



Comparación de métodos de estimación de datos faltantes en series de precipitación diaria en el Valle del Cauca

Jessica Russi Peña

Eider Larrahondo

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2017

Comparación de métodos de estimación de datos faltantes en series de precipitación diaria en el Valle del Cauca

Jessica Russi Peña

Eider Larrahondo

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico(a)

Director:

Johann Alexis Ospina Galindez, MSc

Codirector:

Gabriel Conde Arango, MSc

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2017

Agradecimientos

Agradecemos a Dios por guiarnos y permitirnos culminar este importante ciclo en nuestras vidas, a nuestros familiares y amigos por su acompañamiento y apoyo incondicional durante este proceso.

Gracias a nuestro director Johann Alexis Ospina por compartir su conocimiento con nosotros, dirigirnos y acompañarnos para culminar este trabajo con éxito terminando de esta forma la última etapa del camino.

Eider y Jessica.

Resumen

Los datos de series climáticas se ven afectadas por la existencia de datos faltantes, por ejemplo en interrupciones ocasionales de estaciones automáticas, mal funcionamiento del instrumento de medición o reorganizaciones de la red de monitoreo. Los resultados de cualquier modelo y análisis estadístico podrían deteriorarse cuando se utilizan registros incompletos como entrada, por esto, la estimación de los datos faltantes se convierte en prioridad en el proceso de preparación de datos, que en conjunto con otras metodologías estadísticas aplicadas dejan los datos listos para su respectivo análisis.

De esta forma, en este trabajo se describe y se comparan cuatro metodologías de estimación estadística, por medio de tres indicadores de desempeño (ECM - Error Cuadrático Medio, RECM - Raíz del Error Cuadrático Medio, CVRECM - Coeficiente de Variación del Error Cuadrático Medio), tres de ellas basadas en datos georreferenciados (K-Vecinos Cercanos, Distancia Inversa Ponderada, Kriging Ordinario) y una en estimaciones a partir de la información con la que se disponga de estaciones climáticas adyacentes (Regresión Lineal), para la estimación de datos faltantes en series diarias de precipitación en el Valle del Cauca, según la zona (valle o montaña) en la que se encuentre la estación climática que se va a estimar y el porcentaje de datos faltantes que se deba estimar (10,15 y 20 %).

De los resultados obtenidos se destaca que el método Kriging Ordinario obtuvo el mejor desempeño en la zona de valle bajo un porcentaje de datos faltantes del 10 %, el método Regresión Lineal se desempeñó mejor en la zona de valle con un 15 % de datos faltantes, el método Distancia Inversa Ponderada se desempeñó mejor en la zona de valle bajo un 20 % de datos faltantes y en la zona de montaña bajo un 10 % y finalmente el método K-Vecinos más cercanos se desempeñó mejor en la zona de montaña bajo un 15 % y 20 % de datos faltantes.

En este sentido, se podría decir que no se tiene un método estándar que como tal tenga un buen comportamiento bajo las distintas condiciones o escenarios expuestos en el marco de este trabajo, sobre los cuales se realizó la comparación de éstos, lo que deja entrever que es necesario realizar una exhaustiva búsqueda al momento de escoger o seleccionar un método de estimación dependiendo la zona geográfica en donde se encuentre, ya que no todos los métodos existentes en la literatura tienen un buen desempeño para condiciones específicas y las restricciones de algunos los pueden volver ineficientes en el que hacer práctico.

Palabras claves: Estimación, Datos faltantes, Distancia Inversa Ponderada, K-Vecinos mas cercanos, Kriging Ordinario, Regresión Lineal.

Contenido

Agradecimientos	5
Resumen	7
1. Introducción	2
2. Planteamiento del problema	4
3. Justificación	7
4. Antecedentes	9
5. Objetivos	14
5.1. Objetivo general	14
5.2. Objetivos Específicos:	14
6. Marco Teórico	15
6.1. Marco conceptual	15
6.1.1. Zona de estudio	15
6.1.2. Precipitación	15
6.1.3. Valle	16
6.1.4. Montaña	16
6.1.5. Estaciones meteorológicas	17
6.2. Marco Estadístico	18
6.2.1. Datos faltantes	18
6.2.2. Correlación espacial	19
6.2.3. Métodos de estimación de datos faltantes	23
6.2.4. Indicadores de desempeño	27
6.2.5. Test de Mann Whitney	29
7. Metodología	31
7.1. Datos de estudio	31
7.2. Control de calidad datos de estudio	32

7.3.	Análisis exploratorio	32
7.3.1.	Análisis exploratorio año 1997	38
7.4.	Procedimiento	39
7.4.1.	K-Vecinos más cercanos (KV)	41
7.4.2.	Distancia inversa ponderada (IDW)	42
7.4.3.	Regresión Lineal (RL):	43
7.4.4.	Kriging Ordinario (KO):	44
8.	Resultados	46
8.1.	K-Vecinos Cercanos (KV)	46
8.2.	Regresión Lineal (RL)	48
8.3.	Distancia Inversa Ponderada (IDW)	50
8.4.	Kriging Ordinario (KO)	51
8.5.	Resultados Generales	55
8.6.	Prueba Mann-Whitney	58
9.	Conclusiones y recomendaciones	60
9.1.	Conclusiones	60
9.2.	Recomendaciones	63
	Bibliografía	65
10.	Anexos	71
10.1.	KV	80
10.2.	Regresión	80
10.3.	IDW	80
10.4.	K-O	81

Lista de Tablas

6.1. Definición de las áreas de alto relieve	17
7.1. Estadísticas descriptivas estaciones completas	33
7.2. Frecuencias absolutas y relativas de datos faltantes para todas las estaciones de cada año en cada zona.	37
7.3. Comportamiento ECM según parámetro κ	42
8.1. Resumen descriptivo para los indicadores del método de K-Vecinos.	47
8.2. Resumen descriptivo para los indicadores del método de Regresión Lineal. . .	49
8.3. Resumen descriptivo para los indicadores del método IDW.	50
8.4. Estimación de parámetros para el modelo seleccionado, Valle.	53
8.5. Estimación de parámetros para el modelo seleccionado, Montaña.	54
8.6. Resumen descriptivo para los indicadores del método Kriging Ordinario. . . .	55
8.7. Indicadores para los métodos de estimación por zona geográfica (Mediana). .	56
8.8. Diferencias entre métodos, según zona y porcentaje de datos faltantes (valor-p).	58
8.9. Diferencias entre porcentajes, según zona y método (valor-p).	59
9.1. Métodos por escenario.	60
10.1. Listado de estaciones que se encuentran en la zona de valle año 1997.	71
10.2. Listado de estaciones que se encuentran en la zona de montaña año 1997. . .	72
10.3. Coeficientes de correlación para la estimación método Regresión Lineal zona Valle	73
10.4. Coeficientes de correlación para la estimación método Regresión Lineal zona Montaña	74
10.5. Suma de Cuadrados del Error (SCE), modelos teóricos de semivarianza (KO).	75

Lista de Figuras

6.1. Estaciones climáticas del Valle del Cauca	18
6.2. Comportamiento típico de un semivariograma acotado con una representación de los parámetros básicos. Fuente: (Giraldo, 2009)	21
7.1. Distribución de las estaciones completas.	34
7.2. Histograma de las estaciones completas.	35
7.3. Ubicación estaciones	36
7.4. Comportamiento estaciones para cada semana del año 1997 por zonas.	38
7.5. Comportamiento estaciones para cada quincena del año 1997 por zonas.	39
7.6. Comportamiento estaciones para cada mes del año 1997 por zonas.	39
7.7. Resumen metodología K-Vecinos mas cercanos.	41
7.8. Resumen metodología Distancia Inversa Ponderada.	42
7.9. Resumen metodología Regresión Lineal.	43
7.10. Resumen metodología Kriging Ordinario.	45
8.1. Localización de datos de precipitación total quincenal, zona Valle año 1997. .	51
8.2. Localización de datos de precipitación total quincenal, zona Montaña año 1997.	51
10.1. Comparación de métodos por zona Valle	76
10.2. Comparación de métodos por medio de Boxplot zona Valle	77
10.3. Comparación de métodos por zona Montaña	78
10.4. Comparación de métodos por medio de Boxplot Montaña	79

Capítulo 1

Introducción

En toda investigación resulta de suma importancia la recolección y procesamiento de los datos para obtener resultados confiables. Las conclusiones en un estudio requieren que la calidad de los datos sea buena, garantizando confiabilidad y validez en la información analizada (Ablan et al., 2008).

Los estudios meteorológicos requieren de datos sin observaciones faltantes, con el objetivo de encontrar cambios en las variables, realizar análisis de homogeneidad, aplicar metodologías de análisis multivariado, entre otros (Amaro et al., 2004). Tener la serie de datos completa, permite realizar inferencias basadas en observaciones igualmente espaciadas, es decir, en observaciones que se encuentran en intervalos de tiempo regulares (Ordoñez, 2001), donde la información estadística del sistema se conserva por su secuencialidad. Sin embargo, los datos de precipitación pueden presentar datos faltantes debido a errores de los equipos de medición, errores humanos, valores extremos y los errores de recolección (Infante et al., 2008). Por lo tanto, si la secuencialidad de las observaciones es interrumpida por cualquier razón, el problema de los datos faltantes debe ser manejado de una manera apropiada (Yozgatligil et al., 2013).

Un método ampliamente utilizado en la literatura para solucionar los problemas de datos faltantes es la estimación a partir del valor medio dentro de una serie. Del mismo, modo otro conjunto importante de enfoques para estimación climatológica se basa en técnicas de correlación, éstas consideran las series temporales de la variable estudiada y por lo tanto, toman en cuenta la "historia" de la estación y sus características climáticas, sin embargo, la dependencia espacial pierde importancia si no se consideran estaciones distintas de la estación objetivo, dicho problema se puede superar mediante el uso de observaciones de estaciones meteorológicas adyacentes (Ramos-Calzado et al., 2008).

En cuanto al cambio climático, este resulta ser un factor importante de estudio para el buen control y la toma de decisiones en políticas ambientales, por lo cual es importante, como se mencionó inicialmente, bases de datos completas y de alta calidad. El Valle del Cauca es uno de los departamentos más afectados en su rendimiento hídrico por su cercanía al Océano Pacífico, que es zona de influencia de fenómenos Oceano-Atmosféricos cuasi periódicos, como el fenómeno del Niño (ENSO) (Carvajal et al., 1998).

De esta forma, la economía del Valle del Cauca, se ve afectada por fenómenos como el Niño, que deben su presencia al rendimiento hídrico que presenta la región (Martínez Gómez, 2011). En esta región, la agricultura está bastante tecnificada, el producto más relevante para la economía departamental es la caña de azúcar, donde se encuentran las plantaciones más grandes e importantes del país (DNP, 2007). La producción agrícola demanda el mayor porcentaje del recurso hídrico (80 %), por lo cual las aguas subterráneas son utilizadas en un 88 % para esta producción (Carvajal et al., 1998). Siendo así, la precipitación es fundamental en los ciclos meteorológicos y responsable de muchos fenómenos hidrometeorológicos, por lo cual resulta necesario entender la variabilidad y la distribución espacial de ésta (Jang et al., 2015). Las mediciones de precipitación en el Valle del Cauca son realizadas por centros meteorológicos como la CVC (Corporación Autónoma Regional del Valle del Cauca) entidad de control ambiental de la región y el IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales), por medio de sus redes de estaciones meteorológicas, buscando entender el comportamiento de variables como la precipitación, para tomar decisiones sobre posibles fenómenos o situaciones que se puedan presentar gracias a esta (Gonzalez & Aristizabal, 2006).

En algunas regiones, la ubicación de las estaciones meteorológicas es mayormente irregular, ya que estas fueron instaladas en muchos casos primando criterios de acceso y seguridad y no tanto de idoneidad. Esto puede determinar, que existan desbalances en la densidad y representatividad de los datos obtenidos de las mismas. Para el caso de Colombia, aparecen zonas con un número excesivo de estaciones, mientras otras carecen por completo de sitios de monitoreo y por tanto de representatividad en los datos (Ortega & Rodríguez, 2010). Por lo cual, se presentan problemas a la hora de recoger estos datos con las estaciones de medición, siendo común encontrar datos faltantes en estas series de precipitación (Medina & Jaramillo, 2008), problemas como el cambio de lugar de los instrumentos de medición, la transformación del espacio físico del entorno donde se encuentra la estación o no registrar el dato en el momento adecuado. En este trabajo, se pretende comparar cuatro metodologías de estimación: tres de ellas basadas en datos georreferenciados y una en estimaciones a partir de la información con la que se disponga de estaciones climáticas para la estimación de datos faltantes en series diarias de precipitación en el Valle del Cauca.

Capítulo 2

Planteamiento del problema

Ante las evidencias de un cambio meteorológico a escala global, los análisis de tendencia de variación climática se han convertido en un tema de gran interés para la mayoría de los países del mundo (Greenpeace, 2009). Los expertos en la materia han señalado que para evidenciar la existencia de un cambio climático es necesario analizar las tendencias de variables climáticas tales como precipitación, brillo solar, temperatura, entre otras, considerando diferentes escalas de tiempo y espacio (Méndez González et al., 2008). Las tendencias son cambios graduales de incremento o disminución en el tiempo de las variables, asociadas comúnmente a factores meteorológicos. Por lo tanto, estudiar la variabilidad del clima regional es fundamental para planificar el manejo de recursos naturales, la forma más fácil es con respecto a medias anuales o estacionales de temperatura o precipitación (Castro-Llanos & Carvajal-Escobar, 2013).

La agricultura en el Valle del Cauca es la más importante y variada del país, con cultivos tecnificados de caña de azúcar, café, algodón, soya y sorgo. Esta zona se ha convertido en uno de los principales centros de desarrollo económico del país dominando todo el sur occidente Colombiano. Los procesos y actividades asociados al sector agropecuario están controlados por diversos factores, dentro de los cuales el clima ejerce un papel destacado y este se ve afectado por las temporadas extremas de alta o baja pluviosidad, así como también por las variaciones de temperatura. En los últimos años, en el Valle del Cauca se ha observado grandes alteraciones de patrones climatológicos (Carvajal et al., 1998) que son estudiados y analizados por diferentes gremios para anticiparse a estos o remediarlos, por lo cual un estudio en esta región sobre métodos de estimación de datos faltantes de precipitación diaria como éste ayudará a futuras investigaciones climatológicas.

Las series climáticas se ven afectadas por la existencia de datos faltantes por ejemplo en interrupciones ocasionales de estaciones automáticas, mal funcionamiento del instrumento de medición o reorganizaciones de la red de monitoreo (Simolo et al., 2010). Una manera de salir de esta dificultad es excluir los períodos con valores perdidos del análisis de

datos, o ignorar el problema si la cantidad de estos no es muy grande. No obstante, este tipo de soluciones generan problemas en la caracterización y modelación de las series de precipitación, generando una pérdida de confiabilidad en las conclusiones de dichos estudios. Los resultados de cualquier modelo y análisis estadístico podrían deteriorarse cuando se utilizan registros incompletos como entrada en el análisis, por esto, la estimación de los datos faltantes se convierte en prioridad en el proceso de preparación de datos, que en conjunto con otras metodologías aplicadas dejan los datos listos para su respectivo análisis (Yang Lenyap et al., 2015).

El objetivo de la estimación de datos faltantes es obtener un archivo de datos completos y consistentes para que puedan ser analizados mediante técnicas estadísticas que permitan generar inferencias validas sobre algún fenómeno estudiado, por lo cual, esto hace parte del proceso de investigación con el propósito de arribar a conclusiones sustentadas en evidencia empírica sólida. Si una muestra aleatoria de la variable (S) contiene datos faltantes, esta situación debe tenerse en cuenta en el proceso de construcción del estimador ($\hat{S}$). Si la estimación que se hace es adecuada el estimador sera cercano al verdadero valor del parametro (S) en muestras repetidas, logrando minimizar el sesgo, la varianza, la desviación estándar y el error cuadrático medio (ECM) del estimador (Galván, 2007).

Di Piazza et al. (2011) plantea dos enfoques principales para el tratamiento de datos faltantes, uno es utilizar sólo un subconjunto de registros continuos en un conjunto de datos y otra es hacer caso omiso de los datos basados en la suposición de que representan una serie continua faltante. En el primer enfoque, la información previa no es representativa impidiendo realizar inferencias, mientras que el segundo enfoque se reduce el período de registro y por lo tanto sobrestima la probabilidad de eventos extremos.

Con el objetivo de superar estos problemas, han sido desarrolladas diferentes técnicas basadas en la interpolación, dirigidas a la estimación de observaciones faltantes en las series de tiempo climáticas (Kaplan et al., 2000, Li et al., 2003, Simolo et al., 2010). Entre éstos, se encuentran métodos basados en Regresión Lineal, Kriging Ordinrio y la Distancia Inversa Ponderada por su siglas en inglés (IDW) (Xu et al., 2013). Actualmente se cuenta con varias metodologías para la solución del problema de los datos faltantes en series diarias de precipitación, sin embargo, algunas de estas metodologías no presentan los suficientes argumentos para su validéz o resultan ser mejores que otras para solucionar problemas específicos. Otros métodos que se utilizan son: el promedio aritmético simple, método de la relación normal, relación normal ponderada con el método de correlaciones, redes neuronales de percepción multicapa, método de imputación múltiple basado en MCMC, entre otros (Yozgatligil et al., 2013).

En este contexto, es de suma importancia evaluar métodos de estimación de datos faltantes en series de precipitación, basados en la información que proporcionan estaciones meteorológicas cercanas; teniendo en cuenta la zona geográfica y el porcentaje de datos faltantes. Por lo tanto, en el presente trabajo se plantea la siguiente pregunta de investigación *¿Es posible tener una metodología estadística confiable para estimar datos faltantes de series históricas de precipitación en el Valle del Cauca?*

Capítulo 3

Justificación

Entidades y grupos de investigación reconocen la preocupación que se tiene respecto a la variabilidad climática y las alteraciones en el régimen de lluvias, temperaturas y otras componentes del clima. La principal actividad de la región es la agricultura que se destaca por ser la más tecnificada del país, sin embargo, sufre de los efectos de los eventos derivados de la variabilidad climática. Para esta región, el análisis de componentes temporales y espaciales es una herramienta útil en la orientación y prevención, así se podrá tomar medidas de forma oportuna para afrontar catástrofes o desastres naturales que generen las variaciones de la precipitación (Hurtado et al., 2001).

Las precipitaciones registradas, con el fin de analizar la variabilidad climática que se mencionó anteriormente, son el insumo para la realización de estudios, investigaciones, actividades de seguimiento y manejo de la información que sirven para fundamentar la toma de decisiones en materia de recursos, costos de producción y política ambiental, entre otros. Es importante resaltar algunas limitaciones en cuanto a la disponibilidad de estos datos y en el caso específico de la información climática, ya que su confiabilidad está en la continuidad y calidad en el registro de ellos, durante períodos grandes (Medina & Jaramillo, 2008).

Las estaciones que toman las mediciones a veces presentan daños, irregularidades o simplemente dejan de funcionar; situaciones que intervienen de forma negativa en la calidad de la información. Con este tipo de información se realizan gran variedad de estudios y análisis, en donde la validez de éstos depende en gran medida de los datos o información manejada y la forma en que fue procesada. La validez de estos estudios se perdería, si un problema como el de los datos faltantes se pasara por alto, ya que conllevaría a omitir información relevante para la investigación, alterando los resultados y conclusiones de los estudios (Dumedah & Coulibaly, 2011).

Por otro lado, las estaciones meteorológicas además de presentar fallas, se encuentran distribuidas en toda el área sobre la cual se desean tomar las respectivas mediciones, en la cual se presentan diferentes formas de relieve que varían las condiciones bajo las cuales se encuentran cada una de las estaciones y que influyen en las observaciones que estas recogen. En particular resulta importante comparar los distintos métodos de estimación encontrados en la literatura, según las condiciones en las que se encuentren las estaciones, ya que los métodos pueden verse influenciados por estas (Ferrari & Ozaki, 2014).

Análogamente, es pertinente evaluar métodos adecuados para la solución del problema de los datos faltantes en series de precipitación diaria, con el objetivo de encontrar la metodología más adecuada que permita tener la mejor estimación, dependiendo de las condiciones en las que se encuentren las estaciones y evaluar los desempeños bajo estas. De esta forma, evitar utilizar métodos inadecuados para el llenado de series de datos bajo distintas condiciones (Agricultura, hidrología, meteorología, entre otros), ya que esto puede causar problemas inferenciales asociados a la ineficiencia, aparición de sesgos, distorsión de la estructura de covarianzas; además de no incorporar la incertidumbre asociada a los datos faltantes, se puede subestimar o sobre estimar la varianza produciendo bases de datos no confiables llevando a interpretaciones erróneas de los resultados por parte del investigador Viada et al. (2016), Penagos Cruz (2014), Lopez Araya (2014), Luna Romero & Lavado Casimiro (2015). En este sentido, se pretende evaluar los métodos de estimación de datos faltantes buscando el más eficiente para completar la serie de datos de precipitación, según la zona geográfica y el porcentaje de datos faltantes.

La escogencia de los métodos a comparar se realizó por medio de una amplia revisión bibliográfica de diferentes metodologías propuestas para la estimación de datos de este tipo, dentro de las cuales se escogieron las cuatro que mejor comportamiento tuvieron en estos antecedentes revisados (Firat et al., 2010, Vicente-Serrano et al., 2010, Chen & Liu, 2012, Mair & Fares, 2010, Ferrari & Ozaki, 2014, Campozano et al., 2015).

Capítulo 4

Antecedentes

A continuación se presenta una revisión preliminar de antecedentes con respecto a la estimación de datos faltantes en mediciones de variables climáticas.

Presti et al. (2010) trabajaron el problema de imputación de datos climatológicos, para esto usaron series diarias de precipitación en la cuenca del río Candelaro (Italia), comprendidas entre enero de 1970 y diciembre de 2001. Los autores compararon cuatro métodos de imputación de datos: sustitución simple, regresión lineal simple, regresión clasificada y el método de Theil. De acuerdo a los resultados obtenidos, los autores concluyeron que el método de Theil presentó el mejor desempeño en la estimación de los datos faltantes.

Firat et al. (2010) realizaron estimaciones de datos faltantes y se hizo un test de homogeneidad, para 267 estaciones de precipitación del servicio meteorológico de Turquía comprendidos en el periodo entre 1968-1998. Los datos faltantes se estimaron por medio de el método del vecino más cercano. Los resultados de la estimación se evaluaron mediante el coeficiente de correlación y se observó que había una buena estimación de estos.

Vicente-Serrano et al. (2010) realizaron una reconstrucción, control de calidad y homogenización de series diarias de precipitación, de una base de datos que comprendía los registros de precipitación diaria de 3016 estaciones del Instituto Nacional de Meteorología de España, en el periodo comprendido entre 1901 y 2001. La estimación de datos faltantes se realizó mediante los métodos del vecino más cercano, la distancia inversa ponderada y el método de regresión lineal. Para determinar el mejor método, se extrajo el 1 % de las series completas y evaluaron el desempeño con la Raíz del Error Cuadrático Medio (RMSE), donde se concluyó que el mejor método fue el del vecino más cercano, con el menor RMSE.

Chen & Liu (2012) estimaron la distribución de las precipitaciones en el centro de Taiwán, usando un total de 46 estaciones climáticas y datos de precipitaciones entre los

años 1981 y 2010. Los autores emplearon el método de la distancia inversa ponderada y evaluaron su desempeño a partir del coeficiente de correlación. Finalmente, concluyeron que las estimaciones obtenidas siempre tuvieron una asociación por encima de 0.95.

Yozgatligil et al. (2013) compararon métodos de imputación de datos faltantes, a partir de series de temperatura obtenidas del Servicio Meteorológico de Turquía. Los autores consideraron dos variables: precipitación total mensual y temperatura media mensual, en el periodo 1965 y 2006 con estaciones ubicadas en siete regiones diferentes. Los autores compararon los métodos de imputación (promedio aritmético simple, método de la relación normal, relación normal ponderada con el método de correlaciones, redes neuronales de percepción multicapa y un método de imputación múltiple basados en MCMC). Para evaluar el desempeño de cada método, los autores utilizaron el coeficiente de variación de la raíz del Error Cuadrático Medio (ECM) y el coeficiente de correlación. Los autores concluyeron que el método con mejor desempeño fue el método de imputación múltiple basados en métodos MCMC.

