones de ajuste, esta elección es únicamente escogida como ejemplo).

$$k(x, x') = \sigma_f^2 exp\left[\frac{-(x - x')^2}{2l^2}\right] \tag{2.4}$$

donde σ_f^2 es la covarianza de la función f que define a nuestra distribución y l es la distancia o gradiente temporal entre los puntos x y x'. En el caso en el que $x \approx x'$ la función de covarianza se hace máxima ya que $f(x)$ es prácticamente idéntica a $f(x')$.
En resumen, la función k nos indica cuánto varían nuestros datos en función de la distancia o gradiente temporal en el que se encuentran.

No obstante, la situación anterior es ideal ya que en general los PG presentan ruido que hace que la distribución sea $y = f(x) + N(0, \sigma_n^2)$ y la función k:

$$k(x, x') = \sigma_f^2 exp\left[\frac{-(x - x')^2}{2l^2}\right] + \sigma_n^2 \delta(x, x') \tag{2.5}$$

donde $\delta(x, x')$ es la delta de Kronecker.

Dadas n observaciones y, nuestro objetivo es predecir y_*. Para poder realizar esta predicción, calculamos la función de covarianza asociada al dominio continuo para todas las posibles combinaciones existentes entre cada observación tal que,

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix} \tag{2.6}$$

$$K_* = \begin{bmatrix} k(x_*, x_1) & k(x_*, x_2) & \dots & k(x_*, x_n) \end{bmatrix} \; K_{**} = k(x_*, x_*) \tag{2.7}$$

Los elementos de la diagonal de K son $\sigma_f^2 + \sigma_n^2$.

Como la premisa de los PG es que nuestros datos pueden ser representados como muestras de una distribución gaussiana multivariante entonces,

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \tag{2.8}$$

Y por tanto, nuestro estimador y_* se calcula tal que,

$$\hat{y}_* = K_* K^{-1} y \tag{2.9}$$

El error de nuestro estimador es dado por su varianza:

$$var(y_*) = K** - K_* K^{-1} K_*^T \tag{2.10}$$

2.2.1. Geoestadística

Históricamente, los PG[29] surgen como respuesta a la resolución de series temporales. No obstante a principios de los años 50 el ingeniero de minas sudafricano Danie Krige[30] sienta las bases de lo que hoy conocemos como geoestadística y en extensión del denominado *Kriging*. Una década después, George Matheron[31] desarrolla toda la teoría matemática asociada a la geoestadística. Ambos son considerados como padres de la geoestadística.

Aunque los orígenes de la geoestadística están fuertemente vinculados a la minería, sus aplicaciones son grandes, como por ejemplo en hidrología, geolo-

gía, agricultura, geografía, meteorología, ecología, biología, etc. La única condición indispensable para poder aplicar este tipo de técnicas es la localización espacial de los valores medidos. El enfoque geoestadístico es probabilístico ya que un marco determinista no podría determinar con exactitud los comportamientos espaciales.

Para poder comprender este concepto y posteriores, vamos a presentar un ejemplo. Cerca de Stein (Países Bajos), a orillas del río Mosa[32], se tomaron datos de concentración (ppm) de distintos minerales, cadmio, cobre, plomo y zinc, en diferentes localizaciones espaciales. Las coordenadas exactas se encuentran en (N50°58′18,443″ E5°44′29,776″).

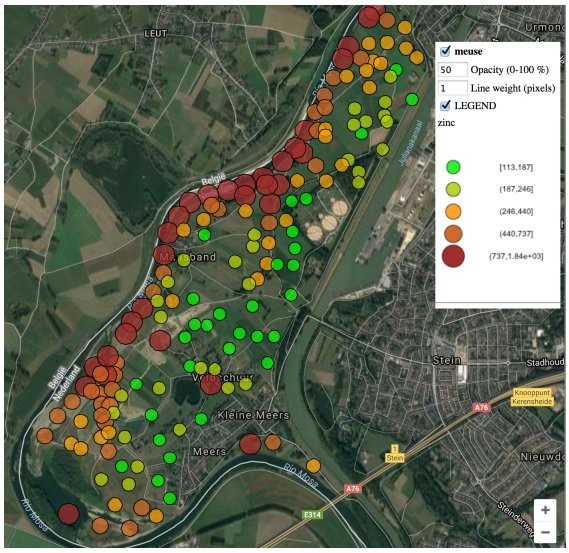


Figura 2.1: Río Mosa (distribución de Zn)

En la Fig. 2.1 se observa la extensión de terreno donde queremos realizar nuestros cálculos. En esta figura se presentan las mediciones de una de las variables, en este caso las partes por millón de zinc. Supongamos por otro lado que la parcela (Fig. 2.1) se subdivide en 3103 fragmentos cuadrados (50 m^2 cada uno), la unión de estas partes nos proporciona un mallado de la parcela. Cada una de estas porciones tiene un valor de zinc asociado que desconocemos a la que denominaremos **variable regionalizada**.

Sin embargo existen ciertos valores que sí conocemos y que siguen un comportamiento aleatorio a priori, a estos valores los llamaremos **variables aleatorias**. Como se puede apreciar, el número de mediciones (155) representa un 5 % del total de posibles observaciones. Nuestro objetivo es predecir el comportamiento del zinc en la parcela global gracias a los valores en los que sí conocemos este valor. Al comportamiento de las observaciones y de nuestro modelo lo designaremos como **función aleatoria**. Existen tantas funciones aleatorias como fragmentos hay en nuestra malla.

En la Tabla 2.1 podemos visualizar nuestra matriz de partida $\boldsymbol{Z}$ en la cual están definidas sus coordenadas espaciales x e y y las partes por millón de Zn que contienen. La única manera de hallar el comportamiento espacial de la variable regionalizada es a través de las mediciones realizadas, ya que es el único valor del que disponemos.

n	x	y	Zn
1	181072.00	333611.00	1022.00
2	181025.00	333558.00	1141.00
3	181165.00	333537.00	640.00
⋮	⋮	⋮	⋮
155	180627.00	330190.00	375.00

Tabla 2.1: Matriz de partida Zn (ppm)

En cualquier sistema del cual queremos conocer su comportamiento, la estimación de la media es fundamental. Podríamos suponer que teniendo una muestra irregularmente distribuida en el espacio, una media aritmética podría ser más que suficiente. No es el caso.

Como se puede apreciar en la Fig. 2.1, los valores más cercanos a la orilla del Mosa presentan valores más altos de zinc que aquellos que se encuentran un poco más alejados. Es obvio por tanto que partir de una hipótesis de independencia no es muy sensato.

No siendo la media aritmética una posibilidad, una segunda opción es la media ponderada. Consideremos $\boldsymbol{Z}(x_\alpha)$ el valor en ppm de zinc de nuestra muestra en los lugares x_α y w_α el vector de ponderaciones correspondiente. De tal modo que la media se define como:

$$m^* = \sum_{\alpha=1}^{n} w_\alpha \boldsymbol{Z}(x_\alpha) \tag{2.11}$$

Suponemos que la media existe en todos los puntos de la región (Fig. 2.1) y es igual a m. Buscamos que la diferencia con la media estimada m^* sea cero en media. El vector de pesos[33] ha de ser calculado de manera imparcial, es por ello que imponemos la condición siguiente $\sum_{\alpha=1}^{n} w_\alpha = 1$, ya que se demuestra:

$$\begin{aligned} E[m^* - m] &= E[\sum_{\alpha=1}^{n} w_\alpha \boldsymbol{Z}(x_\alpha) - m] \\ &= \sum_{\alpha=1}^{n} w_\alpha E[\boldsymbol{Z}(x_\alpha)] - m \\ &= m \sum_{\alpha=1}^{n} w_\alpha - m = 0 \end{aligned} \tag{2.12}$$

Por otra parte $\boldsymbol{Z}(x)$ es una función estacionaria de segundo orden por lo que existe una función de covarianza espacial $\boldsymbol{C}(h) = E[\boldsymbol{Z}(x) * \boldsymbol{Z}(x+h)] - m^2$ que explica el comportamiento espacial entre cualquier par de puntos existente en la región (siendo h el intervalo entre dos puntos cualesquiera). Este concepto se explica con más detalle posteriormente.

Al igual que hemos realizado con la media, calculamos la estimación del error de la varianza.

$$var(m^* - m) = E[(m^* - m)^2] - (E[(m^* - m)])^2 = E[(m^* - m)^2] \tag{2.13}$$

Si lo expresamos en términos de la función de covarianza espacial, entonces tenemos:

$$
\begin{aligned}
var(m^* - m) &= E[m^{*2} - 2mm^* + m^2] = \\
&= \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} w_\alpha w_\beta E[\boldsymbol{Z}(x_\alpha)\boldsymbol{Z}(x_\beta)] - 2m \sum_{\alpha=1}^{n} w_\alpha E[\boldsymbol{Z}(x_\alpha)] + m^2 \\
&= \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} w_\alpha w_\beta \boldsymbol{C}(x_\alpha - x_\beta) \qquad (2.14)
\end{aligned}
$$

Siendo x_α y x_β puntos cualesquiera de la muestra y $\boldsymbol{C}$ la matriz de covarianza espacial.

La estimación de la varianza es la suma de los productos cruzados de los pesos asignados a las muestras y la covarianza espacial entre los distintos puntos de la muestra.

Existen numerosas maneras de calcular el vector de pesos w, no obstante a continuación presentamos los tres tipos de *kriging* más comunes: ordinario, simple y universal.

Kriging Ordinario

Para poder estimar los distintos parámetros anteriormente definidos hemos de buscar minimizar el error de la varianza respetando $\sum_{\alpha=1}^{n} w_\alpha = 1$. Como la varianza es una función cuadrática positiva, la obtención de un mínimo estará condicionada a la primera derivada parcial igual a cero. Además incluimos el término de los pesos gracias al método de Lagrange.

Una función objetivo φ está constituida por una parte cuadrática más un término que contiene un multiplicador de Lagrange μ:

$$\varphi(w_\alpha, \mu) = var(m^* - m) - 2\mu(\sum_{\alpha=1}^{n} w_\alpha - 1) \qquad (2.15)$$

Tenemos por lo tanto, dos ecuaciones tal que:

$$\begin{cases} \frac{\partial \varphi(w_\alpha, \mu)}{\partial w_\alpha} = 0 \text{ para } \alpha = 1, ..., n \\ \\ \frac{\partial \varphi(w_\alpha, \mu)}{\partial \mu} = 0 \end{cases} \tag{2.16}$$

Que derivan en $n+1$ ecuaciones. La solución de este sistema nos proporciona los pesos óptimos para la estimación de la media.

$$\begin{cases} \sum_{\beta=1}^{n} w_\beta \boldsymbol{C}(x_\alpha - x_\beta) - \mu = 0 \text{ para } \alpha = 1, ..., n \\ \\ \sum_{\beta=1}^{n} w_\beta = 1 \end{cases} \tag{2.17}$$

Por otra parte, se demuestra que el multiplicador de Lagrange μ es la varianza del sistema.

$$\begin{aligned} var(m^* - m) &= \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} w_\alpha w_\beta \boldsymbol{C}(x_\alpha - x_\beta) \\ &= \sum_{\alpha=1}^{n} w_\alpha \mu = \mu \end{aligned} \tag{2.18}$$

Kriging Simple

Matemáticamente es el método de cálculo más sencillo. Asume el conocimiento de la variable aleatoria a través de su función de covarianza espacial. Sin embargo, en la mayoría de las aplicaciones reales ni la variable aleatoria es conocida totalmente ni su función de covarianza espacial.

$$E[Z(x+h)] = E[Z(x)] \tag{2.19}$$

$$cov[Z(x+h), Z(x)] = C(h) \tag{2.20}$$

El valor $E[Z(x)] = m$ es igual en todo el dominio de nuestra muestra y la covarianza espacial únicamente depende del valor h que separa cada par de puntos en el espacio.

Este procedimiento predictivo recuerda a una regresión múltiple en la cual $\boldsymbol{Z}(x_\alpha)$ juega el rol de regresora (x_α son las coordenadas de los valores muestrales) y $\boldsymbol{Z}(x_0)$ son los valores que queremos estimar (x_0 son las coordenadas de los valores a estimar).

$$\boldsymbol{Z}^*(x_0) = m + \sum_{\alpha=1}^{n} w_\alpha (Z(x_\alpha) - m) \tag{2.21}$$

donde w_α son los pesos ligados a los residuales $Z(x_\alpha) - m$. Asumiendo estacionalidad, m es igual en todas las localizaciones. El error estimado es igual a la diferencia $\boldsymbol{Z}^*(x_0) - \boldsymbol{Z}(x_0)$. Con lo que asumimos un error estimado igual a cero en media.

$$\begin{aligned} E[\boldsymbol{Z}^*(x_0) - \boldsymbol{Z}(x_0)] &= m + \sum_{\alpha=1}^{n} w_\alpha (E[\boldsymbol{Z}(x_{-\alpha})] - m) - E[\boldsymbol{Z}(x_0)] \\ &= m + \sum_{\alpha=1}^{n} w_\alpha (m - m) - m = 0 \end{aligned} \tag{2.22}$$

La varianza del error estimado, σ_E^2 es:

$$\sigma_E^2 = var(\boldsymbol{Z}^*(x_0) - \boldsymbol{Z}(x_0)) = E[(\boldsymbol{Z}^*(x_0) - \boldsymbol{Z}(x_0))^2] \tag{2.23}$$

Y en consecuencia, su función de covarianza espacial es $cov[\boldsymbol{Z}(x_\alpha), \boldsymbol{Z}(x_\beta)] = \boldsymbol{C}(x_\alpha - x_\beta)$ gracias a la asunción de estacionariedad.

Finalmente el sistema de ecuaciones se reduce a:

$$\sum_{\beta=1}^{n} w_\beta \boldsymbol{C}(x_\alpha - x_\beta) = \boldsymbol{C}(x_\alpha - x_0) \; \alpha = 1, ..., n \tag{2.24}$$

Kriging Universal

El Kriging Universal asume una tendencia polinómica asociada al modelo. Esta tendencia depende directamente de la comprensión del modelo.
Pongamos como ejemplo lo visto en la Fig. 2.1, las cantidades de zinc son mayores en aquellas zonas más cercanas al río, luego es de esperar que añadiendo

un valor dependiente de la distancia de un punto analizado con respecto al río sea una posible tendencia a añadir al modelo.

Añadiendo una función de segundo orden estacionaria y aleatoria al modelo inicial tendremos $\boldsymbol{Z}(x) = m(x) + Y(x)$, siendo $E[\boldsymbol{Z}(x)] = m(x)$. Suponemos que $m(x)$ puede ser representada como combinación lineal de funciones f_l deterministas con coeficientes a_l distintos de cero tal que $m(x) = \sum_{l=0}^{L} a_l f_l(x)$. Si $l = 0$ la función $f_l = 1$.
Para el kriging lineal usamos la clásica combinación lineal:

$$\boldsymbol{Z}^*(x_0) = \sum_{\alpha=1}^{n} w_\alpha \boldsymbol{Z}(x_\alpha) \tag{2.25}$$

Al igual que en el resto de krigings descritos, queremos obtener $E[\boldsymbol{Z}(x_0) - \boldsymbol{Z}^*(x_0)] = 0$ lo que hace que $m(x_0) - \sum_{\alpha=1}^{n} w_\alpha m(x_\alpha) = 0$ y por tanto,

$$\sum_{l=0}^{L} a_l (f_l(x_0) - \sum_{\alpha=1}^{n} w_\alpha f_l(x_\alpha)) = 0 \tag{2.26}$$

Como a_l son distintos de cero,

$$\sum_{\alpha=1}^{n} w_\alpha f_l(x_\alpha) = f_l(x_0)\ l = 0, ..., L \tag{2.27}$$

que constituyen las *condiciones universales* por las que se conoce a este método como kriging universal. Y al igual que hicimos en el kriging ordinario, incluyendo el parámetro Lagrangiano μ_l y minimizándolo, obtenemos el sistema de ecuaciones siguiente:

$$\begin{cases} \sum_{\beta=1}^{n} w_\beta \boldsymbol{C}(x_\alpha - x_\beta) - \sum_{l=0}^{L} \mu_l f_l = \boldsymbol{C}(x_\alpha - x_0)\ \alpha = 1, ..., n \\ \\ \sum_{\beta=1}^{n} w_\beta f_l(x_\beta) = f_l(x_0)\ l = 0, ..., L \end{cases} \tag{2.28}$$

Una vez descritos algunos de los tipos de kriging, debemos de describir los patrones espaciales de los datos.

Para entender el comportamiento espacial de la muestra de zinc hemos de conocer como interactúan los distintos valores medidos entre sí. Para ello hemos de analizar cómo se parecen, o no, cada par de valores. Cada pareja posee dos características, el valor en partes por millón de zinc y la distancia existente entre ambos. No obstante, la distancia es un componente intrínseco de la muestra.

Para poder ver exactamente cómo varían los valores bastaría con realizar la sustracción al cuadrado de ambos valores y dividirlo entre dos. Este valor es la medida de *disimilaridad* o *semivarianza*, a la que denominaremos a partir de ahora γ^*. Supongamos que dos valores distintos de zinc z_α y z_β se sitúan en las posiciones x_α y x_β respectivamente (Fig. 2.2).

$$\gamma^* = \frac{(z_\alpha - z_\beta)^2}{2} \tag{2.29}$$

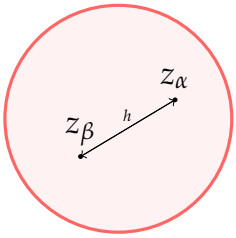


Figura 2.2: Valores de α y β

Como hemos indicado anteriormente, cada uno de los pares de puntos acarrea una distancia h, con lo cual podemos intuir que la medida de disimilaridad es función de ésta $\gamma^*(h)$. En consecuencia, $x_\beta = x_{\alpha+h}$ y por tanto $z_\alpha = z(x_\alpha)$. Podremos definir $\gamma^*(h)$ como:

$$\gamma^*(h) = \frac{1}{2}(z(x_\alpha + h) - z(x_\alpha))^2 \tag{2.30}$$

Si realizamos este cálculo para cada uno de los pares, obtenemos un gráfico denominado *variograma en nube* (Fig. 2.3).

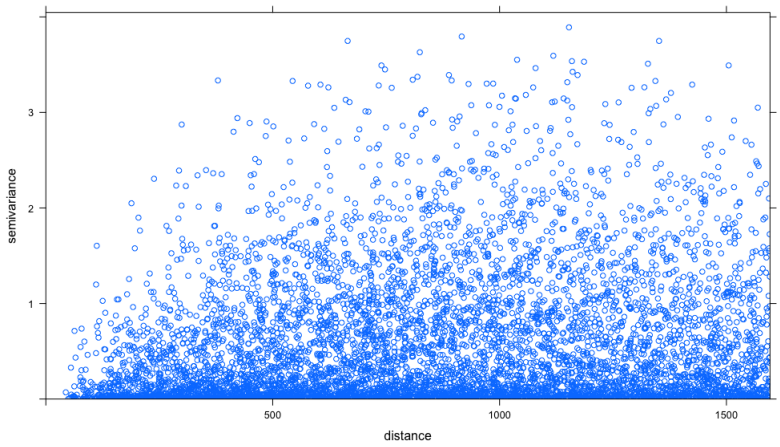


Figura 2.3: Variograma en nube

Aunque este tipo de representación no es del todo esclarecedora, se puede intuir que la disimilaridad crece con la distancia. No obstante, este tipo de gráficos pueden presentar varias nubes de puntos que actúan con diferente distribución o bien nos sirven para encontrar anomalías.

Para poder interpretar la correlación espacial existente, debemos construir un modelo comprensible e identificable: un variograma experimental. El variograma experimental no es más que el promedio, según el número de pares n, de las disimilaridades medidas en el variograma en nube a escalas concretas de la distancia h.

$$\gamma^*(h) = \frac{1}{2n} \sum_{\alpha=1}^{n} (z(x_\alpha + h) - z(x_\alpha))^2 \tag{2.31}$$

Para poder ilustrar la argumentación anterior supongamos que fijamos una distancia igual a 500 metros y buscamos todos los pares de puntos que se encuentran aproximadamente a esta distancia con un margen de $\pm$ 20 m es decir, entre 480 y 520 metros. Sumamos todos los valores de disimilaridad y los dividimos entre el número total de pares, que en este caso es 197. El valor obtenido en este caso es 6910.51 ppm de zinc.

Si repetimos este proceso con distintas distancias h comprendiendo un rango de 0 a 1400 m cada 100 m, $h = [0, 100, 200, 300, 400, ..., 1400]$. Obtendremos 15 valores distintos de disimilaridad para cada una de las distancias h y por tanto el variograma tendrá la forma que se aprecia en la Fig. 2.4.

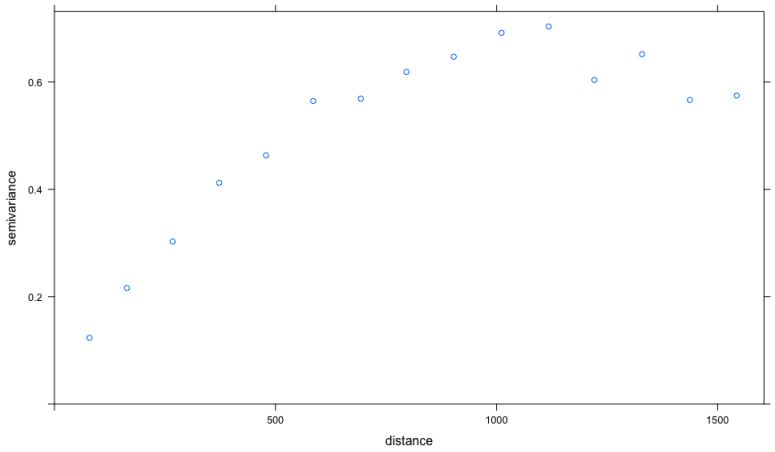


Figura 2.4: Variograma experimental

A diferencia del variograma en nube, el variograma experimental esclarece el comportamiento de la muestra con respecto a la distancia. A distancias mayores, el valor promedio de la disimilaridad aumenta.

Una vez calculado el variograma experimental, el desafío está en ajustar una función teórica que describa el comportamiento observado.

Antes de comenzar a desgranar las distintas funciones teóricas que se ajustan al variograma experimental, es necesario conocer las partes que componen esta función $\gamma(h)$.
Estas partes son las siguientes:

Nugget El valor nugget nug se define como el valor de la disimilaridad para distancias cercanas a cero $\gamma(h = 0)$.

$$\gamma_{nug}(h) = \begin{cases} 0 \text{ para } h = 0 \\ \\ nug \text{ para } h > 0 \end{cases} \tag{2.32}$$

Sill El *sill* b es el valor constante que alcanza el variograma a una distancia h determinada. En el ejemplo anterior, cuando superamos los 800 m alcanzamos valores muy similares. El variograma teórico tendrá que ajustarse de tal modo que estos valores queden lo más cercanos al valor que estimemos.

Rango El rango a es sencillamente el valor de h al que alcanzamos el *sill*. En nuestro ejemplo es un valor cercano a los 800 m ya mencionado.

Las funciones teóricas que vamos a introducir a continuación tendrán presentes estos tres valores. En la Fig. 2.4 los podemos ver representados.

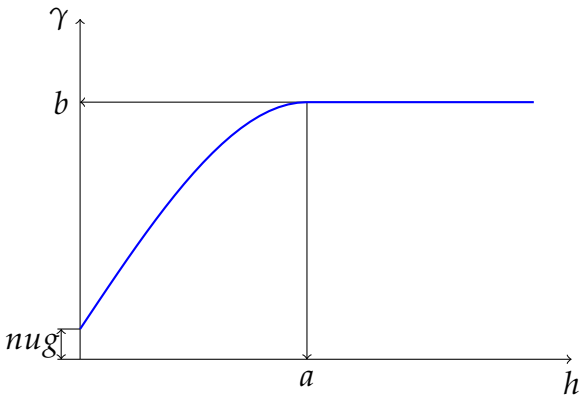


Figura 2.5: Variograma teórico

En la Fig. 2.5 tenemos una representación de un variograma esférico más la componente *nugget* respectiva. Sin embargo no es la única, existen un amplio número de funciones con las que podemos estimar el comportamiento de la semivarianza como las distribuciones exponencial, gaussiana, cuadrática o lineal entre otras.

Existen otras muchas funciones teóricas, todas ellas siguen el mismo patrón general, ser una multiplicación del *sill* b y una función genérica de h, $g_u(h)$.

$$\gamma(h) = bg_u(h) \tag{2.33}$$

Una vez introducidas las distintas funciones hemos de establecer una relación directa entre el método kriging y el cálculo del variograma.

Continuando con el ejemplo del río Mosa, supongamos que queremos hallar la distribución del zinc en la parcela definida en la Fig. 2.1. Es de suponer que la distribución del zinc depende de una serie de fenómenos independientes que actúan a distinta escala a lo largo de la superficie. Cada uno de estos fenómenos está asociado a una función aleatoria que a su vez se compone de un variograma o función de covarianza asociada.

Hemos hablado ya del variograma, y de la función de covarianza cuando explicamos el método *Kriging*, la relación entre ellas se describe tal que:

$$\gamma(h) = \boldsymbol{C}(0) - \boldsymbol{C}(h) \tag{2.34}$$

Los variogramas ajustados no tienen porqué ser función de un único variograma teórico, normalmente son resultado de la suma de una conjunción finita de ellos. Supongamos por un momento que el variograma que mejor se ajusta a las ppm del zinc es la suma de un variograma gaussiano y otro esférico con distintos valores de sill tal que:

$$\gamma_{zinc}(h) = b_{nug}g_{nug}(h) + b_{gau}g_{gau}(h) + b_{sph}g_{sph}(h) \tag{2.35}$$

Por otro lado, el zinc tiene una función de covarianza asociada $C_{zinc}(h)$. Esta función es el resultado de la multiplicación del valor *sill* global b por su correspondiente función de correlación $\rho_{zinc}(h)$ tal que $\boldsymbol{C_{zinc}}(h) = b\rho_{zinc}(h)$. Hay que

recordar, por otra parte que $\boldsymbol{C}(0) = b$, con lo cual podemos deducir la relación entre función de covarianza y variograma:

$$\begin{aligned} \gamma_{zinc}(h) &= b(1 - \rho_{zinc}(h)) \\ &= b_{nug}g_{nug}(h) + b_{gau}g_{gau}(h) + b_{sph}g_{sph}(h) \end{aligned} \tag{2.36}$$

siendo b

$$b = b_{nug} + b_{gau} + b_{sph} \tag{2.37}$$

O de manera más general

$$\gamma(h) = b(1 - \rho(h)) = \sum_{u=0}^{S} b_u g_u(h) \tag{2.38}$$

siendo b

$$b = \sum_{u=0}^{S} b_u \tag{2.39}$$

Hay que tener en cuenta que cada función tiene un rango asociado que difiere con respecto al de las demás funciones. Por ejemplo, el rango asociado al variograma gaussiano es de 450 metros y el asociado a la esférica de 980 metros.

Al igual que sucede con los variogramas, la función $\boldsymbol{Z_{zinc}}(x)$ es resultado de la adición de las distintas subfunciones, correspondientes a los distintos variogramas anidados, más la media. De tal modo que:

$$\boldsymbol{Z_{zinc}}(x) = \boldsymbol{Z_{nug}}(x) + \boldsymbol{Z_{gau}}(x) + \boldsymbol{Z_{sph}}(x) + m \tag{2.40}$$

O de forma más general:

$$\boldsymbol{Z(x)} = \sum_{u=0}^{S} \boldsymbol{Z_u}(x) + m \tag{2.41}$$

Las distintas subfunciones no están correlacionadas entre sí. Al disponer de

una conjunción de funciones, tendremos que hallar cada uno de los vectores pesos correspondientes w^S o en el caso de nuestro ejemplo w^{nug}, w^{gau} y w^{sph}. Es decir, cada función $\boldsymbol{Z}_S^*(x)$ estimada será igual a:

$$\boldsymbol{Z}_S^*(x) = \sum_{\alpha=1}^{n} w_\alpha^S \boldsymbol{Z}(x_\alpha) \tag{2.42}$$

Hasta el momento hemos descrito el modelo univariante, si queremos extender el análisis a más de una variable debemos explicar en qué consiste el Cokriging.

2.2.2. Cokriging

El método *cokriging* es la versión o extensión multivariante del *kriging*. La técnica del cokriging basa su principio en la correlación entre muestras debido a su localización espacial. Para poder entender mejor los conceptos, volvamos al ejemplo sobre el río Mosa. Hasta el momento nos hemos concentrado en las ppm del zinc exclusivamente, sin embargo disponemos de datos de otros minerales como plomo, cobre o cadmio.

Pueden presentarse distintos casos en la distribución de nuestras muestras:

- *Heterotopia*: Las variables han sido medidas en puntos no coincidentes.
- *Heterotopia parcial*: Algunas variables comparten localizaciones.
- *Isotopia*: Todas las variables se representan en los mismos puntos.