Lee & Kang (2015) realizaron un estudio en la cuenca Imha situada en Corea del Sur con datos de precipitación diaria en el periodo 2008-2010, utilizando cinco diferentes funciones Kernel para estimar los datos de precipitación diaria: Epanechnikov, cuártica, biweight, Tricube y coseno, las cuales se compararon con la estimación del vecino mas cercano y se evaluó su desempeño en la simulación del caudal usando el suelo Water Assessment Tool (SWAT) por medio de un modelo hidrológico. Obteniendo que los enfoques de núcleo proporcionan una mejor calidad en la estimación de datos de precipitación con respecto al método del vecino más cercano, tanto en términos estadísticos como en el rendimiento de modelación hidrológica.

Teegavarapu (2009) realizó estudio en la zona oriental del estado de Kentucky, con datos de precipitación anual promedio de 15 estaciones en el periodo de 1971-2013. Propusieron una nueva regla de asociación minera (ARM), un enfoque basado en la interpolación espacial, combinado con la potencia de las técnicas de minería de datos para mejorar las estimaciones de precipitación obtenidos, encontrando que el uso de la minería de reglas de asociación en conjunto con una técnica de interpolación espacial puede mejorar las estimaciones de la precipitación.

Kim & Pachepsky (2010) realizaron una reconstrucción de series diarias de precipitación en 39 estaciones meteorológicas ubicadas en la bahía de Chesapeake, de la National Climatic Data Center de Estados Unidos, en un periodo de Siete años comprendido entre los 2001 y 2007. Se compararon tres métodos: red neuronal artificial (ANN), árbol de regresión (RT) y red neuronal artificial después de árbol de regresión (RT+ANN). Se evaluó los métodos por medio de la raíz del error cuadrático medio y el coeficiente de correlación de

Pearson y se observó que el árbol de regresión (RT) proporcionó una mejor estimación que la red neuronal artificial (ANN), pero el uso de la red neuronal luego del árbol de regresión (RT+ANN) obtuvo el mayor coeficiente de correlación y la raíz del error cuadrático medio más baja.

Mair & Fares (2010) realizaron una comparación de métodos de estimación de datos faltantes de precipitación en las montañas de Waianae de la isla de Oahu, Hawai. Los autores trabajaron con los registros de 23 estaciones en el período comprendido entre los años 2005 y 2008. Los métodos que se compararon fueron: El Polígono de Thiessen, la Distancia Inversa Ponderada, Regresión Lineal, Kriging Ordinario (KO), donde el método con el mejor desempeño fue el Kriging ordinario con el menor Error Cuadrático Medio y el de Mayor Error Cuadrático medio fue con el método del Polígono de Thiessen.

Ferrari & Ozaki (2014) trabajaron con series de precipitación diarias provenientes de 484 estaciones meteorológicas, ubicadas en el estado de Paraná (Brasil) en el período enero de 1975 y diciembre de 2009. En este trabajo, los autores compararon tres métodos: el método del Vecino mas Cercano, Distancia Inversa Ponderada y el método de Regresión Lineal de datos faltantes, como medida de desempeño utilizaron la Raíz del Error Cuadrático medio (RECM), finalmente concluyeron que el método con mejor desempeño fue la Distancia Inversa Ponderada.

Fernandez & Lavado (2014) evaluaron la eficiencia estadística de los métodos de interpolación: IDW, Kriging Ordinario, Co-kriging y Kriging con deriva externa, con el fin de identificar el mejor modelo que representa la variabilidad espacial de la precipitación en el entorno regional de las cuencas Chillón, Rimac, Lorín y parte alta de Mantaro en Perú para elaborar mapas climáticos de la precipitación mensual y anual, los datos de precipitación mensual utilizados para el estudio comprendían el periodo 1981-2010 al que se le realizó un control de calidad y fueron homogeneizados, la mejor técnica para caracterizar la distribución espacial de la precipitación fue la técnica geoestadística multivariada Co-kriging que hace uso de la covariable elevación.

Teegavarapu (2014) realizó un estudio en dos regiones, la primera en el estado de Kentucky para estimar datos faltantes de precipitación diaria de la época 1971-2001 con información de 15 pluviómetros, la segunda en el suroeste de Florida con datos del periodo 1994-1999 con 43 pluviómetros. Para esto, utilizaron un ejercicio integral llevado a cabo para comparar el rendimiento de métodos de interpolación espacial deterministas y estocásticos, los autores propusieron un nuevo método basado en la proximidad óptima: la clasificación por el K-means clustering y K-vecinos más cercanos (K-NN). Los resultados de estas comparaciones indican que los métodos propuestos realizaron mejores estimaciones que los métodos existentes.

Campozaño et al. (2015) evaluaron 17 métodos de estimación de datos faltantes, en 15 estaciones de la cuenca del río Paute en los Andes Ecuatorianos, en el período de enero 1981 a diciembre 2010, de forma anual. Algunos de los 17 métodos evaluados fueron: estaciones más cercanas por medio de valores distancia, valor promedio de las k-estaciones mas cercanas, regresión lineal con la estación mas cercana, entre otros, se evaluaron por medio del error cuadrático medio, error medio y error absoluto. Los resultados mostraron que para completar series temporales de precipitación diaria el método de regresión lineal múltiple ponderada es el mejor, y para la temperatura fue la media meteorológica del día.

Membreño & Zavala (2015) aplicaron métodos estadísticos para obtener un análisis de la relación entre las variables climáticas y otras variables mediante los modelos de regresión simple y múltiple, se analizó la precipitación como variable dependiente y las variables independientes de la latitud, longitud y elevación (usando las curvas de nivel) y un análisis de series temporales para una estación meteorológica ubicada en Chinandega Nicaragua, analizando el comportamiento de la variable precipitación a lo largo del tiempo. Se realizaron diferentes cartografías continuas de la precipitación promedio mensual para el periodo 1984-2005, aplicando el método de interpolación IDW luego de verificar que los datos no seguían una distribución normal, por lo cual no se aplicó el método Kriging y se analizó la calidad de las diferentes cartografías continuas. El uso de las técnicas estadísticas y geoestadísticas IDW mostraron una forma para concebir el análisis de la información espacial en zonas de difícil acceso y con pocos puntos discretos.

Kim & Ryu (2015) buscaron un umbral de datos faltantes, donde se pueda decidir si la serie es viable para estimar o no, trabajando con una estación meteorológica (COOP ID: 101022), cerca del aeropuerto Boise en el condado de Ada, Idaho-EE.UU, en el periodo comprendido entre 1981-2013. Para lo anterior, emplearon un modelo auto regresivo con diferentes porcentajes de datos faltantes (5, 10, 15, 16, 17, y 18 %). El resultado indica que alrededor del 15 % de nivel de datos que falte es plausible para la reconstrucción de la serie de precipitación diaria. El modelo se evaluó por medio del R^2 , donde se observó que a partir de una cantidad de datos faltantes del 15 % no se presenta mucha variación en este indicador.

Medina & Jaramillo (2008) utilizaron series históricas de precipitación diaria de 10 estaciones climáticas de la red meteorológica de la FNC distribuidas en toda la geografía cafetera, en un periodo con más de 25 años de información. Para esto, los autores utilizaron el método de regresión acumulada, encontrando que el método es flexible a las características propias de cada estación y su estimación depende únicamente de su información histórica.

Urrutia et al. (2010) presentaron la metodología que se debe seguir para la imputación de datos en series de precipitación y/o temperatura en estaciones de Colombia en un tramo

de 10 años. El procedimiento consistió en hacer uso de correlaciones parciales, modelos de regresión, ajustes de los datos por medio del método de doble masa y verificación de la tendencia a través del el test de Kendal. Se pudo evaluar que para un porcentaje de datos faltantes del 20 % ésta metodología resulta apropiada, sin embargo, para porcentajes de datos faltantes de más del 20 % se recomienda usar los modelos AMI bivariada para la imputación de estos.

Carvajal & Marco (2011) efectuaron un análisis de Funciones Ortogonales Empíricas (FOES) o Componentes Principales (CP) a 50 estaciones de precipitación mensual (1972-1998) y a 8 estaciones de caudal mensual (1951-2000) del Valle del Cauca-Colombia. Previamente se aplicaron 5 medidas de adecuación para verificar la conveniencia del análisis: (i) La evaluación del grado de significancia de la correlación entre variables; (ii) KMO; (iii) La prueba de esfericidad de Bartlett; (iv) MSA; y, (v) El porcentaje de residuos no redundantes con valores absolutos > 0.05 . Para seleccionar las CP significantes de cada conjunto de variables se utilizaron 7 criterios de selección: El método gráfico, el porcentaje de varianza explicada, la raíz promedio, y las pruebas de Velicer, Bartlett, Broken Stick y de validación cruzada, escogiéndose esta última como la mejor, por ser más robusta y cuantitativa.

Se agruparon las estaciones de precipitación en 3 grupos homogéneos, aplicando un análisis de datos anomalos jerarárquicos que fue verificado con el método geográfico y el análisis discriminante de las 4 primeras FOES de precipitación. Se destacan las ventajas del método FOES para reducir la dimensionalidad de datos multivariantes, calcular datos faltantes, evaluar y reducir multicolinealidad, conformar grupos homogéneos y detectar datos anómalos. Con las primeras cuatro CP significantes se explica el 60.34 % de la varianza total de la precipitación mensual del Valle del Cauca y el 94 % de la varianza total de los registros de caudal seleccionados.

Castro-Llanos & Carvajal-Escobar (2013) analizaron el comportamiento de la precipitación mensual y total anual en 47 estaciones climáticas de la CVC en el Valle del Cauca y el Norte del Cauca, en el periodo comprendido entre 1971 y 2009. Para solucionar el problema de datos faltantes, los autores utilizaron el método de proporciones, el método de imputación por regresión lineal, el método de imputación por la media o método de sustitución. Posteriormente, evaluaron la tendencia de la precipitación mediante los métodos de Mann-Kendall y el estimador de la pendiente de Sen. Finalmente, encontraron que las tendencias crecientes en la precipitación mensual y total anual surgen en el centro y norte del Valle del Cauca, por lo que hay una tendencia de aumento en las lluvias, mientras que las tendencias negativas se presentan en el sur del Valle y el Norte del Cauca.

Capítulo 5

Objetivos

5.1. Objetivo general

Comparar algunas metodologías estadísticas clásicas o basadas en información adyacente con mejor desempeño para estimar datos faltantes en series de precipitación diaria en el Valle del Cauca, cuando se presentan cambios en el porcentaje de datos faltantes y la zona geográfica.

5.2. Objetivos Específicos:

- Caracterizar la estimación de datos faltantes en cada metodología, cuando se presentan cambios en el porcentaje de datos faltantes y la zona geográfica.
- Comparar el desempeño de las metodologías de estimación de datos faltantes (K-Vecinos más cercanos, Distancia Inversa Ponderada, Regresión lineal, Kriging Ordinario), cuando se presentan cambios en el porcentaje de datos faltantes y la zona geográfica.

Capítulo 6

Marco Teórico

A continuación se muestran las definiciones teóricas de algunos de los conceptos y metodologías, que se emplearán para el desarrollo de este trabajo.

6.1. Marco conceptual

Para empezar, se definirán los conceptos que describen la problemática a desarrollar en cuanto a su sentido aplicado, proporcionando de esta forma un conocimiento más profundo sobre el contexto en que se desarrollará este trabajo.

6.1.1. Zona de estudio

Según El Fondo Nacional de Desarrollo FONADE (2016), el departamento del Valle del Cauca, se encuentra situado al sur-occidente del país, formando parte de las regiones andina y pacífica; localizado entre los $05^{\circ}02'08''$ y $03^{\circ}04'02''$ de latitud norte y a los $72^{\circ}42'27''$ y $74^{\circ}27'13''$ de longitud oeste. El área plana del Valle del Cauca tiene un régimen de precipitación bimodal y presenta dos períodos de mayores lluvias, de marzo a mayo, y de septiembre a noviembre; el área montañosa de ladera tiene períodos más amplios y de mayores lluvias que van de marzo a junio y de septiembre a diciembre. Además, cuenta con gran variedad de cultivos, como la caña de azúcar, entre otros, y buena producción de ganado, como el vacuno, porcino, etc.

6.1.2. Precipitación

Se denomina precipitación al agua que proviene de la humedad atmosférica y cae a la superficie terrestre, principalmente en estado líquido, (lluvia) o sólido (nieve o granizo) (UNESCO, 1982).

Cualquier producto de la condensación del vapor de agua atmosférico que se deposita en la superficie de la Tierra. Ocurre cuando la atmósfera (que es una gran solución gaseosa) se satura con el vapor de agua, y el agua se condensa y cae de la solución (es decir, precipita). El aire se satura a través de dos procesos: por enfriamiento y añadiendo humedad. Este elemento se convierte en una variable climatológica cuando se obtienen valores cuantitativos o cualitativos, producto de sus registros y/o mediciones. El volumen de precipitación se mide en milímetros. Un milímetro equivale a un litro de agua por metro cuadrado Maderey Rascon et al. (2005). Los dos principales dispositivos de medición de la precipitación, son el pluviómetro y el pluviógrafo. El primero mide el volumen total de lluvia caída durante el día meteorológico (7 A.M. a 7 A.M. del día siguiente). El segundo es un aparato de registro continuo que permite determinar la intensidad de la precipitación en un intervalo de tiempo dado (Moreno, 2009).

Precipitación en el Valle del Cauca

El Valle del Cauca presenta un régimen de precipitación bimodal como resultado de la Zona de Convergencia Intertropical (ZCIT), con los trimestres secos (diciembre-febrero y junio-agosto) y dos lluviosos (marzo-mayo y septiembre-noviembre), la precipitación anual de la zona interandina del valle oscila entre 4000 mm y 1200 mm en alta cordillera y plana respectivamente. Del total de precipitación, el 70 % corresponden a los meses lluviosos y el 30 % a los secos (Carvajal et al., 1998).

6.1.3. Valle

El valle está constituido por el cauce del río, la planicie de inundación activa y abandonada y está limitado por las montañas o colinas circundantes. En la formación de los valles se incluyen los procesos de meteorización, erosión y movimientos tectónicos condicionados por la litología y la resistencia de los suelos a los procesos de erosión (Rosgen & Silvey, 1996).

Los valles fluviales, y en su generalidad, regiones enteras, progresan a través de una serie de etapas, mediante las cuales se desarrolla del paisaje. El ciclo de erosión brinda una descripción cualitativa de los valles de los ríos y de las áreas (Gracia & Maza, 1997).

6.1.4. Montaña

Es un área en declive en todas las direcciones desde una ubicación determinada, es una de las seis formas básicas de relieve en la superficie de la tierra. Un sistema de montaña se

interpreta como una forma de relieve de orden superior o una agregación de un número de formas de relieve individuales (montaña) que se muestran en la **Tabla 6.1**.

Varios autores se han referido a la elevación entre las cumbres y los fondos de valle, por lo general superior a 1,000 *m*, para caracterizar las zonas de alta montaña, con frecuencia más de 2,000 *m* de relieve que ocurre en tan sólo una distancia horizontal pocos kilómetros. Una distinción cuantitativa entre “sistemas de alta montaña” y “sistemas de montaña sigue siendo difícil”. El criterio para separar los dos grupos se basa en una diferencia de altura de por lo menos 2,000 *m* entre paso de montaña y la cumbre adyacente (Bishop & Shroder, 2004).

Tabla 6.1: Definición de las áreas de alto relieve

Tipo	Relieve (sobre 5km de distancia)	Relieve relativo
Sistema de alta montaña	<1,000 <i>m</i>	500 <i>m/km</i> ²
Sistema de montaña	500-1,000 <i>m</i>	200 <i>m/km</i> ²
Terreno montañoso	100-500 <i>m</i>	100 <i>m/km</i> ²
Terreno colina	50-100 <i>m</i>	50 <i>m/km</i> ²

6.1.5. Estaciones meteorológicas

La red meteorológica del IDEAM está compuesta por un conjunto de estaciones de observación, medición y registro de diferentes fenómenos atmosféricos, con diferentes finalidades, distribuidas en el territorio nacional (IDEAM, 2016).

Las estaciones meteorológicas de esta entidad se clasifican en: Estación Pluviométrica (PM), Estación Pluviográfica (PG), Estación Climatológica Principal (CP), Estación Climatológica Ordinaria (CO), Estación Sinóptica Principal (SP), Estación Sinóptica Suplementaria (SS), Estación Agrometeorológica (AM), Estación de Radiosonda (RS), Estación Meteorológica Especial y Estación Meteorológica Marina.

El modelo digital de elevaciones (DEM), describe la altimetría de una zona mediante un conjunto de cotas. Es una estructura numérica de datos que representa la distribución espacial de la altitud de la superficie del terreno tal como se muestra en la **figura 6.1** (Felicísimo, 1994).

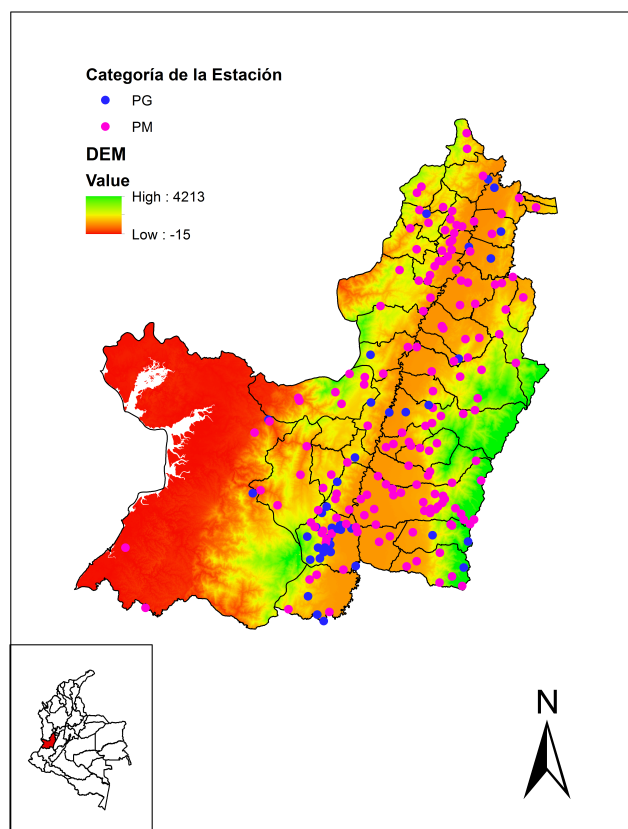


Figura 6.1: Estaciones climáticas del Valle del Cauca

6.2. Marco Estadístico

Ahora, se proporcionará un conjunto de definiciones estadísticas que hacen parte de la teoría que le da sentido a este trabajo y que son fundamentales en el entendimiento de cada una de las partes del proyecto.

6.2.1. Datos faltantes

En un sentido amplio, el término datos faltantes significa que falta algún tipo de información sobre los fenómenos en los que se está interesado (McKnight et al., 2007). Existen situaciones inherentes que pueden afectar la calidad de los datos o registros de las variables de estudio, tal es el caso de la información de tipo climático, específicamente, las series históricas de precipitación, construidas a partir de las mediciones realizadas en distintas estaciones meteorológicas, las cuales pueden verse afectadas por el cambio de lugar o movimiento de instrumentos, el cambio del espacio físico del entorno donde se encuentra la estación, o no registrar el dato en el momento adecuado, lo que conlleva a la inconsistencia o ausencia de datos (Medina & Jaramillo, 2008).

6.2.2. Correlación espacial

Definida de manera simple, la correlación espacial es la concentración o dispersión de los valores de una variable en un mapa. Dicho de otra manera, refleja el grado en que objetos o actividades en una unidad geográfica son similares a otros objetos o actividades en unidades geográficas próximas (Goodchild, 1987). Este tipo de correlación prueba la primera ley geográfica de Tobler (1970) que afirma: todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las distantes (Vilalta y Perdomo, 2005). En este contexto se busca establecer la correlación espacial bajo continuidad en el espacio, un indicador de esto es el semivariograma.

Variable regionalizada

Una variable medida en el espacio de forma que presente una estructura de correlación, se dice que es una variable regionalizada. De manera más formal se puede definir como un proceso estocástico con dominio contenido en un espacio euclidiano d -dimensional R^d , $Z(x) : x \in D \subset R^d$. Si $d = 2$, $Z(x)$ puede asociarse a una variable medida en un punto x del plano. En términos prácticos $Z(x)$ puede verse como una medición de una variable aleatoria (por ejemplo: concentración de un contaminante) en un punto x de una región de estudio. Un proceso estocástico es una colección de variables aleatorias indexadas; esto es, para cada x en el conjunto de índices D , $Z(x)$ es una variable aleatoria. En el caso de que las mediciones sean hechas en una superficie, entonces $Z(x)$ puede interpretarse como la variable aleatoria asociada a ese punto del plano (x representa las coordenadas, planas o geográficas, y Z la variable en cada una de ellas). Estas variables aleatorias pueden representar la magnitud de una variable ambiental medida en un conjunto de coordenadas de la región de estudio (Díaz-Francés, 1993).

Variograma y semivariograma

La función básica que describe la variabilidad espacial y estructura de correlación de un fenómeno de interés se conoce como semivariograma, resulta necesario ajustar un modelo teórico que permita determinar los parámetros del semivariograma para ser usados en la predicción.

Como se asume que la varianza de los incrementos de la variable regionalizada es finita. A esta función denotada por $2\gamma(h)$ se le denomina variograma. Utilizando la definición teórica de la varianza en términos del valor esperado de una variable aleatoria, tenemos:

$$\begin{aligned} 2\gamma(h) &= V(Z(x+h) - Z(x)) \\ &= E((Z(x+h) - Z(x))^2) - (E(Z(x+h) - Z(x)))^2 \\ &= E((Z(x+h) - Z(x))^2) \end{aligned} \tag{6.1}$$

La mitad del variograma $\gamma(h)$, se conoce como la función de semivarianza y caracteriza las propiedades de dependencia espacial del proceso. Dada una realización del fenómeno, la función de semivarianza es estimada, por el método de momentos, a través del semivariograma experimental, que se calcula mediante (Wackernagel, 2013):

$$\bar{\gamma}(h) = \frac{\sum (Z(x+h) - Z(x))^2}{2n} \quad (6.2)$$

donde $Z(x)$ es el valor de la variable en un sitio x , $Z(x+h)$ es otro valor muestral separado del anterior por una distancia h y n es el número de parejas que se encuentran separadas por dicha distancia. La función de semivarianza se calcula para varias distancias h . En la práctica, debido a irregularidad en el muestreo y por ende en las distancias entre los sitios, se toman intervalos de distancia $[0, h]$, $(h, 2h]$, $(2h, 3h]$, ... y el semivariograma experimental corresponde a una distancia promedio entre parejas de sitios dentro de cada intervalo. El número de parejas de puntos n dentro de los intervalos no es constante (Giraldo, 2009).

Modelos teóricos de semivarianza

Existen diversos modelos teóricos de semivarianza que pueden ajustarse al semivariograma experimental. En general dichos modelos pueden dividirse en no acotados (lineal, logarítmico, potencial) y acotados (esférico, exponencial, gaussiano). Los del segundo grupo garantizan que la covarianza de los incrementos es finita, por lo cual son ampliamente usados cuando hay evidencia de que presentan buen ajuste. Todos estos modelos tienen tres parámetros comunes (Giraldo, 2009):

- **Efecto Pepita** Se denota por C_0 y representa una discontinuidad puntual del semivariograma en el origen. Puede ser debido a errores de medición en la variable o a la escala de la misma. En algunas ocasiones puede ser indicativo de que parte de la estructura espacial se concentra a distancias inferiores a las observadas.
- **Meseta** Es la cota superior del semivariograma. También puede definirse como el límite del semivariograma cuando la distancia h tiende a infinito. La meseta puede ser o no finita. Los semivariogramas que tienen meseta finita cumplen con la hipótesis de estacionariedad fuerte; mientras que cuando ocurre lo contrario, el semivariograma define un fenómeno natural que cumple solo con la intrínseca.
- **Rango** En términos prácticos corresponde a la distancia a partir de la cual dos observaciones son independientes. El rango se interpreta como la zona de influencia. Existen algunos modelos de semivariograma en los que no existe una distancia finita para la cual dos observaciones sean independientes; por ello se llama rango efectivo a la distancia para la cual el semivariograma alcanza el 95 % de la meseta. Entre más pequeño sea el rango, más cerca se está del modelo de independencia espacial. El rango no siempre aparece de manera explícita en la fórmula del semivariograma.