El cokriging se sustenta en la correlación entre variables debidas a su localización espacial, por tanto la heterotopia completa no se podría aplicar. Sin embargo, una heterotopia parcial o isotopia sí serían aplicables.

Volvamos de nuevo al ejemplo del río Mosa modificando las condiciones. Disponemos de muestras de igual tamaño para el zinc, plomo y cobre en las mismas localizaciones. Supongamos que el caso del cadmio es diferente, la muestra está distribuida en parte de las localizaciones, pero existen puntos en los que

no existe valor disponible del mineral.

Este caso es muy típico en tipo de muestras parecidas. Pongamos otro ejemplo desarrollado por investigadores de la Universidad de Brescia[34]. La ciudad de Milán dispone de un número de estaciones de medición de contaminación del aire, que no son suficientes para determinar la distribución de ppm de partículas contaminantes en toda la ciudad. A su vez, la *Direzione Generale della Motorizzazione* (Dirección General de Tráfico italiana) dispone de mediciones de densidad de vehículos por tramos y momentos del día en todo el área urbana. Los investigadores usaron los datos de tráfico (variable auxiliar del modelo) altamente correlacionados espacialmente con la calidad del aire para el cálculo de la contaminación en Milán.

Hemos visto que para un caso de heterotopia parcial, el cokriging es una técnica práctica. En el caso isotópico solo es interesante aplicar esta técnica si existe correlación espacial entre todas las variables analizadas. Si una de las variables no está correlacionada espacialmente, es preferible realizar un kriging sobre la misma.

Al igual que en el caso del kriging, el cokriging tiene múltiples variantes. A continuación explicamos dos de ellas.

Cokriging Ordinario

El cokriging ordinario es una combinación lineal del vector de pesos w_α^i con los datos de las distintas variables localizadas en los lugares muestrales. Imaginemos un lugar cualquiera x_0 no perteneciente a nuestra matriz inicial, podremos estimar su valor en función de las muestras vecinas.

$$Z_i^*(x_0) = \sum_{i=1}^{P} \sum_{\alpha=1}^{n_i} w_\alpha^i \boldsymbol{Z_i}(x_\alpha) \tag{2.43}$$

El índice i se refiere a la variable que estudiamos entre las P variables totales. El número n_i depende del número de muestras que dispone nuestra variable,

en el caso de isotopia el valor sería el mismo para todas las variables.

Partiendo del principio que queremos estimar una variable en concreto de las P totales, debemos de tender a un error nulo en media, para ello escogemos un vector de pesos tal que valga la unidad para la variable de interés y cero para las variables auxiliares.

$$\sum_{\alpha=1}^{n_i} w_\alpha^i = \delta_{ii_0} = \begin{cases} 1 \text{ para } i = i_0 \\ \\ 0 \text{ para } i \neq i_0 \end{cases} \tag{2.44}$$

Procediendo de igual modo que hicimos para el kriging, llegamos al sistema de ecuaciones que hemos de resolver para calcular los vectores de peso w_β^j y el correspondiente parámetro de Lagrange μ_i.

$$\begin{cases} \sum_{j=1}^{p} \sum_{\beta=1}^{n_j} w_\beta^j \gamma_{ij}(x_\alpha - x_\beta) + \mu_i = \gamma_{ii_0}(x_\alpha - x_0) \text{ para } i = 1, ..., p; \alpha = 1, ..., n_i \\ \\ \sum_{\beta=1}^{n_i} w_\beta^i = \delta_{ii_0} \text{ para } i = 1, ..., p \end{cases} \tag{2.45}$$

y la varianza:

$$\sigma_{ck}^2 = \sum_{i=1}^{P} \sum_{\alpha=1}^{n_i} w_\alpha^i \gamma_{ii_0}(x_\alpha - x_0) + \mu_{i_0} - \gamma_{i_0 i_0}(0) \tag{2.46}$$

En ambas ecuaciones γ_{ii_0} es el variograma cruzado el cual detallaremos más adelante. Como observamos, a diferencia del kriging, las ecuaciones se entrelazan debido a las P variables que analizamos. Por tanto, la complejidad es mucho mayor.

Cokriging Simple

La única diferencia entre el cokriging ordinario y el simple es la inclusión de la media m_{i_0}. El cokriging simple se apoya en el conocimiento de las medias intrínsecas de las variables analizadas. En aquellos espacios en los cuales no

dispongamos de valores se calculará un valor estimado.

$$Z^*_{i_0}(x_0) = m_{i_0} + \sum_{i=1}^{P} \sum_{\alpha=1}^{n_i} w^i_\alpha (Z_i(x_\alpha) - m_i) \tag{2.47}$$

Una vez definidos los sistemas de ecuaciones, al igual que hicimos con el kriging, hemos de estimar el comportamiento espacial de nuestras variables regionalizadas. Para ello debemos de definir el variograma cruzado.

Variograma Cruzado

Un variograma cruzado se define como la mitad del producto del incremento de dos variables.

$$\gamma_{ij}(h) = \frac{1}{2} E[(Z_i(x+h) - Z_i(x))(Z_j(x+h) - Z_j(x))] \tag{2.48}$$

Volviendo de nuevo al ejemplo del río Mosa, supongamos el caso de isotopia, es decir tenemos el mismo número de muestras para las cuatro variables P (zinc,plomo,cobre y cadmio) en las mismas localizaciones. Queremos calcular todos los variogramas posibles.

Los variogramas posibles son $\frac{P(P+1)}{2}$, de los cuales P son directos (resultado de las variables principales) y $\frac{P(P-1)}{2}$ son cruzados (resultado de las interacciones dobles entre las variables). Como resultado del cálculo de cada variograma obtenemos una matriz $\mathbf{\Gamma}(h)$ de tamaño $P \times P$.

Un variograma cruzado experimental para distintas clases de distancias h para las diferentes n_c pares de puntos x_α, x_β correspondientes a $x_\alpha - x_\beta = h \in h$, se define como:

$$\gamma^*_{ij}(h) = \frac{1}{2n_c} \sum_{\alpha=1}^{N} (z_i(x_\beta) - z_i(x_\alpha)) \cdot (z_j(x_\beta) - z_j(x_\alpha)) \tag{2.49}$$

Una vez que tenemos los valores experimentales de los diferentes variogramas (directos y cruzados) debemos proceder del mismo modo que para el mo-

delo univariante y ajustar funciones a cada uno de ellos.

Modelo Lineal Corregionalizado

Un conjunto multivariante de funciones aleatorias puede representarse a través de modelos lineales espaciales multivariantes. Debemos ajustar cada modelo lineal a los variogramas anidados experimentales. Este ajuste, en realidad consiste en hallar las matrices de corregionalización que describen los fenómenos espaciales a los diferentes rangos preestablecidos y que guardan relación directa con la matriz clásica de varianzas-covarianzas.

Un conjunto real de funciones estacionarias aleatorias de segundo orden $\{\boldsymbol{Z}_i(x); i = 1, ..., P\}$ se descompone en conjuntos espacialmente no correlacionados $\{\boldsymbol{Z}_u^i(x); u = 1, ..., S\}$ tal que:

$$\boldsymbol{Z}_i = \sum_{u=0}^{S} \boldsymbol{Z}_u^i(x) + m_i \tag{2.50}$$

donde, para cualquier valor $i, j = 1, ..., P$ y $u, v = 1, ..., S$

$$E[\boldsymbol{Z}_i(x)] = m_i \tag{2.51}$$
$$E[\boldsymbol{Z}_u^i(x)] = 0 \tag{2.52}$$

y en consecuencia:

$$cov(\boldsymbol{Z}_u^i(x), \boldsymbol{Z}_u^j(x+h)) = E[\boldsymbol{Z}_u^i(x)\boldsymbol{Z}_u^j(x+h)] = \boldsymbol{C}_{ij}^u(h) \tag{2.53}$$
$$cov(\boldsymbol{Z}_u^i(x), \boldsymbol{Z}_v^j(x+h)) = 0 \text{ cuando } u \neq v \tag{2.54}$$

Siendo $\boldsymbol{C}_{ij}^u(h)$ la función de covarianza cruzada. Esta función es resultado del producto entre su sill propio b_{ij}^u y la función de correlación $\rho_u(h)$.

$$\boldsymbol{C}_{ij}(h) = \sum_{u=0}^{S} \boldsymbol{C}_{ij}^{u}(h) = \sum_{u=0}^{S} b_{ij}^{u} \rho_u(h) \tag{2.55}$$

Las matrices de corregionalización $\boldsymbol{B}_u$ de orden $P \times P$ semidefinidas positivas pueden incluirse en el modelo.

$$\boldsymbol{C}(h) = \sum_{u=0}^{S} \boldsymbol{B}_u \rho_u(h) \tag{2.56}$$

El modelo de corregionalización lineal consiste en hallar la matriz B_S. Existen diversos métodos para realizar este ajuste, para nuestro algoritmo hemos empleado el método de mínimos cuadrados generalizados.

Capítulo 3

Regresión Multivariante Gaussiana Subespacial

Hasta ahora la mezcla de procesos gaussianos y técnicas de reducción dimensional han sido muy limitadas y específicas[35, 36]. En estas ocasiones, las técnicas de reducción dimensional tenían como objetivo escoger aquellas variables mejor representadas, decidir el valor de interespaciado para la formación de los variogramas o conocer mejor la estructura de los datos. No obstante, en todos los casos están aplicadas en dominios continuos.

Otros casos son en los que existiendo componente geográfica se aplican técnicas avanzadas de reducción dimensional obviando la correlación espacial directa[37, 38]. Sin embargo el objetivo de estos análisis no es predictivo sino descriptivo o de búsqueda de diferencias entre elementos.

Ambos casos en los que se han mezclado técnicas o aplicado procedimientos de reducción dimensional, se han hecho sobre un dominio continuo. Nuestra propuesta se concentra sobre dominios no continuos sobre los cuales es muy frecuente aplicar técnicas factoriales, pero sobre las cuales no se aplican procesos gaussianos por su carencia de valores temporales o geoposicionados.

Como hemos visto anteriormente, los procesos gaussianos permiten obtener predicciones de una manera bastante precisa aprovechándose de la covarianza espacial o temporal en distribuciones normales. Por otro lado, los procedimientos de reducción dimensional nos permiten proyectar dominios no continuos sobre subespacios y así interpretarlos con mayor claridad. Estas coordenadas subespaciales son nuestro punto de partida.

3.1. Coordenadas Subespaciales

Las técnicas factoriales nos proporcionan una serie de coordenadas subespaciales que representan las proyecciones de los elementos de nuestra matriz de partida. Según la técnica que empleemos obtendremos diferentes proyecciones pero en cualquier caso, sea cual sea el procedimiento empleado, el objetivo es

describir nuestros datos sobre espacios de dimensión reducida.

Partamos de una matriz $\boldsymbol{X}_{N\times P}$ con N elementos y P variables, ninguna de ellas de con dominio continuo y estandarizada por columnas. Aplicando un procedimiento cualquiera de reducción dimensional obtenemos una serie de coordenadas subespaciales $\boldsymbol{R}_{N\times S}$ para cada uno de los N elementos en T subespacios.

Gracias a ambas matrices ($\boldsymbol{X}$ y $\boldsymbol{R}$) conformamos la matriz $\boldsymbol{Z}(u) = [\boldsymbol{X}(\boldsymbol{R})]$ que no es más que la matriz original estandarizada por columnas $\boldsymbol{X}$ con sus coordenadas subespaciales añadidas intrínsecamente. Esta matriz $\boldsymbol{Z}(u)$ es semejante a la empleada en procesos gaussianos espaciales y por tanto, aún sin disponer de una componente continua, podemos emplear la técnica cokriging.

3.1.1. Variogramas Cruzados Subespaciales

Antes de comenzar la interpolación subespacial, se ha de determinar la correlación subespacial de nuestras coordenadas. Para ello procedemos de igual modo que si estuviéramos en un dominio continuo calculando los variogramas experimentales.

Para simplificar nuestro caso supongamos que $T \leq 3$, permitiéndonos así una representación interpretable. Como disponemos de P variables en nuestra matriz de partida $\boldsymbol{X}$, tenemos que calcular $\frac{P(P+1)}{2}$ variogramas, P variogramas directos y $\frac{P(P-1)}{2}$ variogramas cruzados.

El vector de distancias intermedias o *lag* h se compone de k elementos tal que $h = \{h_1, h_2, ..., h_k\}$ y por tanto el variograma experimental se define como el conjunto de matrices $\boldsymbol{\Gamma}^*(h_k) = \gamma^*_{ij}(h_k)$ donde $i, j = 1, ..., P$.
En el caso en el que $i = j$ calculamos un variograma directo, siendo cruzado en el caso contrario $i \neq j$.

Para poder calcular todos los posibles variogramas, procedemos con la multiplicación y postmultiplicación de las funciones aleatorias $\boldsymbol{z}(x_\alpha)$, en cualquier coordenada subespacial α, por las k matrices $\boldsymbol{A}(h_k)$[31] de tal forma que:

$$\gamma^*_{ij}(h_k) = \boldsymbol{z}^T \boldsymbol{A}(h_k)\boldsymbol{z} \tag{3.1}$$

La matriz $\boldsymbol{A}(h_k)$ tiene dimensión $N \times N$ y se define como:

$$\boldsymbol{A}(h_k) = \frac{\boldsymbol{\eta}(h_k) - \boldsymbol{M}(h_k)}{J(h_k)} \tag{3.2}$$

Donde N es el número de muestras, $\boldsymbol{M}(h)$ es una matriz binaria $N \times N$ tal que sus componentes serán igual a uno si la distancia entre las localizaciones es igual a la distancia h_k y cero en caso contrario. La matriz $N \times N$ diagonal $\boldsymbol{\eta}(h_k)$ tiene como componentes el número de pares que poseé cada localización a la distancia h_k. Por último, $J(h_k)$ es el número total de pares que se encuentran a una distancia h_k. $\boldsymbol{A}$ representa la relación subespacial de nuestras coordenadas independientemente de los valores asociados a ellas.

La ecuación 3.2 reproduce de forma algebraica la ecuación 2.49.

$$\begin{aligned}\boldsymbol{A}(h_k) = \frac{\boldsymbol{M}(h_k) - \boldsymbol{\eta}(h_k)}{J(h_k)} = \\ = \frac{1}{J(h_k)} \begin{bmatrix} m_1(h_k) & -\eta_{12} & \dots & -\eta_{1N} \\ -\eta_{12} & m_2(h_k) & \dots & -\eta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -\eta_{1N} & -\eta_{2N} & \dots & m_N(h_k) \end{bmatrix}\end{aligned}$$

donde m_i es el número de pares que tiene el elemento i a la distancia h_k y

$$\eta_{ij} = \begin{cases} 1 \text{ para pares \{i,j\} que se encuentren en distancias } h_k \\ \\ 0 \text{ para pares \{i,j\} que no se encuentren en distancias } h_k \end{cases}$$

con lo cual obtenemos K matrices $\boldsymbol{A}$, cada una de ellas multiplicada y postmultiplicada por las diferentes combinaciones de vectores z_p donde $p = \{1, ..., P\}$. Y por tanto obtenemos K matrices $\boldsymbol{\Gamma}^*$ de dimensión $P \times P$ tal que

$$\boldsymbol{\Gamma}^*(h_k) = \begin{bmatrix} z_1\boldsymbol{A}(h_k)z_1^T & z_1\boldsymbol{A}(h_k)z_2^T & \dots & z_1\boldsymbol{A}(h_k)z_P^T \\ z_2\boldsymbol{A}(h_k)z_1^T & z_2\boldsymbol{A}(h_k)z_2^T & \dots & z_2\boldsymbol{A}(h_k)z_P^T \\ \vdots & \vdots & \ddots & \vdots \\ z_P\boldsymbol{A}(h_k)z_1^T & z_P\boldsymbol{A}(h_k)z_2^T & \dots & z_P\boldsymbol{A}(h_k)z_P^T \end{bmatrix}$$

Para formar un variograma γ_{ij}^* bastará con tomar los K valores correspondientes a la posición h_k tal que $\gamma_{ij}^* = [\Gamma_{ij}^*(h_1), ..., \Gamma_{ij}^*(h_K)]$

Una vez que tenemos los valores experimentales de los diferentes variogramas (directos y cruzados) debemos ajustar funciones teóricas a cada uno de ellos para poder modelizar nuestro subespacio.

3.2. Modelo Lineal Corregionalizado

Recordando del modelo univariante la estructura general del variograma teórico obtenemos $\gamma(h) = bg_u(h)$ y extrapolando a cada uno de los $P(P+1)/2$ variogramas resultantes obtenemos una matriz $\boldsymbol{\Gamma}(h)$ de dimensión $N \times N$ tal que:

$$\boldsymbol{\Gamma}(h) = \sum_{s=0}^{S} \boldsymbol{B}_s g_s(h) \tag{3.3}$$

Siendo $\boldsymbol{B}_s$ la matriz de corregionalización $N \times N$ de *sills* y $g_s(h)$ las funciones ajustadas con rango fijo para cada una de ellas. S es el número de anidaciones. Las matrices de corregionalización serán los únicos valores a determinar para ajustar la matriz $\boldsymbol{\Gamma}(h)$ ya que las funciones $g_s(h)$ con su respectivo rango deberán ser determinadas por el especialista basándose en la observación.

Para poder ajustar el modelo global, hemos de hacer mínima la función:

$$WSS(\boldsymbol{b}_{ij}) = (\boldsymbol{\gamma}_{ij}^* - \boldsymbol{G}\boldsymbol{b}_{ij})^T \boldsymbol{Cov}(\boldsymbol{b}_{ij})(\boldsymbol{\gamma}_{ij}^* - \boldsymbol{G}b_{ij}) \tag{3.4}$$

Siendo $i, j = 1, ..., P$; $\boldsymbol{b}$ los $S \times 1$ sills correspondientes a los variogramas anidados teóricos $\boldsymbol{\gamma}_{ij}$, por otra parte $\boldsymbol{\gamma}_{ij}^*$ corresponde al variograma experimental de dimensión $K \times 1$ y $\boldsymbol{G}$ es la matriz $K \times S$ correspondiente a las distintas funciones de los variogramas anidados. La matriz $\boldsymbol{Cov}(\boldsymbol{b}_{ij})$ es la matriz de varianza-covarianza subespacial de los errores aleatorios. Y WSS representa la suma de cuadrados ponderados.

Para el modelo global tendremos que minimizar[39] la siguiente ecuación:

$$WSS(\boldsymbol{B}) = \sum_{i=1}^{p} \sum_{j=1}^{p} WSS(\boldsymbol{b}_{ij}) \tag{3.5}$$

3.2.1. Matriz Varianza-Covarianza Subespacial

La matriz de varianza-covarianza es fundamental para el cálculo iterativo de las matrices $\boldsymbol{B}$. En nuestro caso $\boldsymbol{Cov}(\boldsymbol{b}_{ij})$[40] representa la variabilidad de nuestra muestra en función de las diferentes distancias h. Supongamos un par de distancias $\{h_k, h_{k'}\}$, este par tiene asociado otro correspondiente a sus variogramas experimentales tal que $\{\gamma_{ij}^*(h_k), \gamma_{ij}^*(h_{k'})\}$ y en consecuencia su matriz de covarianza $\boldsymbol{Cov}(\boldsymbol{b}_{ij})$ es equivalente a:

$$\begin{aligned} \boldsymbol{Cov}(\gamma_{ij}^*(h_k), \gamma_{ij}^*(h_{k'})) = & tr(\boldsymbol{A}(h_k)\boldsymbol{C}(b_{ij})\boldsymbol{A}(h_{k'})\boldsymbol{C}(b_{ij})) + \\ & tr(\boldsymbol{A}(h_k)\boldsymbol{C}(b_{ii})\boldsymbol{A}(h_{k'})\boldsymbol{C}(b_{jj})) \end{aligned} \tag{3.6}$$

donde $\boldsymbol{C}(b_{ij})$ es la matriz de covarianzas $N \times N$ entre $\boldsymbol{z}_i$ y $\boldsymbol{z}_j$ y $\boldsymbol{C}(b_{ii})$ la matriz varianza-covarianza $N \times N$ de $\boldsymbol{z}_i$. A su vez, la matriz $\boldsymbol{C}(b_{ij})$ puede calcularse tal que:

$$\boldsymbol{C}(b_{ij}) = \sum_{s=0}^{S} b_{ij,s}\boldsymbol{\rho}_s \tag{3.7}$$

$\boldsymbol{\rho}_s$ es la matriz $N \times N$ de correlación o correlación cruzada subespacial. Teniendo en cuenta que $\boldsymbol{\rho}_s$ es la misma para todos los semivariogramas, en consecuencia:

$$\boldsymbol{Cov}(\gamma^*_{ij}(h_k), \gamma^*_{ij}(h_{k'})) = \sum_{r=1}^{S}\sum_{q=1}^{S}[(b_{ij,r}b_{ij,q} + b_{ii,r}b_{jj,q})tr(\boldsymbol{\rho}_r \boldsymbol{A}(h_k)\boldsymbol{\rho}_q \boldsymbol{A}(h_{k'}))] \quad (3.8)$$

Por tanto, la traza sólo ha de calcularse una vez. Esto supone una ventaja para poder iterar más rápidamente y con la misma efectividad.

3.2.2. Algoritmo LMC Subespacial

Una vez definidas las matrices de varianza-covarianza $\boldsymbol{Cov}(b_{ij})$ y la función de control $WSS(\boldsymbol{B})$ podemos proceder a definir el sistema iterativo para el cálculo de las matrices *sill* $\boldsymbol{B}$[41].

Paso 0 - Inicializamos τ a 0 y evaluamos $WSS(\hat{\boldsymbol{B}}^\tau)$.

Paso 1 - Entre las S estructuras espaciales (variograma anidado) escogemos una s_0 y la sustraemos del variograma experimental dejando las $S - s_0$ estructuras restantes:

$$\boldsymbol{\Gamma}^*_{s_0}(h_k) = \boldsymbol{\Gamma}^*(h_k) - \sum_{s \neq s_0}^{S} \hat{\boldsymbol{B}}^\tau_s g_s(h_k),\ k = 1, ..., K \quad (3.9)$$

Paso 2 - Ajustamos cada $\boldsymbol{\gamma}^*_{ij,s_0}$ individualmente.

$$\hat{\boldsymbol{b}}^{\tau+1}_{ij,s_0} = (\boldsymbol{g}^T_{s_0}\boldsymbol{Cov}(\hat{b}^\tau_{ij})^{-1}\boldsymbol{g}_{s_0})^{-1}\boldsymbol{g}^T_{s_0}\boldsymbol{Cov}(\hat{b}^\tau_{ij})^{-1}\boldsymbol{\gamma}^*_{ij,s_0} \quad (3.10)$$

donde $\boldsymbol{g}_{s_0}$ es el vector $K \times 1$ correspondiente al vector función del variograma teórico con estructura espacial s_0 para la distancia específica $h_k, k = 1, ..., K$.

Paso 3 - Realizamos la descomposición espectral de la matriz $N \times N$ $\hat{\boldsymbol{b}}_{ij,s_0}^{\tau+1}$ para obtener $\boldsymbol{Q\Lambda Q}^T$, en la que $\boldsymbol{Q}$ es la matriz de vectores propios y $\boldsymbol{\Lambda}$ la matriz diagonal que contiene los valores propios correspondientes. Para garantizar que la matriz resultante $\hat{\boldsymbol{B}}_{s0}^{\tau}$ sea semidefinida positiva, eliminamos los valores propios negativos transformándolos en 0 obteniendo $\boldsymbol{\Lambda}^+$. La matriz $\hat{\boldsymbol{B}}_{s0}^{\tau}$ se sustituye por $\boldsymbol{Q\Lambda}^+\boldsymbol{Q}^T$. Posteriormente el nuevo vector $\boldsymbol{b}_{ij}^{\tau}$ de la estructura s_0 se sustituye en la matriz $\boldsymbol{Cov}(\hat{b}_{ij}^{\tau})$.
Repetimos los pasos 1 a 3 hasta que las S estructuras hayan sido recorridas al menos una vez cada una.

Paso 4 - Calculamos $WSS(\hat{\boldsymbol{B}}^{\tau+1})$ y comprobamos que la diferencia con respecto a $WSS(\hat{\boldsymbol{B}}^{\tau})$ es mayor que un valor predeterminado. Si es así, proseguimos la iteración hasta que se alcance el valor mínimo buscado.

En el algoritmo de iteración influyen enormemente los valores iniciales de los rangos y los sills. La descomposición en diferentes funciones también conlleva cambios importantes en el ajuste. Se recomienda encarecidamente estudiar previamente los variogramas experimentales, así como el gradiente estándar de h, es decir K.

En el caso de los rangos y sills podemos efectuar un ajuste previo como hacíamos para el caso de kriging, con el objetivo de poder calcular los valores iniciales. Por otra parte, si tenemos una composición de funciones anidadas, según como repartamos el peso de b_s, la iteración se realizará con esa condición. En resumen, las condiciones iniciales son muy importantes para el correcto desarrollo del ajuste.

3.2.3. Distribución Cuadrática

Para matrices $\boldsymbol{X}$ de dimensión $N \times P$ con variables continuas a las que aplicamos técnicas simples de reducción dimensional[10] (véase ACP o Biplot Clásicos) el ajuste de sus variogramas experimentales subespaciales suele seguir una distribución cuadrática $\gamma(h) = bh^a$ con $0 < a < 2$[42].

Las distribuciones lineales o cuadráticas, a diferencia de las distribuciones gaussianas, no alcanzan un valor *sill* límite y según las consideraciones geoestadísticas son adecuadas para aquellas distribuciones con propiedades aleatorias o de muy pequeña escala[43]. Sin embargo, estas respuestas no son adecuadas para la representación subespacial que proyectamos, ya que podemos modificar la escala a nuestro antojo. Por otro lado, cuando disponemos de muestras grandes nuestra distribución se aproxima de manera más clara hacia una función cuadrática, lo cual desmonta el supuesto de aleatoriedad.

La principal diferencia entre nuestras distribuciones subespaciales y las espaciales es la ausencia de barreras físicas. En una proyección subespacial no tendremos nunca que lidiar con accidentes geográficos que limitan nuestras observaciones. Al contrario, esta ausencia conlleva más ventajas que desventajas como veremos a continuación.

Esta carencia de obstáculos físicos provoca que no se alcance un valor límite de *sill* y por tanto que éste sea extensible en nuestro subespacio. Esta condición no afecta al ajuste LMC, sino al contrario, ya que hace que la convergencia sea más rápida y el número de iteraciones menor para mayor número de variables si comparamos con funciones anidadas. Podemos por tanto ajustar un número mayor de variables de manera efectiva y sin perder información en el ajuste.

3.3. Mallado

La construcción del mallado es básica para poder estimar nuevas combinaciones de variables. Para ello, a través de la nube de puntos generada por la técnica de reducción dimensional elegida, generamos una estructura de la misma dimensión T que nuestra proyección y la envolvemos con una estructura mayor. Esta estructura constituirá los límites de nuestra malla.

Dependiendo de la estructura de nuestras proyecciones podemos extender los límites a nuestra voluntad. No obstante, aquellos valores que se encuentren

más cercanos a nuestra zona de puntos tendrán menor error. La extensibilidad de nuestro mallado es otra de las ventajas de no encontrarnos con un medio físico continuo.

Por otra parte es necesario proporcionar un interespaciado adecuado entre puntos para no sobrecargar los cálculos computacionales ni tampoco dejar un intervalo demasiado amplio pues perderíamos precisión en nuestras estimaciones.

3.4. Cokriging Simple

Ya hemos definido el cokriging anteriormente. A continuación vamos a presentar la adaptación subespacial del método.