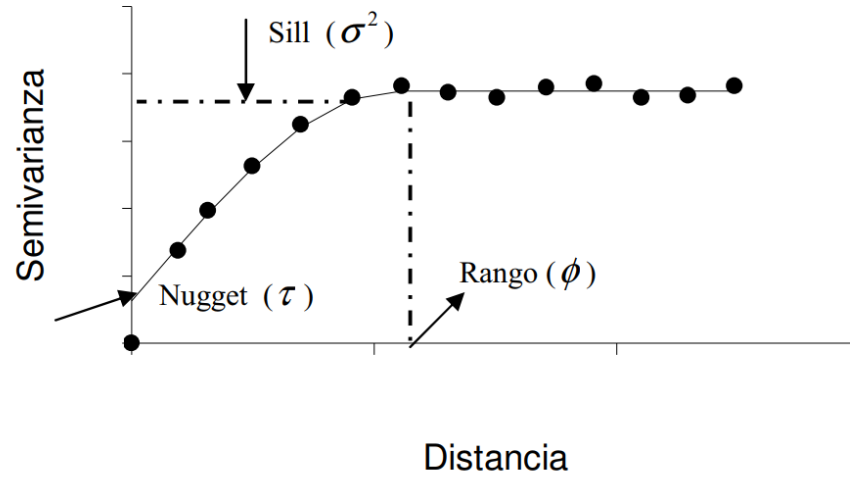


Figura 6.2: Comportamiento típico de un semivariograma acotado con una representación de los parámetros básicos. Fuente: (Giraldo, 2009)

A continuación se presentarán los modelos más comunes de semivarianza tomados de Giraldo (2009), Cressie (1993):

Modelo Esférico

Tiene un crecimiento rápido cerca al origen, pero los incrementos marginales van decreciendo para distancias grandes, hasta que para distancias superiores al rango los incrementos son nulos. Su expresión matemática es la siguiente:

$$r(h) = \begin{cases} C_o + C_1 \left(\frac{3}{2} \left(\frac{h}{a} \right) - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right) & h \leq a \\ C_o + C_1 & h > a \end{cases} \quad (6.3)$$

En donde C_1 representa la meseta, a el rango y h la distancia.

Modelo exponencial

Este modelo se aplica cuando la dependencia espacial tiene un crecimiento exponencial respecto a la distancia entre las observaciones. El valor del rango es igual a la distancia para la cual el semivariograma toma un valor igual al 95 % de la meseta. Este modelo es ampliamente usado. Su expresión matemática es la siguiente:

$$\gamma(h) = C - 0 + C_1(1 - \exp(\frac{-3h}{a})) \quad (6.4)$$

Modelo gaussiano

Al igual que en el modelo exponencial, la dependencia espacial se desvanece solo en una distancia que tiende a infinito. El principal distintivo de este modelo es su forma parabólica cerca al origen. Su expresión matemática es:

$$\gamma(h) = C - 0 + C_1(1 - \exp(\frac{-h}{a^2})) \quad (6.5)$$

Modelo Matern

Esta familia lleva el nombre de Matern Bertil, quien lo introdujo en su tesis doctoral sueca en 1960, posteriormente renombrado como Matern (1986). Se define por:

$$\rho(u; \phi, \kappa) = 2^{\kappa-1} \Gamma(\kappa)^{-1} (u/\phi)^\kappa K_\kappa(u/\phi) \quad (6.6)$$

donde (ϕ, κ) son parámetros y $K_\kappa(\cdot)$ denota la función modificada de Bessel de tercera clase de orden κ .

Esta familia es valida para $\phi > 0$ y $\kappa > 0$. El caso $\kappa = 0.5$ es el mismo de la función de correlación exponencial, $\rho(u) = \exp(-u/\phi)$. La función de correlación Gaussiana, es el caso donde $\kappa \rightarrow \infty$.

Una característica atractiva de este modelo es que el parámetro κ controla la diferenciabilidad del proceso subyacente $S(x)$ de forma directa; la parte entera de κ da el número de veces que $S(x)$ es diferenciable en media cuadrática. Por lo tanto, como ya hemos visto, $\kappa = 0.05$ corresponde a un proceso continuo, pero no diferenciable, $\kappa = 1$ es diferenciable, y mayores valores de κ se pueden elegir para dar derivados adicionales, lo cual es deseable.

Estimación por mínimos cuadrados

Esta técnica se puede aplicar a la estimación del variograma de la siguiente manera. Las coordenadas del variograma empírico u_{ij} son un estimador insesgado de $V(u_{ij})$, para cualquiera que sean las propiedades de distribución del proceso $Y(\cdot)$. Por lo tanto, un estimador consistente de θ en un modelo paramétrico, $V(u) = V(u, \theta)$, se puede obtener al

minimizar la cantidad: (Giraldo, 2009, Cressie, 1993)

$$MCO_1(\theta) = \sum_{j>1} v_{ij} - V(u_{ij}; \theta)^2 \quad (6.7)$$

o, es decir, la contraparte muestral del variograma

$$MCO_2(\theta) = \sum_k u_{ij} - V(v_{ij}; \theta)^2 \quad (6.8)$$

El variograma muestral (6.8) es más comúnmente utilizado en la práctica, pero aparte de la comodidad computacional, los autores no pueden pensar en ninguna buena razón por las cuales este debe ser así, y desde un punto de vista estadístico no es una desventaja obvia: para una distribución irregular de los lugares de muestreo, el promedio de las ordenadas del variograma empírico dentro de los rangos discretos significa que v_k solo es aproximadamente insesgado para $V(u_k; \theta)$.

6.2.3. Métodos de estimación de datos faltantes

A continuación, se ilustra los métodos de estimación de datos faltantes utilizados en el presente trabajo de investigación.

El método del K Vecino más cercano (KV)

Estima el valor de un dato desconocido a partir de las características del dato más próximo, según una medida de similitud o distancia Cover & Hart (1967). Esta regla tiene propiedades estadísticas bien establecidas y facilidad de aplicación a sistemas reales. El método del vecino más cercano se puede extender utilizando no uno, sino un conjunto de datos más cercanos para predecir el valor de los nuevos datos, en lo que se conoce como los k-vecinos más cercanos (k-NN o k-Nearest Neighbors). Al considerar más de un vecino, se brinda inmunidad ante ruido y se suaviza la curva de estimación (Flórez et al., 2014).

Este método es ampliamente utilizado en casos de estimación de registros faltantes en series de datos climáticos. Además considera la distancia más pequeña entre pares de estaciones. Por lo tanto, los datos que faltan se estiman directamente en la serie observada con los datos de las estaciones meteorológicas más cercanas (Ferrari & Ozaki, 2014).

El método KV se adapta fácilmente a la regresión de funciones con valores continuos. El algoritmo asume que todos los datos pertenecen a R^p y mediante una medida de distancia en ese espacio se determinan los k datos más cercanos al nuevo dato x_q para aproximar una

función $f : R^p \rightarrow R$, a partir de los k valores ya seleccionados. Esta función corresponde al promedio de los k valores más cercanos; si se considera el promedio aritmético (todos los datos dentro del grupo tienen igual relevancia), la función aproximación tiene la forma que se presenta en la expresión (6.9) (Flórez et al., 2014).

$$\hat{f}(x_q) \leftarrow \frac{1}{k} \sum_{i=1}^k f(x_i) \quad (6.9)$$

Todos los datos deben estar normalizados, para evitar que las características en el conjunto de entrada con valores más altos dominen el cálculo de la distancia (Cover & Hart, 1967).

El método de la Distancia Inversa Ponderada (IDW)

Este método consiste en calcular un promedio ponderado, donde los pesos para las observaciones son calculados de acuerdo a su distancia con la ubicación de interpolación, como se muestra en la expresión (6.10) y (6.11) (Bivand et al., 2013):

$$\hat{Z}(S_0) = \frac{\sum_{i=1}^n w(s_i) z(s_i)}{\sum_{i=1}^n w(s_i)} \quad (6.10)$$

$$w(s_i) = \|s_i - s_0\|^{-p} \quad (6.11)$$

Donde $\|s_i - s_0\|$ indica la distancia euclidiana y p una fuente de ponderación de la distancia inversa, en su defecto, 2. Si s_0 coincide con una observación, el valor observado es retornado para evitar pesos infinitos (Bivand et al., 2013).

La potencia de la distancia inversa determina el grado en el punto o los puntos más cercanos son preferibles a los puntos más distantes, para valores grandes de IDW converge a la interpolación por vecino mas cercano (Bivand et al., 2013).

El método de Regresión Lineal (RL)

Para este método se debe seleccionar una serie de datos con un comportamiento similar (misma área de influencia topo - climática a la serie que tiene los datos faltantes), la serie de datos de referencia debe contener el registro de datos completos en los periodos para los cuales faltan datos en la serie de estudio (Allen et al., 2006).

Los valores de la serie de referencia se denotan como x_i y los de la serie de estudio y_i cuyos datos no están completos, para caracterizar los registros de las series se toman aquellos periodos en los cuales los datos en ambas series están presentes, obteniéndose la media y desviación estándar para cada serie. Se estiman los coeficientes de la regresión de y con respecto a x para los periodos donde ambas series están completas, esto con las ecuación **6.12** que se observa a continuación (Medina & Jaramillo, 2008).

$$b = \frac{\text{cov}(xy)}{S_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad a = \bar{y} - b\bar{x} \quad (6.12)$$

De ésta forma, Si:

- El coeficiente de regresión es diferente de cero estadísticamente según una prueba de t, al 95 %.
- El coeficiente de determinación R^2 es mayor del 75 %.
- Cumple los supuestos: (Linealidad del modelo, varianza constante, independencia y normalidad errores).

Se tendría un método para estimar los datos faltantes de la serie de estudio, remplaceando la expresión obtenida (Medina & Jaramillo, 2008).

Kriging Ordinario (KO)

El método Kriging Ordinario busca primero efectuar un análisis de la correlación espacial de los datos. Para esto se crea un variograma (o semivariograma) que expresa cuánto varían los datos de las mediciones en función de la distancia entre ellas. Con los datos obtenidos del variograma se resuelve un sistema de ecuaciones lineales para encontrar los pesos óptimos. Este método busca, entonces, obtener los pesos con un conocimiento de cómo se relacionan los datos según la distancia (Bromberg & Pérez, 2012).

Según Cressie (1993) el Kriging es un método de estimación que da la mejor estimación lineal insesgada de los valores de los puntos, esto es, elegir el promedio ponderado de los valores de las muestras la cual tenga la mínima varianza.

Sea $z(x_0)$ un valor a estimar en x_0 , usando los valores de los datos $z(x_\alpha)$, de n puntos muestrales vecinos x_α y combinándolos linealmente con pesos w_α , i.e.

$$Z^*(x_0) = \sum_{\alpha=1}^n w_\alpha Z(x_\alpha) \quad (6.13)$$

Los pesos son elegidos de tal manera que el estimador sea:

1. Insesgado:

$$E[Z^*(x_0) - Z(x_0)] = 0 \quad (6.14)$$

2. Varianza mínima:

$$Var[Z^*(x_0) - Z(x_0)] \quad (6.15)$$

La propiedad 1 (estimador insesgado), es garantizada con la suma unitaria de los pesos, esto es:

$$\sum_{\alpha=1}^n w_{\alpha} = 1 \quad (6.16)$$

$$E[Z^*(x_0) - Z(x_0)] = E\left[\sum_{\alpha=1}^n w_{\alpha} Z(x_{\alpha}) - Z(x_0) \cdot \sum_{\alpha=1}^n w_{\alpha}\right] = \sum_{\alpha=1}^n w_{\alpha} E[Z(x_{\alpha}) - Z(x_0)] = 0 \quad (6.17)$$

Dado que las esperanzas de los incrementos es cero.

La varianza de la estimación es la varianza de la combinación lineal:

$$\sigma_E^2 = Var[Z^*(x_0) - Z(x_0)] \quad (6.18)$$

$$Z^*(x_0) - Z(x_0) = \sum_{\alpha=1}^n w_{\alpha} Z(x_{\alpha}) - 1 \cdot Z(x_0) = \sum_{\alpha=0}^n w_{\alpha} Z(x_{\alpha}) \quad (6.19)$$

con un peso w_0 igual a -1,

$$\sum_{\alpha=0}^n w_{\alpha} = 0 \quad (6.20)$$

La varianza de la estimación

$$\sigma_E^2 = E[(Z^*(x_0) - Z(x_0))^2] = 2 \sum_{\alpha=1}^n w_{\alpha} \gamma(x_{\alpha} - x_0) - \sum_{\alpha=1}^n \sum_{\beta=1}^n w_{\alpha} w_{\beta} \gamma(x_{\alpha} - x_{\beta}) - \gamma(x_0 - x_0)$$

$$(6.21)$$

Minimizando la estimación de la varianza con la restricción sobre los pesos

$$Var[Z^*(x_0) - Z(x_0)] - 2\mu^{OK}(\sum w_\alpha^{OK} - 1) \quad (6.22)$$

obtenemos el sistema de kriging ordinario (OK)

$$\begin{cases} \sum_{\beta=1}^n w_\beta^{OK} \gamma(x_\alpha - x_\beta) + \mu^{OK} = \gamma(x_\alpha - x_0); \text{ Para } \alpha = 1, \dots, n \\ \sum_{\beta=1}^n w_\beta^{OK} = 1 \end{cases} \quad (6.23)$$

Se puede expresar el sistema anterior en forma matricial como,

$$\begin{pmatrix} \gamma(x_1 - x_1) & \dots & \gamma(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n - x_1) & \dots & \gamma(x_n - x_n) & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1^{OK} \\ \vdots \\ w_n^{OK} \\ \mu^{OK} \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix} \quad (6.24)$$

La varianza de la estimación del kriging ordinario es,

$$\sigma_{OK}^2 = \mu^{OK} - \gamma(x_0 - x_0) + \sum_{\alpha=1}^n w_\alpha^{OK} \gamma(x_\alpha - x_0) \quad (6.25)$$

El estimador kriging ordinario es un interpolador exacto en el sentido que si x_0 es idéntico con una locación de los datos entonces el valor estimado es idéntico con el valor del dato en ese punto

$$Z^*(x_0) = Z(x_\alpha); x_0 = x_\alpha. \quad (6.26)$$

6.2.4. Indicadores de desempeño

En este trabajo, se emplearán un conjunto de indicadores para evaluar el desempeño de cada uno de los métodos de estimación y poder compararlos entre ellos. A continuación, se definirán estos indicadores y describirá su funcionamiento.

Error Cuadrático Medio (ECM)

Según De la Horra Navarro (2003) el Error Cuadrático Medio de un estimador T , para estimar θ , se define como:

$$ECM(T) = E[(T - \theta)^2] = E[(T(X1, \dots, Xn) - \theta)^2] \quad (6.27)$$

A continuación, se muestran los pasos a considerar:

- $T(X1, \dots, Xn) - \theta$ mide el error que se comete al estimar θ mediante $T(X1, \dots, Xn)$.
- Consideramos el cuadrado de ese error para evitar que las diferencias positivas se compensen con las negativas.
- Finalmente calculamos cuánto vale en promedio este error cuadrático.

$$ECM(\hat{\theta}) = \frac{\sum_{i=1}^k (\hat{\theta}_i - \theta_i)^2}{k} \quad (6.28)$$

$\hat{\theta}_i$ =Estimador del parámetro.

θ =Parámetro o valor observado.

k= Numero de observaciones.

Coefficiente de Variación de la Raíz del Error Cuadrático Medio (CVRECM)

En el trabajo de Yozgatligil et al. (2013) emplean el *CVRECM* como indicador de comparación de los métodos de imputación de datos, porque la región de estudio presenta diferencias entre ellas considerables. En este sentido, se usará el *CVRECM* como indicador de comparación de los métodos, puesto que ayuda a eliminar las dependencias de escala.

Sea $y_{a_1t}, \dots, y_{a_kt}, \dots, y_{a_mt}$ los valores observados de las observaciones asumidas como faltantes, $y_{miss} = (y_{m_1t}, \dots, y_{m_kt}, \dots, y_{m_mt})$ y $\bar{y}_{.a}$ será el promedio de los valores reales para el período que falta creado artificialmente. De esta forma se tiene que el *CVRECM* se calcula:

$$CVRECM = \frac{\sqrt{\sum_{i=1}^k (y_{a_{kt}} - y_{m_{kt}})^2 / m}}{\bar{y}_{.a}} \quad (6.29)$$

Raíz del Error Cuadrático Medio (RECM)

El indicador *RECM* resulta útil porque está expresado en las mismas unidades que la variable dependiente, por lo cual da una idea del error medio que se comete con el método, además de que es ampliamente usado para evaluar los resultados de predicción en el análisis de series de tiempo, de acuerdo con el principio general de simplicidad científica (Diéguez-Aranda et al., 2009).

$$RECM = \sqrt{\frac{\sum_{i=1}^k (\hat{\theta}_i - \theta_i)^2}{k}} \quad (6.30)$$

$\hat{\theta}_i$ =Estimador del parámetro.

θ =Parámetro o valor observado.

κ = Numero de observaciones.

Evidentemente, cuanto menor sea el valor del estadístico *RECM*, mayor será la exactitud del modelo.

6.2.5. Test de Mann Whitney

Los datos consisten de 2 muestras aleatorias independientes. Sea X que denota la muestra aleatoria de tamaño n de la población 1, y sea Y que denota la muestra aleatoria de tamaño m de la población 2. Se asignan los rangos 1 a $n + m$ a las observaciones de menor a mayor. $R(X_i)$ y $R(Y_j)$ denotan los rangos asignados a X_i y Y_j para todo i para todo j . Sea $N = n + m$, si varios valores de la muestra combinada son iguales (empates), asigne a cada uno el promedio de los rangos.

Supuestos:

1. Ambas muestras son muestras aleatorias de sus respectivas poblaciones.
2. Además de la independencia dentro de cada muestra, las 2 muestras aleatorias son independientes.
3. La escala de medida es al menos ordinal.

Se usa cuando se quiere comparar dos poblaciones usando muestras independientes, es decir; es una prueba alterna a la prueba de t para comparar dos medias usando muestras

independientes. También es conocida como la prueba de suma de rangos de Wilcoxon.

La hipótesis nula es que la mediana de las dos poblaciones son iguales y la hipótesis alterna puede ser que la mediana de la población 1 sea mayor (menor ó distinta) de la mediana de la población 2.

Cuando tanto n sean mayores que 10, se puede demostrar que si no hay empates, entonces $W = \sum_{i=1}^n R(X_i)$ se distribuye aproximadamente como una normal con media $n_1(n_1 + n_2 + 1)/2$ y varianza $n_1 n_2 (n_1 + n_2 + 1)/12$.

$$z = \frac{W - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \sim N(0, 1) \quad (6.31)$$

Cuando hay empates entonces, la varianza es modificada y se obtiene:

$$z = \frac{W - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2}{12} [n_1 + n_2 + 1 \sum_{i=1}^g \frac{t_i^3 - t_i}{(n_1 + n_2)(n_1 + n_2 - 1)}]}} \sim N(0, 1) \quad (6.32)$$

donde, g es el grupo empatado y t_i es el tamaño del i -ésimo grupo empatado, tienen el mismo significado dado anteriormente.

Capítulo 7

Metodología

En este capítulo se propone la metodología para la comparación de métodos de estimación de datos faltantes de series de precipitación diaria en el Valle del Cauca, en diferentes escenarios y usando diferentes metodologías de estimación.

Teniendo en cuenta que el objetivo de este trabajo es evaluar los métodos a emplear para llevar a cabo la estimación de datos faltantes, se hace necesario involucrar diferentes escenarios de altura (valle, montaña) de las estaciones climatológicas que toman las mediciones de la precipitación diaria que se presenta en el departamento del Valle del Cauca, dado que las formas de relieve influyen en las condiciones climatológicas de la zona lo que implica un comportamiento diferente en la variable de análisis (precipitación mm) de acuerdo a la zona en que se mida y proponer diferentes escenarios de datos faltantes que permitan comparar los métodos según estas condiciones, Yozgatligil et al. (2013) propone tres escenarios de datos faltantes artificiales en series temporales 10, 20 y 50 % en el período de expansión de los datos utilizados, los cuales se estiman utilizando los métodos. De la misma forma Radi et al. (2015) estimó datos faltantes de precipitación dividiendo el análisis en tres porcentajes diferentes 5 %, 10 % y 20 % con el fin de representar varios casos de los datos que faltan, por lo que en este estudio se proponen períodos artificiales de datos faltantes del 10, 15 y 20 % para evaluar y comparar los métodos de estimación, dado que los métodos podrían desempeñarse distinto de acuerdo al porcentaje de datos faltantes bajo el cual se trabaje. La metodología para llevar a cabo el estudio se presenta a continuación.

7.1. Datos de estudio

Se cuenta con datos de precipitación (mm) diaria en el Valle del Cauca-Colombia correspondientes al periodo de 1993-2011 con registros de aproximadamente 106 estaciones

meteorológicas, de las cuales 97 cuentan con observaciones para todos los años comprendidos en este período. Las estaciones se encuentran en coordenadas geograficas entre longitudes y latitudes (3.137534326, 4.900536267) y (-77.3575914, -75.85113927) respectivamente y hacen parte de la red nacional de estaciones meteorológicas del IDEAM (Instituto Nacional de Hidrología, Meteorología y Estudios Ambientales) y de la CVC (Corporación Regional de Medio Ambiente para el Valle del Cauca, Colombia).

El comportamiento de la precipitación se encuentra determinado con respecto a la posición geográfica y las características que esto implica (Pabón-Caicedo et al., 2001), por lo cual las estimaciones de los datos y las comparaciones de los métodos de estimación se realizarán por separado para las estaciones meteorológicas ubicadas en la zona de valle y montaña, para lograr identificar si el comportamiento de la precipitación diaria se ve o no influenciada de acuerdo a la altitud de las estaciones. Las estaciones meteorológicas con altitudes hasta 1,239 m.s.n.m se considera que están ubicadas en la zona valle y las estaciones con altitudes mayores a 1,240 m.s.n.m se consideran que están ubicadas en montaña, por lo cual se separarán las 97 estaciones meteorológicas utilizadas para esta investigación.

7.2. Control de calidad datos de estudio

El objetivo del control de calidad es identificar datos incorrectos o registros no verificados de los datos Vicente-Serrano et al. (2010). La base de datos con la cual se va a trabajar contiene la información de precipitación diaria con respecto a la fecha, pero los días en los que no se tienen los registros la base de datos simplemente salta hasta el próximo día con observación, por lo que se utilizó el software estadístico R 3.3.1 (2016-06-21) R Core Team (2012) para organizar la base de datos y encontrar todas los días con datos faltantes.

7.3. Análisis exploratorio

La finalidad del Análisis Exploratorio de Datos (AED) es examinar los datos previamente a la aplicación de cualquier técnica estadística, de esta forma conseguir un entendimiento básico de estos y de las relaciones existentes entre las variables analizadas. Con el AED logramos organizar y preparar los datos, detectar fallos en el diseño y recogida de estos también se puede realizar tratamiento y evaluación de datos ausentes e identificación de casos atípicos Salvador Figueras & Gargallo (2003). Esta resulta ser una fase importante

en toda clase de investigación, puesto que permite conocer y familiarizarse con los datos a utilizar y obtener información preliminar para encontrar adecuadas metodologías estadísticas para aplicar, además de cómo trabajarlas de acuerdo a los resultados obtenidos en esta fase. Con respecto a esto, a continuación, se muestra un resumen del análisis exploratorio de la base de datos de las 97 estaciones que se encuentran comprendidas entre los años 1993 y 2011, ubicadas en el departamento del Valle del Cauca, registros brindados por la CVC.