Partiendo de la matriz $\boldsymbol{Z}(u_\alpha)$ siendo u_α las coordenadas subespaciales de nuestra muestra, queremos hallar $\boldsymbol{Z}^*(u_0)$ siendo u_0 las coordenadas subespaciales de nuestro mallado.

$$Z_i^*(u_0) = m_i + \sum_{i=1}^{P} \sum_{\alpha=1}^{N} w_\alpha^i (Z_i(u_\alpha) - m_i) \tag{3.11}$$

Hemos de calcular los distintos vectores de pesos w_α^i. Para ello hemos de recordar la ecuación 2.24 correspondiente al kriging simple y la ecuación 3.7 que nos proporciona la relación entre la matriz $\boldsymbol{B}$ ajustada en el LMC y la matriz de covarianza subespacial $\boldsymbol{C}$. En forma matricial obtenemos

$$\boldsymbol{C}(u_\alpha - u_\beta)\boldsymbol{W}_s = \boldsymbol{B}_s \otimes \rho_s(u_0 - u_\alpha) \tag{3.12}$$

De la cual hemos de despejar la matriz de pesos $\boldsymbol{W}_s$ de dimensión $P \times (N \times N_0)$, siendo N_0 el número total de puntos de nuestro mallado, sabiendo que la

matriz $\boldsymbol{C}$ es igual a

$$\boldsymbol{C}(u_\alpha - u_\beta) = \sum_{s=1}^{S} \boldsymbol{B}_s \otimes \rho_s(u_\alpha - u_\beta) \tag{3.13}$$

y que $\boldsymbol{B}$ es

$$\boldsymbol{B} = \sum_{s=1}^{S} \boldsymbol{B}_s \tag{3.14}$$

De forma matricial[44]

$$\left(\sum_{s=1}^{S} \begin{bmatrix} b_{11,s} & \dots & b_{P1,s} \\ \vdots & \ddots & \vdots \\ b_{1P,s} & \dots & b_{PP,s} \end{bmatrix} \otimes \begin{bmatrix} \rho_s(u_{\alpha_1} - u_{\alpha_1}) & \dots & \rho_s(u_{\alpha_1} - u_{\alpha_N}) \\ \vdots & \ddots & \vdots \\ \rho_s(u_{\alpha_1} - u_{\alpha_N}) & \dots & \rho_s(u_{\alpha_N} - u_{\alpha_N}) \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{W}_{u_{01},s} \\ \vdots \\ \boldsymbol{W}_{u_{0N_0},s} \end{bmatrix} =$$

$$= \begin{bmatrix} b_{11,s} & \dots & b_{P1,s} \\ \vdots & \ddots & \vdots \\ b_{1P,s} & \dots & b_{PP,s} \end{bmatrix} \otimes \begin{bmatrix} \rho_s(u_{0_1} - u_{\alpha_1}) & \dots & \rho_s(u_{0_{N_0}} - u_{\alpha_N}) \\ \vdots & \ddots & \vdots \\ \rho_s(u_{0_1} - u_{\alpha_N}) & \dots & \rho_s(u_{0_{N_0}} - u_{\alpha_N}) \end{bmatrix}$$

Una vez calculada la matriz $\boldsymbol{W}$ y recordando que nuestra matriz de partida $\boldsymbol{X}$ está estandarizada por columnas $m_i = 0$, obtenemos

$$Z_i^*(u_0) = \sum_{i=1}^{P} \sum_{\alpha=1}^{N} w_\alpha^i Z_i(u_\alpha) \tag{3.15}$$

El error de predicción se calcula

$$var(\boldsymbol{Z} - \boldsymbol{Z}^*) = \boldsymbol{C}(0) - \sum_{s=1}^{S} \boldsymbol{c}_s^T \boldsymbol{C} \boldsymbol{c}_s \tag{3.16}$$

donde $\boldsymbol{c}_s = \boldsymbol{B}_s \otimes \rho_s(u_0 - u_\alpha)$

3.5. Validación Cruzada

Con objeto de validar nuestro modelo y de esta forma comprobar que el ajuste LMC, los valores introducidos o su posición en el subespacio proyectado

son correctos, empleamos el siguiente procedimiento corroborador[45].

Suprimimos un elemento $Z_i(u_\alpha)$ de la matriz $\boldsymbol{Z}_i$ quedándonos con una matriz de dimensiones $(N-1) \times P$. A continuación estimamos el valor $Z_i^*(u_{[\alpha]})$ con los $N-1$ valores restantes. Los corchetes alrededor de α simbolizan el hecho de que la estimación se realiza en la localización u_α excluyendo su valor $Z_{i\alpha}$.

La diferencia entre el valor real $Z_i(u_\alpha)$ y el estimado $Z_i^*(u_{[\alpha]})$ nos indica cómo de bien ajusta nuestro modelo en los alrededores de este punto. Este proceso se repite N veces, tantas como puntos disponemos en nuestra muestra. De esta forma comprobamos la bondad de nuestro modelo.

Para medir la eficiencia del modelo calculamos la desviación del valor cuadrático medio

$$\frac{1}{N}\sum_{\alpha=1}^{N}(Z_i(u_\alpha) - Z_i^*(u_{[\alpha]}))^2 \cong 0 \tag{3.17}$$

y la media cuadrática del error estimado

$$\frac{1}{N}\sum_{\alpha=1}^{N}\frac{(Z_i(u_\alpha) - Z_i^*(u_{[\alpha]}))^2}{\sigma_{[\alpha]}^2} \cong 1 \tag{3.18}$$

Siendo $\sigma_{[\alpha]}^2$ la desviación típica del error predicho en la posición u_α.

3.6. Predicciones

Dada una matriz $\boldsymbol{Y}$ de dimensión $N^* \times (P - P^*)$, siendo P^* el número de variables a predecir, con $P^* < P$ y N^* los elementos que queremos predecir, queremos hallar la matriz $\boldsymbol{Y}^*$ de dimensiones $N^* \times P^*$ a través de las proyecciones subespaciales calculadas $\boldsymbol{Z}^*$.

$$\boldsymbol{Y} \cup \boldsymbol{Y^*} = \left[\begin{array}{ccc|ccc} y_{11} & \cdots & y_{1(P-P^*)} & y^*_{1(P-P^*+1)} & \cdots & y^*_{1P^*} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{N^*1} & \cdots & y_{N^*(P-P^*)} & y^*_{N^*(P-P^*+1)} & \cdots & y^*_{N^*P^*} \end{array}\right]$$

$$\boldsymbol{Z^*} = \left[\begin{array}{ccc|ccc} z^*_{11} & \cdots & z^*_{1(P-P^*)} & z^*_{1(P-P^*+1)} & \cdots & z^*_{1P^*} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ z^*_{N_0 1} & \cdots & z^*_{N_0(P-P^*)} & z^*_{N_0(P-P^*+1)} & \cdots & z^*_{N_0P^*} \end{array}\right]$$

Sea $\boldsymbol{M}$ la matriz $(N_0 \times (P - P^*)) \times (N^* \times (P - P^*))$ de distancias entre $\boldsymbol{Z^*}$ e $\boldsymbol{Y}$ para sus variables comunes, el mínimo de cada columna de $\boldsymbol{M}$ nos proporciona el valor al cual corresponden las coordenadas u_0 (filas de $\boldsymbol{M}$) de nuestro mallado y con ellas hallamos la matriz $\boldsymbol{Y^*}$.

$$\boldsymbol{M} = \begin{bmatrix} m_{11} & \dots & m_{1N^*} \\ \vdots & \ddots & \vdots \\ m_{N_0 1} & \dots & m_{N_0N^*} \end{bmatrix} = \begin{bmatrix} d(z^*_{11}, y_{11}) & \dots & d(z^*_{1(P-P^*)}, y_{1(P-P^*)}) \\ \vdots & \ddots & \vdots \\ d(z^*_{N_0 1}, y_{N^*1}) & \dots & d(z^*_{N_0(P-P^*)}, y_{N^*(P-P^*)}) \end{bmatrix}$$

En resumen, buscamos aquellos valores predichos en el cokriging sobre el mallado más similares a los valores $\boldsymbol{Y}$ para las $P - P^*$ variables comunes o explicativas.

La elección de variables P^* es flexible según las condiciones de nuestro modelo. Al igual que en modelos clásicos de regresión, hemos de evitar colinealidad entre las variables para poder predecir correctamente.

3.7. Software

El algoritmo MGSR ha sido computado íntegramente en el lenguaje de programación R. Un paquete está abierto y disponible en el repositorio de Github

(victorvicpal/MGSR). Su primera versión está también disponible en zenodo (10.5281/zenodo.264102).

En el anexo se incluye un tutorial del uso del programa.

Capítulo 4

Aplicaciones

Las aplicaciones o casos en los que la Regresión Multivariante Gaussiana Subspacial (MGSR) puede ser aplicada son múltiples. Las dos condiciones principales son la carencia de dominio continuo, esencialmente espacial y una base de datos multivariante.

Aunque en el marco teórico hemos explorado diferentes técnicas con distintos tipos de variables, las dos aplicaciones que se van a introducir poseen exclusivamente variables continuas. No obstante, y aunque no se haya desarrollado, es posible aplicar estas técnicas más allá de este tipo de variables.

Los dos casos que se presentan a continuación provienen de materias dispares. Si bien el primero versa sobre operaciones quirúrgicas de implantación de stent, el segundo lo hace sobre cambios de color en piedras silíceas. Gracias a la inestimable ayuda del Instituto de Recursos Naturales y Agrobiología de Salamanca (IRNASA) y el equipo de Cardiología del Hospital Universitario de Salamanca tuvimos acceso a ambas bases de datos.

4.1. El modelo MGSR aplicado a la predicción de los efectos del stent (Postoperatorio y Follow-Up) en enfermos de infarto de miocardio

4.1.1. Introducción

En la actualidad la mayor parte de pacientes con enfermedades coronarias que requieren una revascularización son sometidos a una intervención percutánea coronaria (PCI) con implantación de stent (Fig. 4.1). El talón de Aquiles de la implantación de stent es su fracaso en mantener la permeabilidad del vaso sanguíneo con el tiempo debido a la aparición de reestenosis. La reestenosis es un problema clínico que normalmente se presenta como angina de pecho recurrente o incluso como infarto de miocardio requiriendo una intervención coronaria adicional[46, 47]. La identificación de pacientes con riesgo de reestenosis es un

desafío importante. Sin embargo, la utilidad clínica del stent no es clara[48, 49] ya que existen muchos factores de riesgo (paciente, procedimiento o aparato).

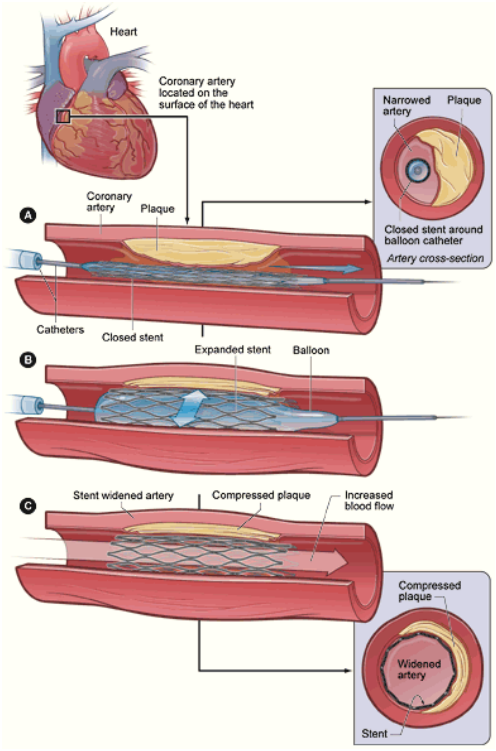


Figura 4.1: Intervención Percutánea Coronaria (PCI)

La reestenosis se define como la reducción en el lumen después de una PCI con o sin implantación de stent (Fig. 4.2). El proceso de reestenosis ocurre de manera gradual, normalmente entre los 3 y 12 meses posteriores a la inserción del stent. Un 10 % de los pacientes intervenidos presentan angina de pecho e incluso infarto de miocardio requiriendo repetidas PCI en estas ocasiones[50]. No obstante, una nueva generación de stents recubiertos de fármacos han reducido la incidencia de reestenosis.

Los datos utilizados en este estudio corresponden a los previamente publicados en el ensayo clínico GRACIA-3[51]. El Ensayo clínico aleatorio, abierto y multicéntrico GRACIA-3 evaluó la eficacia del stent liberador de fármaco (paclitaxel) frente a un stent convencional metálico. Los datos fueron recogidos de

20 hospitales españoles.

Para determinar la incidencia de reestenosis, se llevaron a cabo angiografías coronarias post-operatorias y a los 12 meses de la operación (follow-up). Todos los angiogramas fueron analizados en un laboratorio base independiente (ICICOR, Valladolid, España) a través de un sistema computacional cuantitativo (Medis, Leesburg, Va).

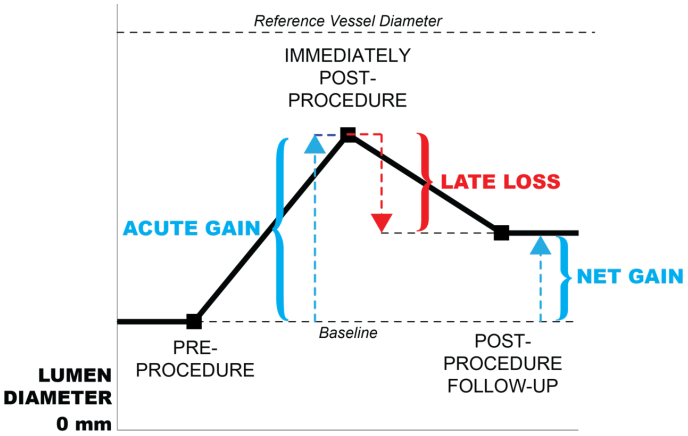


Figura 4.2: Ilustración esquemática de la reestenosis

La medición de reestenosis fue realizada por un lector experimentado, el cual no tenía conocimiento del tipo de stent implantado en el paciente. Las angiografías a los 12 meses fueron realizadas aleatoriamente a 299 de los 346 pacientes analizados (86 %).

La reestenosis en stent fue definida siguiendo el ratio de reestenosis binaria a los 12 meses de la operación. La angiografía de reestenosis binaria fue definida como un estrechamiento mayor del 50 % del lumen en el segmento obstruido (definido como la porción del vaso que recibe tratamiento con stent, incluyendo márgenes proximales y distales de 5 mm)[50].

La incidencia de reestenosis en el ensayo GRACIA-3 fue similar en ambos tipos de stent, recubierto y desnudo (11.3 % frente a 10.1 %; riesgo relativo, 1.06;

95 % intervalo de confianza 0.74 a 1.52; p=0.89). Aunque la pérdida de lumen (0.04±0.055mm frente a 0.27±0.057mm, p=0.003) se redujo en los grupos con recubrimiento (paclitaxel), no existen evidencias significativas entre el uso de tirofabina y alguna mejoría en la PCI.

Las mediciones angiográficas realizadas (Pre-operación, Post-operación y tras 12 meses) sirvieron para construir nuestro modelo predictivo. Las variables analizadas fueron la longitud del segmento afectado, el diámetro de referencia del segmento obstruido, las obstrucciones del segmento en las tres situaciones (previo, posterior y a los 12 meses de la PCI).

Como hemos enunciado anteriormente, 299 pacientes fueron sometidos a una angiografía a los 12 meses de la intervención, 119 de ellos (40 %) tenían datos ausentes. Es por ello que nuestra matriz de partida se redujo a 180 individuos.

Nuestro principal objetivo es predecir el comportamiento del lumen en estados posteriores a la PCI, tanto el post-operatorio como a los 12 meses de la angioplastia (FU). Dado que existen datos ausentes en cada una de las variables analizadas, el objetivo secundario fue predecir estos datos mediante el algoritmo MGSR.

La matriz de partida se compone de 180 individuos y 5 variables. La nomenclatura empleada para las variables es la siguiente:

- **Seg Length PRE :** longitud del segmento obstruido antes de la angioplastia (mm).
- **Diam Ref PRE :** diámetro de referencia en el segmento obstruido antes de la angioplastia (mm).
- **Diam Obs PRE :** diámetro del lumen en el segmento obstruido antes de la angioplastia (mm).

- **Diam Obs POST :** diámetro del lumen tras la angioplastia (mm).
- **Diam Obs FU :** diámetro del lumen a los 12 meses de la angioplastia (mm).

La Tabla 4.1 muestra las medias, errores estándar, máximos y mínimos de cada una de las 5 variables analizadas.

PCI	**PRE**			**POST**	**FU**
	Seg. Length	Obs. Diam.	Ref. Diam.	Obs. Diam.	Obs. Diam.
Media	43.42	0.78	2.89	2.35	2.22
S.E.	1.16	0.03	0.04	0.04	0.04
Máx.	111.98	1.91	4.32	3.68	3.53
Mín.	8.86	0.00	1.55	1.09	0.46

Tabla 4.1: Descriptiva básica de las variables analizadas.

4.1.2. Resultados

Debido a que todas las variables son cuantitativas, aplicamos un JK-Biplot a la matriz de partida. Además se realizó una estandarización por columnas. La Tabla 4.2 muestra los resultados del JK-Biplot. Como se observa, la inercia absorbida por el primer plano principal es 69 %. Para simplificar los análisis posteriores (sistemas computacionales muy pesados), y como la pérdida de información es relativamente pequeña, escogemos el primer plano para realizar nuestro análisis.

Dimensión	1	2	3	4	5
Valores propios	401.04	215.59	109.30	92.07	76.98
Inercia	44.81	24.09	12.21	10.29	8.60
Inercia acumulada	44.81	68.90	81.11	91.40	100.00

Tabla 4.2: Resultados del análisis JK-Biplot

En la Fig. 4.3 se puede ver la representación conjunta de filas y columnas del análisis JK-Biplot en el primer plano principal. Este análisis nos permite observar la estructura de las variables analizadas. Como se aprecia, las variables más

correlacionadas son el diámetro de referencia en el segmento obstruido antes de la angioplastia (Diam Ref PRE) y el diámetro del lumen post-operatorio (Diam Obs POST). Además, la longitud del segmento en el segmento obstruido antes de la angioplastia (Seg Length PRE) tiene cierto grado de independencia lineal con respecto a las dos anteriores.

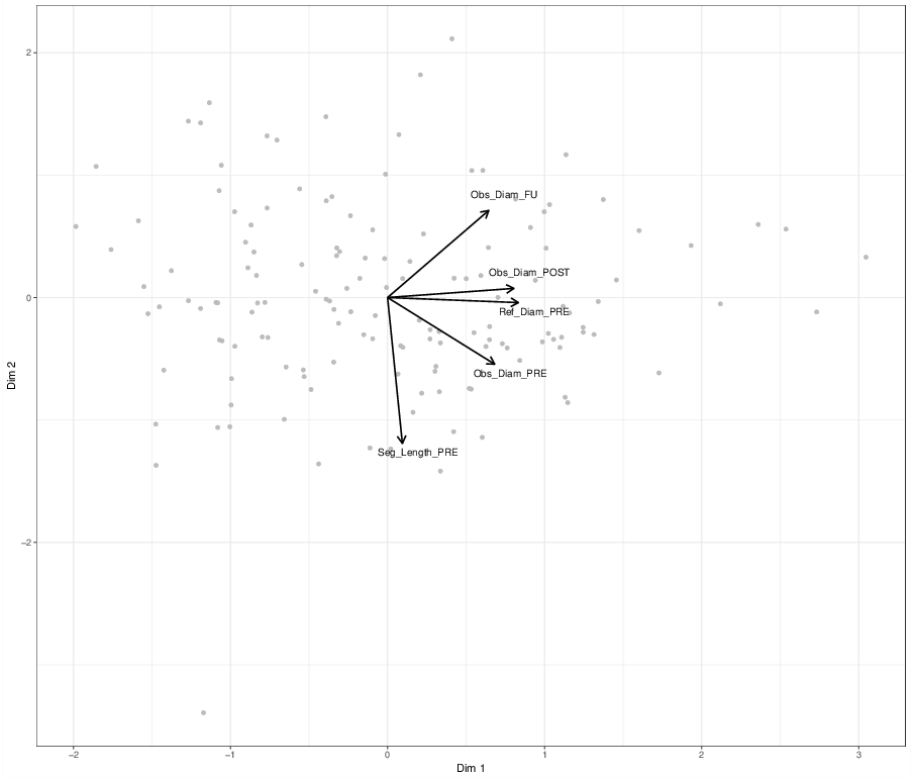


Figura 4.3: Representación JK-Biplot en el plano 1-2.

Agregamos las coordenadas subespaciales obtenidas del JK-Biplot a nuestra matriz de partida para aplicar nuestro modelo MGSR. A continuación, construimos los 5 variogramas directos y los 10 cruzados. Por último, aplicamos un modelo de corregionalización lineal (LMC). En la Fig. 4.4 se pueden ver los resultados del ajuste.

Los variogramas directos y cruzados (Fig. 4.4) tienen correlación subespacial positiva en su mayoría. La única excepción es el variograma cruzado formado

entre las variables Seg Length PRE y Obs Diam Fu. Esta correlación subespacial es apreciable en el comportamiento de las mismas variables en el JK-Biplot (Fig. 4.3).

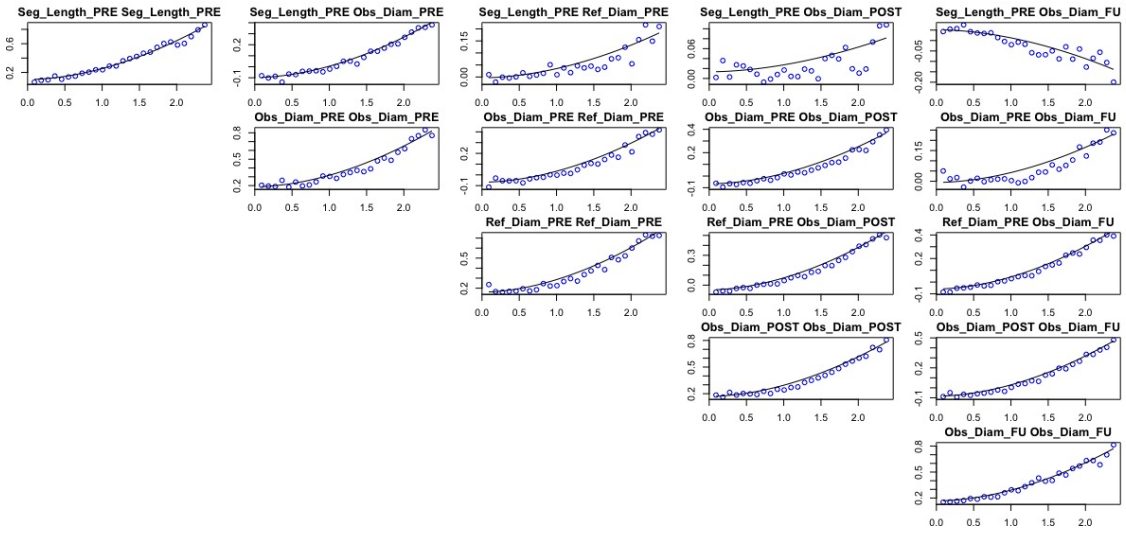


Figura 4.4: Variogramas directos y cruzados del modelo.

En la Tabla 4.3 se muestran los valores ajustados de los sill (Nugget y Power) para cada uno de los variogramas. Son destacables los valores casi despreciables de los variogramas cruzados correspondientes a la longitud del segmento en el segmento obstruido antes de la angioplastia (Seg Length PRE) y el resto. En las Fig. 4.4 se puede apreciar la falta de correlación subespacial existente entre el resto de variables y Seg Length PRE.

En la Tabla 4.4 observamos los resultados de la validación cruzada. La raíz del error cuadrático medio (RMSE) es cercana a los 0.3 mm para los datos correspondientes a la predicción del diámetro de lumen (Obstrucción del diámetro post-operatorio y a los 12 meses de la PCI). Además, el pseudo-R^2 es cercano a 1 en todas las variables, rubricando la hipótesis defendida por Wackernagel[45].

Variograma	Variables	Nug	Pow
Directo	Seg. Len. - Seg. Len.	0.10	0.15
	Obs. D. PRE - Obs. D. PRE	0.19	0.13
	Ref. D. - Ref D.	0.16	0.13
	Obs. D. POST - Obs. D. POST	0.17	0.13
	Obs D. FU - Obs D. FU	0.17	0.12
Cruzado	Seg. Len. - Obs. D. PRE	-0.10	0.10
	Seg. Len. - Ref. D.	-0.00	0.04
	Seg. Len. - Obs. D. POST	0.01	0.01
	Seg. Len. - Obs. D. FU	0.05	-0.04
	Obs D. PRE - Ref D.	-0.07	0.10
	Obs D. PRE - Obs. D. POST	-0.07	0.09
	Obs D. PRE - Obs. D. FU	-0.01	0.05
	Ref D. PRE - Obs. D. POST	-0.05	0.12
	Ref D. PRE - Obs. D. FU	-0.06	0.10
	Obs. D. POST - Obs. D. FU	-0.08	0.11
		Rango	1.83

Tabla 4.3: Sills LMC

Una vez hecha la validación cruzada, construimos un mallado adecuado a la distribución subespacial formada por el JK-Biplot. En este caso, el mallado se construyó dejando un espacio de 0.1 unidades en cada uno de los márgenes de los primeras dimensiones. Este espectro es suficiente ya que nuestros pacientes representan bien la población que pretendemos abarcar. Además, el espacio entre puntos del mallado se estableció en 0.1 unidades. Este interespaciado es adecuado tanto para su computación como para su análisis.

	PRE			**POST**	**FU**
	Seg. Length	Obs. Diam.	Ref. Diam.	Obs. Diam.	Obs. Diam.
RMSE	0.19	0.40	0.33	0.35	0.32
R^2	0.99	0.99	0.99	0.99	0.99

Tabla 4.4: Validación cruzada

El siguiente procedimiento fue realizar un cokriging simple cuyos resultados se observan en la Fig. 4.5. La estructura que observamos en las distintas

variables recuerda a la representada en la Fig. 4.3, ya que direcciones y sentidos coinciden en ambas.

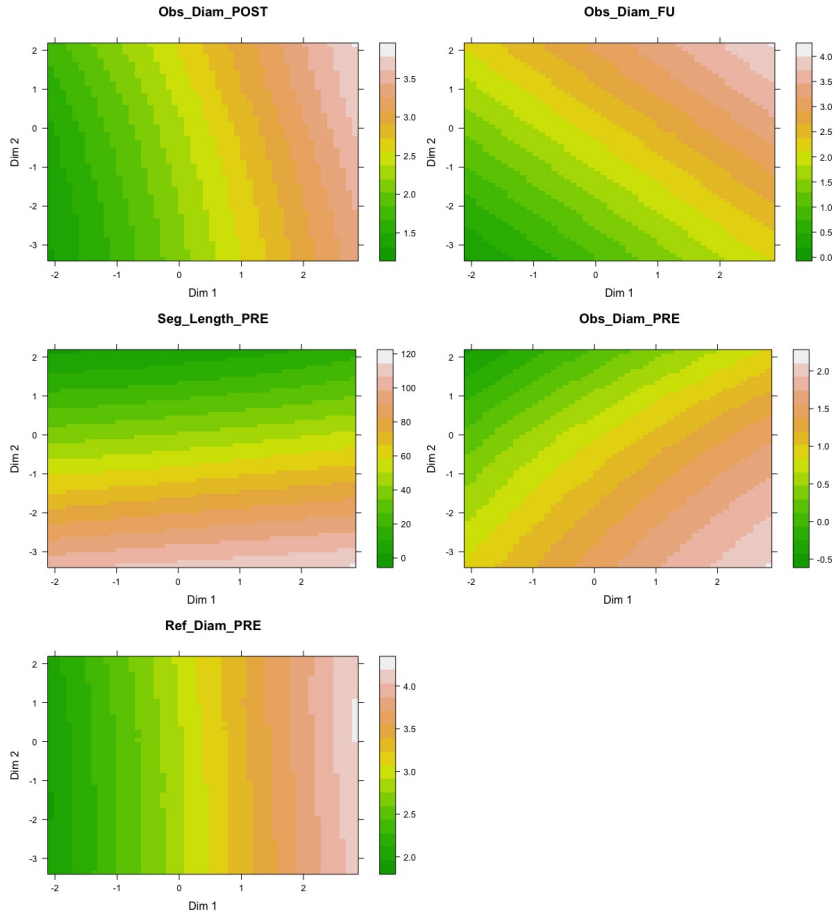


Figura 4.5: Resultados del cokriging simple.

En la Fig. 4.6 se muestra la varianza residual proyectada sobre nuestro mallado. Podemos observar un comportamiento descendiente en su distribución. Los puntos cercanos al centro presentan un máximo positivo que disminuye hasta llegar a cotas negativas de manera concéntrica (los valores negativos presentan un gradiente mayor que los positivos). Las zonas que comprenden el valor máximo y su opuesto negativo representan las zonas de confianza de nuestro modelo.

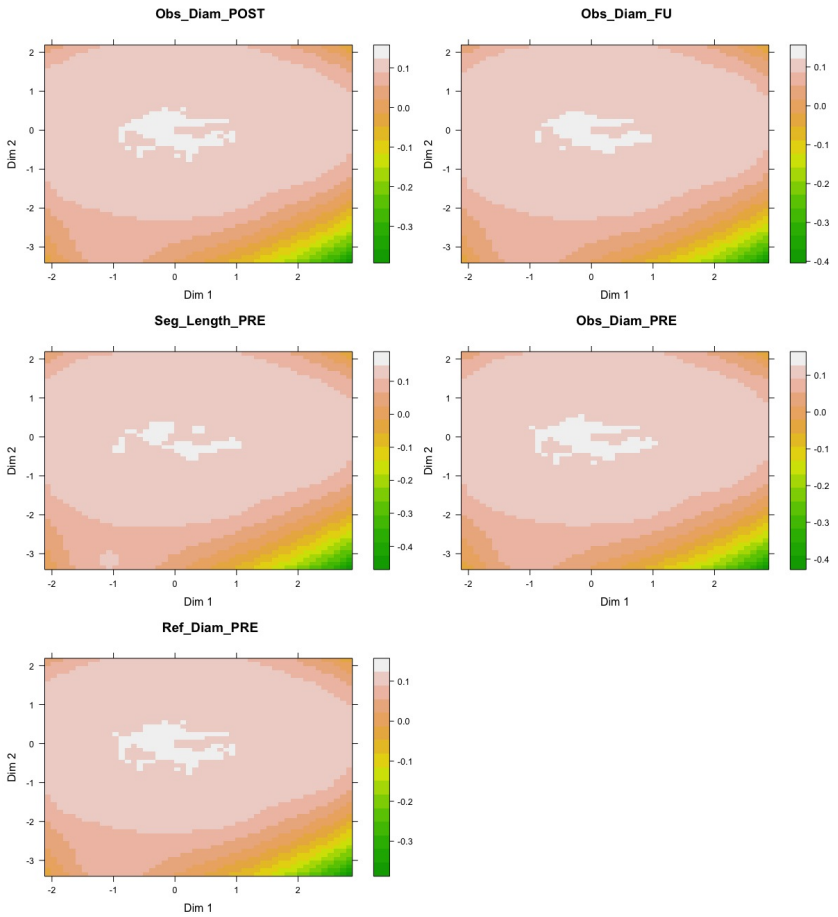


Figura 4.6: Varianza residual proyectada sobre el mallado.

Como ejemplo, observemos la obstrucción del diámetro tras la angioplastia (Obs Diam POST). El valor máximo en este caso se concentra en la zona cercana al origen y tiene un valor de +0.15 luego nuestra zona de confianza será la comprendida entre -0.15 y +0.15. Este mismo símil puede aplicarse al resto de predicciones.

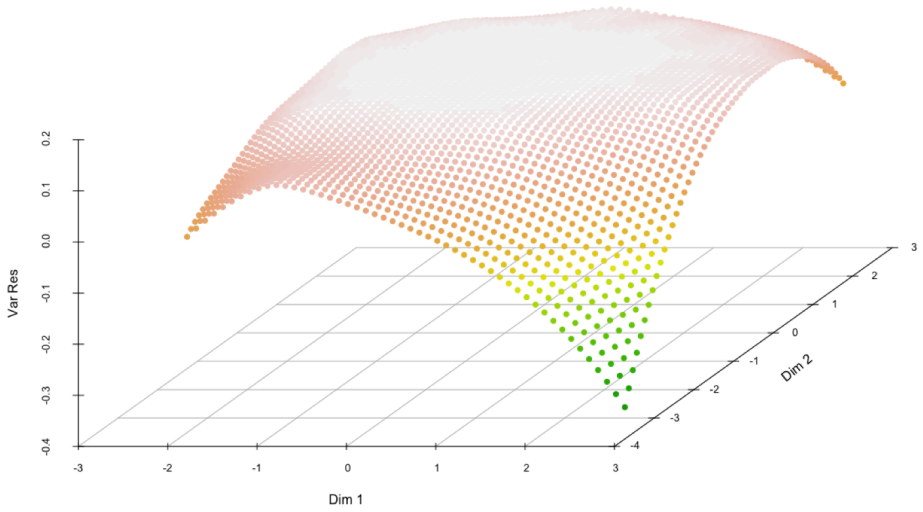


Figura 4.7: Varianza residual de la variable Obs Diam POST.

En la Fig. 4.7 se observa la reproducción en tres dimensiones de la varianza residual calculada para la obstrucción del diámetro del vaso tras la angioplastia (Obs Diam POST).

4.1.3. Validez y estabilidad del modelo

Para comprobar la validez de nuestro modelo, hemos dividido la matriz de partida en dos submatrices (Train/Test), para ello extraemos aleatoriamente de la matriz de partida un 20 % de los pacientes (Test). Con el restante 80 % (Train) construimos un nuevo modelo. Una vez aplicado el algoritmo MGSR sobre el Train, calculamos las predicciones para los pacientes extraídos y analizamos sus residuales. Además, para confirmar la estabilidad de nuestro modelo, hemos realizado 10 análisis como el anteriormente descrito.

Nuestras matrices Train se componen de 144 pacientes y 5 variables (longitud del segmento, diámetro de referencia y lumen en el segmento obstruido antes de la angioplastia, y el diámetro del lumen post-operatorio y a los 12 me-

ses de la intervención). La matriz Test la constituyen los 36 pacientes restantes.

Cada uno de los 10 modelos sigue la misma estructura que el modelo general. Debido a la gran computación necesaria para realizar los 10 modelos, se utilizó un clúster de computación para los cálculos.

Media	S.E.	Mín	Máx
69.11	0.37	67.15	70.64

Tabla 4.5: Media, error estándar (S.E.), máx. y mín. inercia de los 10 modelos.

Aplicamos un JK-Biplot a cada uno de los modelos. Obtenemos una inercia media cercana al 70 % en el primer plano en todos los análisis (Tabla 4.5). Al igual que en el modelo general, escogemos el primer plano para realizar nuestro análisis.

A modo de ejemplo en la Fig. 4.8 podemos observar uno de los 10 JK-Biplots calculados.

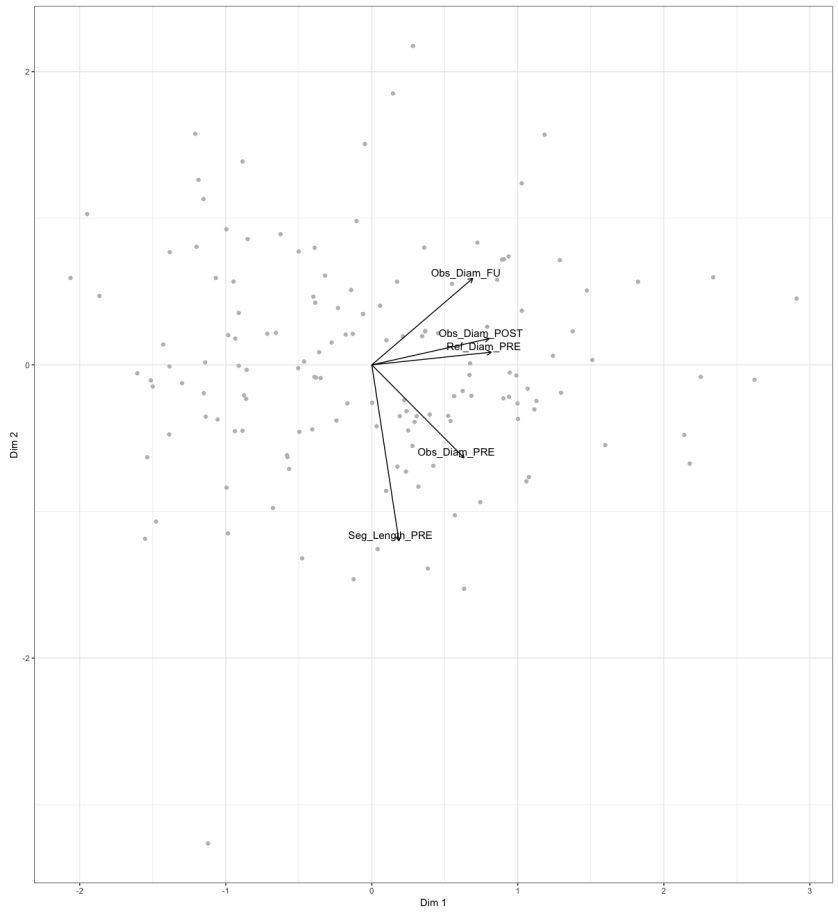


Figura 4.8: Representación JK-Biplot de uno de los 10 modelos.

Gracias a las coordenadas subespaciales obtenidas en los 10 JK-Biplot, podemos agregarlas a nuestras matrices de partida. Esta conjunción de coordenadas y matriz de partida estandarizada nos permite aplicar los modelos de corregionalización lineal (LMC) a cada una de las matrices.

Variograma	**Media**		**S.E.**		**Mín**		**Máx**	
Directo	Nug	Pow	Nug	Pow	Nug	Pow	Nug	Pow
Seg Len. - Seg Len.	0.10	0.16	0.00	0.01	0.09	0.14	0.12	0.19
Obs. D. PRE - Obs D. PRE	0.17	0.14	0.01	0.00	0.15	0.12	0.20	0.17
Ref. D. - Ref. D.	0.15	0.15	0.00	0.00	0.12	0.13	0.17	0.17
Obs. D. POST - Obs. D. POST	0.17	0.14	0.01	0.00	0.15	0.12	0.20	0.16
Obs. D. FU - Obs. D. FU	0.17	0.14	0.00	0.00	0.14	0.11	0.19	0.16
Cruzado	Nug	Pow	Nug	Pow	Nug	Pow	Nug	Pow
Seg. Len. - Obs. D. PRE	-0.11	0.11	0.00	0.00	-0.12	0.10	-0.10	0.13
Seg. Len. - Ref. D. PRE	0.01	0.03	0.00	0.00	-0.00	0.01	0.02	0.04
Seg. Len. - Obs. D. POST	0.01	0.01	0.00	0.00	-0.00	-0.01	0.02	0.02
Seg. Len. - Obs. D. FU	0.06	-0.05	0.00	0.00	0.04	-0.06	0.07	-0.03
Obs. D. PRE - Ref. D.	-0.07	0.10	0.00	0.00	-0.10	0.08	-0.05	0.12
Obs. D. PRE - Obs. D. POST	-0.05	0.08	0.00	0.00	-0.07	0.08	-0.04	0.09
Obs. D. PRE - Obs. D. FU	0.01	0.04	0.00	0.00	-0.02	0.02	0.02	0.05
Ref. D. - Obs. D. POST	-0.05	0.14	0.01	0.00	-0.08	0.12	-0.03	0.16
Ref. D. - Obs. D. FU	-0.06	0.12	0.00	0.00	-0.09	0.09	-0.05	0.13
Obs. D. POST - Obs. D. FU	-0.09	0.13	0.01	0.00	-0.12	0.10	-0.07	0.14

Tabla 4.6: Media, error estándar (S.E.), máx. y mín. de los ajustes LMC.

En la Tabla 4.6 podemos ver el resumen de los resultados de los 10 ajustes LMC para cada modelo. Los errores estándar son nulos en su práctica mayoría, lo cual indica la estabilidad de nuestro modelo.

Modelo	1	2	3	4	5	6	7	8	9	10
Rango	1.81	1.67	1.67	1.80	1.80	1.80	1.60	1.60	1.80	1.60

Tabla 4.7: Rangos de los 10 modelos.

Los rangos de los modelos ajustados oscilaron entre 1.6 y 1.8 como se observa en la Tabla 4.7.

El mallado se realizó del mismo modo que el empleado en el modelo general. En cada modelo se tuvo en cuenta la distribución subespacial de cada uno de los JK-Biplot.

A continuación llevamos a cabo cada uno de los cokriging simples correspondientes. En la Fig. 4.9 se muestra el resultado del modelo seleccionado como ejemplo. Como se aprecia claramente, la estructura es muy parecida a la presentada en el modelo general (Fig. 4.9).

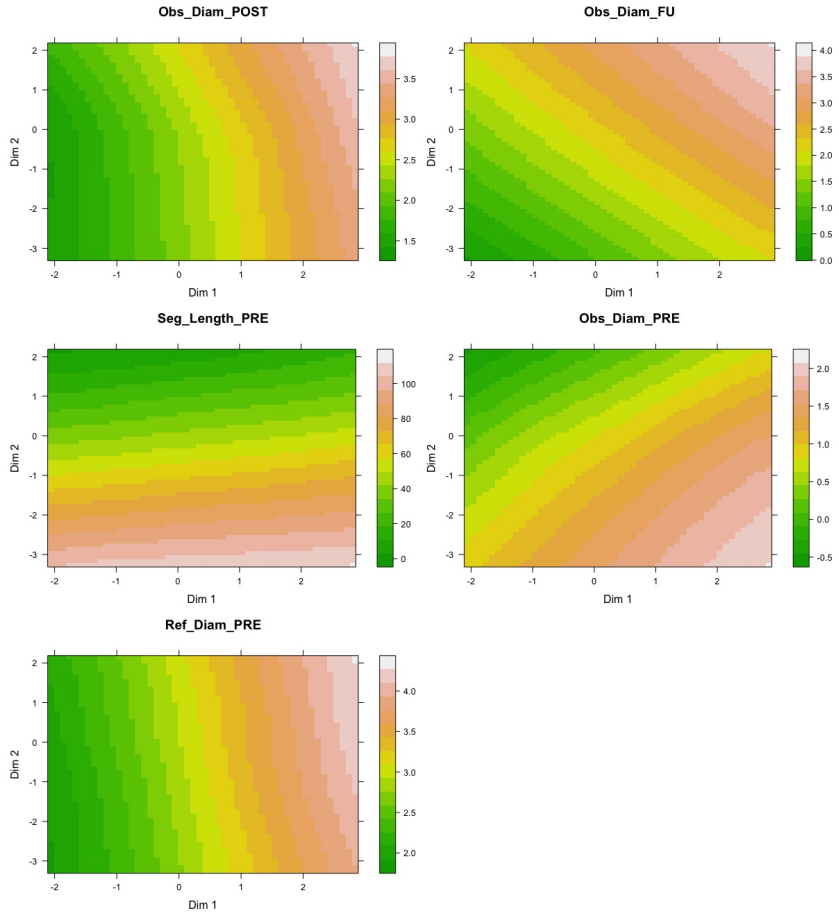


Figura 4.9: Cokriging Simple aplicado a uno de los modelos

En la Fig. 4.10 se muestra la varianza residual del modelo seleccionado. Los resultados son semejantes a los ya visualizados en el modelo general (Fig. 4.6).

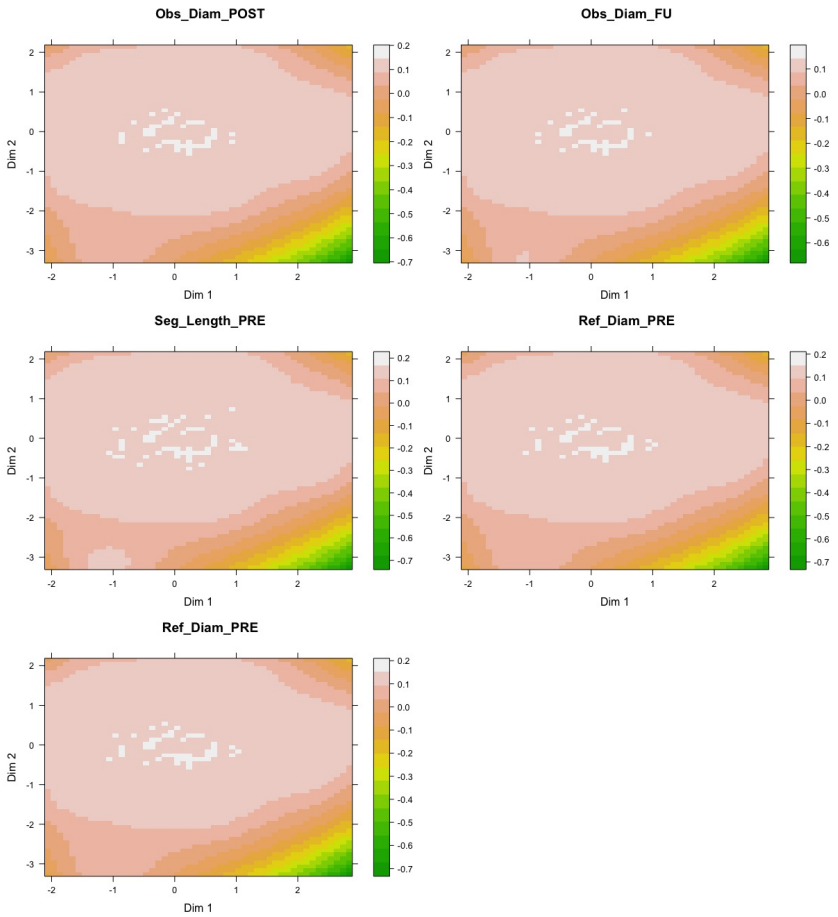


Figura 4.10: Varianza residual de uno de los modelos.

Se verificó la validez de los modelos en función de los valores extraídos en las matrices Test. Partiendo de los valores previos a la PCI (longitud del segmento, diámetro de referencia y lumen en el segmento obstruido antes de la angioplastia), calculamos los valores del diámetro de lumen post-operatorio y a los 12 meses de la angioplastia. El resumen de los residuales calculados se puede ver en la Tabla 4.8. Los resultados son estables lo que confirma las hipótesis de estabilidad y validez.

			PRE		POST	FU
Modelo		Seg. Len.	Obs D.	Ref D.	Obs D.	Obs D.
1	Media	0.03	0	-0.03	-0.01	0
	S.E.	0.01	0.05	0.04	0.09	0.11
	Mín	-0.16	-0.89	-0.52	-1.14	-1.42
	Máx	0.18	0.6	0.63	1.1	1.71
2	Media	-0.01	0.11	-0.03	0.2	-0.08
	S.E.	0.01	0.06	0.03	0.07	0.08
	Mín	-0.15	-0.6	-0.32	-0.51	-1.27
	Máx	0.15	0.97	0.43	1.32	1.14
3	Media	-0.02	-0.01	-0.05	0.12	-0.02
	S.E.	0.02	0.06	0.04	0.07	0.09
	Mín	-0.37	-0.65	-0.54	-0.64	-1.25
	Máx	0.19	0.61	0.51	1.18	1.01
4	Media	-0.01	0.03	-0.04	-0.05	-0.09
	S.E.	0.01	0.06	0.03	0.08	0.1
	Mín	-0.19	-0.57	-0.63	-1.07	-1.15
	Máx	0.16	1.14	0.35	1.15	1.6
5	Media	-0.03	-0.05	0.01	-0.05	-0.02
	S.E.	0.02	0.06	0.04	0.08	0.09
	Mín	-0.29	-0.84	-0.47	-1.01	-1.21
	Máx	0.15	0.71	0.46	1.09	1.15
6	Media	-0.01	0.07	-0.03	-0.07	0.09
	S.E.	0.02	0.07	0.03	0.09	0.1
	Mín	-0.22	-0.87	-0.35	-1.39	-0.85
	Máx	0.18	1.25	0.37	0.95	1.46
7	Media	0	0.11	-0.03	-0.03	0.02
	S.E.	0.02	0.06	0.03	0.1	0.1
	Mín	-0.42	-0.67	-0.49	-1.32	-1.57
	Máx	0.26	1.1	0.28	1.09	1.7
8	Media	-0.01	0.07	-0.02	-0.17	-0.2
	S.E.	0.02	0.06	0.04	0.08	0.08
	Mín	-0.28	-0.76	-0.44	-1.23	-1.1
	Máx	0.27	1.06	0.69	0.84	0.94
9	Media	-0.07	-0.01	0	-0.01	-0.08
	S.E.	0.07	0.07	0.04	0.09	0.1
	Mín	-2.08	-0.76	-0.7	-1.36	-1.13
	Máx	0.52	1.06	0.64	1.39	1.23
10	Media	0	0.04	-0.02	-0.09	0.05
	S.E.	0.02	0.06	0.03	0.09	0.1
	Mín	-0.3	-0.82	-0.47	-0.99	-0.96
	Máx	0.19	0.88	0.45	1.23	1.37

Tabla 4.8: Media, error estándar (S.E.), máx. y mín. de los residuales.

Para poder ilustrar cómo funcionaría el modelo en un caso real utilizamos como ejemplo un paciente al que se le va a someter a una PCI. Este paciente presenta una longitud de segmento de vaso obstruido de 68 mm, un diámetro de referencia del segmento de 2.59 mm y un diámetro de lumen de 1 mm (Tabla 4.9). Con estos valores iniciales, nuestro modelo calcula que el paciente en cuestión tras la angioplastia habrá incrementado 1.47 mm el diámetro de lumen (2.47 mm) y a los 12 meses pierde 0.29 mm.

PCI		**PRE**		**POST**	**FU**
	Seg. Len.	Obs. D.	Ref D.	Obs. D.	Obs. D.
Real	68.28	1.00	2.59	2.06	1.58
Predicho	68.35	0.87	2.47	1.83	1.54
Residual	-0.07	0.13	0.12	0.23	0.04

Tabla 4.9: Ejemplo de predicción del modelo para el paciente seleccionado.

En la Tabla 4.9 se muestran los valores reales, predichos y la diferencia entre ellos. La Fig. 4.11 representa los estados de predicción y reales de este paciente.

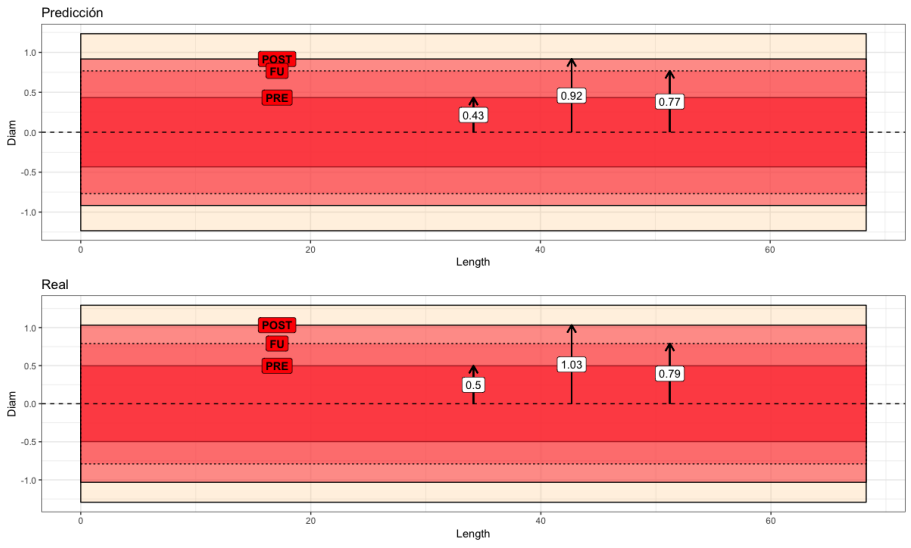


Figura 4.11: Representación de la predicción para el paciente seleccionado.

Nuestro modelo a su vez, nos permite detectar aquellos pacientes que se sa-

len de un patrón normal. Como ejemplo se ha seleccionado otro paciente, que sufrió reestenosis. En la tabla 4.10 se recogen los resultados obtenidos por nuestro modelo. Como se aprecia, los resultados de la predicción en el estado pre-angioplastia se alejan mucho de la realidad. No obstante, nuestra predicción apunta hacia una posible reestenosis futura, que como vemos en los valores reales sí se produjo.

PCI		**PRE**		**POST**	**FU**
	Seg. Len.	Obs. D.	Ref D.	Obs. D.	Obs. D.
Real	111.98	0.94	2.32	1.88	0.59
Predicción PRE	111.74	1.86	3.58	2.82	1.65

Tabla 4.10: Predicción de un paciente que sufrió restenosis.

En la Fig. 4.12 se pueden observar tanto los valores medidos en las angiografías como la predicción de nuestro modelo. El hecho de que los valores de partida estén alejados de los predichos nos indica que el paciente no se ajusta al comportamiento normal de la población. Este factor nos ayuda a detectar pacientes con posibles anomalías.

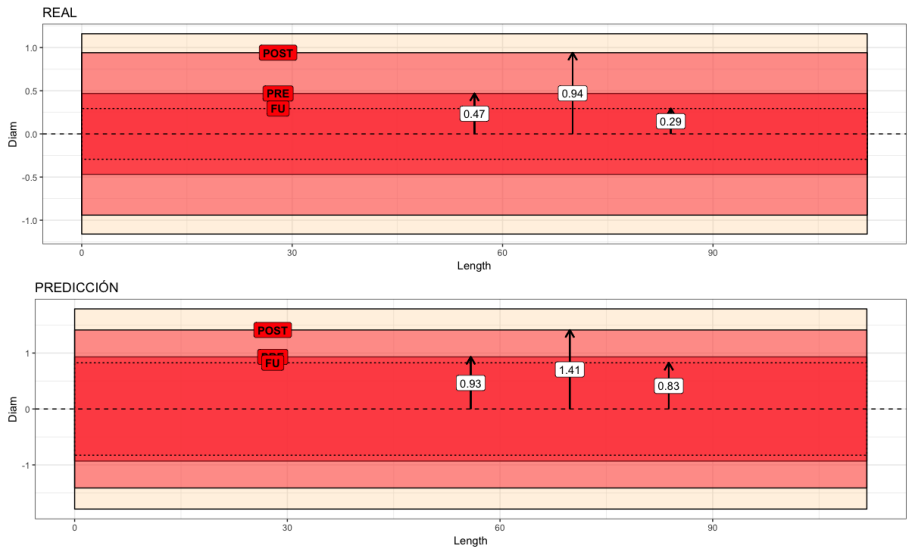


Figura 4.12: Predicción Reestenosis PRE y POST

Otra ventaja de nuestro modelo es la autocorrección. Pongamos como ejemplo un tercer paciente cuyas características aparecen en la tabla 4.11. Supongamos que dicho paciente va a ser intervenido y nuestro modelo prevé que la ganancia será de 1.15 mm de diámetro. Sin embargo, tras la implantación del stent esta mejora es únicamente de 0.8 mm (1.74-0.94 mm). Este resultado de por sí nos puede indicar que el paciente no ha respondido como se esperaba.

Para corregir el error presente en nuestro modelo, introducimos el valor medido en la angiografía post-operatoria y calculamos de nuevo el valor que obtendremos a los 12 meses. Como se aprecia en la Fig. 4.13, la predicción POST en el seguimiento es inferior que la calculada previamente, acercándose más a la realidad.

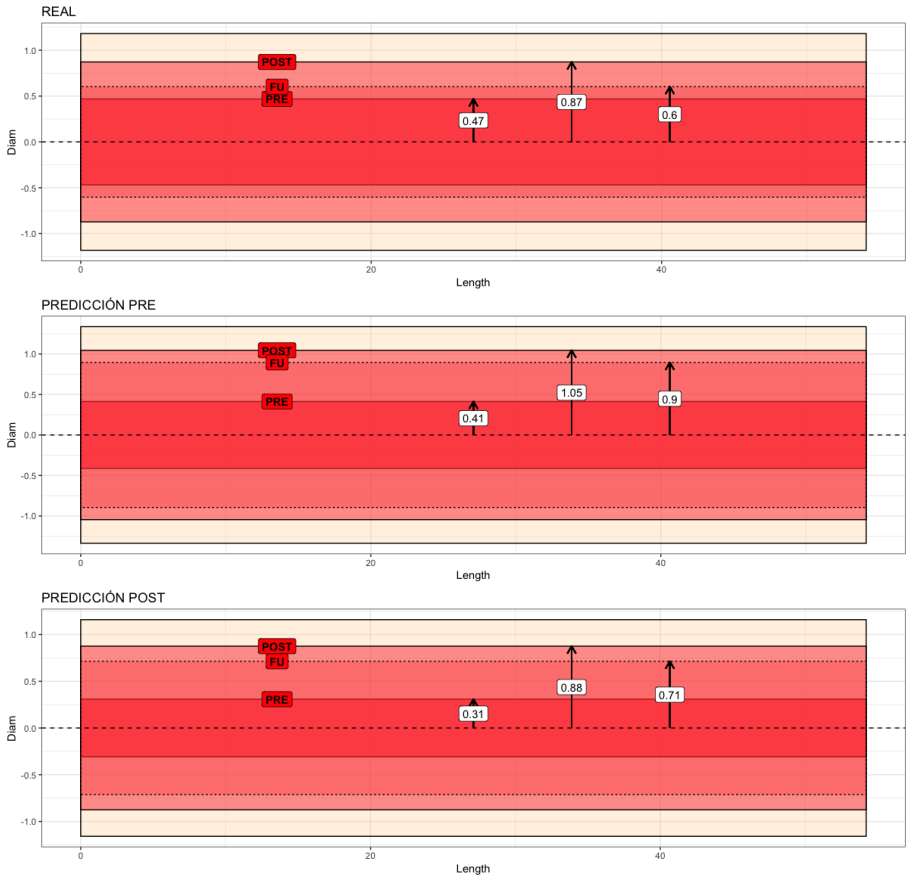


Figura 4.13: Predicción para nuestro paciente de referencia.

PCI		**PRE**		**POST**	**FU**
	Seg. Len.	Obs. D.	Ref D.	Obs. D.	Obs. D.
Real	54.10	0.94	2.36	1.74	1.20
Predicción PRE	54.16	0.83	2.67	2.09	1.79
Predicción POST	54.20	0.62	2.32	1.75	1.43

Tabla 4.11: Predicción Restenosis PRE y POST

Analicemos por último un paciente al que no se le realizó una angiografía a los 12 meses de la angioplastia (4.12). Este paciente presentó un incremento de diámetro de lumen tras la PCI de 2.2 mm (3.6-1.4). Nuestro modelo predice que a los 12 meses este paciente sufrirá un decremento de 0.72 mm (3.6-2.88) respecto a la angioplastia. En la Tabla 4.12 se pueden ver los valores exactos de las variables analizadas y en la Fig. 4.14 una representación del estado final del paciente.