El año que de forma general incluyendo todas las estaciones posee el mayor porcentaje de datos faltantes es el 2011, con un 18 %, lo que refleja que los años que se incluyen en la base de datos no se encuentran en tan mal estado. La ubicación de las estaciones, puede ser el motivo para que se presente una variación considerable en estas. Por otro lado, el menor porcentaje de datos faltantes para los años, se presenta en 1997 con un 4 %, lo que resulta ser bastante bajo teniendo en cuenta la cantidad de datos con los que se cuenta. En este periodo, aunque tenga poca cantidad de valores faltantes, se observó mucha variación en la precipitación para sus estaciones, esto puede ser porque, aunque sean estaciones de la misma región, se presentan diferentes formas de relieve y climas en cada una de las posiciones donde se encuentran.

Al realizar un análisis sobre el comportamiento de los datos faltantes para los 19 años que se van a trabajar, se encontró que algunos tienen comportamientos estables para estaciones en común, de igual forma que los picos que se presentan oscilan en porcentajes bajos de valores faltantes, lo que nos indica que algunas estaciones están en malas condiciones para algunos años pero que en general la base de datos se encuentra en buen estado y es una base de datos trabajable y reconstruible con los métodos propuestos.

Tabla 7.1: Estadísticas descriptivas estaciones completas

Zona	Etiqueta	Nombre	Municipio	Lon	Lat	Altura	Media	Desviación	Cte. variación
Valle	1	Puerto Molina	Obando	-76.02	4.62	906	5.10	9.94	194.83
	2	Acueducto Buga	Guadalajara de Buga	-76.28	3.89	1025	3.17	7.93	250.36
	3	Bugalagrande	Bugalagrande	-76.14	4.21	997	3.87	9.41	243.52
	4	El Lucero	La Unión	-76.11	4.51	1084	3.55	8.72	245.92
Montaña	5	Bella Vista	Palmira	-76.00	3.63	3656	3.59	7.25	202.28
	6	El Retiro	Roldanillo	-76.23	4.40	1695	4.56	9.92	217.67
	7	Patio Bonito	Toro	-76.13	4.59	1647	3.64	7.71	212.00
	8	Moravito	bolívar	-76.38	4.30	1548	4.13	8.08	195.99

Analizando todas las estaciones, se observa que hay ocho estaciones que poseen todos sus registros para todos los años. La ubicación de estas estaciones y la zona a la cual pertenecen, se encuentra de forma detallada en la **Tabla 7.1** en donde de igual forma se puede observar que la estación con mayor porcentaje de variación es la Acueducto Buga, sin embargo, las

demás también poseen un porcentaje de variación bastante elevado. Además, estas estaciones presentan una precipitación media similar, aunque en la estación Puerto Molina es un poco más elevada que en las demás.

Por otro lado, la estación La Ceja es la que más porcentaje de datos faltantes presenta para un año en particular, con un 89 % para el año 2002, lo que hace que este periodo sea inconstruible por medio de las metodologías. En la estación Yurumangui se presentó la mayor precipitación promedio, siendo esta de 44.85 mm para el año 2009, de igual forma presenta la variación más alta para el periodo 2010. La precipitación más alta se presentó en el año 2004 siendo de 432 mm, registrada por la estación El Bosque. La precipitación media más baja se presentó en el año 2009, registrada por la estación Llorede Grasas y la de menor variación fue la presentada por la estación Los Bancos en el año 2009 igualmente.

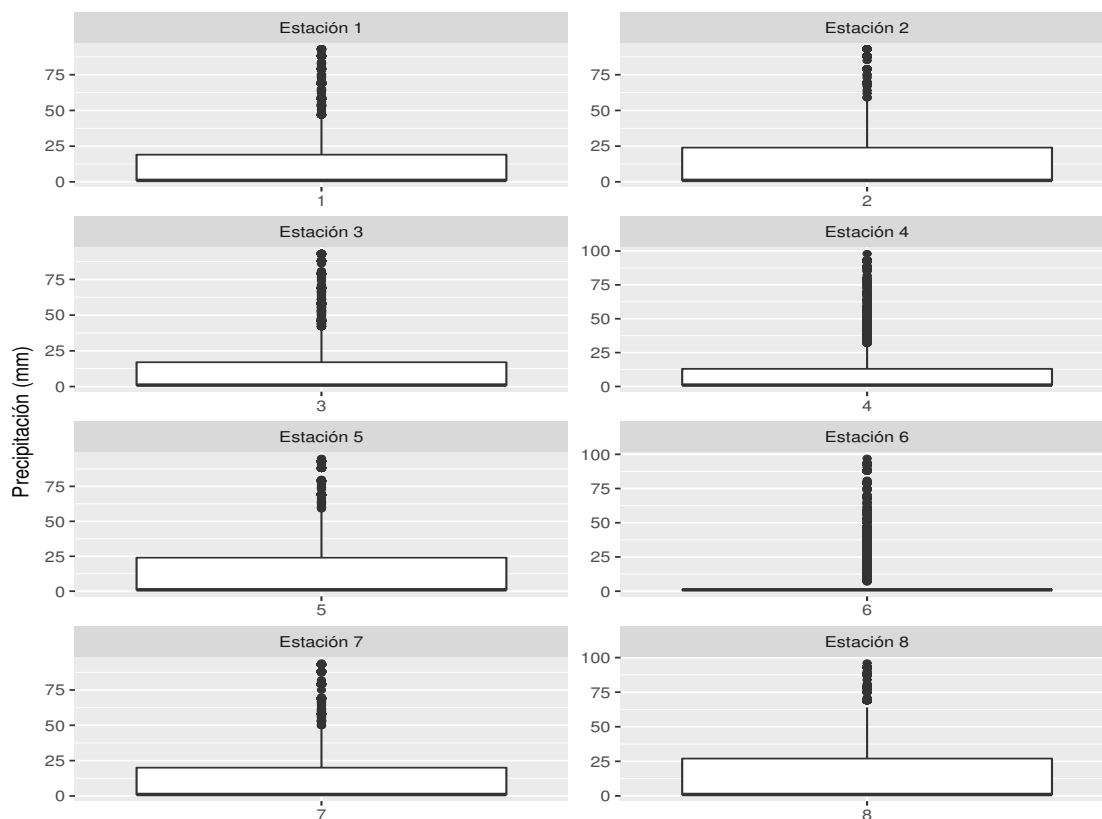


Figura 7.1: Distribución de las estaciones completas.

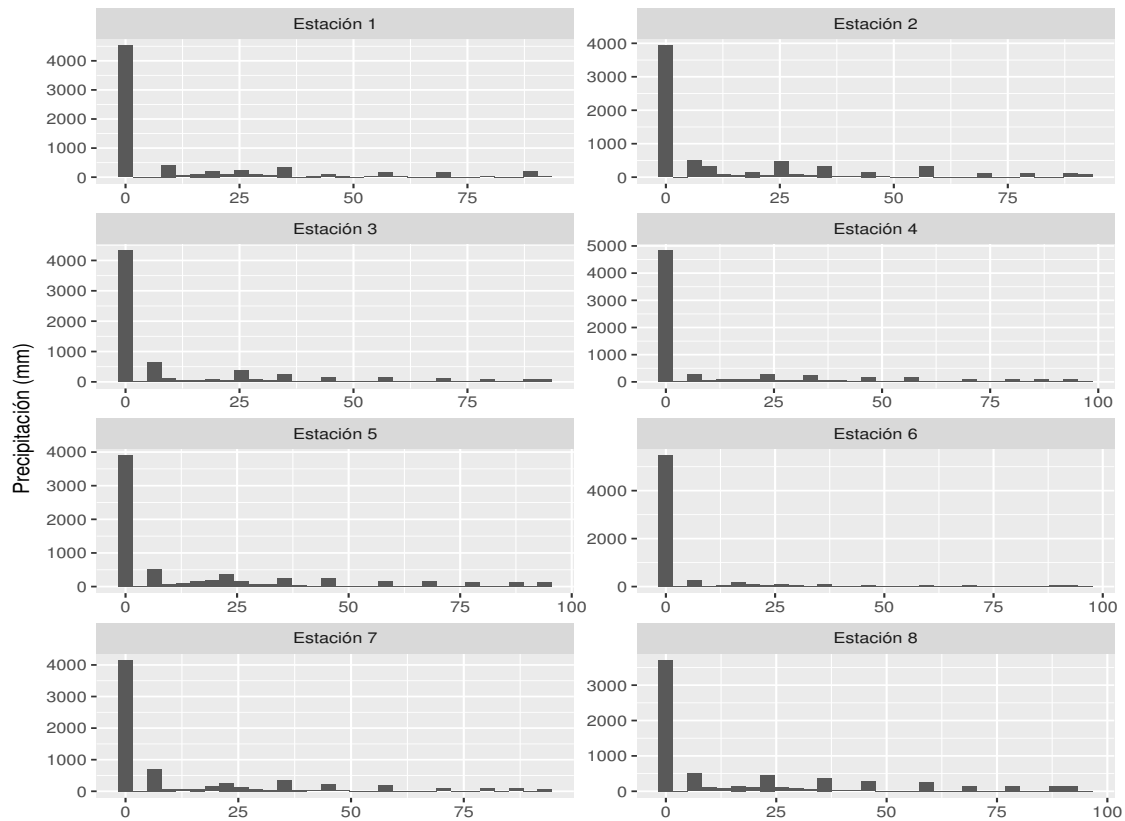


Figura 7.2: Histograma de las estaciones completas.

En general, en las **Figuras 7.1 y 7.2** se observa que las estaciones que se encuentran completas tienen en si una distribución similar en la precipitación (mm), mostrando su respectivo Boxplot e Histograma. Además en la **Figura 7.3** se puede observar la ubicación de las estaciones según la zona valle (verde) y montaña (Amarillo), de igual forma la altura que tienen cada una de estas estaciones.

Revisando la literatura, se observó que no existe un porcentaje de datos faltantes como tal establecido para determinar si una serie se puede reconstruir o no según Trewin et al. (2007), en este sentido, consideramos prudente realizar la estimación para series que como máximo su porcentaje de datos faltantes sea del 25 %. De esta forma, observando en la **Tabla 7.2** para cada año la cantidad de datos faltantes que tienen todas las estaciones y su respectivo porcentaje, buscando el año que tenga menos cantidad y porcentaje de datos faltantes en este caso el año 1997, para realizar la comparación de los métodos de estimación y de igual forma que sea posible reconstruir la base de datos de precipitación diaria periodo 1993-2011 ya que los porcentajes de datos faltantes para cada año resultan ser bajos lo que indica que esta puede reconstruirse.

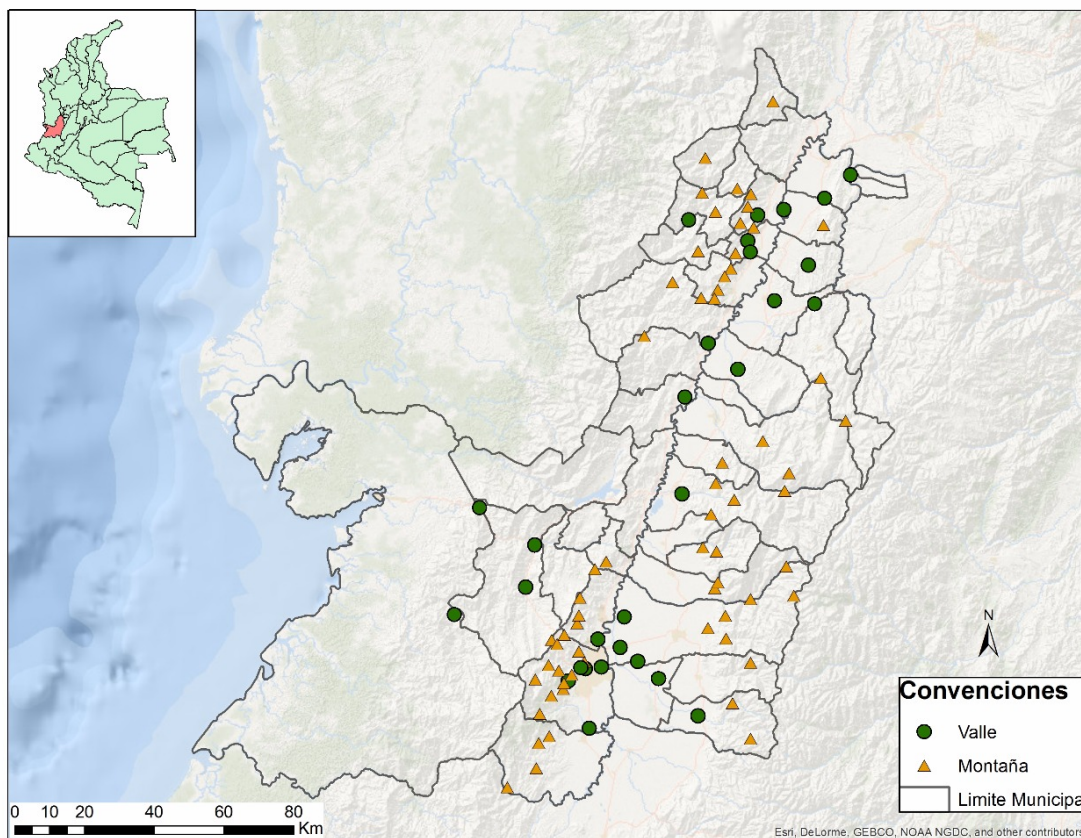


Figura 7.3: Ubicación de estaciones.

De las 97 estaciones, 31 se encuentran ubicadas en la zona de valle y comprenden alturas entre los 26 y 1,097 msnm. En general, las estaciones que se encuentran en la zona de valle no tienen más del 25 % en datos faltantes. Sin embargo, la estación con mayor porcentaje de datos faltantes es la estación Yurumangui, ubicada en el municipio de Buenaventura a una longitud de -77.3575914, latitud de 3.373195479 y con altura de 26 msnm. De esta forma, ésta se descartará del análisis, ya que se encuentra bastante retirada de las demás estaciones y presenta un comportamiento totalmente diferente debido a su cercanía con la costa pacífica, dejando un total de 30 estaciones para la zona de valle. En general, la cantidad de periodos que tienen más de 25 % de datos faltantes en la zona de valle no es considerablemente alta, aproximadamente 14 del total, los cuales poseen un porcentaje elevado de datos faltantes, eso por encima del 25 %, lo que deja ver que en general esta zona se puede reconstruir.

Tabla 7.2: Frecuencias absolutas y relativas de datos faltantes para todas las estaciones de cada año en cada zona.

Año	Valle		Montaña		Total	
	n	%	n	%	n	%
1993	52	0.5	320	1.3	372	1.1
1994	162	1.4	375	1.6	537	1.5
1995	240	2.1	222	0.9	462	1.3
1996	57	0.5	260	1.1	317	0.9
1997	178	1.6	83	0.3	261	0.7
1998	257	2.3	83	0.3	340	1.0
1999	128	1.1	416	1.7	544	1.5
2000	172	1.5	280	1.2	452	1.3
2001	64	0.6	320	1.3	384	1.1
2002	273	2.4	810	3.4	1083	3.1
2003	385	3.4	235	1.0	620	1.8
2004	209	1.8	218	0.9	427	1.2
2005	420	3.7	375	1.6	795	2.2
2006	414	3.7	346	1.4	760	2.1
2007	253	2.2	587	2.4	840	2.4
2008	287	2.5	392	1.6	679	1.9
2009	215	1.9	412	1.7	627	1.8
2010	396	3.5	213	0.9	609	1.7
2011	727	6.4	526	2.2	1253	3.5

Por otro lado, 66 estaciones están ubicadas en la zona de montaña, con alturas comprendidas entre los 1,106 y 3,699 msnm. Siendo así, contamos con un total de 96 estaciones que contienen registros para todos los años de estudio, la estación que mayor porcentaje de datos faltantes presenta dentro de sus registros es El Paraíso, tanto para la zona Valle como para la zona de Montaña y como se ha mencionado anteriormente, no es considerable la proporción de estaciones que poseen mas del 25 % de datos faltantes.

7.3.1. Análisis exploratorio año 1997

Se realizará un análisis exploratorio para el año 1997, ya que es el año con menor cantidad de datos faltantes ver (**Tabla 7.2**), por lo que se tomó como referente para realizar la comparación de los métodos de estimación por medio de los indicadores planteados.

Este se realizó para buscar si existían patrones que permitieran agrupar de alguna forma la información disponible, ya que específicamente para el método de Kriging Ordinario encontrar un modelo que explique la variación diaria de la precipitación es posible pero no aseguraría una estimación optima debido a la alta variabilidad que tiene esta información.

Al analizar las **Figuras 7.6,7.5,7.4** se evidencia una estructura climatológica Bimodal con dos temporadas de alta pluviosidad en las tres figuras para las mismas épocas del año, por lo cual para el desarrollo de la metodología para la que se necesita modelar la variación de la variable de estudio (precipitación) se usará la suma quincenal de la precipitación de las estaciones según la zona como se muestra en la **Figura 7.5**, así se lograra reducir la variación de este y de la misma forma la variación de las estimaciones por el método de KO.

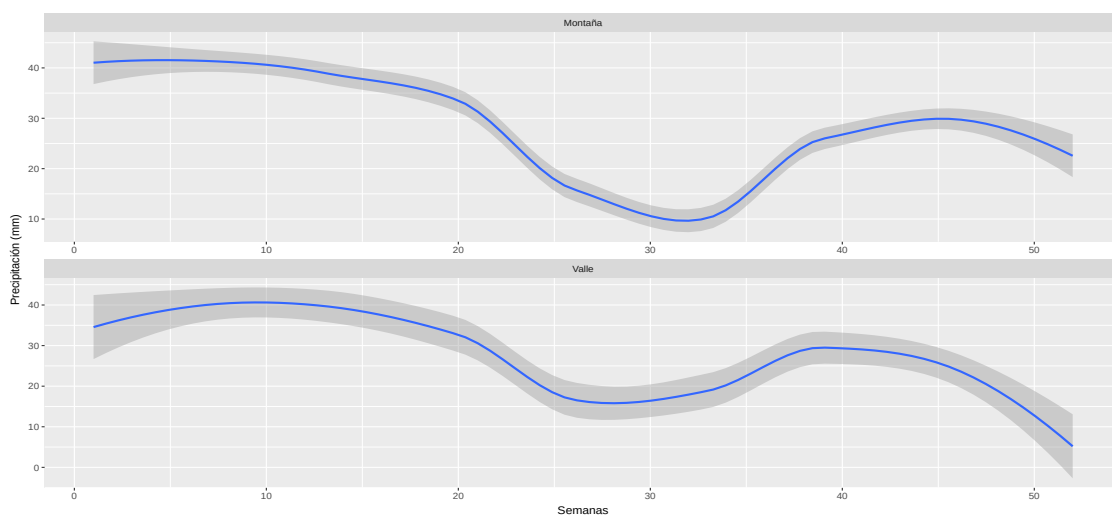


Figura 7.4: Comportamiento estaciones para cada semana del año 1997 por zonas.

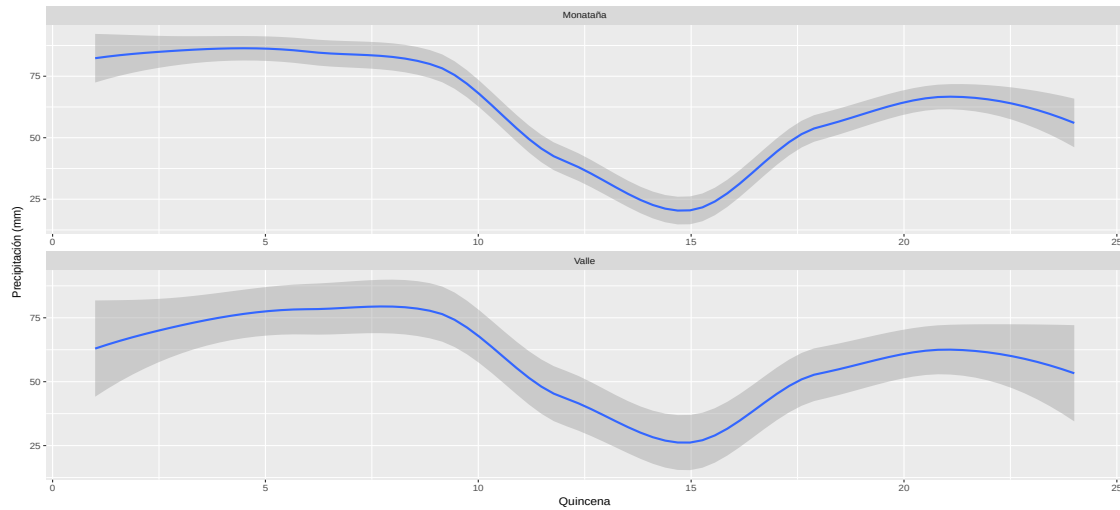


Figura 7.5: Comportamiento estaciones para cada quincena del año 1997 por zonas.

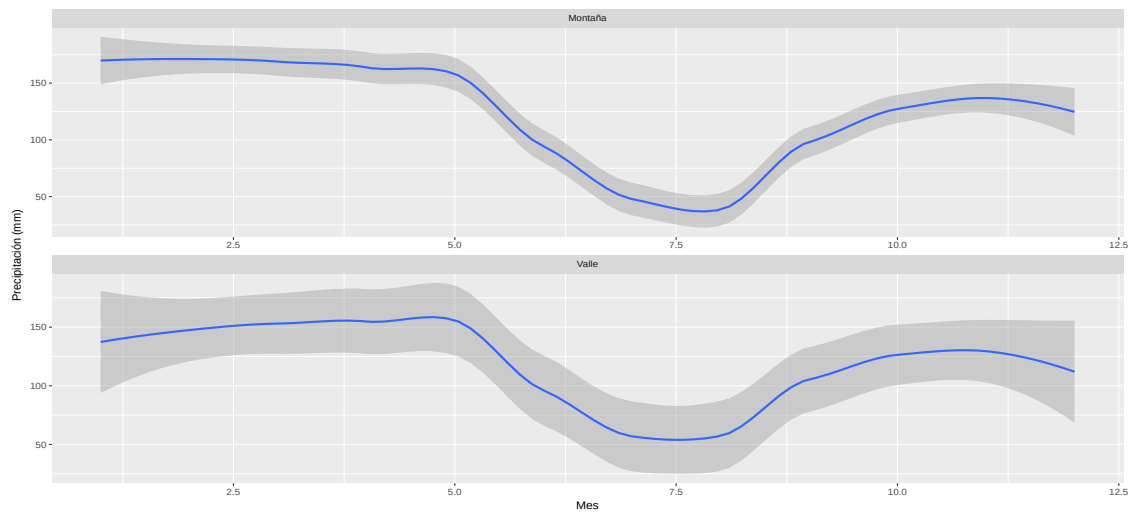


Figura 7.6: Comportamiento estaciones para cada mes del año 1997 por zonas.

7.4. Procedimiento

Teniendo en cuenta que se realizará la comparación de métodos por medio de validación cruzada y se necesita tener los datos completos, se tendrán en cuenta solo las estaciones que para este año posean todos sus registros. Siendo así, de las 30 estaciones de la zona de valle se excluyen del análisis cinco estaciones quedando con un total de 25 para esta zona. Por otro lado, de las 66 estaciones de la zona de montaña se excluyen tres estaciones, quedando con un total de 63 para esta. Por lo tanto, se tiene un total de 88 estaciones con las que se

trabajaré para realizar la comparación.

Posteriormente, se procederá a separar la base de datos en dos partes, en donde una solo cuenta con estaciones de la zona de valle (Ver **Anexo 10.1**) y la otra de montaña (Ver **Anexo 10.2**). Para cada zona se tienen las 365 observaciones de cada estación con su respectiva longitud y latitud. Luego, se realizara la extracción de datos, para crear períodos faltantes artificiales en la base de datos para su posterior estimación con los métodos a comparar, esto de forma aleatoria para cada estación.

La extracción de los datos se hará por cada una de las estaciones, esto teniendo en cuenta cada porcentaje mencionado anteriormente (10, 15 y 20 %). De esta forma, para cada uno de los escenarios de datos faltantes en las estaciones por cada zona se procede a hacer la estimación de los registros según las otras estaciones en el mismo día, aplicando así cada uno de los métodos planteados para realizar la comparación.

Cada uno de los métodos realiza la estimación de los datos de forma diferente, y es por ello que se comparará para poder tener una idea de cuál de ellos dentro de su forma de ejecutarse, realiza una mejor estimación teniendo en cuenta los escenarios mencionados. Por consiguiente, para cada método se obtendrían tres bases de datos con las estimaciones según la zona y el porcentaje de datos faltantes creados, con las cuales se calcularán los indicadores para poder comparar los métodos teniendo en cuenta las dos bases de datos iniciales con datos observados de valle y montaña. La comparación de los métodos se realizará por cada escenario de datos faltantes teniendo en cuenta las zonas geográficas de estudio. Se emplearán tres indicadores para la comparación en los escenarios mencionados, estos son:

- Error Cuadrático Medio (ECM).
- Raíz del Error Cuadrático Medio (RECM).
- Coeficiente de Variación Raíz del Error Cuadrático Medio (CVRECM).

El ECM se emplea para conocer el estimador, en este caso el método, que menor error de estimación comete en promedio. Por otra parte, la RECM es empleada como medida de exactitud de uso común para evaluar los resultados de predicción en series de tiempo y el CVRECM es empleado para eliminar las dependencias de escala en tal caso de que estas existan.

Luego de obtener los indicadores en los diferentes escenarios, se compararán los métodos escogiendo el valor mínimo de cada indicador por cada método. De esta forma, se hará una amplia evaluación de los métodos en donde por medio de tres diferentes indicadores se

escogerá el mejor método de estimación de datos faltantes de acuerdo a las condiciones bajo las cuales se encuentra (porcentaje de datos faltantes y zona geográfica).

7.4.1. K-Vecinos más cercanos (KV)

En este método se tomarán las estaciones más cercanas a la estación con el punto a estimar o dato faltante a estimar, específicamente las k más cercanas que posean información, en donde este k se escoge por medio de validación cruzada. En este caso se trabajará con un $k = 5$ para ejecutar el método en cada una de las estaciones donde se creen datos faltantes, ya que es el número de estaciones cercanas que generó un menor ECM en los escenarios planteados, tal como se muestra en la **Tabla 7.3**, en donde se observa el comportamiento del ECM variando el parámetro k (Cantidad de estaciones) y se tiene que este se estabiliza en $k = 5$ para los distintos porcentajes de datos faltantes.

Posterior a esto, se calculará el promedio de los valores correspondientes a estas estaciones mas cercanas según las longitudes y las latitudes de la estación de referencia y las demás estaciones que cuentan con información de la variable en ese mismo día y dicho promedio sería la estimación correspondiente de la variable en ese punto para esa estación. La estimación se realizará diaria, puesto que el método trabaja por día, pero teniendo por referencia la estación (Ver **Figura 7.7**).

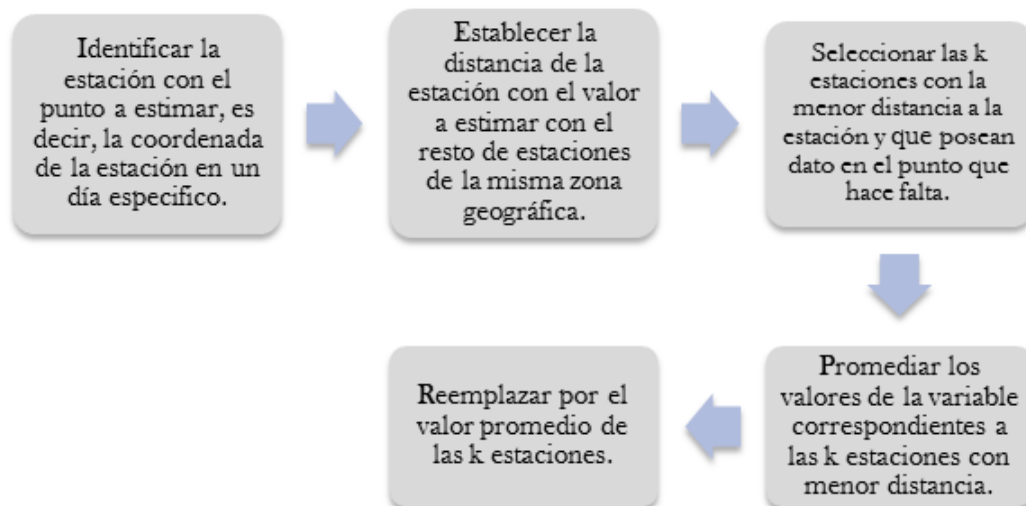


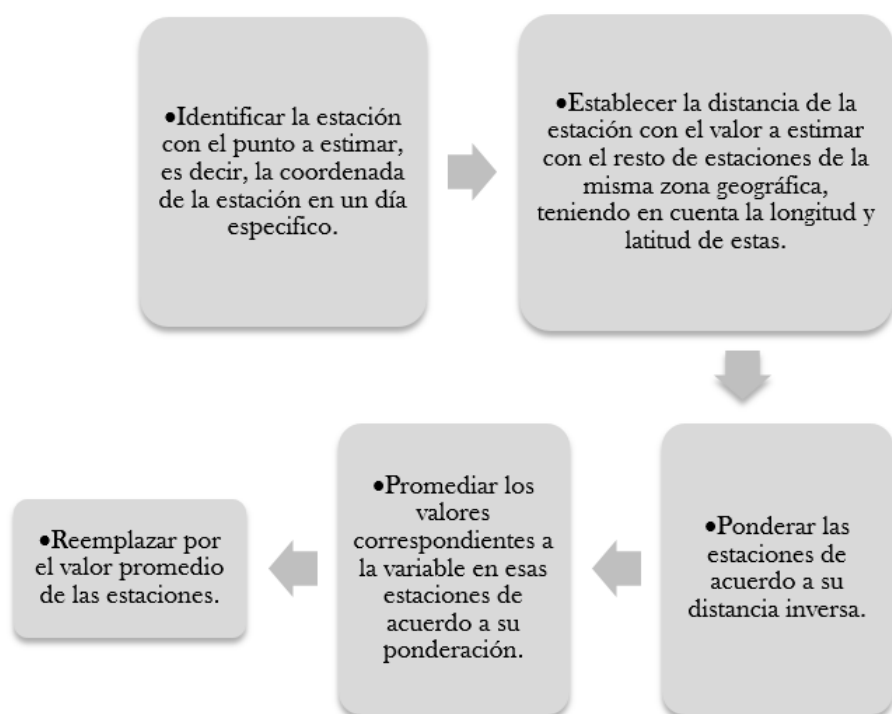
Figura 7.7: Resumen metodología K-Vecinos mas cercanos.

Tabla 7.3: Comportamiento ECM según parámetro κ .

Montaña				Valle			
κ	10 %	15 %	20 %	κ	10 %	15 %	20 %
1	79.92	88.18	81.96	1	107.59	145.13	100.31
5	52.10	57.76	58.96	5	81.73	118.63	87.12
10	53.69	58.76	59.50	10	84.96	119.71	87.06
15	56.55	59.95	63.26	15	86.09	125.29	89.17
20	59.37	61.67	66.5	20	84.00	124.78	89.04
25	59.42	62.25	67.82	25	84.48	124.75	89.54
30	59.84	62.95	69.18				
35	59.92	63.67	70.29				

7.4.2. Distancia inversa ponderada (IDW)

En este método se tomarán todas las estaciones que posean registro o los valores para la variable en cuestión, que se encuentren alrededor de la estación en la cual se encuentra el punto a estimar según las longitudes y las latitudes de las estaciones.

**Figura 7.8:** Resumen metodología Distancia Inversa Ponderada.

Luego de tener estas estaciones, se realizará un promedio ponderado por la distancia inversa de las estaciones con registro y la estación con el dato a estimar, dándole de esta forma mayor peso en la estimación a los valores de las estaciones que se encuentren más cercanas a la estación del dato faltante. De esta forma, el resultado de ese promedio ponderado sería la estimación del valor faltante de dicha estación. La estimación se realizará diaria, al igual que el anterior método por la misma razón.

7.4.3. Regresión Lineal (RL):

En este método hay una estación que posee al menos un dato para estimarse y se toma una estación de la misma área topo climática, es decir, cercana a esta y que se encuentre bajo las mismas condiciones de relieve y clima, con la cual se realizará un modelo de regresión para estimar los valores faltantes de la primera. Esta es una de razones por la cual las estaciones de la base de datos se han separado considerando su altura y relieve, ya que es un factor influyente en la estimación de los datos.

En este caso, como se cuenta con muchas estaciones, para determinar cuál es la estación que podría arrojar un mejor modelo para realizar las estimaciones de los datos faltantes, se calculará el coeficiente de correlación con cada una de las estaciones que se encuentran en la misma zona geográfica de la estación con valores a estimar. Luego, se realizaría el modelo con la estación que arroje el coeficiente de correlación más alto para hacer la estimación.

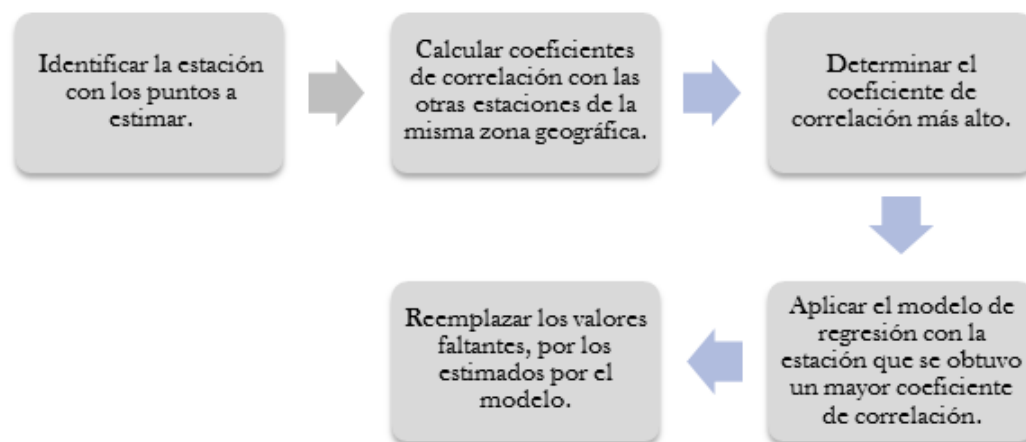


Figura 7.9: Resumen metodología Regresión Lineal.

De esta forma se tienen 88 coeficientes de correlación que involucran las otras estaciones, se escoge la estación con la cual se realizará el modelo para estimar los valores faltantes, teniendo en cuenta la que tenga un mayor coeficiente de correlación con la estación a la que se le va a estimar el dato. Cabe resaltar que es necesario que la estación que se va a tomar como referencia para realizar la estimación debe tener información en el punto donde la estación que posee el dato faltante no tiene.

7.4.4. Kriging Ordinario (KO):

Luego de haber realizado el análisis exploratorio de la base de datos, se observó que la variable objeto de estudio posee poca variabilidad en rangos de quincena para el periodo escogido, de esta forma para efectos prácticos se selecciona esta temporalidad para poder aplicar el método y realizar la respectiva comparación con los otros.

Se realizará el análisis descriptivo para encontrar la distancia máxima con la que se trabajarán los semivariogramas empíricos que ayudaran a escoger cuáles serán los modelos para cada quincena, se determinará el número de bins por medio de la regla de Sturges para realizar el semivariograma empírico según la suma de la lluvia para cada quincena del año 1997 por zona geográfica, con esto se observara hasta qué punto existe una correlación espacial según los bins y se obtendrán los valores iniciales para la estimación de los parámetros del modelo seleccionado (meseta, pepita y el rango).

Se considerarán dentro de este método cuatro modelos de semivarianza (matern, exponencial, gaussiano y esférico) para escoger el que mejor modela la precipitación en la correspondiente quincena, según la menor suma de cuadrados como criterio de decisión. Posteriormente se hará la estimación de los parámetros para cada modelo por medio del método de Mínimos Cuadrados Ordinarios. Seguidamente se realizará la estimación por medio del método Kriging Ordinario teniendo claro el modelo que se usará para cada quincena en cada zona y sus respectivos parámetros.

En este sentido, se calcula la matriz de distancias y ya conociendo la estructura del modelo y los parámetros para cada día se procederá a construir la matriz Γ y el vector γ usando el semivariograma escogido con los parámetros estimados para cada quincena. Luego, se calcularía el vector W , que serían los pesos w_α usados en la estimación del valor faltante al implementarse la expresión **6.13** mencionada en la **Sección 6.2.3**. De esta forma, se obtendrían las estimaciones para cada uno de los valores faltantes generados de forma aleatoria para el año 1997 por medio de este método.

Al obtener los pesos para la estimación por medio del método kriging Ordinario, resulta

necesario realizar una corrección de estos, debido a que esta es una técnica de estimación no convexa. Los pesos negativos del método aplicados a valores extremos pueden conducir a cálculos fuera del rango de los datos observados. Esta característica puede ser deseable cuando se trabaja con variables suaves tales como elevaciones o temperaturas Deutsch (1996). En otras situaciones el estimador no convexo causa problemas como estimaciones no físicas (probabilidades negativas, probabilidades mayores que uno, porosidades negativas o espesores negativos) en nuestro caso estimaciones negativas de precipitación diaria. La corrección equivale a restablecer los pesos negativos a cero y normalizar los pesos restantes a sumar a uno (normalizar a (1,0)) Deutsch (1996). Para finalmente, proceder a realizar la estimación de la precipitación en las estaciones donde se crearon artificialmente datos faltantes.

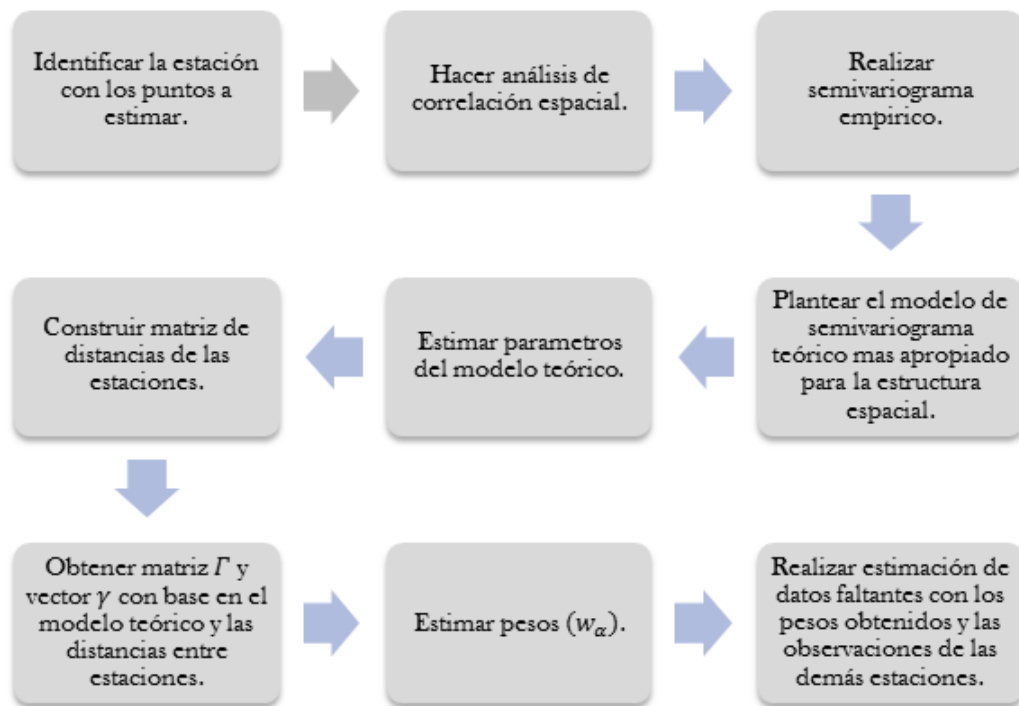


Figura 7.10: Resumen metodología Kriging Ordinario.

Luego de realizar el cálculo de los indicadores y las respectivas comparaciones por porcentaje de datos faltantes existentes y zona geográfica, se procederá a probar si verdaderamente estas diferencias obtenidas son significativas ó por el contrario, no lo son, de esta forma obtener un resultado bajo un criterio estadístico. En el caso de que alguna diferencia de significativa, se procederá a escoger el método que está por encima de los demás en cuanto a desempeño, por otro lado, observar si no es significativo el desempeño de los indicadores bajo las condiciones que se realizó la comparación de los métodos. Se mostrarán los valores-p de las pruebas realizadas entre cada uno de los escenarios, esto bajo un nivel de significancia de 0.05.

Capítulo 8

Resultados

En este capítulo se presentan los resultados correspondientes a la estimación de datos faltantes en datos climatológicos diarios de precipitación en el Valle del Cauca, para el año 1997 con un total de 88 estaciones climatológicas divididas en las zonas de valle (25) y montaña (63) obtenidos por medio de los métodos de estimación K-Vecinos Cercanos, Regresión Lineal, Distancia Inversa Ponderada y Kriging Ordinario, teniendo en cuenta la distribución espacial de los datos de estudio, cada uno de los métodos empleados será evaluado a través de medidas de desempeño como: el Error Cuadrático Medio, La Raíz del Error Cuadrático Medio y el Coeficiente de Variación de la Raíz del Error Cuadrático Medio. Por último, se comparan los métodos utilizados para la estimación de estos datos faltantes. El análisis se realiza por medio del software estadístico R 3.3.1 (2016-06-21) y Excel (2016).

8.1. K-Vecinos Cercanos (KV)

En la **Tabla 8.1** se presenta el resumen descriptivo para los indicadores del método de K-Vecinos Cercanos. Para la zona valle en el indicador de la RECM se obtuvo el menor promedio 8.14 mm para el 10 % de datos faltantes con una desviación estándar de 3.75 mm, con respecto a la mediana el mejor comportamiento se encontró para el 20 % de datos faltantes de 6.70 mm con una desviación estándar de 4.01 mm. Para la zona de montaña el mejor comportamiento con respecto al indicador RECM tanto para el promedio como para la mediana lo obtuvo el 10 % de datos faltantes con 6.75 mm y 6 mm respectivamente con una desviación estándar de 3.12 mm.

A pesar de esto, se puede observar que para la zona de valle, los indicadores no presentan diferencias relevantes en cuanto al cambio en el porcentaje de datos faltantes, sin embargo, se puede ver que hay un incremento en el valor del indicador al momento de aumentar el porcentaje de datos faltantes, a excepción del CVRECM, que disminuye al aumentarse el porcentaje a 15 % y vuelve a disminuir en 20 %. En cuanto al comportamiento de la mediana

de los indicadores, se observa que el ECM y RECM disminuye en la medida que se aumenta el porcentaje de datos faltantes, con el CVRECM sucede lo mismo que con la media. No obstante, se presenta un gran porcentaje de variación en los indicadores, en mayor medida en el ECM, en la RECM y CVRECM el porcentaje de variación es menor, pero igualmente de gran magnitud.

Tabla 8.1: Resumen descriptivo para los indicadores del método de K-Vecinos.

Zona	Porcentaje	Indicador	Media	Mediana	Desviación	C. Variación
Valle	10 %	ECM	79.73	50.70	76.43	96 %
		RECM	8.14	7.12	3.75	46 %
		CVRECM	4.49	4.29	1.92	43 %
	15 %	ECM	83.31	50.38	90.56	109 %
		RECM	8.18	7.10	4.14	51 %
		CVRECM	4.41	4.11	1.86	42 %
	20 %	ECM	84.38	44.88	83.09	98 %
		RECM	8.30	6.70	4.01	48 %
		CVRECM	4.52	4.14	1.88	42 %
Montaña	10 %	ECM	55.18	36.00	62.54	113 %
		RECM	6.75	6.00	3.12	46 %
		CVRECM	3.38	3.38	0.91	27 %
	15 %	ECM	54.92	39.14	56.55	103 %
		RECM	6.81	6.26	2.93	43 %
		CVRECM	3.40	3.43	0.85	25 %
	20 %	ECM	58.64	41.19	54.66	93 %
		RECM	7.08	6.42	2.94	42 %
		CVRECM	3.43	3.49	0.82	24 %

Para la zona de montaña, se presenta un panorama similar al que sucede con los indicadores en la zona de valle, sin embargo, se resalta que la media y la mediana en los tres indicadores tomados bajos los diferentes porcentajes de datos faltantes en esta zona son más bajos que los de la zona de valle, pero en cuanto al porcentaje de variación sigue teniendo aproximadamente el mismo comportamiento. Se podría decir, que la razón por la cual la media y la mediana de los indicadores resultan ser más bajos en esta zona que en la de valle, es por el hecho de que hay más estaciones, lo que podría ayudar a una mejor estimación de los datos faltantes.

Se observó para la zona valle en el indicador ECM el 10 % de datos faltantes fue el porcentaje que obtuvo mayores valores por lo cual el método no se comportó bien para este según este indicador y por el contrario las observaciones obtenidas con un 20 % de datos faltantes arrojaron los menores valores para este indicador indicando así que este método se comporta bien bajo el 20 % de datos faltantes siempre y cuando las estaciones se encuentren en la zona valle. En el comportamiento del indicador RECM en la zona valle este presentó

un comportamiento similar al ECM para esta misma zona, para el CVRECM se observó que los valores obtenidos para este indicador por el 10 % de datos faltantes obtuvo el mejor comportamiento arrojando los menores valores para este en la zona valle.

El comportamiento del indicador ECM en la zona montaña mostró que el 10 % de datos faltantes fue el que obtuvo mayores valores lo que indica que este método no se comporta bien para este porcentaje en la zona montaña, contrario a lo que ocurrió con el 15 % de datos faltantes que resulta ser el que obtuvo el mejor comportamiento para este método según este indicador en la zona montaña. El comportamiento del indicador RECM resulta muy similar al encontrado en el ECM para esta zona. Para el CVRECM se observa que el 15 % obtuvo los mejores resultados, por lo que se podría concluir que el 15 % de datos faltantes fue el mejor % de datos faltantes para la zona montaña según el método K-Vecinos mas cercanos.

8.2. Regresión Lineal (RL)

En este método, se seleccionó para cada estación a estimar una estación de referencia con la cual se va a estimar, la que haya tenido un mayor coeficiente de correlación (γ) con la estación a estimar según la información que si contenía esta, para cada zona se obtienen los resultados, mostrando que en la zona Valle se obtuvieron coeficientes de correlación mas bajos que en la zona Montaña, de igual forma (Ver anexos **10.3** y **10.4**) se muestran los detalles de los modelos de regresión aplicados para la estimación de datos faltantes en cada estación climática y el respectivo coeficiente de correlación obtenido por el modelo de regresión. Cabe resaltar que los supuestos de estos modelos de regresión en su mayoría no se cumplieron y que los coeficientes de correlación en general no fueron buenos lo que indica por ambos lados que los resultados de esta estimación no son confiable o eficientes.

Del mismo modo, en la **Tabla 8.2** se presenta el resumen descriptivo para los indicadores del método de Regresión Lineal. En la zona valle se encontró el mejor comportamiento de la RECM para el 10 % de datos faltantes con un promedio de 8.58 mm y una desviación estándar del 5.64 mm, con respecto a la mediana el mejor comportamiento lo obtuvo el 15 % de datos faltantes con 6.77 mm y una desviación de 5.30 mm; La zona de montaña obtuvo el mejor resultado para el 10 % de datos faltantes tanto para el promedio como para la mediana con 7.45 mm y 6.83 mm respectivamente y una desviación estándar de 3.4 mm de la RECM.

En la **Tabla 8.2** se puede observar que para la zona de valle, los indicadores no presentan diferencias relevantes en cuanto al cambio en el porcentaje de datos faltantes, pero se puede ver que para este método la media y mediana de los indicadores en esta zona son un poco mas altos que en el de K-Vecinos. No obstante, también se presenta un gran porcentaje

de variación en los indicadores, en mayor medida en el ECM, en la RECM y CVRECM el porcentaje de variación es menor, pero igualmente de gran magnitud.