PCI		**PRE**		**POST**	**FU**
	Seg. Len.	Obs. D.	Ref D.	Obs. D.	Obs. D.
Real	76.50	1.40	3.71	3.60	-
Predicción	76.60	1.60	3.97	3.16	2.88
Residuales	-0.11	-0.20	-0.26	0.44	-

Tabla 4.12: Paciente con valor ausente.

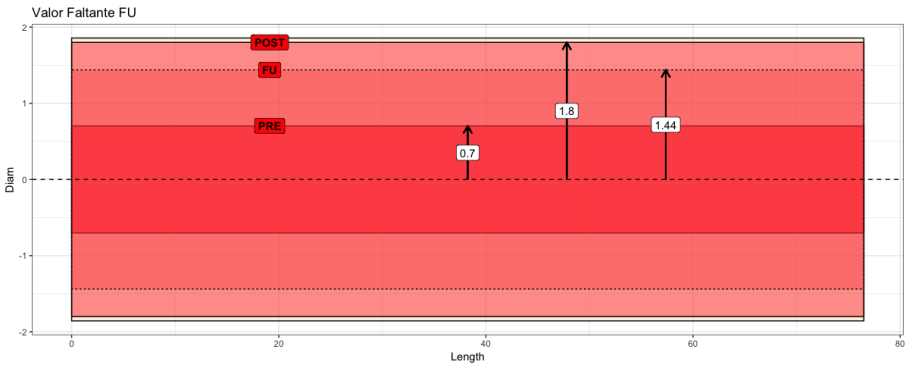


Figura 4.14: Ejemplo de predicción de valores ausentes.

Como conclusiones finales de nuestro análisis se ha conseguido reproducir el comportamiento del diámetro obstruido del lumen tras angioplastia. Este

modelo permite conocer el estado futuro de un paciente antes de siquiera ser intervenido, lo cual puede resultar de gran ayuda para poder analizar con mayor detalle la evolución del paciente. Además, es posible detectar pacientes cuyas características se salgan de lo normal y gracias a ello poder realizar un seguimiento más exhaustivo.

Una de las ventajas del MGSR es su flexibilidad. Como hemos indicado anteriormente, existen numerosos valores perdidos en las distintas variables analizadas. No obstante, podemos estimar estos valores en función de los datos disponibles.

La estructura de los 10 modelos Train/Test ha demostrado que el modelo propuesto es estable y válido.

El modelo MGSR es capaz de autocorregirse e igualmente reajustarse con la incorporación de nuevos pacientes. Este reajuste no supondría un esfuerzo computacional grande ya que la iteración se realizaría en base a lo ya conocido.

4.2. Predicción de los efectos de la cristalización de fosfatos en el envejecimiento de conglomerados silíceos

4.2.1. Introducción

Antes de realizar una determinada intervención de restauración y/o conservación sobre un monumento de interés histórico-artístico construido en piedra, hay que realizar el levantamiento del monumento. Para que la intervención sea adecuada en las zonas que se va a intervenir (zonas degradadas) se ha de realizar de la manera más precisa.

El conocimiento sobre la naturaleza y el comportamiento ante las condiciones ambientales de distintos materiales pétreos silíceos empleados en la construcción y ornamentación de monumentos del Patrimonio Histórico es fundamental. Hemos de analizar la respuesta de estos materiales ante los agentes externos (clima, contaminación ambiental, procedencia de sales de aguas subterráneas, ascensión capilar, etc.) y frente a productos de conservación (hidrofugantes, consolidantes, etc.) mediante los procesos fisicoquímicos que tienen lugar en ambos casos.

Las rocas, que forman parte de un monumento de interés cultural, están expuestas a condiciones ambientales que pueden sufrir alteraciones en su estructura y color. Para reproducir las patologías observadas se emplean técnicas de envejecimiento acelerado en cámaras climáticas[52] bajo condiciones controladas.

El color es uno de los parámetros para determinar la calidad de un tratamiento de conservación de piedras ornamentales. Las variaciones de color se producen por cambios en las condiciones climáticas, contaminación u otros agentes. En este trabajo se analizan las variaciones cromáticas producidas sobre un conglomerado silíceo blanco de Zamora. Este tipo de conglomerado se em-

plea en la mayoría de las construcciones de interés histórico-artístico de Zamora (Catedral de Zamora, Fig. 4.15).

Figura 4.15: Catedral de Zamora

Zamora presenta condiciones climatológicas de un clima mediterráneo de tendencia continental (termoclastia, gelifracción y haloclastia).

Figura 4.16: Muestra de cantera

Para poder reproducir estos fenómenos climatológicos se emplearon 25 ciclos de hielo/deshielo junto a frio/calor (-20 a 110 ºC) sobre cubos cortados de roca de cantera (Fig. 4.16), de dimensiones de ($6 \times 6 \times 6$ cm) siguiendo la normativa clásica de Tiano and Pecchioni[53]:

- **Tratamiento 1 (T1):**
  Después de un periodo de secado a 60ºC al que se alcanza un peso constante, los cubos son sumergidos en agua destilada por un periodo de 16 horas, después del cual son enfriados a -20ºC y conservados a esta temperatura durante 3 horas. A continuación, la temperatura se aumenta hasta los 110ºC, conservándose otras 3 horas. Finalmente, los cubos se dejan 2 horas a temperatura ambiente y el proceso se comienza de nuevo[53].

Los monumentos también se ven afectados por otro tipo de condiciones que el tratamiento anterior no describe. En nuestro caso estas condiciones son: baja contaminación ambiental y haloclastia. No existe una normativa específica para reproducir estas circunstancias.

Para poder simular estas condiciones se aplicó a un segundo tratamiento.

- **Tratamiento 2 (T2):**
  Ensayo combinado de hielo/deshielo junto a frio/calor + cristalización de sales (fosfatos), siguiendo las recomendaciones modificadas de Tiano y Pecchioni[53], utilizando intervalos de temperaturas mas suaves (-20 a 110ºC en lugar de -28 a 160ºC) y usando una concentración mas baja de sales (1 % en lugar del 14 %) de $Na_3PO_4 \times 10H_2O$ debido a la poca solubilidad de esta sal en agua a temperatura ambiente.

En este segundo tratamiento la diferencia es que en lugar de sumergir los cubos en agua destilada, se realiza en una solución de 1 % de $Na_3PO_4 \times 10H_2O$. Este tratamiento pretende observar los efectos de la disolución de minerales en las rocas como consecuencia de la deposición de excrementos de aves en monumentos.

En ambos tratamientos se realizaron un total de 25 ciclos midiendo en los 5 primeros (1,2,3,4,5) y de cinco en cinco a partir del quinto (10,15,20 y 25).

Para las mediciones de color se utilizó un colorímetro Minolta CR-310 (Fig. 4.17) para sólidos. El sistema de medida contiene una lámpara de arco de xenón

dentro de una cámara mixta, la cual proporciona un área de medición difusa por encima de los 50 mm de diámetro. Para el análisis de color se recoge, por cables de fibra óptica, la luz reflejada perpendicular a la superficie de la muestra.

Tres fotocélulas de silicio de alta sensibilidad controlan la salida de luz de la lámpara de arco de xenón. Estas fotocélulas poseen un filtro coincidente con los estándares de la Comisión Internacional de Iluminación (CIE) para curvas colométricas (Fig. 4.18).

Figura 4.17: Minolta CR-310

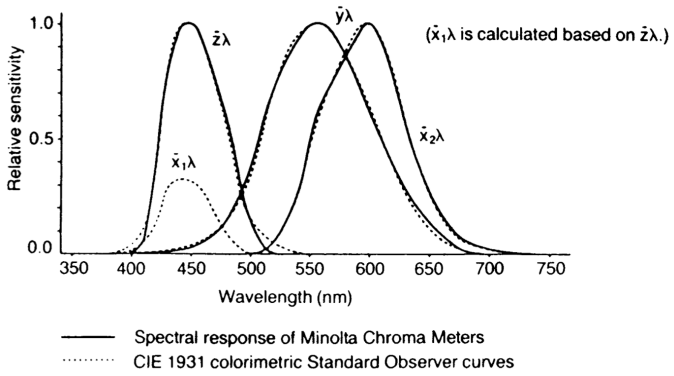


Figura 4.18: Respuesta Espectral Colorímetro

Los colores fueron medidos en las tres coordenadas cromáticas CIELAB (CIE L*a*b*, Fig. 4.19). Estas coordenadas representan un espacio de color tridimensional, en donde L* representa la luminosidad de negro ($L^* = 0$) a blanco difuso ($L^* = 100$), a* comprende valores entre verde ($a^* < 0$) y magenta ($a^* > 0$) y b* abarca desde azul ($b^* < 0$) a amarillo ($b^* < 0$).

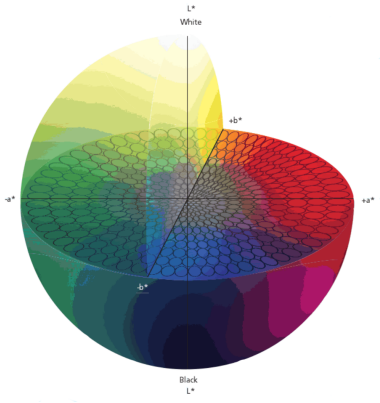


Figura 4.19: Gama Coordenadas Cromáticas CIELAB

En el tratamiento T1, se parte de una matriz de 155 filas y 4 columnas (ciclos, L*, a* y b*) y en T2 55 filas y 4 columnas (ciclos, L*, a* y b*).
Los resultados de las mediciones se muestran en las tablas 4.13, 4.14 y 4.15.

		ciclo 0	ciclo 1	ciclo 2	ciclo 3	ciclo 4	ciclo 5	ciclo 10	ciclo 15	ciclo 20	ciclo 25
T1	Media N=155	74.43	74.37	74.26	73.96	73.77	73.96	73.15	72.36	71.61	70.49
	S.E.	0.20	0.10	0.09	0.11	0.09	0.11	0.10	0.11	0.17	0.17
T2	Media N=55	74.13	73.37	73.32	73.13	73.27	73.29	72.63	72.26	71.83	71.23
	S.E.	0.32	0.54	0.57	0.55	0.56	0.52	0.50	0.50	0.52	0.47

Tabla 4.13: Media y Error Estándar (S.E.) de L^* por ciclo

		ciclo 0	ciclo 1	ciclo 2	ciclo 3	ciclo 4	ciclo 5	ciclo 10	ciclo 15	ciclo 20	ciclo 25
T1	Media N=155	0.58	0.59	0.68	0.84	0.91	0.75	1.04	1.14	1.17	1.39
	S.E.	0.02	0.04	0.04	0.03	0.04	0.03	0.04	0.05	0.06	0.06
T2	Media N=55	0.52	0.35	0.47	0.60	0.48	0.42	0.53	0.61	0.55	0.62
	S.E.	0.03	0.07	0.08	0.07	0.06	0.06	0.06	0.08	0.07	0.06

Tabla 4.14: Media y Error Estándar (S.E.) de a^* por ciclo

El objetivo de nuestro análisis no es otro que predecir el comportamiento de las coordenadas cromáticas (L*,a*,b*) en función de los diferentes ciclos aplicados. Una desventaja de este tipo de tratamientos es la imposibilidad de evaluar el color en fases intermedias o más allá de los ciclos medidos.

		ciclo 0	ciclo 1	ciclo 2	ciclo 3	ciclo 4	ciclo 5	ciclo 10	ciclo 15	ciclo 20	ciclo 25
T1	Media N=155	6.33	6.31	6.43	6.51	6.78	6.67	7.22	7.27	7.38	7.59
	S.E.	0.14	0.13	0.15	0.14	0.16	0.19	0.20	0.18	0.19	0.19
T2	Media N=55	6.59	6.65	6.55	6.42	6.75	6.35	7.03	6.85	6.94	6.89
	S.E.	0.22	0.25	0.34	0.31	0.26	0.31	0.19	0.17	0.23	0.24

Tabla 4.15: Media y Error Estándar (S.E.) de b^* por ciclo

Aplicando el algoritmo MGSR pretendemos predecir el comportamiento de las muestras sin por ello tener que aplicar nuevos tratamientos y mediciones. La flexibilidad que nos proporciona el MGSR nos permite trabajar con diferentes combinaciones de valores y obtener el resto con facilidad.

4.2.2. Resultados

Al disponer de una base de datos compuesta por cuatro variables (ciclo y coordenadas cromáticas L*,a*,b*) continuas, son varios los posibles métodos de reducción dimensional que pueden ser aplicados. Sin embargo, debido a lo ventajoso que resulta poder visualizar elementos y variables en el mismo gráfico, elegimos un JK-Biplot. Asimismo, la distribución de las variables en el biplot nos muestran las correlaciones lineales entre las variables permitiéndonos discernir la elección de variables para nuestras predicciones.

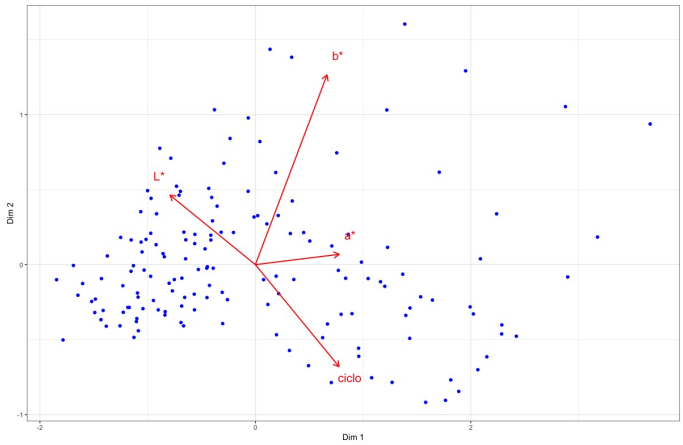


Figura 4.20: JK-Biplot T1

En las Figs. 4.20 y 4.21 aparecen las representaciones simultáneas en los planos principales del JK-Biplot para los tratamientos T1 y T2. Ambas figuras muestran una representación hipotética en el primer plano principal para nuestras bases de datos. La absorción de inercia de los dos primeros ejes factoriales para T1 y T2 son 93.26 % y 86.22 % respectivamente. En vista de la mínima pérdida de información, escogemos estos ejes para desarrollar nuestro algoritmo.

Observando el comportamiento de las variables en la Fig. 4.20 se aprecia una correlación negativa entre ciclos y luminosidad (L*) y es que a mayor número de ciclos aplicados, la muestra tiende a oscurecer. La interpretación es sencilla: dado que los vectores correspondientes poseen la misma dirección pero sentido

opuesto o lo que es lo mismo, tienen un ángulo cercano a 180º entre ellos, las variables tienen un comportamiento opuesto.

La situación es distinta para las coordenadas cromáticas a* y b*. El ángulo que forman es cercano a los 60º lo que indica una interdependencia leve. Lo mismo sucede entre los pares (L*,b*) y (a*,ciclo). No obstante, esta situación es idónea para poder construir nuestro modelo como veremos más adelante.

La Fig. 4.21 muestra los resultados correspondientes a los valores experimentales del T2. Se observa, al igual que en el caso del T1, una correlación negativa entre luminosidad (L*) y ciclos. Sin embargo, a diferencia del T1, las coordenadas a* y b* poseen una correlación directa. Además, la relación de este par de variables (a*,b*) y la variable *ciclo* es de cierta independencia ya que el ángulo existente es cercano a 90º.

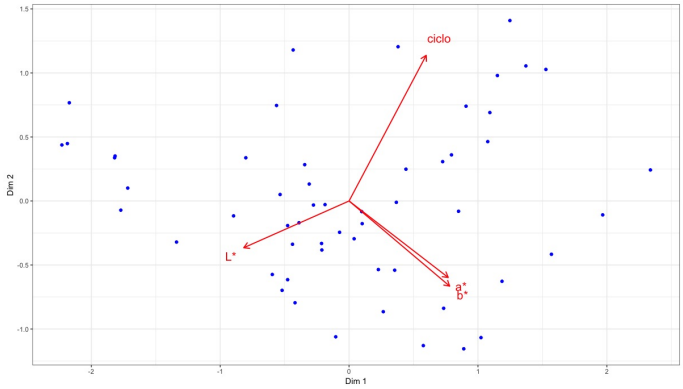


Figura 4.21: JK-Biplot T2

Ligando los datos estandarizados de nuestras muestras y las coordenadas subespaciales obtenidas de los Biplot, aplicamos un LMC para obtener los variogramas cruzados correspondientes. Las Figs. 4.22 y 4.23 muestran los variogramas cruzados experimentales y ajustados (LMC) para ambos tratamientos. En la Tabla 4.16 se muestran los valores ajustados para las distribuciones cuadráticas.

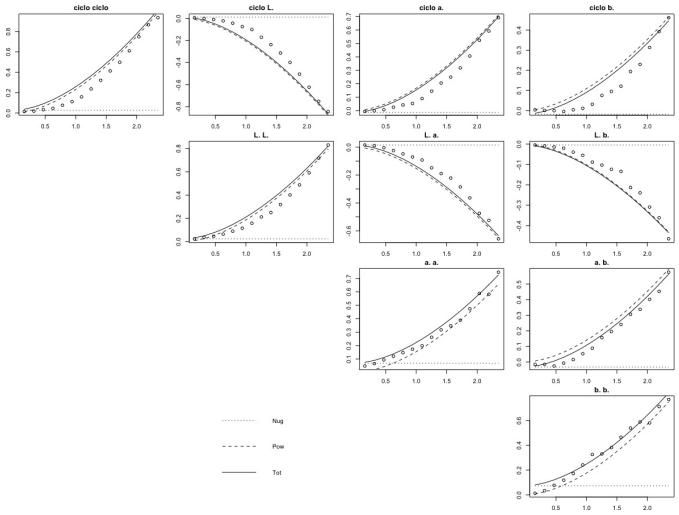


Figura 4.22: Variograma cruzado T1

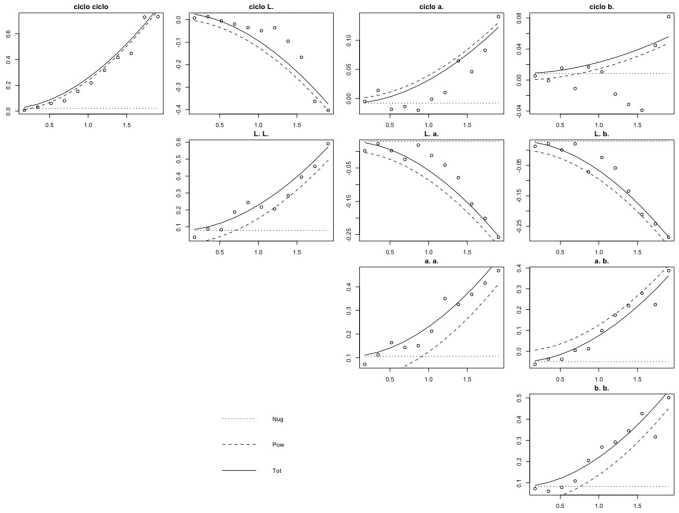


Figura 4.23: Variograma cruzado T2

Ambos ajustes presentan una estructura semejante y acorde con lo descrito en los modelos Biplot. Tenemos un total de 4 variogramas directos y 6 variogramas cruzados para ambos tratamientos. En el T1 los variogramas cruzados son decrecientes para los casos ciclo-L*, L*-a* y L*-b* y positivos en el resto. En el caso del T2, las relaciones decrecientes se presentan en los mismos variogramas cruzados. Además el variograma ciclo-b* parece no presentar una correlación

subespacial alta.

	WC T1 sills		WC T2 sills	
CrossVar	Nug	Pow	Nug	Pow
ciclo	0.03	0.23	0.02	0.24
L*	0.02	0.19	0.08	0.15
a*	0.07	0.15	0.11	0.13
b*	0.07	0.18	0.08	0.14
ciclo L*	0.01	-0.21	0.03	-0.12
ciclo a*	-0.01	0.17	-0.01	0.04
ciclo b*	-0.02	0.11	0.01	0.01
L* a*	0.02	-0.15	0.03	-0.09
L* b*	-0.00	-0.10	0.03	-0.10
a* b*	-0.03	0.14	-0.05	0.13
	Rango: 1.7		Rango: 1.85	

Tabla 4.16: resultados del LMC

Analizando los valores *sill* obtenidos gracias al LMC, podemos ver ciertas semejanzas entre tratamientos en los variogramas directos de ciclo y en los dos variogramas cruzados L*-b* y a*-b*. Sin embargo, los valores de rango, aunque próximos, difieren en 0.15 puntos. Otro hecho destacable es el poco peso que tienen los valores *sill* correspondientes al *Nugget*.

Una vez disponemos del variograma cruzado podemos efectuar nuestro método de validación cruzada para comprobar que el ajuste es correcto. En la Tabla 4.17 se observan los resultados del mismo.

	WC T1		WC T2	
	RMSE	R^2	RMSE	R^2
ciclo	0.04	0.99	0.04	0.97
L*	0.06	0.99	0.15	0.97
a*	0.14	0.99	0.17	0.97
b*	0.03	0.99	0.12	0.97

Tabla 4.17: Validación Cruzada

Los resultados obtenidos son acordes a lo enunciado por Wackernagel[45] ya que los valores RMSE están cercanos a cero y los denominados pseudo-R^2 cercanos a la unidad.

Debido a que nuestro objetivo es predecir más allá de los 25 ciclos, construimos unos mallados extendiendo sus márgenes en la dirección del vector *ciclos* mostrado en las Figs. 4.20 y 4.21. La distancia entre nodos es de dos centésimas de unidad. Esta elección tiene que realizarse en función de los objetivos del investigador y del poder computacional disponible. Un mallado demasiado fino puede resultar muy costoso y poco productivo si no precisamos unas predicciones excesivamente exactas. En el caso que nos compete 0,02 es un valor óptimo ya que las diferencias entre ciclos son centesimales.

En el caso del mallado para el T1, aumentamos nuestro mallado un 18.2 % en sentido creciente para los valores de la dimensión 1 y en sentido decreciente un 47.6 % para los valores de la dimensión 2. Esta elección nos permite predecir valores de hasta 40 ciclos. La extensión podría ser aún mayor, pero la lejanía con respecto a los valores experimentales provocaría errores de predicción.

En la Fig. 4.24 se observa el mallado del T1. Las líneas rojas representan los límites impuestos por los valores experimentales.

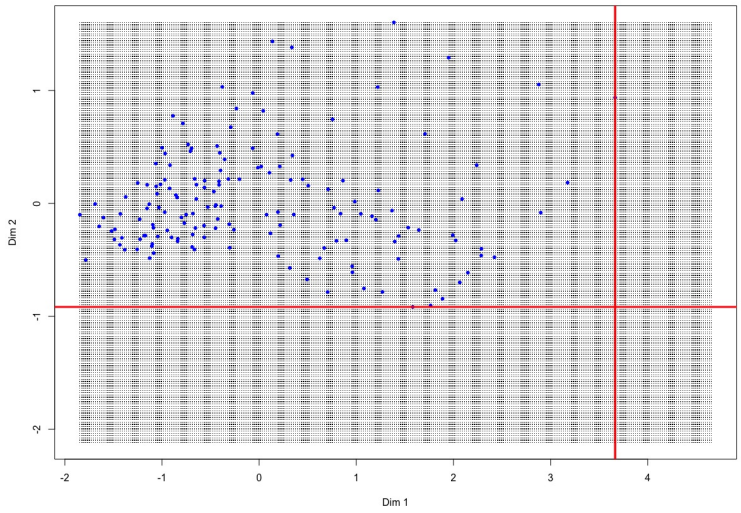


Figura 4.24: Mallado para T1

En el caso del T2 (Fig. 4.24), aumentamos un 39.1 % los márgenes en dirección positiva de la dimensión 2. En este caso también somos capaces de alcanzar los 40 ciclos de predicción.

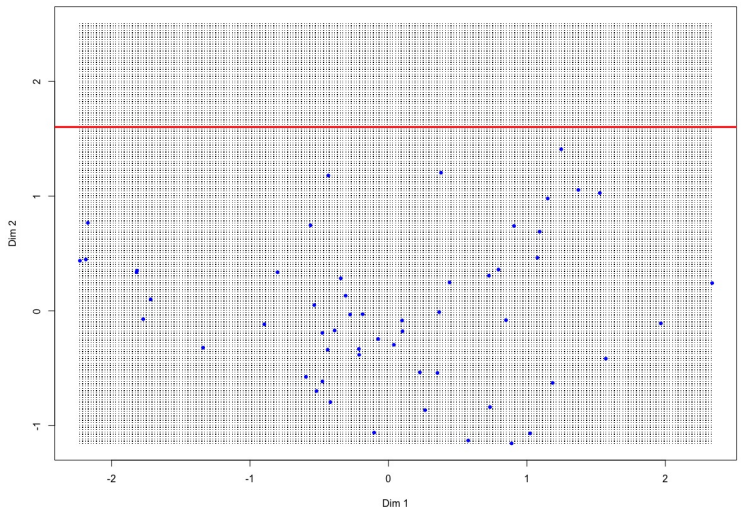


Figura 4.25: Mallado para T2

A continuación aplicamos un Cokriging Simple para ambos tratamientos

con el fin de hallar los valores de las variables correspondientes en cada nodo del mallado.

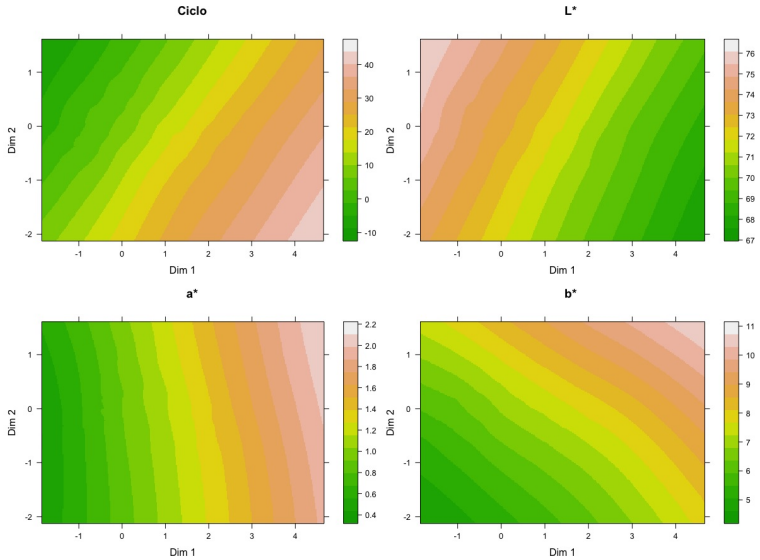


Figura 4.26: Cokriging Simple aplicado a T1

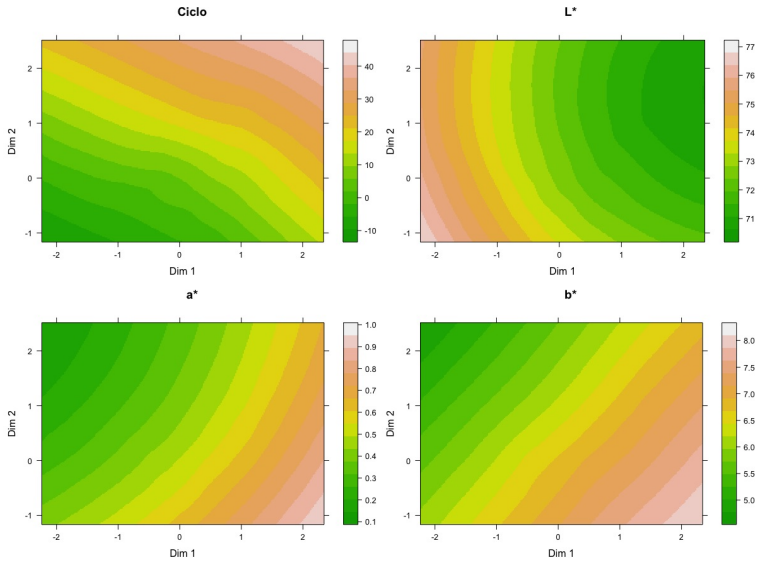


Figura 4.27: Cokriging Simple aplicado a T2

En las Figs. 4.26 y 4.27 se pueden observar los resultados de la iteración. Si

comparamos las tendencias de la Fig. 4.26 con la dirección de los vectores de la Fig. 4.20 vemos la concordancia direccional de las variables entre ambas figuras. Lo mismo sucede para el T2 (Fig. 4.21 y 4.27).