Tabla 8.2: Resumen descriptivo para los indicadores del método de Regresión Lineal.

Zona	Porcentaje	Indicador	Media	Mediana	Desviación	C. Variación
Valle	10 %	ECM	104.24	49.06	162.07	155 %
		RECM	8.58	7.00	5.64	66 %
		CVRECM	4.41	4.36	1.70	38 %
	15 %	ECM	101.28	45.86	147.73	146 %
		RECM	8.62	6.77	5.30	61 %
		CVRECM	4.42	4.06	1.50	34 %
	20 %	ECM	103.65	46.84	143.62	139 %
		RECM	8.74	6.84	5.33	61 %
		CVRECM	4.50	4.28	1.54	34 %
Montaña	10 %	ECM	66.89	46.58	61.27	92 %
		RECM	7.45	6.83	3.40	46 %
		CVRECM	3.75	3.77	1.44	38 %
	15 %	ECM	66.03	53.24	52.06	79 %
		RECM	7.57	7.30	2.98	39 %
		CVRECM	3.78	3.89	0.89	23 %
	20 %	ECM	65.87	54.80	51.44	78 %
		RECM	7.57	7.40	2.95	39 %
		CVRECM	3.66	3.69	0.83	23 %

En el análisis para la zona de montaña, se presenta un panorama distinto al que sucede con los indicadores en la zona de valle ya que la media y la mediana de estos resultan ser mayores con respecto al método K-vecinos, sin embargo, la media y la mediana en los tres indicadores tomados bajos los diferentes porcentajes de datos faltantes en esta zona son más bajos que los de la zona de valle, pero en cuanto al porcentaje de variación sigue teniendo aproximadamente el mismo comportamiento.

Para la zona valle el indicador ECM tuvo el mejor comportamiento con el 15 % de datos faltantes con los valores mas bajos, el comportamiento del indicador RECM para el 10 % de datos faltantes presentó los valores más bajos y por lo tanto tuvo el mejor comportamiento en la estimación; Lo mismo ocurrió con el comportamiento del CVRECM donde el 10 % tuvo el mejor comportamiento.

En la zona montaña los indicadores ECM y CVRECM presentaron el mismo comportamiento con respecto a que el 20 % de datos faltantes fue el que presentó el mejor comportamiento obteniendo los resultados mas bajos para estos dos indicadores, por otro lado el indicador RECM cuyo comportamiento evidencio que el mejor porcentaje de datos faltantes fue el 10 %.

8.3. Distancia Inversa Ponderada (IDW)

En la **Tabla 8.3** se presenta el resumen descriptivo para los indicadores del método de la Distancia Inversa Ponderada. Para la zona valle se encontró que el menor promedio de la RECM se obtuvo con un 10 % de datos faltantes 7.97 mm con una desviación estándar de 3.8 mm, para la mediana el mejor comportamiento se obtuvo para el 20 % 6.21 mm con una desviación estándar de 3.91 mm; En la zona de montaña el mejor comportamiento de las estimaciones se encontró para 10 % de datos faltantes con un promedio de la RECM de 6.62 mm y una desviación estándar de 3.15 mm, la mediana se comportó mejor para el 10 % de datos faltantes con 5.54 mm y desviación estándar 3.15 mm.

Del mismo modo en la **Tabla 8.3** se puede observar que para la zona de valle, los indicadores no presentan diferencias relevantes en cuanto al cambio en el porcentaje de datos faltantes, pero se puede ver que para este método la media y mediana de los indicadores en esta zona son más bajos con respecto a los métodos de K-vecinos y Regresión Lineal. De igual forma, se presenta un gran porcentaje de variación en los indicadores, en mayor medida para el ECM, en la RECM y CVRECM el porcentaje de variación es menor, pero igualmente de gran magnitud.

Tabla 8.3: Resumen descriptivo para los indicadores del método IDW.

Zona	Porcentaje	Indicador	Media	Mediana	Desviación	C. Variación
Valle	10 %	ECM	77.33	49.74	76.99	100 %
		RECM	7.97	7.05	3.80	48 %
		CVRECM	4.34	4.30	1.61	37 %
	15 %	ECM	80.96	48.08	95.24	118 %
		RECM	8.00	6.93	4.20	53 %
		CVRECM	4.27	4.18	1.63	38 %
	20 %	ECM	81.14	38.55	82.64	102 %
		RECM	8.15	6.21	3.91	48 %
		CVRECM	4.40	4.00	1.60	36 %
Montaña	10 %	ECM	53.67	30.66	58.04	108 %
		RECM	6.62	5.54	3.15	48 %
		CVECM	3.31	3.31	0.98	30 %
	15 %	ECM	53.98	42.14	52.21	97 %
		RECM	6.75	6.49	2.92	43 %
		CVRECM	3.36	3.36	0.88	26 %
	20 %	ECM	56.85	44.71	51.03	90 %
		RECM	6.96	6.69	2.93	42 %
		CVRECM	3.37	3.44	0.85	25 %

En la zona de montaña, se presenta un panorama distinto a lo que sucedió en el método de Regresión Lineal, ya que el método IDW obtuvo indicadores menores con respecto a la media y la mediana tanto a K-vecinos como a Regresión Lineal, de la misma forma la media y la mediana en los tres indicadores tomados bajos los diferentes porcentajes de datos faltantes en esta zona son más bajos que los de la zona de valle, pero en cuanto al porcentaje de variación sigue teniendo aproximadamente el mismo comportamiento.

Para la zona valle se observó que el porcentaje de datos faltantes que tuvo un mejor comportamiento para el indicador ECM fue el 20 %, además, los resultados obtenidos para el indicador RECM mostraron que el 10 % fue el que tuvo el mejor comportamiento, finalmente se encontró que el 15 % de datos faltantes fue el que mejor se comportó con respecto al indicador CVRECM. Para la zona montaña los resultados fueron iguales con los tres indicadores, en donde el 10 % de datos faltantes es el que tuvo el mejor comportamiento.

8.4. Kriging Ordinario (KO)

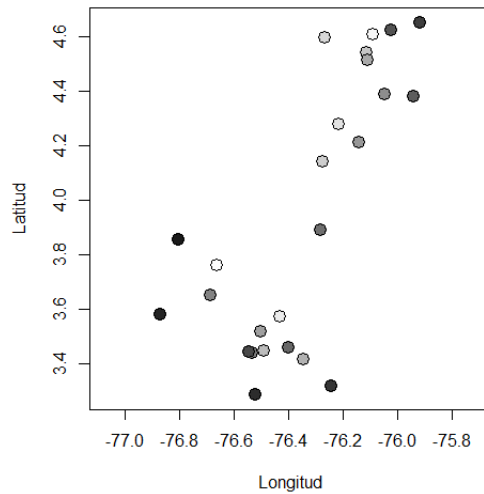


Figura 8.1: Localización de datos de precipitación total quincenal, zona Valle año 1997.

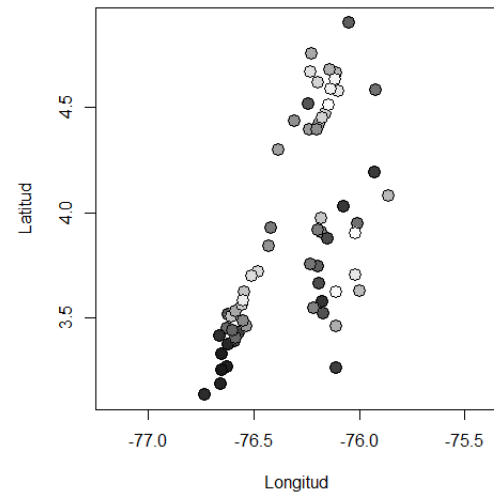


Figura 8.2: Localización de datos de precipitación total quincenal, zona Montaña año 1997.

En las **Figuras 8.1** y **8.2** se observa la localización de las estaciones para cada una de las zonas de estudio, valle y montaña respectivamente, en donde a escala de grises se muestra la intensidad de la medición total de precipitación para el año 1997 en dicho punto, es decir, que a mayor cantidad en milímetros se haya registrado, más oscuro será el punto que represente dicha estación. De esta forma, se puede ver que en la zona de montaña se presenta mayor precipitación en comparación con la zona de valle, y en mayor medida en la parte sur de esta zona geográfica. De igual forma, también se observa la distancia que hay entre las estaciones en ambas zonas, donde las de montaña se encuentran mas próximas unas de otras, por ser mas, mientras que en la zona de valle se encuentran un poco mas distantes, al ser pocas las estaciones que se encuentran en esta zona.

En el (Anexo **10.5**) se muestran las sumas de cuadrados para cada uno de los modelos considerados a seleccionar para realizar la estimación de la precipitación. En esta se resalta el menor valor por zona geográfica para cada una de las quincenas, el cual indica el modelo mas conveniente para realizar la estimación en esa zona geográfica para esa quincena.

Se puede observar que estas son muy elevadas para ambas zonas geográficas en todas las quincenas del año 1997, lo que me indica que de entrada los modelos de semivarianza propuestos para realizar la estimación de los valores faltantes creados artificialmente no se ajustan muy bien a los datos. Sin embargo, se selecciona el modelo que menor suma de cuadrados arroje para cada quincena. De igual forma, en ambas zonas geográficas los modelos de semivarianza Exponencial y Matern presentan sumas de cuadrados similares, aunque siempre resulta ser menor la del modelo Exponencial.

En la **Tabla 8.4** se muestra el modelo de semivarianza seleccionado para la zona de valle de acuerdo a la menor suma de cuadrados, dentro de los cuatro posibles para cada quincena. Además, se observan las estimaciones de los parámetros para el modelo seleccionado en cada una de las quincenas. De igual forma se muestra en la **Tabla 8.5** para la zona de montaña.

En la zona de valle se observa que el modelo de semivarianza con mayor recurrencia o que mejor se ajusta para mayor cantidad de quincenas es el Exponencial, mientras que en la zona de montaña no hay un modelo con una recurrencia marcada, sin embargo los que más se repiten son el Exponencial y Gaussiano.

Tabla 8.4: Estimación de parámetros para el modelo seleccionado, Valle.

Quincena	Modelo	C_0	C_1	a
1	Exponencial	0.00	4619.49	0.20
2	Exponencial	41.51	8355.91	0.17
3	Esférico	236.26	770.27	0.59
4	Exponencial	15.67	1113.71	0.01
5	Exponencial	2716.80	6736.88	12.88
6	Exponencial	2280.37	9087.77	4.43
7	Exponencial	7952.12	24571.48	52.33
8	Exponencial	3930.36	10984.28	5.53
9	Exponencial	0.00	3484.87	0.16
10	Exponencial	2517.28	8723.32	8.21
11	Exponencial	5920.61	18384.43	14.87
12	Exponencial	2203.96	7050.38	56.95
13	Exponencial	6.00	402.36	0.10
14	Exponencial	48.33	933.38	0.30
15	Exponencial	2302.55	8130.00	83.12
16	Exponencial	0.00	756.75	0.19
17	Exponencial	0.00	3383.98	0.17
18	Esférico	1874.15	11367.67	40.46
19	Esférico	395.28	3846.61	0.62
20	Exponencial	6994.44	23615.10	149.14
21	Exponencial	0.00	7402.97	0.14
22	Exponencial	0.00	22932.61	1.31
23	Exponencial	4906.63	17941.16	34.81
24	Exponencial	7447.19	25990.34	188.35

Tabla 8.5: Estimación de parámetros para el modelo seleccionado, Montaña.

Quincena	Modelo	C_0	C_1	a
1	Exponencial	1481.01	2956.09	4.69
2	Esférico	3044.17	825689.95	269.76
3	Gaussiano	1438.52	2302304.70	35.89
4	Gaussiano	567.31	129213.75	27.99
5	Exponencial	1659.47	3077.17	12.36
6	Esférico	651.30	2257.70	0.57
7	Esférico	1115.26	2254.34	9.56
8	Gaussiano	1976.24	1391493.40	28.07
9	Exponencial	546.48	1488.01	0.21
10	Gaussiano	1019.06	2441270.22	30.11
11	Gaussiano	1950.76	730617.56	55.78
12	Gaussiano	579.55	353735.46	26.94
13	Exponencial	10.88	74.83	0.10
14	Exponencial	29.20	551306.99	6895.23
15	Exponencial	2.74	7.30	0.20
16	Exponencial	2.85	42.52	0.07
17	Exponencial	579.85	272.73	0.34
18	Matern	978.36	3077.32	0.35
19	Gaussiano	989.50	2018198.95	28.19
20	Esférico	640.06	348010.70	761.96
21	Exponencial	7172.72	12426.28	2.59
22	Gaussiano	1499.79	639884.06	15.58
23	Exponencial	187.10	682.80	0.00
24	Exponencial	103.04	175.04	0.09

En la **Tabla 8.6** se presenta el resumen descriptivo para los indicadores del método kriging ordinario. Para la zona valle se encontró que el menor promedio del ECM se obtuvo con un 20 % de datos faltantes 88.52 mm con una desviación estándar de 150.03 mm, para la mediana el mejor comportamiento se obtuvo para el 10 % 44.35 mm con una desviación estándar de 197.14 mm como se puede observar las desviaciones estándar de estos indicadores fueron bastante altas lo que esta afectando directamente el promedio de estas por lo cual resulta ser mas representativo la mediana y así se obtiene que el porcentaje de datos faltantes que tuvo mejor comportamiento para este indicador fue el del 10 %.

En la zona de montaña las desviaciones estándar obtenidas para los diferentes métodos fueron mas pequeñas, el mejor comportamiento de las estimaciones se encontró para 15 % de datos faltantes tanto para el promedio 7.14 mm como para la mediana de la RECM y

una desviación estándar de 2.99 mm. En la **Tabla 8.6** se puede observar que para la zona de valle, los resultados de los indicadores dieron mucho mas altos, se puede decir que el método estima mejor para la zona de montaña, de igual forma esto se le puede asignar a que esta zona tiene mas estaciones para realizar la estimación.

Tabla 8.6: Resumen descriptivo para los indicadores del método Kriging Ordinario.

Zona	Porcentaje	Indicador	Media	Mediana	Desviación	C. Variación
Valle	10 %	ECM	100.48	44.35	197.14	196 %
		RECM	8.27	6.66	5.78	70 %
		CVRECM	4.08	3.71	1.48	36 %
	15 %	ECM	93.14	54.35	168.17	181 %
		RECM	8.27	7.37	5.08	61 %
		CVRECM	4.31	3.67	1.78	41 %
	20 %	ECM	88.52	64.36	150.03	169 %
		RECM	8.16	8.02	4.78	59 %
		CVRECM	4.32	3.96	1.46	34 %
Montaña	10 %	ECM	63.96	45.76	62.37	98 %
		RECM	7.33	6.76	3.22	44 %
		CVECM	3.50	3.50	0.97	28 %
	15 %	ECM	59.71	45.10	61.80	103 %
		RECM	7.14	6.72	2.99	42 %
		CVRECM	3.49	3.36	0.88	25 %
	20 %	ECM	59.64	48.13	59.23	99 %
		RECM	7.21	6.94	2.80	39 %
		CVRECM	3.50	3.45	0.78	22 %

De igual forma el comportamiento de los indicadores ECM y RECM para los diferentes porcentajes de datos faltantes en la zona valle, indicó que el 20 % de datos faltantes se comporto mejor, evidenciándose en los resultados con los valores mas bajos. El CVRECM tuvo el mejor comportamiento para el 10 % de datos faltantes. Para la zona de montaña el ECM se comporto mejor para el 20 % de datos faltantes obteniendo los menores valores en los resultados obtenidos, por otro lado los indicadores RECM y CVRECM se comportaron mejor para el 15 % de datos faltantes.

8.5. Resultados Generales

La sensibilidad es una cualidad deseable en un indicador, puesto que ello implica qué cambios producidos en la muestra pueden ser detectados por este; pero mucha sensibilidad en un indicador puede ser contraproducente, puesto que cambios irrelevantes en la muestra pueden producir grandes cambios en el indicador, lo cual puede prestarse para interpretaciones equivocadas, esto ocurre con el promedio, cuando la distribución es asimétrica, es decir

cuando hay unos pocos valores muy grandes o muy pequeños, la media es muy afectada por ellos Behar G. & Yepes (1996). En este caso se presentaron datos atípicos muy altos lo que genero gran variabilidad en los valores de los indicadores por lo cual la mejor opción para analizar los resultados es la mediana como se mostrara a continuación.

En la **Tabla 8.7** se muestran los resultados de las estimaciones realizadas por cada uno de los métodos de estimación de datos faltantes en ambas zonas geográficas en las que fueron aplicados, teniendo en cuenta el porcentaje de datos faltantes creados artificialmente para cada zona. Para cada método, se calcularon dos indicadores, RECM y CVRECM, que se explican en la **Sección 6.2.4**.

Observando la RECM, se tiene que el método con mejor desempeño en la estimación para la zona geográfica de valle, bajo un porcentaje de datos faltantes del 10 % es el Kriging Ordinario, bajo un porcentaje del 15 % es el método de regresión, con el 20 % resulta ser el método IDW el de mejor desempeño para esta zona, esto para cada uno de los escenarios de datos faltantes propuestos. Sin embargo, el CVRECM es menor en el Kriging Ordinario independientemente del porcentaje de datos faltantes, para la zona de valle.

Tabla 8.7: Indicadores para los métodos de estimación por zona geográfica (Mediana).

		Valle			Montaña		
Método	Indicador	10 %	15 %	20 %	10 %	15 %	20 %
K-V	RECM	7.12	7.10	6.70	6.00	6.26*	6.42*
	CVRECM	4.29	4.11	4.14	3.38	3.43	3.49
IDW	RECM	7.05	6.93	6.21*	5.54*	6.49	6.69
	CVRECM	4.30	4.18	4.00	3.31	3.36	3.44
RL	RECM	7.00	6.77*	6.84	6.83	7.30	7.40
	CVRECM	4.36	4.06	4.28	3.77	3.89	3.69
KO	RECM	6.66*	7.37	8.02	6.76	6.72	6.94
	CVRECM	3.71	3.67	3.96	3.50	3.36	3.45

Cabe resaltar que para esta zona no hay mucha variación de los indicadores por método al cambiar el porcentaje de datos faltantes, pero si lo hay entre métodos, donde esta diferencia se observa de forma mas marcada; Además, que el método de K-Vecinos es el de menor desempeño en todos los escenarios de datos faltantes.

Por otro lado, observando la RECM para escoger el método con mejor desempeño en la zona de montaña, bajo un 10 % de datos faltantes se tiene que es el IDW, bajo el 15 % y 20 % resulta ser el método de K-Vecinos. Sin embargo, el CVRECM es menor en el método IDW, independientemente del porcentaje de datos faltantes. Cabe destacar que para esta zona no hay mucha variación de los indicadores por método al cambiar el porcentaje de

datos faltantes, pero si lo hay entre métodos, donde esta diferencia se observa de forma mas marcada. Por otro lado, es evidente que los métodos presentan un mejor desempeño en la zona de montaña, el cual es mucho mejor que en la zona de valle, lo que podría deberse a que en esta zona se cuenta con mayor número de estaciones climatologicas, que podría conllevar a mejores estimaciones de los valores faltantes.

Zona Valle

Según la **Tabla 8.7** y las **Figuras 10.1, 10.2** se observa la RECM, en donde se tiene que el Kriging Ordinario es el que método que mejor se comporta y los demás presentan un comportamiento similar entre ellos, a excepción del método de Regresión que es el que mayor valores para el RECM presenta en la mayoría de las estaciones de la zona de valle. Con el CVRECM se puede observar que no hay mucha variabilidad en los estimadores encontrados, además de que el Kriging Ordinario sigue siendo el que mejor se comporta con respecto a los otros métodos. Sin embargo, resulta ser la regresión y el Kriging Ordinario los métodos que mejor comportamiento tienen y el K-Vecinos e IDW, tienen un comportamiento similar y es menos eficiente que los dos ya mencionados.

Zona Montaña

Analizando el comportamiento de los indicadores calculados para comparar los métodos de estimación por porcentaje de datos faltantes creados artificialmente para la zona de montaña mostrados en la **Tabla 8.7** y las **Figuras 10.3, 10.4**, al igual que en la zona de valle, comparando los métodos por medio del indicador RECM, el Kriging Ordinario resulta ser el que mejores estimaciones de la precipitación hace para los diferentes escenarios de porcentajes de datos faltantes, a excepción de dos estaciones en las cuales este no se comporta mejor que los otros tres métodos.

También se tiene que las estimaciones realizadas por medio del método de K-Vecinos mas cercanos son muy similares a la del IDW en los tres escenarios de datos faltantes. Por otra parte, el método de Regresión es el que arroja las estimaciones mas alejadas del valor real de la observación, siendo el método que peores estimaciones da en todos los casos, aunque en algunas estaciones este por debajo de los demás. Sin embargo, se podría decir que son muy próximos los valores de los indicadores para cada uno de los métodos en la zona de montaña.

Con el CVRECM se puede observar que no hay mucha variabilidad en los estimadores encontrados. Sin embargo, ahora resulta ser la regresión y el Kriging Ordinario los métodos que mejor comportamiento tienen y el K-vecinos e IDW, tienen un comportamiento similar y menos eficiente que los dos ya mencionados. El método de Regresión parece tener un comportamiento diferente a los demás, mas marcado que en la zona de valle, indicando que sus estimaciones no son tan buenas como la de los demás métodos.

8.6. Prueba Mann-Whitney

Como se puede observar en la **Tabla 8.8** el valor-p en todos los casos se encuentra por encima del nivel de significancia dado de 0.05, por lo que no se puede afirmar de que existen o se encontraron diferencias significativas entre los indicadores de desempeño de cada uno de los métodos, esto de acuerdo a la zona geográfica bajo la cual fueron calculados.

Tabla 8.8: Diferencias entre métodos, según zona y porcentaje de datos faltantes (valor-p).

Porcentaje %	Método		Valle	Montaña
10	KV	IDW	0.8462	0.6891
	KV	Regresión	0.8009	0.2515
	KV	KO	0.6004	0.2618
	IDW	Regresión	0.8843	0.1367
	IDW	KO	0.892	0.1173
	Regresión	KO	0.8461	0.9844
15	KV	IDW	0.8159	0.8224
	KV	Regresión	0.9536	0.1074
	KV	KO	1	0.5549
	IDW	Regresión	0.8462	0.0947
	IDW	KO	0.771	0.4041
	Regresión	KO	1	0.3616
20	KV	IDW	0.9073	0.8453
	KV	Regresión	1	0.267
	KV	KO	0.8614	0.6855
	IDW	Regresión	0.9923	0.1742
	IDW	KO	0.8614	0.5227
	Regresión	KO	1	0.4762

Lo mismo sucede con los indicadores al observar por porcentaje de datos faltantes bajo los cuales se realizó la comparación, no existen diferencias significativas entre ellos, tal como se muestra en la **Tabla 8.9**.

Tabla 8.9: Diferencias entre porcentajes, según zona y método (valor-p).

Método	Porcentaje %		Valle	Montaña
KV	10	15	0.892	0.7997
	10	20	0.9073	0.4096
	15	20	0.9536	0.5847
IDW	10	15	0.892	0.6325
	10	20	0.8386	0.3959
	15	20	0.8009	0.7162
Regresión	10	15	0.9845	0.5797
	10	20	0.9536	0.5864
	15	20	0.9845	0.9883
KO	10	15	0.8009	0.7978
	10	20	0.698	0.9922
	15	20	1	0.7623

Capítulo 9

Conclusiones y recomendaciones

9.1. Conclusiones

En este capítulo se presentan las conclusiones correspondientes a la estimación de datos faltantes en datos climatológicos diarios de precipitación en el Valle del Cauca, para el año 1997 con un total de 88 estaciones climatológicas divididas en las zonas de valle (25) y montaña (63) obtenidos por medio de los métodos de estimación K-vecinos Cercanos, Regresión Lineal, Distancia Inversa Ponderada y Kriging Ordinario, teniendo en cuenta la distribución espacial de los datos de estudio, cada uno de los métodos empleados fue evaluado a través de medidas de desempeño como: el Error Cuadrático Medio, La Raíz del Error Cuadrático Medio y el Coeficiente de Variación de la Raíz del Error Cuadrático Medio.

A manera de conclusiones específicas por zona y porcentaje de datos faltantes tenemos que para cada uno de los escenarios bajo los cuales fueron comparados los métodos de estimación, los más recomendados para cada uno, como se puede observar en la **Tabla 9.1**, son:

Tabla 9.1: Métodos por escenario.

Porcentaje %	Valle	Montaña
10	KO	IDW
15	Regresión	K-V
20	IDW	K-V

Se puede observar que el KO se desempeñó mejor en la zona de valle que cuenta con pocas estaciones a comparación de la zona de montaña, pero bajo un porcentaje de datos faltantes muy bajo, mientras que el IDW y el K-V parecen desempeñarse mejor bajo porcentajes más altos de datos faltantes en ambas zonas geográficas.

A pesar de que algunos métodos presentan mejor desempeño que otros, según los indicadores analizados, las diferencias entre estos no resultan ser significativas en ninguno de los escenarios planteados para este año. En este sentido, si se desea obtener una mayor precisión en las estimaciones que se vayan a realizar, se debe considerar los métodos que mayor desempeño obtuvieron de acuerdo a los resultados obtenidos en el proceso de comparación.

En cuanto al método de K-Vecinos más cercanos, se considera un número óptimo de estaciones para poder realizar buenas estimaciones, un número K que logre minimizar el RECM para posteriormente aplicar el promedio de estas K estaciones escogidas según las distancias mínimas hacia la estación a estimar, lo que implica que este no presenta tantas restricciones al llevarse a cabo y es de fácil aplicación, sin embargo, obvia la estructura espacial de la variable.

El método de Regresión es de los complejos a la hora de implementarse, ya que como se observó, es necesario determinar primero una estación que tenga la misma información que la estación a estimar y con la cual se tenga una mayor correlación en cuanto a la variable de estudio, y que no siempre será la estación mas cercana, por lo que requiere tener a consideración todas las estaciones disponibles para la zona geográfica que se esté analizando, además que este tiene varias restricciones (como los supuestos sobre los errores) que aseguran una buena estimación tal como se indicó en la **sección 6.2.3** y que el no cumplimiento de estas puede llevar a malas estimaciones o a información no confiable con coeficientes de correlación no tan altos indicando que la estación escogida a pesar de tener la mayor correlación no representa de la mejor forma la variabilidad de la estación a estimar, tal como se presentó en este trabajo. Además, este método toma en cuenta la historia de la estación y sus características climáticas y la información espacial entra a través del procedimiento de selección de estaciones.

El método de la Distancia Inversa Ponderada resulta ser uno de los métodos que mejor se desempeña a la hora de estimar los valores faltantes, puesto que requiere en su implementación la ponderación de las estaciones dándosele importancia a las distancias a las que se encuentran estas dentro del valor a estimar. Su complejidad radica en considerar todas las estaciones, puesto que se obtienen estimaciones que están representadas por todas las observaciones disponibles para la zona geográfica, además este método no considera la posible presencia de anisotropías ya que ocurre en áreas no homogéneas.

En el trabajo se observó, que a pesar de que el Kriging Ordinario es uno de los métodos que mejor desempeño tiene en las estimaciones, es el que mayor cantidad de restricciones tiene a la hora de implementarse y que tiene un mayor grado de complejidad frente a los demás métodos observados, entre las restricciones que se encontraron y que pueden dificultar la

implementación de este método están:

- En el variograma o semivariograma se debe de presentar un comportamiento creciente para poder realizarse una modelación, lo que en ocasiones no ocurre y por lo cual no se podría realizar ninguna estimación debido a la falta de correlación espacial en los datos. Sin embargo, este problema no se presenta en los otros métodos, ya que esta estructura no es tomada en cuenta.
- Dado que este método funciona con todas las estaciones que se tiene en la zona geográfica y se alimenta de la abundancia de estaciones que se tengan para estimar, la escasez o la poca existencia de estas sobre las cuales se pueda basar la estimación es un problema, dado que no se obtendrían buenas estimaciones y en el peor de los casos no se podría realizar la estimación de los valores faltantes ya que no se lograría tener una estructura de correlación espacial adecuada, tal como ocurrió en este trabajo. En los otros métodos, se necesitan como mínimo al menos una estación para poder estimar los valores faltantes deseados, en este método sería muy complicado realizarlo de esta forma.
- El método presenta inconsistencias al momento de realizar las estimaciones, ya que como se observó, es posible que para algunas variables obtenga valores inconsistentes en las estimaciones, como en este caso valores negativos para la precipitación, lo que debe corregirse por medio de la regla Deutsch (1996) que se expuso en este trabajo, situación que hace de este método aún mas complejo que los demás, ya que con los otros no se tiene esta posibilidad.
- Los pesos para la implementación del método tienen que ser transformados en algunos casos, cuando las estimaciones para la variable objeto de estudio no sean consistentes.

En general no se tiene un método estándar que tenga un buen comportamiento para todas las distintas condiciones o escenarios expuestos en el trabajo sobre los cuales se realizó la comparación de estos, por el contrario, bajo las distintas condiciones que se presentaron hay un método el cual tiene un mejor desempeño, lo que deja entrever que es necesario realizar una exhaustiva búsqueda al momento de escoger o seleccionar un método de estimación dependiendo la zona geográfica en donde se encuentre, ya que no todos los métodos existentes en la literatura tienen un buen desempeño para condiciones específicas y las restricciones de algunos los pueden volver inefficientes en el que hacer práctico.

Se puede decir también que los métodos resultan ser más adecuados para áreas o zonas geográficas en donde la variable a consideración o de análisis tenga poca variabilidad, lo que claramente no ocurre en este caso, ya que como se observó, la precipitación cuenta con gran variabilidad en todos los casos por tratarse de precipitación diaria, aún cuando se considera la altura y se separa la región por zonas, por lo cual sería recomendable

trabajar esta estimación de datos de manera mensual lo que reduciría la variabilidad de la precipitación. Sin embargo, a pesar de hallarse gran variabilidad en los datos estudiados como se mencionó anteriormente, esta resultó ser menor y mas homogenea en la zona de montaña, por lo que en este caso resultan tener un mejor desempeño los métodos más prácticos y menos complejos de aplicar como el IDW y KV, lo que resulta ser factible para los diferentes gremios que implementan estas metodologías en datos de precipitación, tal como se encontró en la revisión de los antecedentes realizada (Membreño & Zavala, 2015, Teegavarapu, 2014, Ferrari & Ozaki, 2014, Chen & Liu, 2012, Vicente-Serrano et al., 2010, Firat et al., 2010).

Por el contrario, los métodos que mejor desempeño obtuvieron en la zona de valle fueron los métodos un poco más complejos y con mayor número de restricciones en su aplicación, como lo son el KO y RL como se observo en la revisión bibliográfica realizada (Mair & Fares, 2010, Campozano et al., 2015, Medina & Jaramillo, 2008), métodos que requieren del cumplimiento de una serie de supuestos que en la realidad es difícil que se cumplan.

En este sentido, si se requieren buenas estimaciones se debe explorar bien los métodos que existen en la literatura y de esta forma obtener bases de datos completas y consistentes que ayuden a realizar y llevar a cabo los análisis que se requieran, de acuerdo a las condiciones bajo las cuales se este trabajando.

9.2. Recomendaciones

- Trabajar con más información ya que el método de Kriging Ordinario promete ser un excelente método pero se alimenta de la abundancia de estaciones adyacentes al punto a estimar para conseguir una buena correlación espacial. De igual forma el método de regresión lineal escoge un modelo de regresión para estimar y si se tuviera mas estaciones adyacentes se podría obtener mejores modelos con coeficientes de correlación más altos y con el respectivo cumplimiento de los supuestos de estos.
- Definitivamente trabajar estimación de datos faltantes de forma mensual ya que según lo trabajado de forma diaria se encuentra demasiada variabilidad lo que impide realizar una estimación eficiente y no permite que los métodos utilizados se desempeñen de la mejor forma.
- Considerar implementar el método de K-Vecinos más cercanos, ponderándolos con la distancia inversa, teniendo así un promedio ponderado para la estimación, con el fin de analizar si se obtienen mejores estimaciones para la precipitación. De igual forma explorar el método de K-Vecinos más cercanos con diferentes valores para k y comparar con los otros métodos.

- Analizar el comportamiento del método de la Distancia Inversa Ponderada determinando el radio dentro del cual se seleccionarán las estaciones para realizar la estimación de los valores faltantes, en comparación con el método cuando se seleccionan todas las estaciones de la zona geográfica en la cual se está trabajando.
- Analizar los métodos con otros porcentajes de datos faltantes aumentando el rango entre los escenarios, para observar si se obtienen diferencias más marcadas entre los indicadores que miden el desempeño de los métodos de estimación.
- Analizar métodos modernos de comparación geofuncional, gam e incluir convariables con datos satelitales.

Bibliografía

- Ablan, M., Rigoberto, A., Vargas, M. P., & Acevedo, M. (2008). Propuesta metodológica para el control de calidad de datos de precipitación (methodological proposal for the control of quality of data of precipitation). *Agronomía Trop*, 58(1), 57–60.
- Allen, R. G. et al. (2006). *Evapotranspiración del cultivo: guías para la determinación de los requerimientos de agua de los cultivos*, volume 56. Food & Agriculture Org.
- Amaro, I. R., Demey, J., & Machiavelli, R. (2004). Aplicación del análisis r/s de hurst para estudiar las propiedades fractales de la precipitación en venezuela. *Interciencia*, 29(11), 617–620.
- Behar G., R. & Yepes, M. (1996). *Estadística: un enfoque descriptivo*. Universidad del Valle, Facultad de Ingeniería, Departamento de Producción e Investigación de Operaciones.
- Bishop, M. & Shroder, J. F. (2004). *Geographic information science and mountain geomorphology*. Springer Science & Business Media.
- Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). Hello world: Introducing spatial data. In *Applied Spatial Data Analysis with R* (pp. 1–16). Springer.
- Bromberg, F. & Pérez, D. (2012). Interpolación espacial mediante aprendizaje de máquinas en vinedos de la provincia de mendoza, argentina. In *13th Argentine Symposium on Artificial Intelligence, ASAI*, volume 41.
- Campoazano, L., Sánchez, E., Avilés, A., & Samaniego, E. (2015). Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the ecuadorian andes. *Maskana*, 5(1).
- Carvajal, Y., Jimenez, H., & Materon, H. (1998). *Incidencia del fenomemo enso en la climatología del valle del rio Cauca-Colombia*. Inst. fretudes andine.
- Carvajal, Y. & Marco, J. B. (2011). Análisis de variabilidad de datos medioambientales aplicando funciones ortogonales empíricas o componentes principales. *Ingeniería de Recursos Naturales y del Ambiente*, 1(2), 4–11.

- Castro-Llanos, D. A. & Carvajal-Escobar, Y. (2013). Análisis de tendencia en la precipitación pluvial anual y mensual en el departamento del valle del cauca. *Memorias*, 11(20), 9–17.
- Chen, F.-W. & Liu, C.-W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (idw) in the middle of taiwan. *Paddy and Water Environment*, 10(3), 209–222.
- Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21–27.
- Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.
- De la Horra Navarro, J. (2003). *Estadística aplicada*.
- Deutsch, C. V. (1996). Correcting for negative weights in ordinary kriging. *Computers & Geosciences*, 22(7), 765–773.
- Di Piazza, A., Conti, F. L., Noto, L., Viola, F., & La Loggia, G. (2011). Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for sicily, italy. *International Journal of Applied Earth Observation and Geoinformation*, 13(3), 396–408.
- Díaz-Francés, E. (1993). Introducción a conceptos básicos de geoestadística. *Memorias Seminario Estadística y Medio Ambiente. Centro de Investigación en Matemáticas, CIMAT. Guanajuato, México*.
- Diéguez-Aranda, U., Rojo Alboreca, A., Castedo-Dorado, F., Álvarez González, J., Barrio-Anta, M., Crecente-Campo, F., González González, J., Pérez-Cruzado, C., Rodríguez Soalleiro, R., López-Sánchez, C., et al. (2009). Herramientas selvícolas para la gestión forestal sostenible en galicia. *Consellería do Medio Rural, Xunta de Galicia*, 268.
- DNP, A. I. p. l. P. (2007). La competitividad del valle del cauca. departamento nacional de planeación.[en línea]. *Bogotá. Colombia*.
- Dumedah, G. & Coulibaly, P. (2011). Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. *Journal of Hydrology*, 400(1), 95–102.
- Felícísimo, Á. M. (1994). *Modelos digitales del terreno*. Pentalfa Oviedo.
- Fernandez, C. & Lavado, W. (2014). Analisis comparativo de metodos de interpolacion espacial para la elaboracion de mapas climaticos de precipitacion. *Servicio nacional de meteorologia e hidrologia del Peru (Senamhi)*.
- Ferrari, G. T. & Ozaki, V. (2014). Missing data imputation of climate datasets: Implications to modeling extreme drought events. *Revista Brasileira de Meteorologia*, 29(1), 21–28.

- Firat, M., Dikbas, F., Koç, A. C., & Gungor, M. (2010). Missing data analysis and homogeneity test for turkish precipitation series. *Sadhana*, 35(6), 707–720.
- Flórez, J. M., España, G. M., & Torres, H. V. (2014). Estrategia de regresión basada en el método de los k vecinos más cercanos para la estimación de la distancia de falla en sistemas radiales. *Revista Facultad de Ingeniería*, (45), 100–108.
- FONADE (2016). Fondo financiero de proyectos de desarrollo. <http://www.fonade.gov.co/> [Fecha de consulta: 12-10-2016].
- Galván, M. (2007). *Imputación de datos: teoría y práctica*, volume 54. United Nations Publications.
- Giraldo, R. (2009). Estadística espacial. *Notas de clase*. Bogotá. Universidad Nacional.
- Gonzalez, H. & Aristizabal, F. (2006). Aplicacion de tecnicas estadísticas en las series climatologicas mensuales totales de precipitacion, evaporacion y brillo solar, con el fin de corregir, complementar y verificar la calidad de la informacion.
- Goodchild, M. F. (1987). A spatial analytical perspective on geographical information systems. *International journal of geographical information system*, 1(4), 327–334.
- Gracia, J. & Maza, A. (1997). Morfología de ríos. *Instituto de Ingeniería UNAM*. México.
- Greenpeace (2009). La crisis del clima. *Evidencias del Cambio Climático en España. Recopilación de la literatura científica existente sobre impactos del cambio climático en España*.
- Hurtado, G., González, O., & Montaña, J. (2001). Tercera parte: Aspectos departamentales. *Atlas climatológico de Colombia*. Bogotá, Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia IDEAM.
- IDEAM (2016). Instituto de hidrología, meteorología y estudios ambientales. <http://www.ideam.gov.co/> [Fecha de consulta: 12-10-2016].
- Infante, S., Ortega, J., & Cedeño, F. (2008). Estimación de datos faltantes en estaciones meteorológicas de venezuela vía un modelo de redes neuronales. *Revista de Climatología, Departamento de Matemáticas*.
- Jang, D., Park, H., & Choi, J. (2015). Selection of optimum spatial interpolation method to complement an area missing precipitation data of rcp climate change scenario. *International Journal of Software Engineering and Its Applications*, 9(8), 179–188.
- Kaplan, A., Kushnir, Y., & Cane, M. A. (2000). Reduced space optimal interpolation of historical marine sea level pressure: 1854-1992*. *Journal of Climate*, 13(16), 2987–3002.

- Kim, J. & Ryu, J. H. (2015). Quantifying a threshold of missing values for gap filling processes in daily precipitation series. *Water Resources Management*, 29(11), 4173–4184.
- Kim, J.-W. & Pachepsky, Y. A. (2010). Reconstructing missing daily precipitation data using regression trees and artificial neural networks for swat streamflow simulation. *Journal of hydrology*, 394(3), 305–314.
- Lee, H. & Kang, K. (2015). Interpolation of missing precipitation data using kernel estimations for hydrologic modeling. *Advances in Meteorology*, 2015, 12.
- Li, X., Cheng, G., & Lu, L. (2003). Comparison study of spatial interpolation methods of air temperature over qinghai-xizang plateau. *Plateau Meteorology*, 22(6), 565–573.
- Lopez Araya, J. (2014). Experiencias en la aplicación operativa de un método multivariado de imputación de datos meteorológicos. *Tecnología en Marcha*, 27.3, 70–79.
- Luna Romero, E. & Lavado Casimiro, W. (2015). Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la cuenca jetepeque, Perú. *Revista Tecnológica ESPOL – RTE*, 28.3, 42–52.
- Maderey Rascon, L. E., Roman, J., et al. (2005). *Principios de hidrogeografía. Estudio del ciclo hidrológico*. UNAM.
- Mair, A. & Fares, A. (2010). Comparison of rainfall interpolation methods in a mountainous region of a tropical island. *Journal of Hydrologic Engineering*, 16(4), 371–383.
- Martínez Gómez, M. T. (2011). Diagnóstico situación de riesgos hidrometeorológicos en Colombia y avances en la zonificación de riesgos. *IDEAM, Instituto de hidrología, meteorología y estudios ambientales*.
- McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.
- Medina, Rd, M. E. & Jaramillo, A. (2008). Estimación estadística de valores faltantes en series históricas de lluvia.
- Membreño, A. M. & Zavala, A. d. J. R. (2015). Aplicación de técnicas estadísticas y geoestadísticas para elaborar cartografía de precipitaciones. departamentos del occidente de nicaragua. *Ciencias Espaciales*, 8(1), 143–159.
- Méndez González, J., Nívar Cháidez, J. d. J., & González Ontiveros, V. (2008). Análisis de tendencias de precipitación (1920-2004) en México. *Investigaciones geográficas*, (65), 38–55.
- Moreno, G. (2009). *Vulnerabilidad y estrategia de adaptación para la agricultura en Colombia*. Subdirección de Estudios Ambientales del Ideam.

- Ordoñez, P. H. (2001). Inferencia estadística, regresión y muestreo. *Manizales: Universidad Nacional de Colombia, sede Manizales*. 2.000. 162 p.
- Ortega, J. H. G. & Rodriguez, L. M. S. (2010). Aplicación de tecnologías de información geográfica para el estudio de la variabilidad climática en la cuenca alta del río cauca.
- Pabón-Caicedo, J. D., Eslava-Ramírez, J. A., & Gómez-Torres, R. E. (2001). Generalidades de la distribución espacial y temporal de la temperatura del aire y de la precipitación en colombia. *Meteorologia colombiana*, 4, 47–59.
- Penagos Cruz, G. (2014). Variables hidrometereológicas asociadas al cambio climático en girardot y la región del alto magdalena. *Ambiente y Desarrollo*, 18.35, 134–147.
- Presti, R. L., Barca, E., & Passarella, G. (2010). A methodology for treating missing data applied to daily rainfall data in the candelaro river basin (Italy). *Environmental monitoring and assessment*, 160(1-4), 1–22.
- R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Radi, N. F. A., Zakaria, R., Azman, M. A.-z., Mohamad, M. S., Wan Yusoff, W. N. S., Md Noar, N. A. Z., Zakaria, R., & Ab Hamid, M. R. (2015). Estimation of missing rainfall data using spatial interpolation and imputation methods. In *AIP Conference Proceedings*, volume 1643 (pp. 42–48).
- Ramos-Calzado, P., Gomez-Camacho, J., Perez-Bernal, F., & Pita-Lopez, M. (2008). A novel approach to precipitation series completion in climatological datasets: application to andalusia. *International Journal of Climatology*, 28(11), 1525–1534.
- Rosgen, D. L. & Silvey, H. L. (1996). *Applied river morphology*, volume 1481. Wildland Hydrology Pagosa Springs, Colorado.
- Salvador Figueras, M. & Gargallo, P. (2003). Análisis exploratorio de datos.[en línea] 5campus.com. *Estadística* <http://www.5campus.com/leccion/aed> [Fecha de consulta: 28-07-2017].
- Simolo, C., Brunetti, M., Maugeri, M., & Nanni, T. (2010). Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *International Journal of Climatology*, 30(10), 1564–1576.
- Teegavarapu, R. S. (2009). Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. *Journal of Hydroinformatics*, 11(2), 133–146.

- Teegavarapu, R. S. (2014). Missing precipitation data estimation using optimal proximity metric-based imputation, nearest-neighbour classification and cluster-based interpolation methods. *Hydrological Sciences Journal*, 59(11), 2009–2026.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1), 234–240.
- Trewin, B., Baddour, O., & Kontongomde, H. (2007). *The role of climatological normals in a changing climate*. World Meteorological Organization.
- UNESCO, R. (1982). Guía metodológica para la elaboración del balance hídrico de América del Sur.
- Urrutia, J. A., Palomino, R., & Salazar, H. D. (2010). Metodología para la imputación de datos faltantes en meteorología. *Scientia et Technica*, 3(46), 44–49.
- Viada, C., Bouza, C., Ballesteros, J., Fors, M., & Robaina, M. (2016). Revisión sistemática de los métodos de imputación de datos faltantes. *Experiencias en la modelación de la toma de decisiones en la salud humana, medio ambiente y desarrollo humano*, 2, 113–130.
- Vicente-Serrano, S. M., Beguería, S., López-Moreno, J. I., García-Vera, M. A., & Stepanek, P. (2010). A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. *International Journal of Climatology*, 30(8), 1146–1163.
- Vilalta y Perdomo, C. J. (2005). Cómo enseñar autocorrelación espacial. *Economía, Sociedad y Territorio*, (18).
- Wackernagel, H. (2013). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.
- Xu, C.-D., Wang, J.-F., Hu, M.-G., & Li, Q.-X. (2013). Interpolation of missing temperature data at meteorological stations using p-bshade. *Journal of Climate*, 26(19), 7452–7463.
- Yang Lenyap, K. U., SHAADAN, N., DENI, S. M., & JEMAIN, A. A. (2015). Application of functional data analysis for the treatment of missing air quality data. *Sains Malaysiana*, 44(10), 1531–1540.
- Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, I. (2013). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and applied climatology*, 112(1-2), 143–167.

Capítulo 10

Anexos

Tabla 10.1: Listado de estaciones que se encuentran en la zona de valle año 1997.

Código	Nombre	Código	Nombre
2610000103	CRUCES	2622400103	PLANTA RIO CALI
2610000105	PUERTO MOLINA	2622500102	LLOREDA GRASAS
2612700105	MATAPALO	2624000102	RIOFRIO
2612710101	FLORIDA	2624600103	LA HERRADURA
2612710103	ING LA QUINTA	2624700103	EL VESUBIO
2612720101	YUNDE CARRIZAL	2624700104	EL PORVENIR
2613600105	ACUEDUCTO BUGA	2624700117	EL LUCERO
2614400102	BUGALAGRANDE	5301400108	ALTO - ANCHICAYA
2614900102	LA ELVIRA	5301500102	LOBOGUERRERO
2615400107	QUEBRADANUEVA	5301500106	TRIANA
2620000103	PLANTA RIO CAUCA	5301500107	DAGUA
2622100102	EL PALACIO	5420410115	PUERTO NUEVO
2622330101	COLEGIO SAN JUAN BOSCO		

Tabla 10.2: Listado de estaciones que se encuentran en la zona de montaña año 1997.