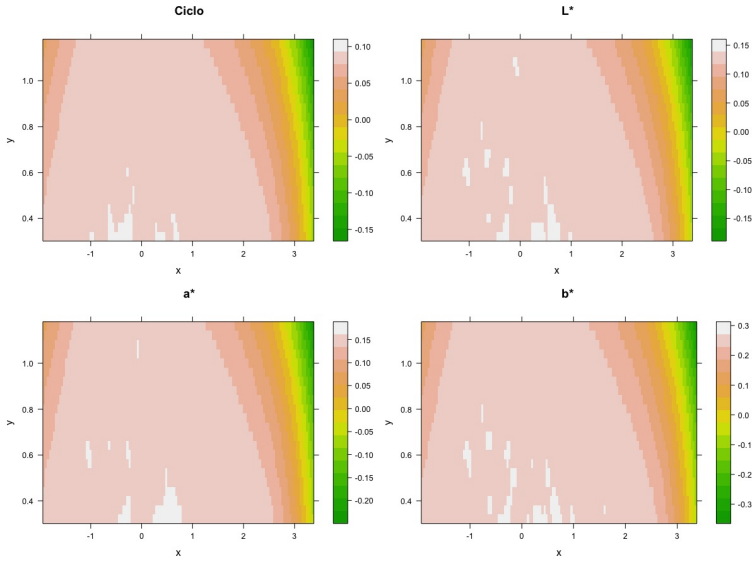


Figura 4.28: Error predicciones en T1

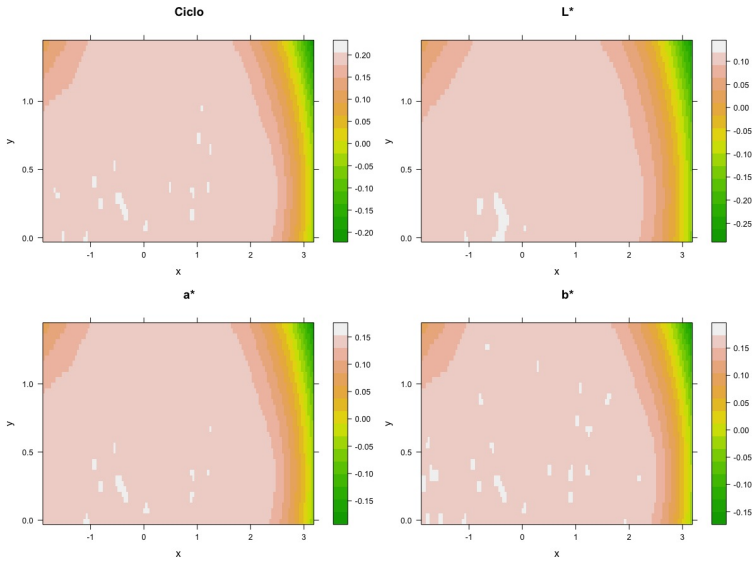


Figura 4.29: Error predicciones en T2

En las Figs. 4.28 y 4.29 se muestra el error de predicción en ambos trata-

mientos. Los resultados, semejantes a los ya expuestos en la Tabla 4.17 muestran valores constantes en buena parte del espectro analizado. Este resultado es coherente con el Cokriging Simple y con la elección de un modelo cuadrático para el ajuste del LMC.

Como se aprecia en las Figs. 4.28 y 4.29, los límites de la abscisa son mayores que los correspondientes al Cokriging. Se escogieron estos rangos para poder observar los márgenes de fiabilidad del modelo. En valores superiores a los descritos en las Figs. 4.26 y 4.27 se observa que el error tiende a cero. Este resultado es irreal ya que el error al aumentar los ciclos no puede ser menor. El motivo de este fenómeno lo provoca la elección de un modelo cuadrático para el ajuste.

Una vez realizado el Cokriging podemos realizar una multitud de predicciones. Si superponemos los cuatro gráficos en uno único, tanto para T1 como para T2 obtendríamos un entramado de líneas con diferentes direcciones. Cada uno de los nodos de nuestros mallados contienen un valor de ciclo y tres coordenadas cromáticas para cada tratamiento. Este amplio espectro nos proporciona una amplia gama de valores.

Supongamos la siguiente situación: Queremos hallar las coordenadas L* y a* para 35 ciclos para un b* fijo igual a 6.8. Esta elección no es arbitraria. En ambos tratamientos las variables ciclo y b* (Figs. 4.20 y 4.21) presentan un ángulo que también es observable en los resultados del Cokriging simple (Figs. 4.26 y 4.27). Este ángulo permite que dos rectas se crucen en un único punto proporcionándonos los valores b* y L*.
Para poder interpretar mejor cómo se realizaría esta predicción observemos la Fig. 4.30.

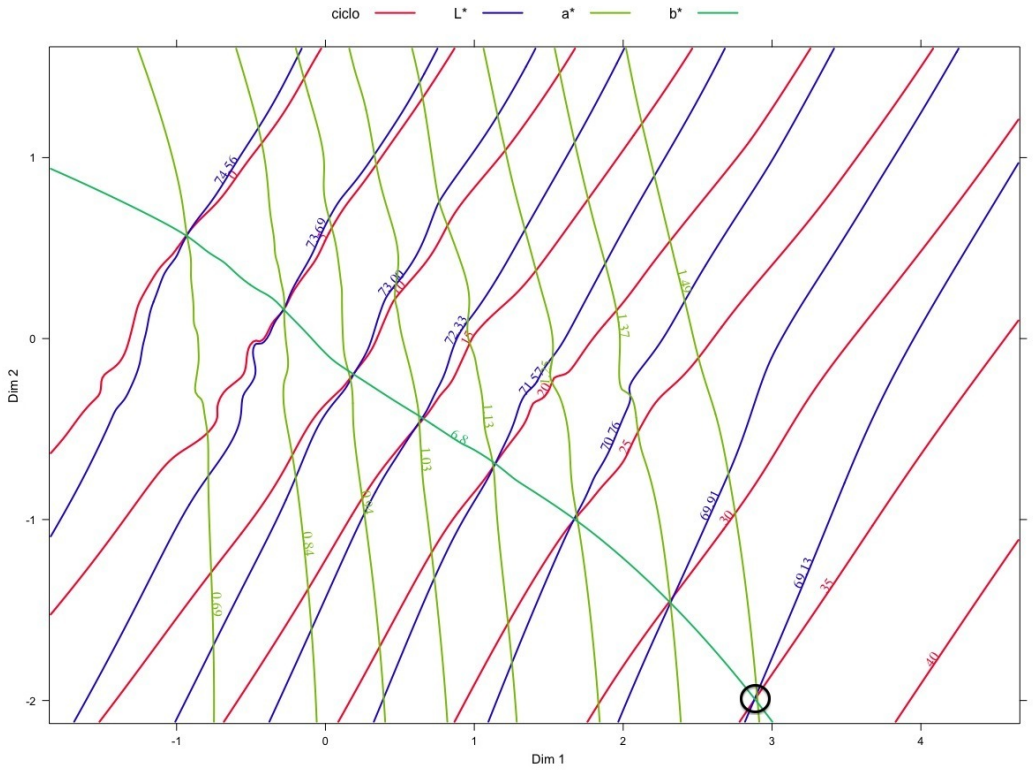


Figura 4.30: Cálculo de a* y L* para b*=6.8 en T1

Cada una de las líneas representadas reproduce un valor de una de las variables. Si fijamos la variable b* ($b^* = 6{,}8$), cada una de las intersecciones entre esta línea y las distintas variables nos proporciona combinaciones de distinta índole. El círculo negro representa la interacción entre el ciclo 35 (no medido) y $b^* = 6{,}8$. El resultado de esta interacción nos proporciona los dos valores restantes, $a^* = 1{,}49$ y $L^* = 69{,}13$. Como podemos ver, en el mismo gráfico se pueden observar muchas otras interacciones.

Para el caso T2 (Fig. 4.31) podemos obtener combinaciones semejantes para el ciclo 35 y $b^* = 6{,}8$, salvo que en esta ocasión los valores $a^* = 0{,}61$ y $L^* = 70{,}83$ difieren con respecto al T1.

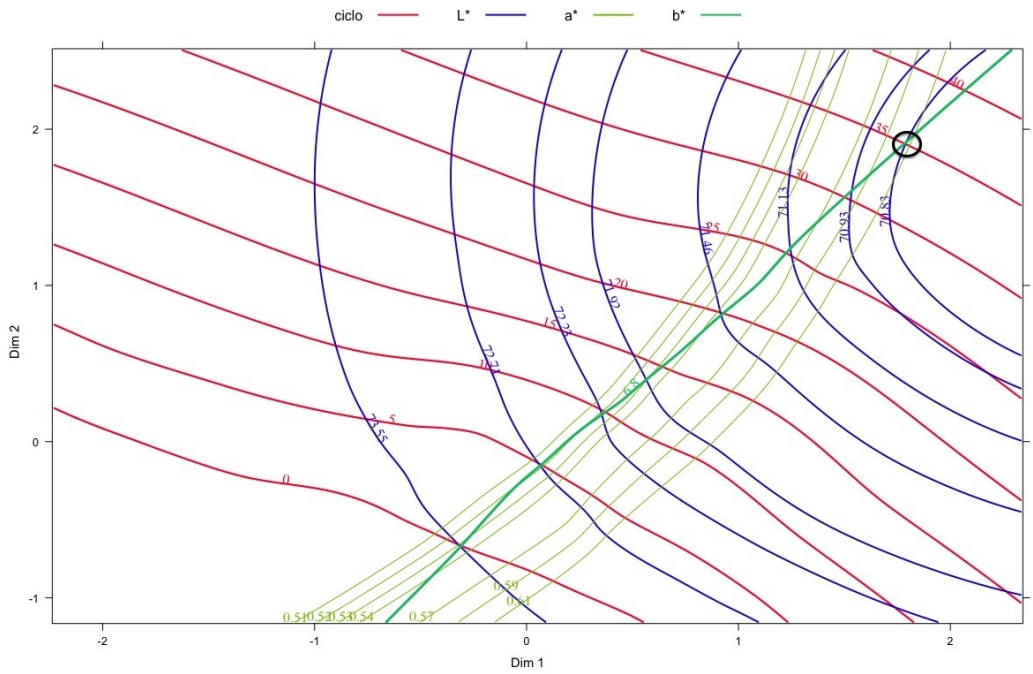


Figura 4.31: Cálculo de a* y L* para b*=6.8 en T2

Es importante matizar que existen posibles combinaciones de variables desaconsejables para predecir. Por ejemplo, en el caso de escoger L* y ciclo como variables independientes, el resultado de la posible predicción sería de mala calidad. El motivo es la alta correlación subespacial que existe entre ambas.

Por último, para poder ver el comportamiento de las coordenadas cromáticas más allá de 25 ciclos presentamos la Fig. 4.32. Esta figura representa la relación entre las distintas coordenadas cromáticas para ciclos superiores a los medidos (26 a 35).

En la Fig. 4.32.1 las coordenadas L* y a* se comportan de manera similar en ambos tratamientos (cuando L* disminuye, a* aumenta). Sin embargo, los cambios que ocurren entre ciclos son menores en el caso del T2. Además, las variaciones en L* son mayores en T2 que en T1 y en caso contrario para a*.

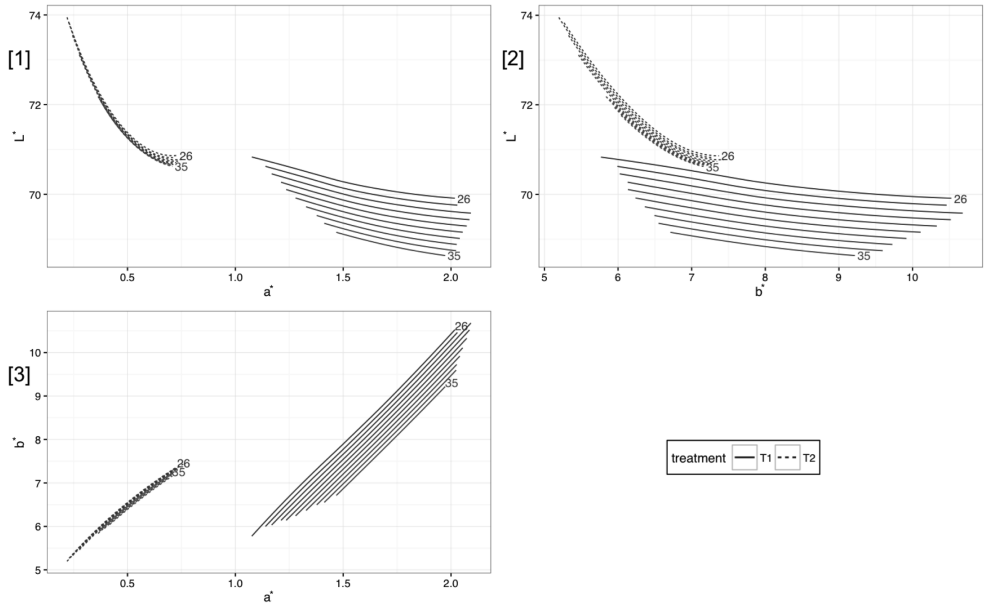


Figura 4.32: Coordenadas cromáticas más allá de los 25 ciclos

En la Fig 4.32.2 las variaciones entre L* y b* son semejantes a las de L* y a*. Estos cambios son mayores para b* (Fig. 4.32.2) que para a* (Fig. 4.32.1).

En la Fig. 4.32.3, las variables entre las coordenadas a* y b* en ambos tratamientos son directas (cuando a* aumenta, b* aumenta). Estas alteraciones, entre ciclos, son mayores en T1. La variabilidad entre ciclos es mayor en el T1 que en el T2 para las coordenadas a* y b*.

Gracias a la aplicación del MGSR hemos podido describir el comportamiento de las coordenadas cromáticas (L*,a*,b*) para un amplio espectro de ciclos. Los datos analizados sólo incluían valores fijos de ciclos. Con el modelo desarrollado podemos predecir que sucede en periodos entre ciclos e incluso más allá de los ciclos medidos.

Estos resultados permiten ahorrar tiempo y dinero en mediciones para ciclos mayores, siendo extensible a muestras con distintas características y experimentos semejantes.

El comportamiento de las predicciones sigue la tendencia ya analizada en otros artículos[54]. En ambos tratamientos se observan tendencias de oscurecimiento, enrojecimiento y amarilleado. El oscurecimiento es menos apreciado en T2 debido a la cristalización de fosfatos.

El modelo propuesto es versátil y flexible a los intereses del investigador.

Capítulo 5

Conclusiones

1 La Regresión Multivariante Gaussiana Subespacial (MGSR) es un método dinámico que combina virtudes de dos disciplinas muy distintas como las técnicas de reducción dimensional y los procesos gaussianos.

2 Se ha demostrado que los procesos gaussianos pueden ser aplicados en dominios hipotéticos, saliéndose de la norma general que indica que únicamente se pueden aplicar en dominios continuos como espacio o tiempo, sin perder sus principales propiedades.

3 El modelo es flexible, permitiendo al investigador usarlo de diferentes modos. Una vez calculadas las proyecciones sobre los subespacios podemos elegir qué variable o variables predecir en función del resto.

4 Cuando logramos ajustar correctamente el modelo lineal de corregionalización (LMC) y escogemos un mallado adecuado, el MGSR obtiene un alto poder predictivo.

5 Se ha probado la validez del algoritmo MGSR en dos aplicaciones de muy distinta naturaleza, lo que demuestra que este algoritmo puede ser llevado a cabo en muy diversos campos científicos.

Bibliografía

[1] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, p. 559–572, 1901.

[2] W. Torgerson, *Theory & Methods of Scaling*. New York: Wiley, 1958.

[3] J. Benzécri, *L'Analyse des Données. Volume II. L'Analyse des Correspondances*. Paris: Dunod, 1973.

[4] M. Greenacre, *Theory and applications of correspondence analysis*. London: Academic Press, 1984.

[5] C. Spearman, "General intelligence, objectively determined and measured," *The American Journal of Psychology*, vol. 15, no. 2, pp. 201–292, 1904.

[6] B. Escofier and J. Pagès, "Multiple factorial analysis: A method to compare groups of variables," *Data Analysis and Informatics*, vol. 3, pp. 41–55, 1984.

[7] A. Gifi, *Nonlinear Multivariate Analysis*. Leiden: Department of Data Theory FSW/RUL, 1981.

[8] E. van der Burg, J. de Leeuw, and G. Dijksterhuis, "Overals: Nonlinear canonical correlation with k sets of variables," *Computational Statistics & Data Analysis*, vol. 18, pp. 141–163, 1994.

[9] E. van der Burg, J. de Leeuw, and R. Verdegaal, "Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling factors," *Psychometrika*, vol. 53, p. 177–197, 1988.

[10] K. Gabriel, "The biplot graphic display of matrices with application to principal component analysis," *Biometrika*, vol. 58, no. 3, pp. 453–467, 1971.

[11] M. Galindo-Villardón, "Una alternativa de representacion simultanea: Hj-biplot," *Qüestiió*, vol. 10, no. 1, 1986.

[12] J. Vicente-Villardón, "Una alternativa a las técnicas factoriales clásicas basada en una generalización de los métodos biplot.," *Doctoral dissertation, Tesis.Universidad de Salamanca. España.*, p. 248, 1992.

[13] J. Gower, S. Lubbe, and N. le Roux, *Understanding Biplots*. London: Chapman & Hall, 2011.

[14] B. Escofier, "Le traitement des variables vectorielles," *Biometrics*, vol. 29, pp. 751–760, 1973.

[15] J. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, pp. 33–51, 1975.

[16] W. Krzanowski, "Between-groups comparison of principal components," *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 703–707, 1979.

[17] L. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, pp. 279–311, 1966.

[18] P. Kroonenberg and J. de Leeuw, "Principal component analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, vol. 45, pp. 69–97, 1980.

[19] P. Kroonenberg, *Three-mode principal components analysis*. Leiden: DSWO Press, 1983.

[20] K. Gabriel, "Analysis of meteorological data by means of canonical decomposition and biplots," *Journal of Applied Meteorology*, vol. 11, 1972.

[21] G. Ramírez, M. Vasquez, A. Camardiel, B. Pérez, and P. Galindo, "Detección gráfica de la multicolinealidad mediante el h-plot de la inversa de la

matriz de correlaciones," *Revista Colombiana de Estadística*, vol. 2, pp. 207–219, 2005.

[22] V. Vairinhos, "Desarrollo de un sistema para minería de datos basado en los métodos biplot," *Doctoral dissertation, Tesis. Universidad de Salamanca. España.*, 2003.

[23] J. Martín-Rodríguez, P. Galindo, and J. L. Vicente-Villardón, "Comparison and integration of subspaces from a biplot perspective," *Journal of Statistical Planning and Inference*, vol. 102, no. 2, pp. 411–423, 2002.

[24] A. Vallejo-Arboleda, J. Vicente-Villardón, and P. Galindo-Villardón, "Canonical statis: Biplot analysis of multi-table group structured data based on statis-act methodology," *Computational Statistics & Data Analysis*, 2006.

[25] J. Vicente-Villardón, M. Galindo-Villardón, and B.-Z. A., *Logistic biplots. Multiple correspondence analysis and related methods*. London: Chapman & Hall, 2006.

[26] J. Hernández-Sánchez and J. Vicente-Villardón, "Logistic biplot for nominal data," *Advances in Data Analysis and Classification*, 2016.

[27] M. Rodriguez-Rosa, "Contribuciones al análisis de la sostenibilidad internacional, desde una perspectiva algebraica multivariante comparada," *Doctoral dissertation, Tesis. Universidad de Salamanca. España.*, 2016.

[28] O. Cárdenas, "Los métodos biplot: Evolución y aplicaciones," *Revista Venezolana de Análisis de Coyuntura*, vol. 13, no. 1, pp. 279–303, 2007.

[29] J. Doob, "The elementary gaussian processes," *The Annals of Mathematical Statistics*, vol. 15, no. 3, pp. 229–282, 1944.

[30] D. Krige, "A statistical approach to some basic mine valuation problems on the witwatersrand," *Journal of Chemical, Metallurgical, and Mining Society of South Africa*, vol. 52, pp. 119–139, 1951.

[31] G. Matheron, *Traité de géostatistique appliquée*. Paris: Éditions Technip, 1962.

[32] P. Burrough and R. McDonnell, *Principles of Geographical Information Systems*. Oxford University Press, 1998.

[33] N. Cressie, *Statistics for Spatial Data*. Wiley Classics Library, 2015.

[34] V. Singh, C. Carnevale, G. Finzi, and E. Pisoni, "A cokriging based approach to reconstruct air pollution maps, processing measurement station concentrations and deterministic model simulations," *Environmental Modelling & Software*, vol. 26, no. 6, p. 778–786, 2011.

[35] U. Mueller and E. Grunsky, "Multivariate spatial analysis of lake sediment geochemical data; melville peninsula, nunavut, canada," *Applied Geochemistry*, vol. 75, p. 247–262, 2016.

[36] L. Salvati, A. Mavrakis, A. Colantoni, G. Mancino, and A. Ferrara, "Complex adaptive systems, soil degradation and land sensitivity to desertification: A multivariate assessment of italian agro-forest landscape," *Science of The Total Environment*, vol. 521–522, p. 235–245, 2016.

[37] J. Carbonero, P. García, C. Avila, O. Arribas, and M. Lizana, "Distribution, habitat characterization and conservation status of iberolacerta martinezricai (arribas, 1996), in the sierra de francia, salamanca, spain (squamata: Sauria: Lacertidae)," *Herpetozoa*, vol. 28, no. 3/4, p. 235–245, 2015.

[38] I. Camargo-Buitrago, P. Quirós-Mc Intire, and R. Gordón-Mendoza, "Identificación de mega-ambientes para potenciar el uso de genotipos superiores de arroz en panamá," *Pesquisa Agropecuária Brasileira*, vol. 46, no. 9, pp. 1061–1069, 2011.

[39] M. Goulard and M. Voltz, "Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix," *Mathematical Geology*, vol. 24, no. 3, p. 269–286, 1992.

[40] S. Searle, *Linear models*. New York: Wiley, 1971.

[41] B. Pelletier, P. Dutilleul, G. Larocque, and J. Fyles, "Fitting the linear model of coregionalization by generalized least squares," *Math Geol*, vol. 36, no. 3, p. 323–343, 2004.

[42] E. Pardo-Igúzquiza, "Mlreml4: A program for the inference of the power variogram model by maximum likelihood and restricted maximum likelihood," *Computers & Geosciences*, vol. 24, no. 6, p. 537–543, 1998.

[43] P. Kitanidis, *Introduction to Geostatistics*. California: Cambridge University Press, 1997.

[44] D. Myers, "Matrix formulation of co-kriging," *Mathematical Geology*, vol. 14, no. 3, p. 249–257, 1982.

[45] H. Wackernagel, *Multivariate Geostatistics: An Introduction with Applications*. Springer-Verlag, 2003.

[46] F. Alfonso, R. Byrne, F. Rivero, and A. Kastrati, "Current treatment of instent restenosis," *Journal of the American College of Cardiology*, vol. 63, no. 24, pp. 2659–2673, 2014.

[47] D. Alexopoulos, "Acute myocardial infarction late following stent implantation: Incidence, mechanisms and clinical presentation," *International Journal of Cardiology*, vol. 152, no. 3, pp. 295–301, 2011.

[48] S. Cassese, R. Byrne, T. Tada, S. Pinieck, M. Joner, T. Ibrahim, L. King, M. Fusaro, K.-L. Laugwitz, and A. Kastrati, "Incidence and predictors of restenosis after coronary stenting in 10.004 patients with surveillance angiography," *Heart*, vol. 100, no. 2, pp. 153–159, 2014.

[49] S. Rathore, M. Terashima, O. Katoh, H. Matsuo, N. Tanaka, Y. Kinoshita, M. Kimura, E. Tuschikane, K. Nasu, M. Ehara, K. Asakura, Y. Asakura, and T. Suzuki, "Predictors of angiographic restenosis after drug eluting stents in the coronary arteries: contemporary practice in real world patients," *EuroIntervention*, vol. 5, no. 3, pp. 349–354, 2009.

[50] R. Kuntz and D. Baim, "Defining coronary restenosis. newer clinical and angiographic paradigms," *Circulation*, vol. 88, no. 3, pp. 1310–1323, 1993.

[51] P. Sánchez, F. Gimeno, P. Ancillo, J. Sanz, J. Alonso-Briales, F. Bosa, I. Santos, J. Sanchis, A. Bethencourt, J. López-Messa, A. Pérez de Prado, J. Alonso, J. San Roman, and F. Fernández-Avilés, "Role of the paclitaxel-eluting

stent and tirofiban in patients with st-elevation myocardial infarction undergoing postfibrinolysis angioplasty. the gracia-3 randomized clinical trial," *Circulation: Cardiovascular Interventions*, vol. 3, pp. 297–307, 2010.

[52] A. Iñigo, V. Rives, and M. Vicente, "Reproducción en cámara climática de las formas de alteración más frecuentes detectadas en materiales graníticos, en clima de tendencia continental," *Materiales de Construcción*, vol. 50, no. 257, pp. 57–60, 2000.

[53] P. Tiano and E. Pecchioni, *Invecchiamento artificiale di materiali lapidei*. Firenze, Italy: Proceedings of the Giornata di Studio "Camera climatiche od ambientali nella ricerca applicata", 1990.

[54] A. Iñigo, J. García-Talegón, and V. Vicente-Tavera, "Canonical biplot statistical analysis to detect the magnitude of the effects of phosphates crystallization aging on the color in siliceous conglomerates," *Color Research and Application*, vol. 39, no. 1, pp. 82–87, 2014.

Anexo A

Tutorial Software MGSR

El algoritmo MGSR ha sido computado íntegramente en R. A continuación se presenta un tutorial de instalación y manejo.

El paquete `MGSR` está incluido en el repositorio de Github victorvicpal/MGSR desde el cual puede instalarse si se dispone de la librería `devtools` instalada. Si no es el caso, puede instalarla cómo se indica aquí:

```
In [1]: install.packages('devtools')
```

Una vez procedido a la instalación. Usando la función `install_github` que permite instalar paquetes de R que estén presentes en `Github`. Podemos instalar el paquete `MGSR`.

```
In [2]: library(devtools)
        install_github("victorvicpal/MGSR")
        library(MGSR)
```

Las dependencias del paquete se instalan conjuntamente.

- `MASS` >> Esta librería se actualiza varias veces al año y es muy utilizada.
- `flexclust` >> Flexclust sigue en funcionamiento, aunque con futuras versiones de R corre riesgo de volverse obsoleta. En futuras versiones de `MGSR` se pretende sustituir esta librería por otra más actualizada.
- `matrixcalc` >> Sucede exactamente igual que con `flexclust`. En futuras versiones se pretende cambiar por otra más reciente.

Ejemplo: Iris Para poder ilustrar el funcionamiento del algoritmo MGSR, vamos a aplicarlo a la base de datos Iris. La matriz de partida se compone de 150 filas y 5 columnas, cuatro de ellas son variables numéricas y la quinta categórica.

```
In [3]: data(iris)
        summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
```

```
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

Como vemos, la base de datos presenta tres clases (Setosa, Versicolor y Virginica). Veamos cómo se comporta cada una en función de las variables.

En el siguiente gráfico vemos la función de densidad de cada variable numérica.

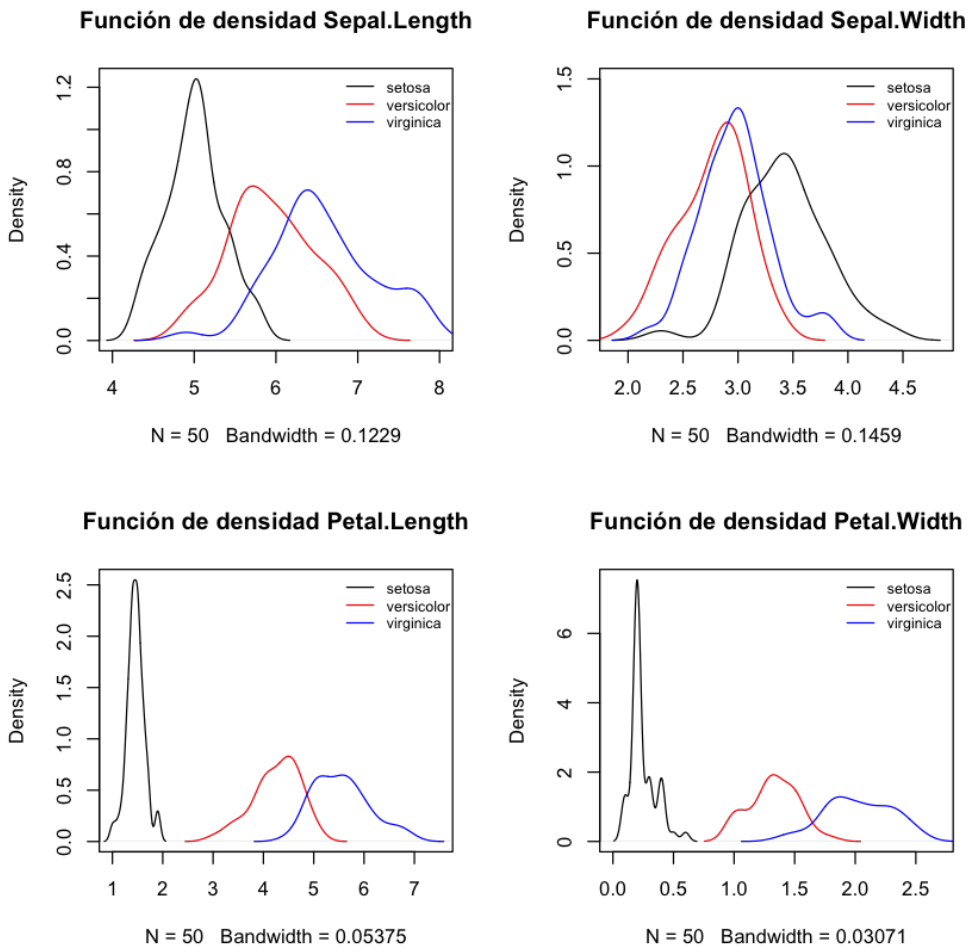
```
In [4]: par(mfrow=c(2,2))
        plot(density(iris$Sepal.Length[which(iris$Species=='setosa')]),
             xlim=c(4,8),main='Función de densidad Sepal.Length')
        lines(density(iris$Sepal.Length[which(iris$Species=='versicolor')]),
        col='red')
        lines(density(iris$Sepal.Length[which(iris$Species=='virginica')]),
        col='blue')
        legend('topright', c('setosa','versicolor','virginica'),lty=1,
               col=c('black', 'red', 'blue'), bty='n', cex=.75)

        plot(density(iris$Sepal.Width[which(iris$Species=='setosa')]),
             ylim=c(0,1.5),main='Función de densidad Sepal.Width')
        lines(density(iris$Sepal.Width[which(iris$Species=='versicolor')]),
        col='red')
        lines(density(iris$Sepal.Width[which(iris$Species=='virginica')]),
        col='blue')
```

```
legend('topright', c('setosa','versicolor','virginica'),lty=1,
       col=c('black', 'red', 'blue'), bty='n', cex=.75)

plot(density(iris$Petal.Length[which(iris$Species=='setosa')]),
     xlim=c(1,7.5),main='Función de densidad Petal.Length')
lines(density(iris$Petal.Length[which(iris$Species=='versicolor')]),
col='red')
lines(density(iris$Petal.Length[which(iris$Species=='virginica')]),
col='blue')
legend('topright', c('setosa','versicolor','virginica'),lty=1,
       col=c('black', 'red', 'blue'), bty='n', cex=.75)

plot(density(iris$Petal.Width[which(iris$Species=='setosa')]),
     xlim=c(0,2.7),main='Función de densidad Petal.Width')
lines(density(iris$Petal.Width[which(iris$Species=='versicolor')]),
col='red')
lines(density(iris$Petal.Width[which(iris$Species=='virginica')]),
col='blue')
legend('topright', c('setosa','versicolor','virginica'),lty=1,
       col=c('black', 'red', 'blue'), bty='n', cex=.75)
```



Separamos la categoría Virginica para poder realizar una predicción concreta de sus posibles valores.

```
In [5]: versicolor <- iris[which(iris$Species=='versicolor'),-5]
```

A.1. Coordenadas Subespaciales

El primer paso esencial del MGSR es obtener unas coordenadas subespaciales provenientes de una técnica de reducción dimensional.

Para simplificar los cálculos y dado que no se requiere una librería adicional. Procedemos a realizar un análisis de componentes principales (PCA).

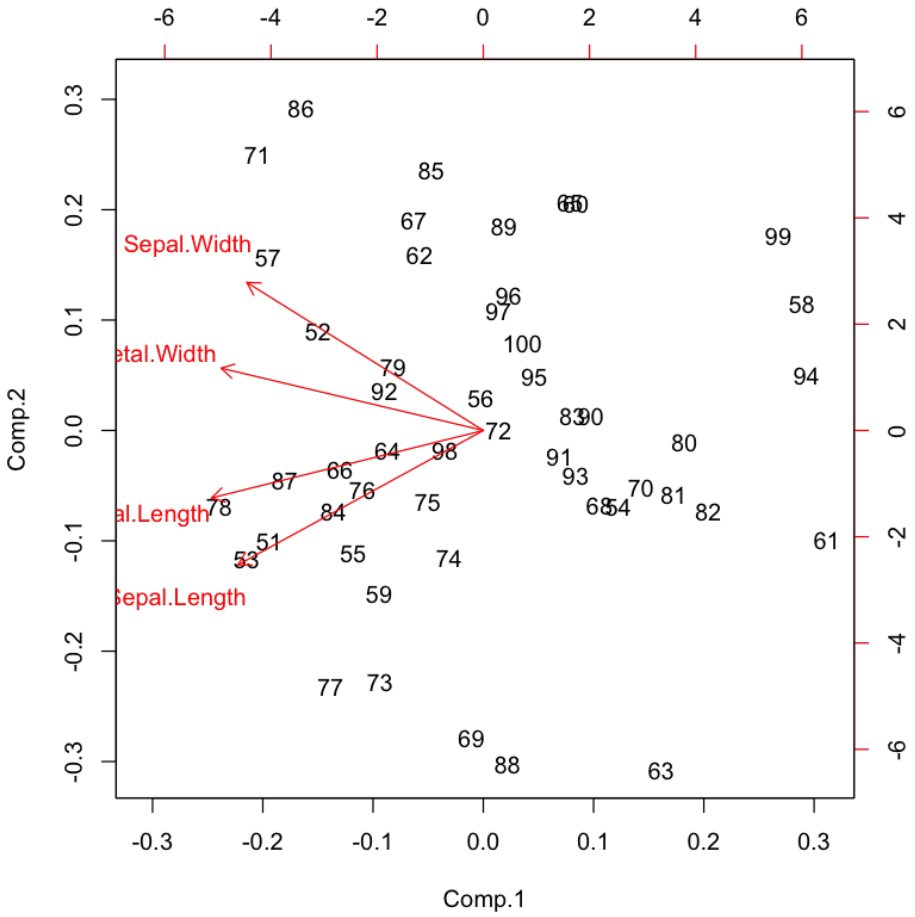
Previo al PCA, estandarizamos por columnas las variables numéricas.

Además posteriormente, representamos el JK-Biplot resultante de aplicar la función `biplot` de `stats`.

```
In [6]: means_versicolor <- apply(versicolor,2,mean)
        sd_versicolor <- apply(versicolor,2,sd)
        versicolor_st <- versicolor

        for (i in 1:4)
        {versicolor_st[,i] <- (versicolor[,i]-
        means_versicolor[i])/sd_versicolor[i]}
```

```
In [7]: PC_versicolor <- princomp(versicolor_st)
         biplot(PC_versicolor)
```



A.2. Variograma Cruzado

La función `crossvariogram` nos calcula los variogramas cruzados según la matriz que le introduzcamos.

- `coord` : data.frame conteniendo las coordenadas subespaciales.
- `values` : data.frame conteniendo los valores estandarizados de nuestra matriz de partida.

- `n` : número de lags.

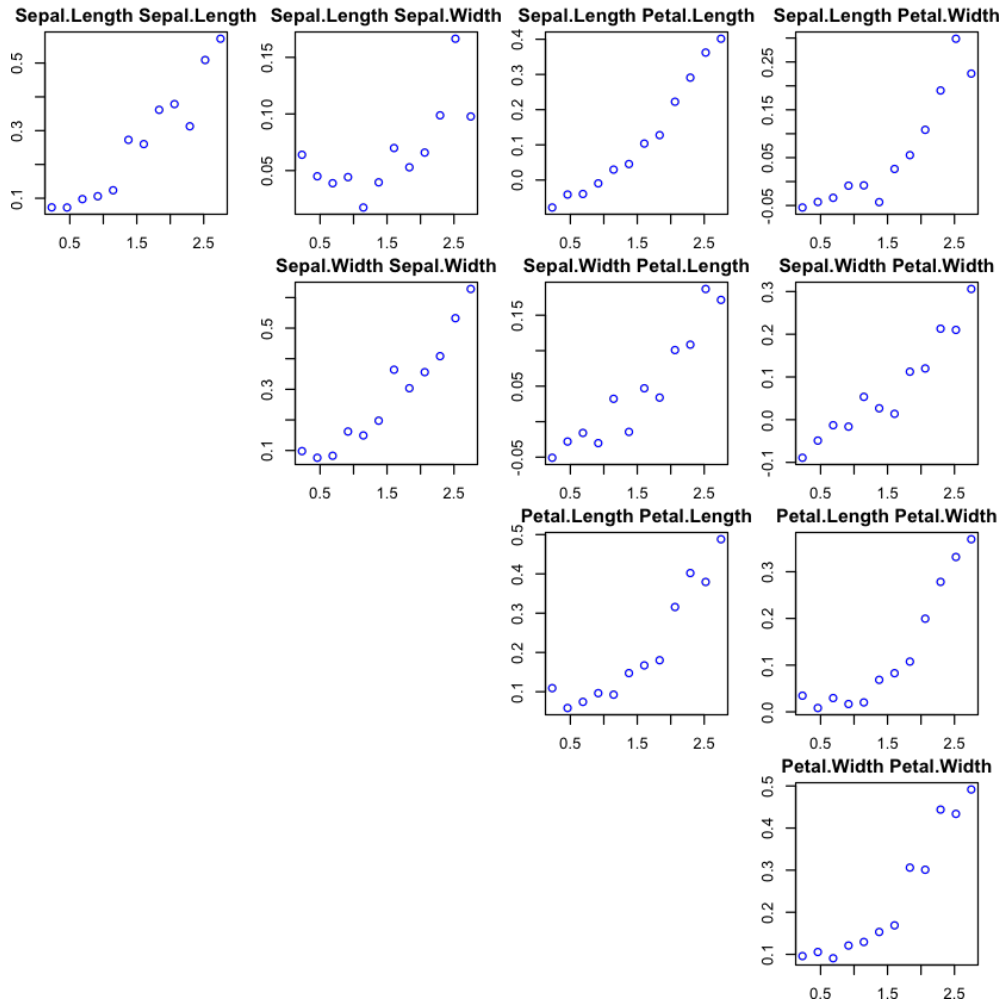
Aplicamos la función a nuestros datos.

```
In [8]: CV <- crossvariogram(coord=as.data.frame(PC_versicolor$scores[,1:2]),
                             values=as.data.frame(versicolor_st),
                             n=12)
```

Para poder visualizar los resultados, el paquete incluye una función `plot.crossvariogram`

- `CV` : variograma cruzado resultado de la función `crossvariogram`.
- `RES` : LMC ajustado con la función `lmc`.

```
In [9]: plot.crossvariogram(CV)
```



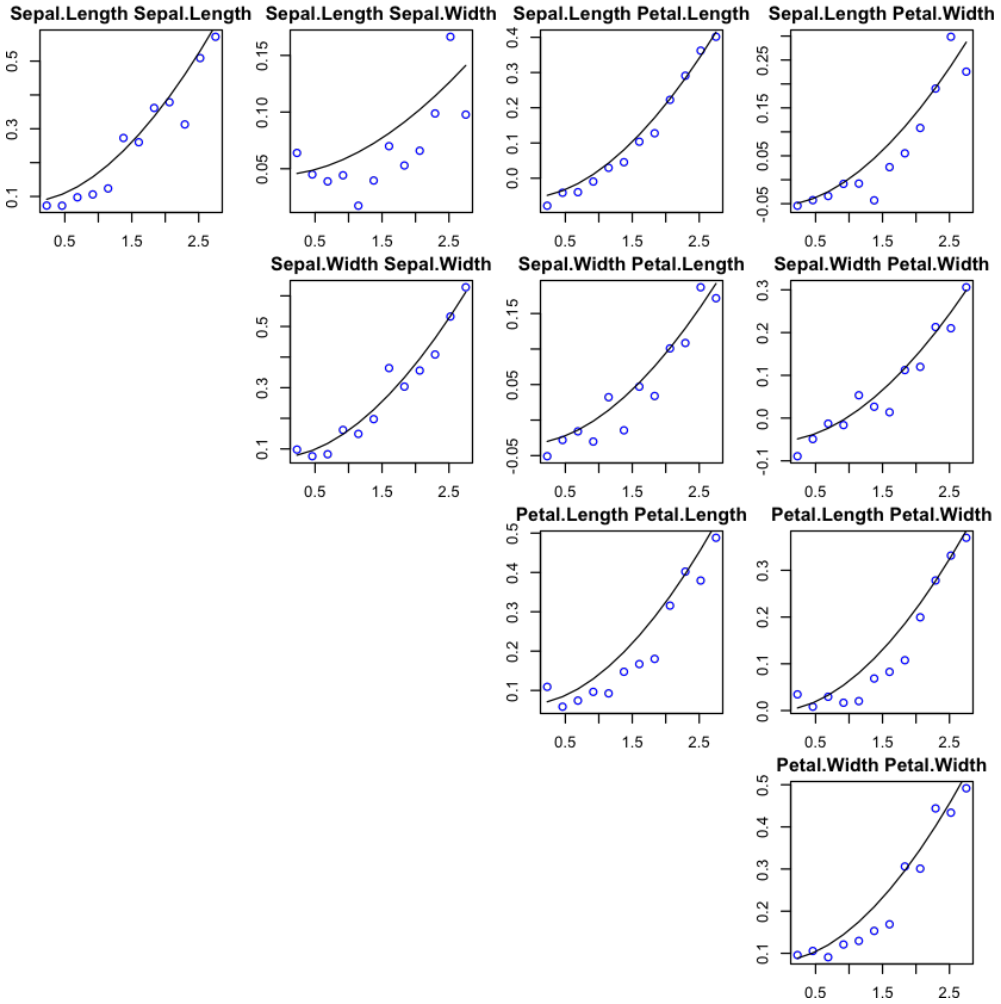
A.3. Modelo Lineal de Corregionalización (LMC)

La función `lmc` nos permite ajustar nuestro variograma cruzado con diversas funciones lineales (lineal, cuadrática, gaussiana o esférica).

- `CV` : Variograma cruzado resultado de la función `crossvariogram`.
- `fun` : Función lineal para el ajuste (`Lin`,`Pow`,`Gau`,`Sph`).
- `a` : Rango para el ajuste.
- `tol` : Tolerancia para la convergencia del algoritmo.
- `mode` : `aut` para el modo automático (recomendado) o `man` para ajuste manual.

En nuestro caso escogemos una función cuadrática (`Pow`) y un rango de 1.8.

```
In [10]: RES <- lmc(CV,fun='Pow',a=1.8)
         plot.crossvariogram(CV,RES)
```



A.4. Mallado

Necesitamos construir un mallado sobre el subespacio representado por el PCA, para ello utilizamos la función `GRID_MGSR`.

- `DAT` : coordenadas subespaciales provenientes de la técnica factorial aplicada.
- `lag` : distancia entre puntos del mallado.

- x1,x2,y1,y2 : extensión de los límites proyectados por nuestros datos. Cero por defecto.

```
In [11]: xygrid <- GRID_MGSR(DAT=as.data.frame(PC_versicolor$scores[,1:2]),
                             lag=0.05)
```

A.5. Cokriging

Una vez tenemos el ajuste LMC y el mallado. Podemos aplicar la función `cokrig` que hallará los valores correspondientes a los puntos del mallado.

- RES : Resultado del LMC obtenido de la función `lmc`.
- xygrid : Mallado obtenido con `GRID_MGSR`.

```
In [12]: Z_versicolor_st <- cokrig(RES,xygrid)
```

Desestandarizamos los resultados.

```
In [13]: Z_versicolor <- Z_versicolor_st

         for (i in 1:length(versicolor[1,]))
         {Z_versicolor[,i+2] <- Z_versicolor_st[,i+2]*sd_versicolor[i]+
         means_versicolor[i]}
         head(Z_versicolor)
```

x	y	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
-2.870986	-1.594691	7.077012	2.797466	5.156145	1.593534
-2.820986	-1.594691	7.065086	2.791310	5.144413	1.586682
-2.770986	-1.594691	7.053189	2.785111	5.132687	1.579860
-2.720986	-1.594691	7.041333	2.778869	5.120976	1.573072
-2.670986	-1.594691	7.029530	2.772589	5.109292	1.566321
-2.620986	-1.594691	7.017791	2.766272	5.097644	1.559612

A.6. Validación Cruzada

Para comprobar la validez de nuestro análisis aplicamos la función `CrossValidation`. El proceso de validación cruzada es idéntico al propuesto por Wackernagel (2003).

- `RES` : Resultado del LMC obtenido de la función `lmc`.

```
In [14]: Val <- CrossValidation(RES)
```

Comprobamos los resultados obtenidos, raíz cuadrada del error cuadrático medio y pseudo-R^2.

El RMSE varía entre 0.11 y 0.19. El pseudo-R^2 entre 0.96 y 0.97.

```
In [15]: cbind.data.frame(R2=Val$R2par,RMSE=Val$RMSE)
```

	R2	RMSE
Sepal.Length	0.9753034	0.1157234
Sepal.Width	0.9709831	0.1269965
Petal.Length	0.9690471	0.1111144
Petal.Width	0.9597514	0.1905291

A.7. Predicciones

Para ver cómo se podrían realizar predicciones supongamos que queremos saber los valores correspondientes a `Petal.Length` y `Petal.Width` dados dos valores de `Sepal`.

Valores iniciales:

- `Sepal.Length` = 6.1
- `Sepal.Width` = 2.6

```
In [16]: Values_init <- data.frame(Sepal.Length=6.1,Sepal.Width=2.6)
         ind_Z <- apply(dist2(Values_init,Z_versicolor[,3:4]),1,which.min)
         Z_versicolor[ind_Z,]
```

x	y	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
0.179	-0.694	6.102752	2.598918	4.320728	1.261553

Nuestro modelo nos dice que los valores del `Petal` serán:

- `Sepal.Length` = 4.32
- `Sepal.Width` = 1.26

Anexo B

Publicación del autor

Received: 9 February 2017 | Revised: 18 April 2017 | Accepted: 25 April 2017

DOI: 10.1002/col.22142

RESEARCH ARTICLE

WILEY

Multivariate Gaussian subspatial regression applied to predict the effect of phosphate crystallization aging on the color in silicious conglomerates

Victor Vicente-Palacios[1] | **Adolo Carlos Iñigo**[2] | **Jacinta García-Talegón**[3]

[1] Department of Statistics, University of Salamanca, Salamanca, Spain

[2] Department of Environmental Degradation Processes and Recovery, IRNASA-CSIC of Salamanca, Salamanca, Spain

[3] Department of Geology, University of Salamanca, Salamanca, Spain

Correspondence
Adolo Carlos Iñigo, Department of Environmental Degradation Processes and Recovery, IRNASA-CSIC of Salamanca, c/ Cordel de Merinas, 40-52, Salamanca 37008, Spain.
Email: adolfo.inigo@irnasa.csic.es

Abstract

A new methodology has been applied to the experimental data obtained about a white siliceous conglomerate from Zamora (Spain), which was subjected to 25 cycles of 2 types of aging [freezing/thawing with cooling/heating (T1) and freezing/thawing with cooling/heating + phosphate crystallization (T2)]. Our model (multivariate Gaussian subspatial regression) allows the behavior and prediction of the chromatic coordinates (L*,a*,b*), including more than 25 cycles, to be analyzed. This model is much more flexible than classical models as it allows multiple variable combinations to be predicted in a dynamic way. The final result showed that the conglomerate experiences darkening, yellowing, and reddening, as the number of cycles increase and that the darkening is much less pronounced in T2 due to phosphate crystallization.

KEYWORDS

multivariate Gaussian subspatial regression, Gaussian process, color difference metrics, preservation, restoration

1 | INTRODUCTION

One of the aesthetic parameters of a building stone is its color, which strongly contributes to its ornamental value. Similar to other properties, monitoring of color is important for assessing the effectiveness of a treatment and to judge the changes that occur when using artificial aging tests.[1–7] To predict the expected decaying processes that occur in stone materials, it is very necessary to anticipate problems that may arise from accelerated aging. To address this, currently climatic chambers are being used for carrying out aging experiments under controlled conditions. Obviously, the aging pattern of stone depends on both the intrinsic properties of the stone and on the specific external conditions to which it is subjected.[8–10]

Several authors have studied the source of phosphate salts in monuments. Borrelli[11] carried out different types of chemical analyses on some samples of deteriorated stone products and has also shown the presence of small amounts of nitrites, nitrates, and phosphates, most probably of biological origin. All of the salts found in monuments are mobilized by the capillary action of water (reaching a visible maximum height of approximately 3 m), which corresponds not only to rainfall but also to water that is used for washing the road which passes before and under the gate. Some of the authors of this article have published several articles regarding color. The studies focused on changes in the color coordinates (L*,a*,b*) in different types of rocks subjected to different conservation and artificial aging treatments. ANOVA,[12] MANOVA-Biplot,[1] and the Canonical Biplot[2,13] were applied. These statistical methods are based on maximizing the variability between treatments, and their aim is to determine the differences in probabilistic terms; although these methods are not predictive models.

The main aim of our study is to discern the interactions between the chromatic coordinates (L*,a*,b*) at unknown cycles for a natural siliceous white conglomerate (WC) for 2 different treatments. The behavior of the chromatic coordinates

TABLE 1 L* mean and SE by cycle

		Cycle 0	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 10	Cycle 15	Cycle 20	Cycle 25
T1	Mean $N = 155$	74.43	74.37	74.26	73.96	73.77	73.96	73.15	72.36	71.61	70.49
	SE	0.20	0.10	0.09	0.11	0.09	0.11	0.10	0.11	0.17	0.17
T2	Mean $N = 55$	74.13	73.37	73.32	73.13	73.27	73.29	72.63	72.26	71.83	71.23
	S.E.	0.32	0.54	0.57	0.55	0.56	0.52	0.50	0.50	0.52	0.47

(L*,a*,b*) is partially known due to the experimental procedure, however, this only includes cycles 1–25. Furthermore, not every cycle is analyzed (each cycle from 1 to 5 is measured, but after that the cycles are measured every 5 cycles until reaching the 25th).[2] If we want to know what would occur afterward 25 cycles or between cycles, we need to build a predictive model, which is the purpose of this study.

2 | MATERIALS AND METHODS

One natural siliceous WC from Zamora, Spain, is used in this study (WC, Z1). This conglomerate is porous and contained strongly reactive components (opal, etc). Its origin, mineralogical, chemical, and petrophysical properties have been described elsewhere.[14] This material is widely used in buildings and is currently used in the reconstruction of the Cathedral and other ornamental buildings in Zamora.

The stone was cut into cubic samples (6 cm × 6 cm × 6 cm) and was subjected to the following accelerated aging treatments under controlled conditions:

T1: Cycles of freezing/thawing and cooling/heating (–20°C to 110°C) in a simulation chamber[5,15]

T2: A combined freezing/thawing and cooling/heating + phosphate crystallization.

We used different conglomerate cubes for T1 and T2.

There is no specific guideline in T2, the samples were immersed in a 1% solution of hydrated sodium phosphate ($Na_3PO_4{\cdot}10H_2O$) instead of distilled water.[5]

Color was measured with a Minolta Model CR-310 colorimeter for solids. The optical system of the measuring head contained a pulsed xenon arc lamp inside a mixing chamber, which provided diffuse and even lighting over the 50 mm diameter measuring area. Only the light that reflected perpendicular to the specimen surface was collected by the fiber-optic cable for color analysis. Three high-sensitivity silicon photocells controlled the light output from the xenon arc lamp. The photocells had a filter so that their spectral response would be in accordance with the standard International Commission on Illumination (CIE) colorimetric curves.

The analyzed cycles were: fresh quarry sandstone (0), 1, 2, 3, 4, 5, 10, 15, 20, and 25. Five measures were taken for each cube. One measure for each face apart from the labeled one of the sample. The experimental chromatic coordinates (L*,a*,b*) obtained are represented in Tables 1–3.

3 | STATISTICAL ANALYSIS

The partial least squares regression (PLS) is a method that can be used to analyze our data. PLS projects the predicted and the observable variables into a new subspace, which explains the maximum multidimensional variance between them. The use of PLS is suitable if the number of predictor variables is larger than the number of observations; however, in our case the use of PLS was not appropriate for our particular dataset.

Our proposal is a combination of factorial techniques and Gaussian processes. Factorial techniques are useful to represent observations in terms of unobserved variables, called factors. All of these techniques provide a set of coordinates

TABLE 2 a* mean and SE by cycle

		Cycle 0	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 10	Cycle 15	Cycle 20	Cycle 25
T1	Mean $N = 155$	0.58	0.59	0.68	0.84	0.91	0.75	1.04	1.14	1.17	1.39
	SE	0.02	0.04	0.04	0.03	0.04	0.03	0.04	0.05	0.06	0.06
T2	Mean $N = 55$	0.52	0.35	0.47	0.60	0.48	0.42	0.53	0.61	0.55	0.62
	SE	0.03	0.07	0.08	0.07	0.06	0.06	0.06	0.08	0.07	0.06

TABLE 3 b* mean and SE by cycle

		Cycle 0	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 10	Cycle 15	Cycle 20	Cycle 25
T1	Mean $N = 155$	6.33	6.31	6.43	6.51	6.78	6.67	7.22	7.27	7.38	7.59
	SE	0.14	0.13	0.15	0.14	0.16	0.19	0.20	0.18	0.19	0.19
T2	Mean $N = 55$	6.59	6.65	6.55	6.42	6.75	6.35	7.03	6.85	6.94	6.89
	SE	0.22	0.25	0.34	0.31	0.26	0.31	0.19	0.17	0.23	0.24

linked to the observations, which reveal information about the variables analyzed. These types of procedures are merely descriptive and have low predictive power. On the other hand, Gaussian processes[16] are statistical methods where observations occur in a continuous domain (mainly time or space). Furthermore, variables have a multivariate normal distribution. Gaussian Processes use similarity between points to predict the value of an unobserved point.

Our dataset lacked a continuous domain, but the projections obtained from the factorial techniques provided a subspatial domain. We used this subspace to simulate a continuous domain that permitted the application of Gaussian processes, such as cokriging. This procedure is called multivariate Gaussian subspatial regression (MGSR). It allows $\mathbf{X}_{N\times P}$ to be the departing data matrix that is composed of P variables and N individuals. Hence, we produced a $\mathbf{R}_{N\times S}$ matrix that is the result of applying a specific factorial technique which proffers the S coordinates. On the basis of these coordinates, we propose a cokriging approach. Before starting the Gaussian process we need to build the proper matrix that will be used later on. To do this, we proceed to standardize by columns the $\mathbf{X}_{N\times P}$ (remove the column means and divide by its standard deviation) and attach the coordinates $\mathbf{R}_{N\times S}$ to our new matrix. $\mathbf{Z}(u) = [\mathbf{x}(\mathbf{R})]$ where $u = \langle r_s \rangle$, $u \in S$ are the set of subspatial locations. The dimension of the resulting matrix was $N \times (S + P)$.

Once we have obtained our $\mathbf{Z}(u)$ matrix, a spatial interpolator such as cokriging[17] was intended to be applied. Instead of using a spatial domain we use the generated subspatial location of the classical Biplot.

These subspatial locations represent how our analyzed samples and variables perform. Hence, we want to describe this subspatial behavior within variograms. Variograms[18] are illustrations of how the semivariance acts in function of the distance. Semivariance is defined as half the expectation between 2 different values at 2 locations (u and u + h), and is used in univariate analyses. To transfer our analysis to a multivariate problem we needed to build crossvariograms.[17]

A crossvariogram γ_{ij} describes the degree of spatial dependence of our projected variables measuring the variation between 2 samples depending on the distance h (also known as lag) between them.

Then we define

$$\begin{cases} E[\mathbf{Z_i}(u+h)-\mathbf{Z_i}(u)]=0 \\ Cov\left[(\mathbf{Z_i}(u+h)-\mathbf{Z_i}(u)),\left(\mathbf{Z_j}(u+h)-\mathbf{Z_j}(u)\right)\right]=2\gamma_{ij} \end{cases} \quad (1)$$

with $i, j = 1, \ldots, N$ and consequently, the semivariogram

$$\mathbf{\Gamma}(\mathbf{h})=\frac{1}{2}E\left[(\mathbf{Z_i}(u+h)-\mathbf{Z_i}(u))\cdot\left(\mathbf{Z_j}(u+h)-\mathbf{Z_j}(u)\right)\right] \quad (2)$$

Using a more practical approach, we needed to build a set of experimental crossvariograms based on our matrix $\mathbf{Z}(u)$. The experimental crossvariogram for different distance classes ϑ gathering n_c pairs of locations u_α, u_β corresponding to the distance between them $u_\alpha - u_\beta = h \in \vartheta$ was

$$\gamma_{ij}^{*}(\vartheta)=\frac{1}{2n_c}\sum_{\alpha=1}^{N}\left(z_i(u_\beta)-z_i(u_\alpha)\right)\cdot\left(z_j(u_\beta)-z_j(u_\alpha)\right) \quad (3)$$

Therefore, we obtained P (P + 1)/2 experimental semivariograms, and subsequently these direct and crossvariograms were fitted.

The different parts of a theoretical semivariogram[18] are:

- Nugget: It represents variability at small distances ($h \approx 0$).
- Sill: The semivariance value b at which the semivariogram levels off.
- Range: The a distance at which the semivariogram reaches the sill value.

Fitting the experimental variogram is an important step. There are many feasible fittings that exist, however, we must choose the one that is the most appropriate. Therefore, sometimes it is necessary to sum functions[19] at different ranges and sills in order to ensure a suitable fit.

The linear model of corregionalization (LMC) permits all of the P × (P + 1)/2 semivariograms to be fitted as linear combinations of F basic semivariogram functions (Gaussian, Exponential, Spherical, Power, etc). The LMC can be expressed as a multivariate nested semivariogram model[17]

$$\mathbf{\Gamma}(\mathbf{h})=\sum_{f=1}^{F}\mathbf{B_f}g_f(h) \quad (4)$$

where $\Gamma(\mathbf{h})$ is the P × P matrix of semivariogram values at lag h, and $\mathbf{B}_f$ is the P × P matrix of sills of the basic semivariogram function $g_f(h)$. $\mathbf{B}_f$ has to be positive semidefinite[19,20]

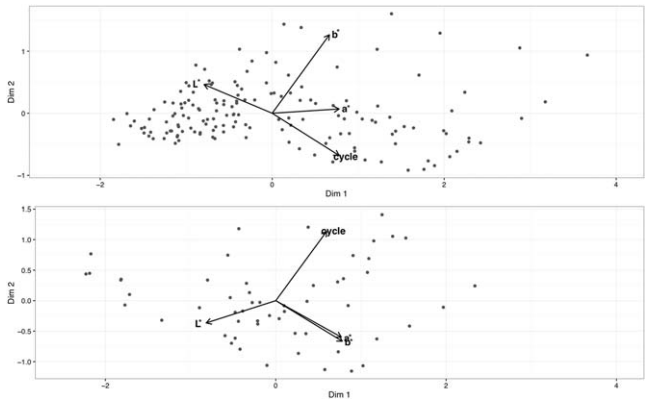


FIGURE 1 Biplot representation for T1 and T2

to assure that the variance-covariance matrix is also positive semidefinite.

Although different approaches of LMC can be found in the literature, we use the following algorithm[21] because of its clarity:

First, sills initial values were estimated by generalized least squares obtaining a $\hat{\mathbf{B}}^{\tau_0}$ matrix.

Next, we selected a f_0 structure among the F subspatial structures. We subtracted the $F-1$ structures from the experimental semivariograms obtaining $\Gamma^*_{f_0}(\vartheta)Z$. Then, we fitted a model to each γ^* individually where an estimated P $\times$ P matrix $\hat{\mathbf{B}}^{\tau+1}_{f_0}$ was obtained.

Then we accomplished a spectral decomposition of $\hat{\mathbf{B}}^{\tau+1}_{f_0}$ as

$$\hat{\mathbf{B}}^{\tau+1}_{\mathbf{f}_0} = \mathbf{Q}\,\mathbf{D}\,\mathbf{Q}^{\mathrm{T}} \tag{5}$$

where $\mathbf{Q}$ is composed of the eigenvectors and $\mathbf{D}$ is the diagonal matrix of eigenvalues. The negative values of $\mathbf{D}$ were replaced by a zero to obtain $\mathbf{D}^{+}$ and we recalculated $\hat{\mathbf{B}}^{\tau+1}$ as $\mathbf{Q}\mathbf{D}^{+}\mathbf{Q}^{\mathrm{T}}$.

We repeated the sequence until all F structures had been completed. Comparing the weighted sum of squares of $\hat{\mathbf{B}}^{\tau+1}$ and $\hat{\mathbf{B}}^{\tau}$ we were able to decide if an additional iteration was needed or not.

Once we obtained our $\Gamma(\mathbf{h})$ set, we needed to create a subspatial grid based on our sample locations. This grid was useful for building our forecast model. Unlike geostatistical

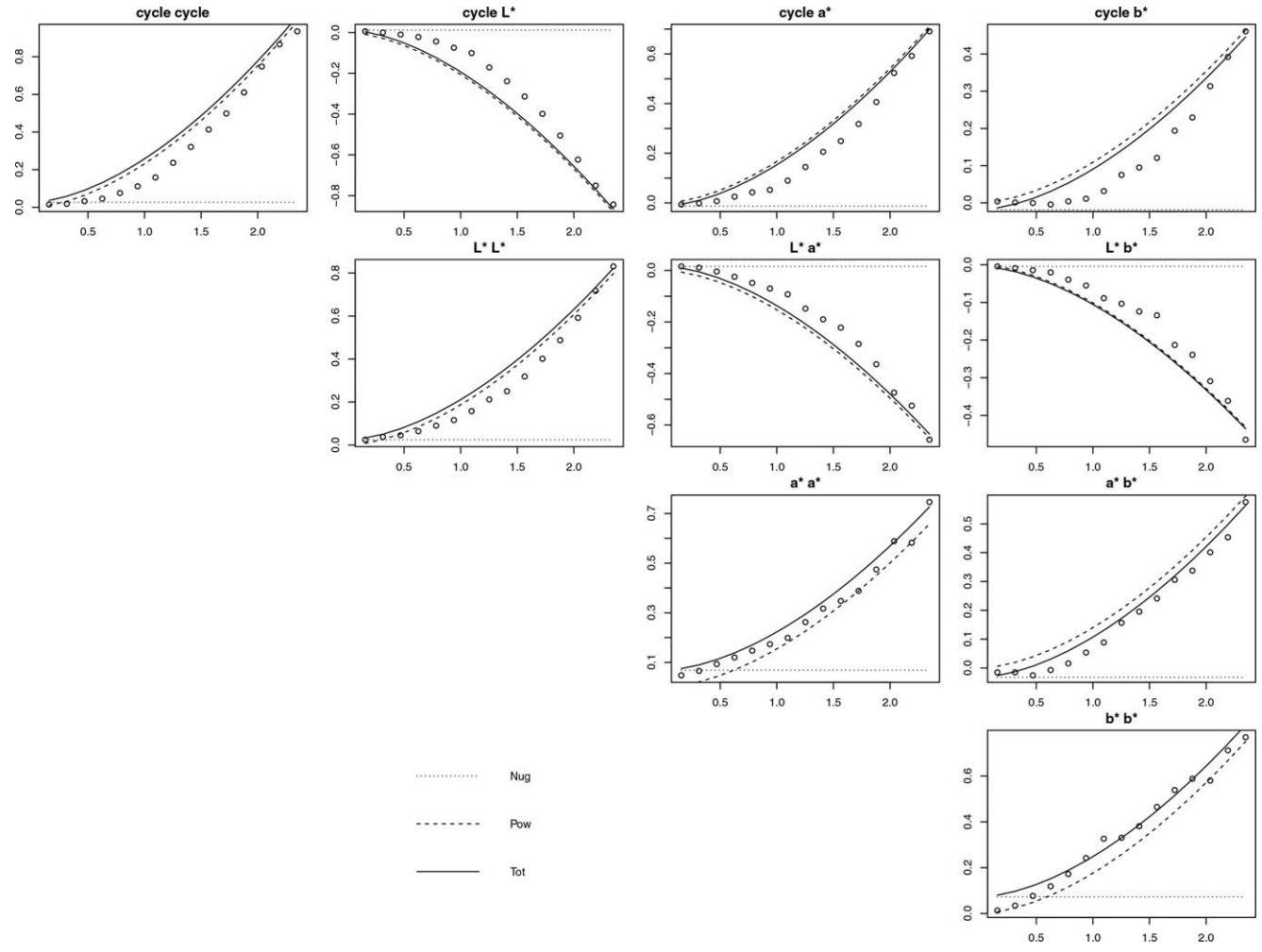


FIGURE 2 T1 Crossvariogram

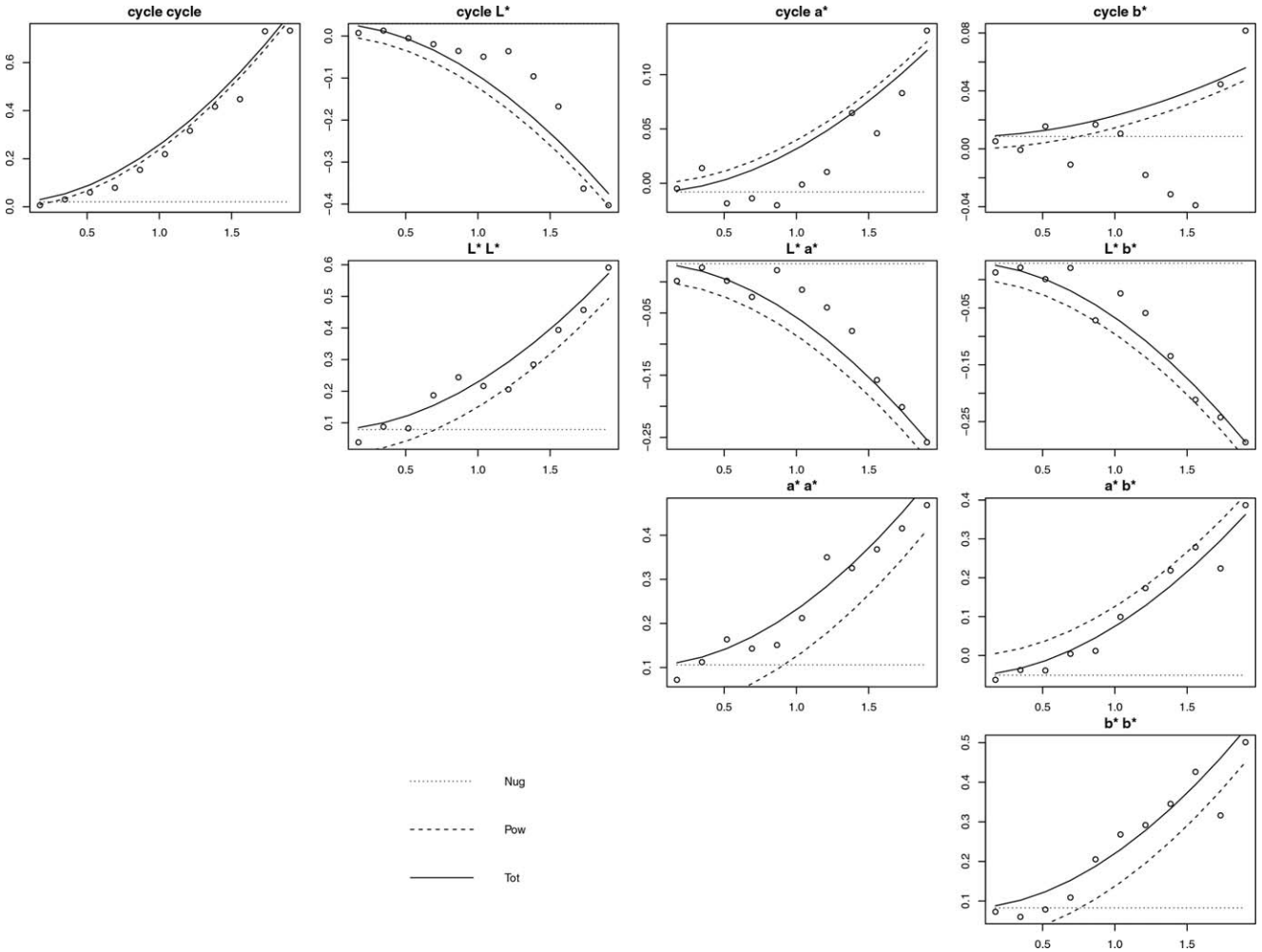


FIGURE 3 T2 Crossvariogram

analyses,[22] there was no real field where boundaries restricted our study. However, this aspect was more positive than negative because we were able to create a simpler grid without losing information. By establishing the interval between the maximum and minimum location in their different S-dimensions we created a frame, which could be extended if necessary. Subsequently, we built the grid choosing a suitable division.

Using cokriging we were able to project our predictions onto the grid and thus were able to compare the results within the variables.

Cokriging is the multivariate extension of kriging, whose main purpose is to compute a weighted average of the sample values in close proximity to the grid point. It searches for the best linear unbiased estimator, based on assumptions on covariances.

There are different procedures such as ordinary, universal, or simple cokriging.[17,23] However, since there were no field constraints, our model was less complex and therefore we were able to apply a simple cokriging which was, mathematically, the simplest and the least general method. Simple cokriging is based on 3 assumptions: stationary, known means and known covariance functions.

The simple cokriging estimator is

$$\hat{\mathbf{Z}}_{\mathbf{i}_0}(u_0) = m_{i_0} + \sum_{i=1}^{P}\sum_{\alpha=1}^{N} \omega_\alpha^i (\mathbf{Z}_{\mathbf{i}}(u_\alpha) - m_i) \tag{6}$$

where u_0 is the grid location and u_α the sample location, ω_α^i is the weight and m corresponds to the means of our variables. We can associate a simple cokriging system[24] to this estimator as $\mathbf{C}_{ij}\,\omega_i = c_{ii_0}$, where $\mathbf{C}_{ij}$ is the N × N covariance matrix, and c_{ii_0} is the N_0 × N covariance matrix between grid and sample locations. Both covariance matrix were determined in a similar way as the theoretical semivariogram, however, we could not assume that both approaches were identical.[17]

$$\mathbf{C}_{ij}(h) = \sum_{f=1}^{F} \mathbf{B_f}\ \ \rho_f(h) \tag{7}$$

$\rho_f(h)$ represents the covariance function. The weights can be calculated as $\mathbf{C}_{ij}\,\omega_i = c_{ii_0}$ and the cokriging error by

TABLE 4 LMC results

	WC T1 sills		WC T2 sills	
CrossVar	**Nug**	**Pow**	**Nug**	**Pow**
Cycle	0.03	0.23	0.02	0.24
L*	0.02	0.19	0.08	0.15
a*	0.07	0.15	0.11	0.13
b*	0.07	0.18	0.08	0.14
Cycle L*	0.01	−0.21	0.03	−0.12
Cycle a*	−0.01	0.17	−0.01	0.04
Cycle b*	−0.02	0.11	0.01	0.01
L*a*	0.02	−0.15	0.03	−0.09
L*b*	−0.00	−0.10	0.03	−0.10
a*b*	−0.03	0.14	−0.05	0.13
	Range: 1.7		Range: 1.85	

$$Var\left(\hat{\mathbf{Z}}_{\mathbf{i}_0}(u_0)-\mathbf{Z}_{\mathbf{i}_0}(u_0)\right)=c_{i_0i_0}^{T}\mathbf{C}_{ij}^{-1}c_{ii_0} \quad (8)$$

Since we were using a standardized matrix, all means were zero, and the estimation of $\hat{\mathbf{Z}}_{\mathbf{i}_0}(u_0)$ was direct.

4 | RESULTS AND DISCUSSION

The MGSR algorithm has been programmed in R,[25] the results were based on the outcomes of MGSR. We chose classical Biplot[26] as the initial factorial technique. We needed to calculate the row metric preserving Biplot for both data sets. Figure 1 shows a hypothetical representation in the first principal plane for WC T1 and T2. The inertia absorption of the first 2 factor axes for WC T1 and T2 was 93.26% and 86.22%, respectively. In view of the minimum loss, we kept 2 first axes to simplify the cokriging iteration.

Figures 2 and 3 show the crossvariogram distribution and its fittings for both data sets. The chosen semivariogram function was a power distribution $g_f(h) = h^{\alpha}$ which was suitable due to the absence of boundaries. The LMC permitted the sill constants for the nugget and power functions to be calculated. Table 4 displays the results of the LMC.

TABLE 5 Crossvalidation results

	WC T1		WC T2	
	RMSE	R^2	**RMSE**	R^2
Cycle	0.04	0.99	0.04	0.97
L*	0.06	0.99	0.15	0.97
a*	0.14	0.99	0.17	0.97
b*	0.03	0.99	0.12	0.97

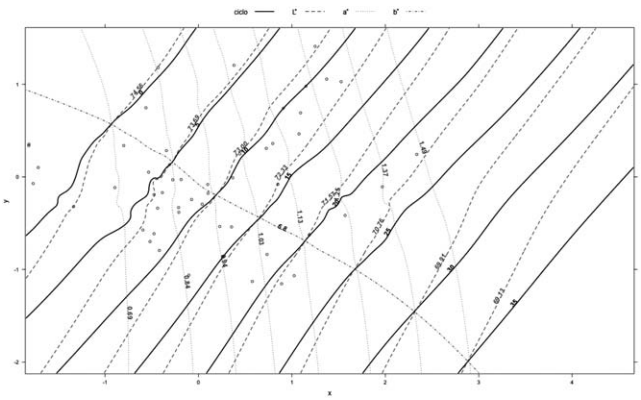


FIGURE 4 Graphical representation of the chromatic coordinates for T1

The crossvalidation method was based on a common resampling. Each sample value $Z(u_\alpha)$ was removed in turn from the dataset and a value $\hat{\mathbf{Z}}\left(u_{[\alpha]}\right)$ was estimated at that location using the $N-1$ other samples.

As a result we determined the residuals $\mathbf{Z}(u_\alpha)-\hat{\mathbf{Z}}\left(u_{[\alpha]}\right)$, and consequently the root mean squared error (RMSE) and the goodness of fit of our model R^2. Table 5 shows the goodness of fit and the RMSE of both models.

Due to the main aim of our research, we built our grid extending our subspatial verges. Then, by applying the simple cokriging iteration we obtained standardized values that must be computed to retrieve the original scale. By virtue of this grid we were able to predict the outcome for 40 cycles.

The principal characteristic of our model is the huge combination of predictions we were able to produce. Figures 4 and 5 illustrate our models. The size of the frame was the result of our grid and each line corresponded to our variables (cycle, L*,a*,b*).

For a better interpretation we fixed b* = 6.8, which was a common value in both models. Hence, we could predict the L* and a* values for each cycle. For example, taking the

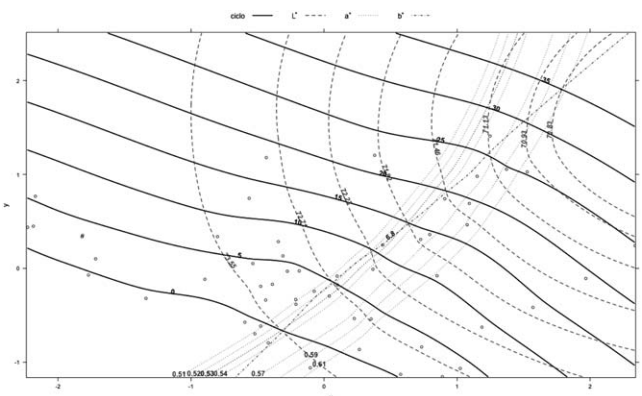


FIGURE 5 Graphical representation of the chromatic coordinates for T2

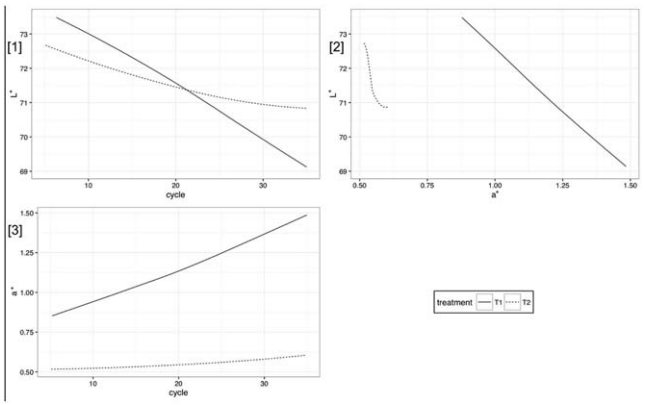


FIGURE 6 Graphical representation of (L*,a*, cycle) for b* = 6.8

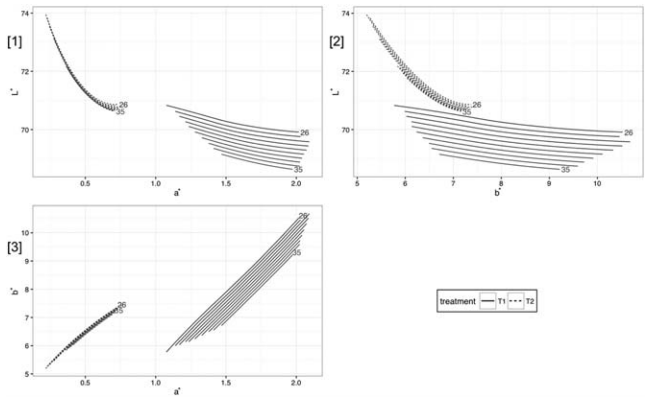


FIGURE 7 Graphical representation of the chromatic coordinates in the predicted cycles

30 cycle for the WC T1 (Figure 4) the resulting intersection with b* = 6.8 is 69.91 for L* and 1.37 for a*. Additionally, for WC T2 (Figure 5) we calculated that L* = 70.93 and a* = 0.59. Other values could be fixed in order to predict variables. The projected points in both figures match with the Biplot coordinates.

As shown in Figure 6, by fixing the value of coordinate b* (b* = 6.8), we could observe the relationships between the rest of the variables (cycles, L* and a*). In Figure 6–1, as the number of cycles of T1 increased, the chromatic coordinate L* decreased. That is to say, the samples became darker until the predicted cycle 35 was reached. However, L* decreased with less intensity in T2 and even became stabilized at cycle 35, with a value of L* = 71. This was due to the crystallization of sodium phosphate on the surface of the cubes.

In Figure 6–2 the relationship between L* and a* was greatly influenced by the values of coordinate L* since it had more variability in T1. In T2 the trend was similar but variability was less than in T1 due to the crystallization of phosphates on the surface of the cubes.

In Figure 6–3 it can be seen that the red color of the sample intensified (a*) as the number of T1 cycles increased. However, this increase was less in T2 due to the same cause as already mentioned above.

In Figure 7 the behavior of the chromatic variables (L*, a*,b*) for the specific prediction of cycles (26–35) is shown. In Figure 7–1 the coordinates L* and a* behaved in a similar manner in both treatments (when L* decreases, a* increases). However, the changes that occurred between cycles were less in T2. Additionally, variations in coordinate L* were greater in T2 than in T1, which was the opposite for coordinate a*. In Figure 7–2 the variations between coordinates L* and b* were similar to those between L* and a*, as shown in Figure 7–1. These changes were greater for b* (Figure 7–2) than in a* (Figure 7–1). In Figure 7–3, the variations between coordinates a* and b* in both treatments were direct (when a* increases, b* increases). These variations, between cycles, were greater in T1. The variability in between cycles of the coordinates a* and b* was greater in T1 than in T2 as shown in the previous figures.

5 | CONCLUSIONS

The proposed methodology (MGSR) has allowed us to describe the behavior of the chromatic coordinates (L*,a*, b*) for a wide range of cycles. The samples analyzed only include fixed value cycles. Using our model we could analyze what occurred between these cycles and even in a greater number of cycles.

The variations analyzed in our model reproduced those already reported in papers previously published by some of the authors of this work. In both treatments a trend that involved darkening, reddening and yellowing was observed; darkening was less appreciable in T2.

The proposed model is very versatile and flexible when studying the behavior of the variables and their predictions. These variables can be combined according to the objectives desired, including fixing a value or interval of the same.

REFERENCES

[1] Iñigo AC, Vicente-Tavera S, Rives V. MANOVA-BIPLOT statistical analysis of the effect of artificial ageing (freezing/thawing) on the colour of treated granite stones. *Color Res Appl.* 2004;*29*:115–120.

[2] Iñigo AC, García-Talegón J, Vicente-Tavera S. Canonical Biplot statistical analysis to detect the magnitude of the effects of phosphates crystallization aging on the color in siliceous conglomerates. *Color Res Appl.* 2014;*39*:82–87.

[3] Johnson JB, Haneef SJ, Hepburn BJ, Hutchinson AJ, Thomson GE, Wood GC. Laboratory exposure system to simulate atmospheric degradation of building stone under dry and wet deposition conditions. *Atmos Environ.* 1990;*24*:2585–2592.

[4] Alonso FJ, Vázquez P, Esbert RM, Ordaz J. Ornamental granite durability: Evaluation of damage caused by salt crystallization test. *Mater Constr.* 2008;*58*(289–290):191–201.

[5] Iñigo AC, García-Talegón J, Vicente-Tavera S, et al. Colour and ultrasound propagation speed changes by different ageing of freezing/thawing and cooling/heating in granitic materials. *Cold Reg Sci Technol.* 2013;*85*:71–78.

[6] Aly N, Gomez-Heras M, Hamed A, Álvarez BM, Soliman F. The influence of temperature in a capillary imbibition salt weathering simulation test on Mokattam limestone. *Mater Constr.* 2015;*65*(317):e044

[7] Vázquez P, Luque A, Alonso FJ, Grossi CM. Surface changes on crystalline stones due to salt crystallisation. *Environ Earth Sci.* 2013;*69*:1237–1248.

[8] Iñigo AC, Rives V, Vicente-Tavera S. Reproducción en cámara climática de las formas de alteración más frecuentes detectadas en materiales graníticos, en clima de tendencia continental. *Mater Constr.* 2000;*50*(257):57–60.

[9] Iñigo AC, Vicente-Tavera S. Different degrees of stone decay on the inner and outer walls of a Cloister. *Build Environ.* 2001; *36*:911–917.

[10] Iñigo AC, Vicente-Tavera S. Surface-Inside (10 cm) thermal gradients in granitic rocks. Effect of environmental conditions. *Build Environ.* 2002;*37*:101–108.

[11] Lazzarini L, Borelli E, Bouabdelli M, Antonelli F. Insight into the conservation problems of the stone building Bab Agnaou, a XII cent. Monumental gate in Marrakech (Morocco). *J Cult Herit.* 2007;*8*:315–322.

[12] Iñigo AC, Vicente-Tavera S, Rives V, Vicente MA. Color changes in the surface of granitic materials by consolidated and/or water repelent treatments. *Color Res Appl.* 1997;*22*: 133–141.

[13] Grossi D, Aparecida LE, García-Talegón J, Iñigo AC, Vicente-Tavera S. Evaluation of colourimetric changes in the Itaquera Granite of the Ramos de Azevedo Monument, São Paulo, Brazil. *Int J Conserv Sci.* 2015;*6*:313–322.

[14] Añorve M. *Valoración Del Deterioro y Conservación En La Piedra Monumental*. Madrid: CEDEX (Centro de Estudios y Experimentación de Obras Publicas); 1997.

[15] Iñigo AC, Vicente-Tavera S, Rives V. Statistical design applied to hydric property behaviour for monitoring granite consolidation and/or water-repellent treatments. *Mater Constr.* 2006;*56*(281):17–28.

[16] Seeger M. Gaussian processes for machine learning. *Int J Neural Syst.* 2004;*14*:69–104.

[17] Wackernagel H. *Multivariate Geostatistics: An Introduction with Applications*. New York, NY: Springer-Verlag; 2003.

[18] Oliver MA, Webster R. *Basic Steps in Geostatistics: The Variogram and Kriging*. New York, NY: Springer International Publishing; 2015.

[19] Journel AG, Huijbregts CJ. *Mining Geostatistics*. London: Academic Press; 1978.

[20] Goovaerts P, Webster R. Scale-dependent correlation between topsoil copper and cobalt concentrations in Scotland. *Eur J Soil Sci.* 1994;*45*:79–95.

[21] Pelletier B, Dutilleul P, Larocque G, Fyles JW. Fitting the linear model of coregionalization by generalized least squares. *Math Geol.* 2004;*36*:323–343.

[22] Chiles JP, Delfiner P. *Geostatistics: Modeling Spatial Uncertainty*. New York, NY: Wiley; 2012.

[23] Cressie N. *Statistics for Spatial Data*. New York, NY: Wiley Classics Library; 2015.

[24] Myers DE. Matrix Formulation of Co-Kriging. *Math Geol.* 1982;*14*:249–257.

[25] Vicente-Palacios V. Multivariate Gaussian Subspatial Regression. victorvicpal/MGSR. http://doi.org/10.5281/zenodo.264102. Accessed 2016.

[26] Gabriel KR. The biplot graphic display of matrices with application to principal component analysis. *Biometrika.* 1971;*58*:453–467.

How to cite this article: Vicente-Palacios V, Iñigo AC, García-Talegón J. Multivariate Gaussian subspatial regression applied to predict the effect of phosphate crystallization aging on the color in silicious conglomerates. *Color Res Appl*. 2017;00:000–000. https://doi.org/10.1002/col.22142