Código	Nombre	Código	Nombre
2612200103	LOS ALPES	2622410102	YANACONAS
2612720103	BOLO BLANCO	2622410104	PEÑAS BLANCAS
2612800103	BELLA VISTA	2622420101	MONTEBELLO
2612800104	AUJI	2622420102	SAN PABLO
2612800106	LA CEJA	2622420103	AGUACATAL
2612810108	TENJO	2622420104	VILLA ARACELLY
2612810112	AUSTRIA	2622500101	DAPA
2612810122	PLANTA NIMA	2622600101	SANTA INES
2613000103	EL CASTILLO	2622600103	YUMBILLO
2613100103	LA SELVA	2622900101	VILLAMARIA
2613100107	COSTA RICA	2622900102	OCACHE
2613600101	DOSQUEBRADAS	2623300101	BUENOS AIRES
2613600102	EL DILUVIO	2623500101	EL CANEY
2613600103	LA PRIMAVERA	2624600102	EL RETIRO
2613800101	ANGOSTURAS	2624700108	LAS PEÑAS
2614100103	LA GITANA	2624700109	MONTECRISTO
2614100104	LOS BANCOS	2624700110	EL AGUACATE
2614200101	VENUS	2624700113	EL OSO
2614400101	ALEGRIAS	2624700114	BUENAVISTA
2614400104	IRLANDA	2624700115	EL OREGANO
2615000102	LA ARBOLEDA	2624700118	PATIO BONITO
2621500101	VILLACOLOMBIA	2624700120	LA DESPENSA
2621500103	EL SILENCIOSO	2624700121	LA QUIEBRA
2622100103	SAN VICENTE	2625500102	EL AGUILA EL GRANARIO
2622100104	PEÑA MONA	5420410107	BETANIA
2622110101	LA ARGENTINA	5420410112	LA MARIA
2622320101	LA FONDA	5420410113	EL BOSQUE
2622320103	ALTO DE LAS IGLESIAS	5420410114	ARABIA
2622330105	LOS CRISTALES	5420410116	EL BALSAL
2622330106	LAS BRISAS	5420410117	EL CAIRO
2622400101	COLEGIO SAN LUIS	5420410118	MORAVITO
2622400104	BRASILIA		

Tabla 10.3: Coeficientes de correlación para la estimación método Regresión Lineal zona Valle

Estación X	10 % Valle		15 % Valle		20 % Valle	
	Estación Y	(γ)	Estación Y	(γ)	Estación Y	(γ)
ACUEDUCTO BUGA	RIOFRIO	0.31	RIOFRIO	0.31	RIOFRIO	0.32
ALTO...ANCHICAYA	TRIANA	0.30	TRIANA	0.30	TRIANA	0.30
BUGALAGRANDE	RIOFRIO	0.35	RIOFRIO	0.36	RIOFRIO	0.35
COLEGIO_SAN_JUAN_BOSCO	PLANTA_RIO_CALI	0.58	PLANTA_RIO_CALI	0.60	PLANTA_RIO_CALI	0.61
CRUCES	PUERTO_MOLINA	0.23	LA_HERRADURA	0.23	LA_HERRADURA	0.23
DAGUA	ALTO...ANCHICAYA	0.27	ALTO...ANCHICAYA	0.24	LA_HERRADURA	0.24
EL_LUCERO	EL_PORVENIR	0.46	EL_PORVENIR	0.46	EL_PORVENIR	0.46
EL_PALACIO	COLEGIO_SAN_JUAN_BOSCO	0.29	FLORIDA	0.29	FLORIDA	0.33
EL_PORVENIR	EL_LUCERO	0.51	EL_LUCERO	0.48	EL_LUCERO	0.49
EL_VESUBIO	EL_PORVENIR	0.27	EL_PORVENIR	0.27	EL_PORVENIR	0.26
FLORIDA	EL_PALACIO	0.23	EL_PALACIO	0.24	EL_PALACIO	0.31
ING.LA.QUINTA	FLORIDA	0.23	FLORIDA	0.27	FLORIDA	0.28
LA_ELVIRA	LA_HERRADURA	0.39	LA_HERRADURA	0.38	LA_HERRADURA	0.36
LA_HERRADURA	LA_ELVIRA	0.36	LA_ELVIRA	0.36	LA_ELVIRA	0.36
LLOREDA_GRASAS	PLANTA_RIO_CAUCA	0.45	PLANTA_RIO_CAUCA	0.44	PLANTA_RIO_CAUCA	0.44
LOBOGUERRERO	RIOFRIO	0.11	RIOFRIO	0.11	RIOFRIO	0.12
MATAPALO	COLEGIO_SAN_JUAN_BOSCO	0.42	COLEGIO_SAN_JUAN_BOSCO	0.42	COLEGIO_SAN_JUAN_BOSCO	0.42
PLANTA_RIO_CALI	COLEGIO_SAN_JUAN_BOSCO	0.62	COLEGIO_SAN_JUAN_BOSCO	0.61	COLEGIO_SAN_JUAN_BOSCO	0.62
PLANTA_RIO_CAUCA	COLEGIO_SAN_JUAN_BOSCO	0.45	COLEGIO_SAN_JUAN_BOSCO	0.43	COLEGIO_SAN_JUAN_BOSCO	0.45
PUERTO_MOLINA	EL_LUCERO	0.33	EL_LUCERO	0.32	EL_LUCERO	0.33
PUERTO_NUEVO	COLEGIO_SAN_JUAN_BOSCO	0.13	YUNDE_CARRIZAL	0.12	QUEBRADANUEVA	0.13
QUEBRADANUEVA	LA_ELVIRA	0.39	LA_ELVIRA	0.36	LA_ELVIRA	0.36
RIOFRIO	BUGALAGRANDE	0.35	ACUEDUCTO_BUGA	0.37	BUGALAGRANDE	0.38
TRIANA	ALTO...ANCHICAYA	0.32	ALTO...ANCHICAYA	0.32	ALTO...ANCHICAYA	0.33
YUNDE_CARRIZAL	COLEGIO_SAN_JUAN_BOSCO	0.19	COLEGIO_SAN_JUAN_BOSCO	0.19	COLEGIO_SAN_JUAN_BOSCO	0.19

Tabla 10.4: Coeficientes de correlación para la estimación método Regresión Lineal zona Montaña

Estacion X	10 % Montaña		15 % Montaña		20 % Montaña	
	Estacion Y	(γ)	Estacion Y	(γ)	Estacion Y	(γ)
AGUCATAL	VILLA_ARACELLY	0.59	VILLA_ARACELLY	0.59	SAN.PABLO	0.60
ALEGRÍAS	LA.GITANA	0.52	LA.GITANA	0.54	LA.GITANA	0.52
ALTO.DE.LAS.IGLESIAS	LA.FONDA	0.62	LA.FONDA	0.62	LA.FONDA	0.61
ANGOSTURAS	LA.GITANA	0.30	LA.GITANA	0.32	LA.GITANA	0.31
ARABIA	EL.OREGANO	0.53	BETANIA	0.55	BETANIA	0.55
AUJI	LA.CEJA	0.42	LA.CEJA	0.41	BELLA.VISTA	0.39
AUSTRIA	TENJO	0.36	TENJO	0.35	TENJO	0.32
BELLA.VISTA	LA.CEJA	0.53	LA.CEJA	0.56	LA.CEJA	0.57
BETANIA	EL.RETIRO	0.56	EL.RETIRO	0.61	EL.RETIRO	0.61
BOLO.BLANCO	BELLA.VISTA	0.50	BELLA.VISTA	0.51	BELLA.VISTA	0.50
BRASILIA	YANACONAS	0.47	YANACONAS	0.43	YANACONAS	0.52
BUENAVISTA	EL.OREGANO	0.61	EL.OREGANO	0.61	EL.OREGANO	0.62
BUENOS.AIRES	PATIO.BONITO	0.26	PATIO.BONITO	0.26	PATIO.BONITO	0.26
COLEGIO.SAN.LUIS	MONTEBELLO	0.56	LOS.CRISTALES	0.50	MONTEBELLO	0.50
COSTA.RICA	LA.FONDA	0.29	ARABIA	0.30	EL.CASTILLO	0.30
DAPA	VILLA_ARACELLY	0.46	VILLA_ARACELLY	0.46	VILLA_ARACELLY	0.44
DOSQUEBRADAS	LA.SELVA	0.38	LA.SELVA	0.38	LA.SELVA	0.39
EL.AGUACATE	MONTECRISTO	0.62	EL.RETIRO	0.61	MONTECRISTO	0.61
EL.AGUILA.EL.GRANARIO	PEÑA.MONA	0.29	PEÑA.MONA	0.28	PEÑA.MONA	0.27
EL.BALSAL	BUENAVISTA	0.48	BUENAVISTA	0.47	BUENAVISTA	0.47
EL.BOSQUE	LAS.PEÑAS	0.29	LAS.PEÑAS	0.28	LAS.PEÑAS	0.29
EL.CAIRO	PATIO.BONITO	0.46	PATIO.BONITO	0.47	PATIO.BONITO	0.45
EL.CANEY	PATIO.BONITO	0.30	BUENOS.AIRES	0.29	BUENOS.AIRES	0.30
EL.CASTILLO	TENJO	0.51	TENJO	0.47	LA.SELVA	0.47
EL.DILUVIO	LA.SELVA	0.31	LA.SELVA	0.32	LA.SELVA	0.33
EL.OREGANO	BUENAVISTA	0.58	BUENAVISTA	0.59	BUENAVISTA	0.59
EL.OSO	PATIO.BONITO	0.53	PATIO.BONITO	0.53	PATIO.BONITO	0.52
EL.RETIRO	EL.AGUACATE	0.60	EL.AGUACATE	0.63	EL.AGUACATE	0.62
EL.SILENCIOSO	VILLACOLOMBIA	0.26	VILLACOLOMBIA	0.29	VILLACOLOMBIA	0.31
IRLANDA	VENUS	0.42	VENUS	0.44	VENUS	0.40
LA.ARBOLEDA	DOSQUEBRADAS	0.14	DOSQUEBRADAS	0.14	DOSQUEBRADAS	0.14
LA.ARGENTINA	ALTO.DE.LAS.IGLESIAS	0.55	ALTO.DE.LAS.IGLESIAS	0.54	ALTO.DE.LAS.IGLESIAS	0.53
LA.CEJA	BELLA.VISTA	0.54	BELLA.VISTA	0.54	BELLA.VISTA	0.55
LA.DESPENSA	BUENAVISTA	0.36	BUENAVISTA	0.37	ARABIA	0.39
LA.FONDA	ALTO.DE.LAS.IGLESIAS	0.64	ALTO.DE.LAS.IGLESIAS	0.64	ALTO.DE.LAS.IGLESIAS	0.64
LA.GITANA	LOS.BANCOS	0.51	BELLA.VISTA	0.50	LOS.BANCOS	0.49
LA.MARIA	BUENAVISTA	0.52	BUENAVISTA	0.52	BUENAVISTA	0.51
LA.PRIMAVERA	DOSQUEBRADAS	0.31	DOSQUEBRADAS	0.30	DOSQUEBRADAS	0.31
LA.QUIEBRA	LAS.PEÑAS	0.51	LAS.PEÑAS	0.52	LAS.PEÑAS	0.56
LA.SELVA	EL.CASTILLO	0.53	EL.CASTILLO	0.53	EL.CASTILLO	0.49
LAS.BRISAS	LA.FONDA	0.61	LA.FONDA	0.63	LA.FONDA	0.63
LAS.PEÑAS	LA.QUIEBRA	0.61	LA.QUIEBRA	0.61	LA.QUIEBRA	0.64
LOS.ALPESES	BELLA.VISTA	0.42	BELLA.VISTA	0.41	BELLA.VISTA	0.40
LOS.BANCOS	LA.GITANA	0.49	LA.GITANA	0.47	LA.GITANA	0.47
LOS.CRISTALES	LA.FONDA	0.62	LA.FONDA	0.63	LA.FONDA	0.64
MONTEBELLO	COLEGIO.SAN.LUIS	0.63	COLEGIO.SAN.LUIS	0.62	COLEGIO.SAN.LUIS	0.61
MONTECRISTO	EL.AGUACATE	0.63	EL.AGUACATE	0.63	EL.AGUACATE	0.63
MORAVITO	EL.RETIRO	0.42	EL.RETIRO	0.42	EL.RETIRO	0.43
OCACHE	VILLAMARIA	0.46	VILLAMARIA	0.47	VILLAMARIA	0.47
PATIO.BONITO	EL.OREGANO	0.52	EL.OREGANO	0.52	BUENAVISTA	0.52
PEÑA.MONA	SAN.VICENTE	0.52	SAN.VICENTE	0.52	SAN.VICENTE	0.53
PEÑAS.BLANCAS	YANACONAS	0.52	YANACONAS	0.50	YANACONAS	0.52
PLANTA.NIMA	EL.CASTILLO	0.39	EL.CASTILLO	0.40	EL.CASTILLO	0.42
SAN.PABLO	VILLA_ARACELLY	0.61	AGUCATAL	0.63	AGUCATAL	0.62
SAN.VICENTE	PEÑA.MONA	0.55	PEÑA.MONA	0.56	PEÑA.MONA	0.55
SANTA.INES	VILLA_ARACELLY	0.32	VILLA_ARACELLY	0.29	VILLA_ARACELLY	0.28
TENJO	EL.CASTILLO	0.57	EL.CASTILLO	0.56	EL.CASTILLO	0.57
VENUS	ALEGRÍAS	0.45	ALEGRÍAS	0.44	ALEGRÍAS	0.44
VILLA_ARACELLY	AGUCATAL	0.59	AGUCATAL	0.60	AGUCATAL	0.60
VILLACOLOMBIA	PEÑA.MONA	0.45	PEÑA.MONA	0.46	PEÑA.MONA	0.46
VILLAMARIA	OCACHE	0.45	OCACHE	0.44	OCACHE	0.43
YANACONAS	BRASILIA	0.59	BRASILIA	0.57	BRASILIA	0.57
YUMBILLO	DAPA	0.39	DAPA	0.42	DAPA	0.46

Tabla 10.5: Suma de Cuadrados del Error (SCE), modelos teóricos de semivarianza (KO).

Quincena	Zona	Modelo			
		Exponencial	Esferico	Gaussiano	Matern
1	Valle	11023.58*	12379.82	12517.02	11023.58
	Montaña	2386.19*	2387.47	2413.70	2386.20
2	Valle	13199.24*	15266.63	15805.71	13199.24
	Montaña	10475.96	10475.59*	10583.44	10475.96
3	Valle	1535.48	1497.27*	1644.94	1535.48
	Montaña	4679.28	4679.02	4206.56*	4679.28
4	Valle	2261.33*	2432.93	2494.73	2261.33
	Montaña	695.75	695.74	689.12*	695.75
5	Valle	8223.77*	8231.40	8259.90	8223.77
	Montaña	2355.14*	2358.36	2393.45	2355.14
6	Valle	8847.83*	8901.50	9185.65	8847.84
	Montaña	3690.98	3674.22*	3728.94	3690.98
7	Valle	32924.6*	32926.28	32932.61	32924.60
	Montaña	1114.11	1114.05*	1114.14	1114.19
8	Valle	13020.29*	13072.10	13309.83	13020.29
	Montaña	5181.94	5181.75	4856.52*	5181.95
9	Valle	6288.02*	7313.59	7429.60	6288.02
	Montaña	1193.67*	1241.09	1315.14	1193.69
10	Valle	10695.04*	10716.03	10804.20	10695.04
	Montaña	4002.90	4002.46	3008.51*	4002.90
11	Valle	21923.89*	21938.86	22002.96	21923.89
	Montaña	2075.57	2075.57	2075.54*	2075.57
12	Valle	8964.52*	8964.92	8966.63	8964.53
	Montaña	1409.29	1409.27	1354.6*	1409.31
13	Valle	11997.08*	11997.08	11997.08	11997.08
	Montaña	2333.21*	2333.64	2333.23	2333.42
14	Valle	14754.42*	14754.45	14754.45	14754.48
	Montaña	12746.75*	12746.75	12746.76	12746.76
15	Valle	1753.46*	1754.46	1783.63	1753.49
	Montaña	6067.67*	6067.92	19187.70	6067.68
16	Valle	2074.09*	2431.77	2261.33	2074.10
	Montaña	765.87*	772.36	767.16	766.44
17	Valle	8226.04*	8231.45	8259.90	8226.04
	Montaña	289.2*	309.27	297.23	289.37
18	Valle	7828.47	7528.79*	8644.24	7828.47
	Montaña	3702.81	3674.22	3818.22	1170.93*
19	Valle	31888.21	30190.57*	31888.21	31888.21
	Montaña	4059.45	4059.03	3176.47*	4059.48
20	Valle	11317.64*	13074.02	13309.96	11317.64
	Montaña	6479.61	5181.75*	6479.61	6479.63
21	Valle	6288.27*	7314.23	7429.65	6288.27
	Montaña	1193.67*	1591.99	1808.46	1193.71
22	Valle	10710.09*	10716.74	10804.20	10710.09
	Montaña	2950.90	2950.19	1944.06*	2950.92
23	Valle	21922.34*	21938.82	22002.96	21922.35
	Montaña	2002.72*	2003.71	2002.74	2002.84
24	Valle	8964.83*	8964.92	8966.63	8964.84
	Montaña	271.34*	298.93	302.32	271.36

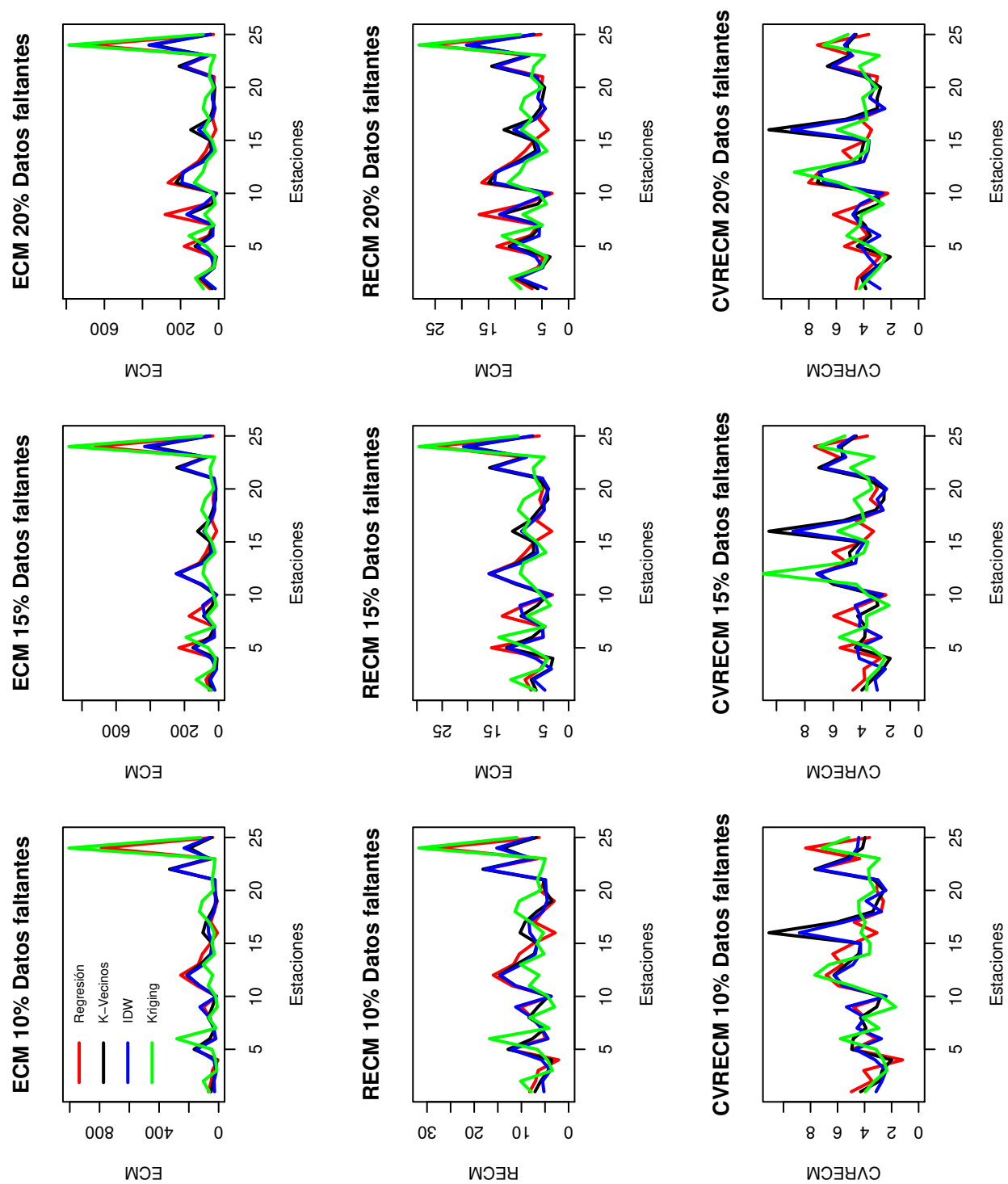


Figura 10.1: Comparación de métodos por zona Valle

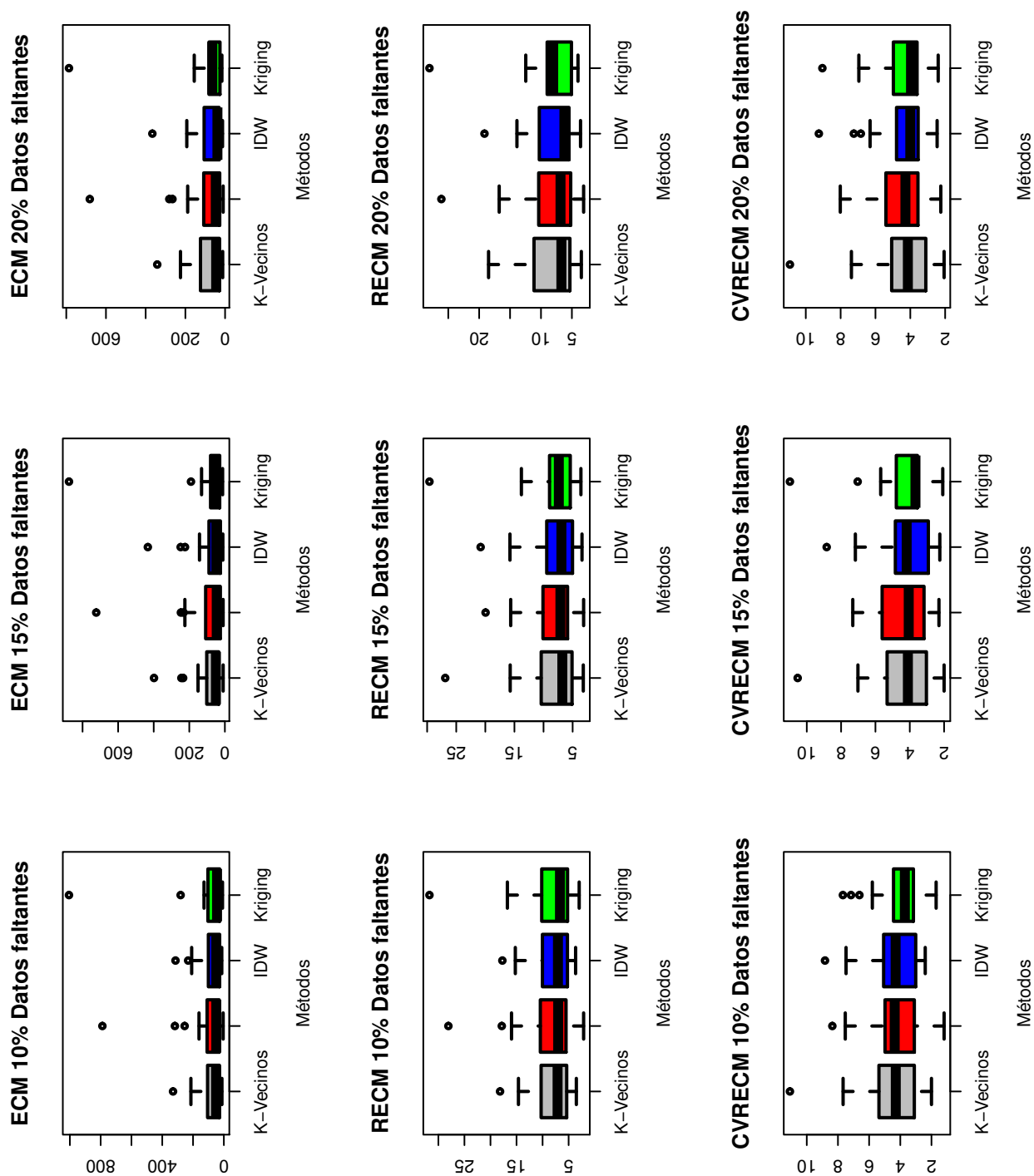


Figura 10.2: Comparación de métodos por medio de Boxplot zona Valle

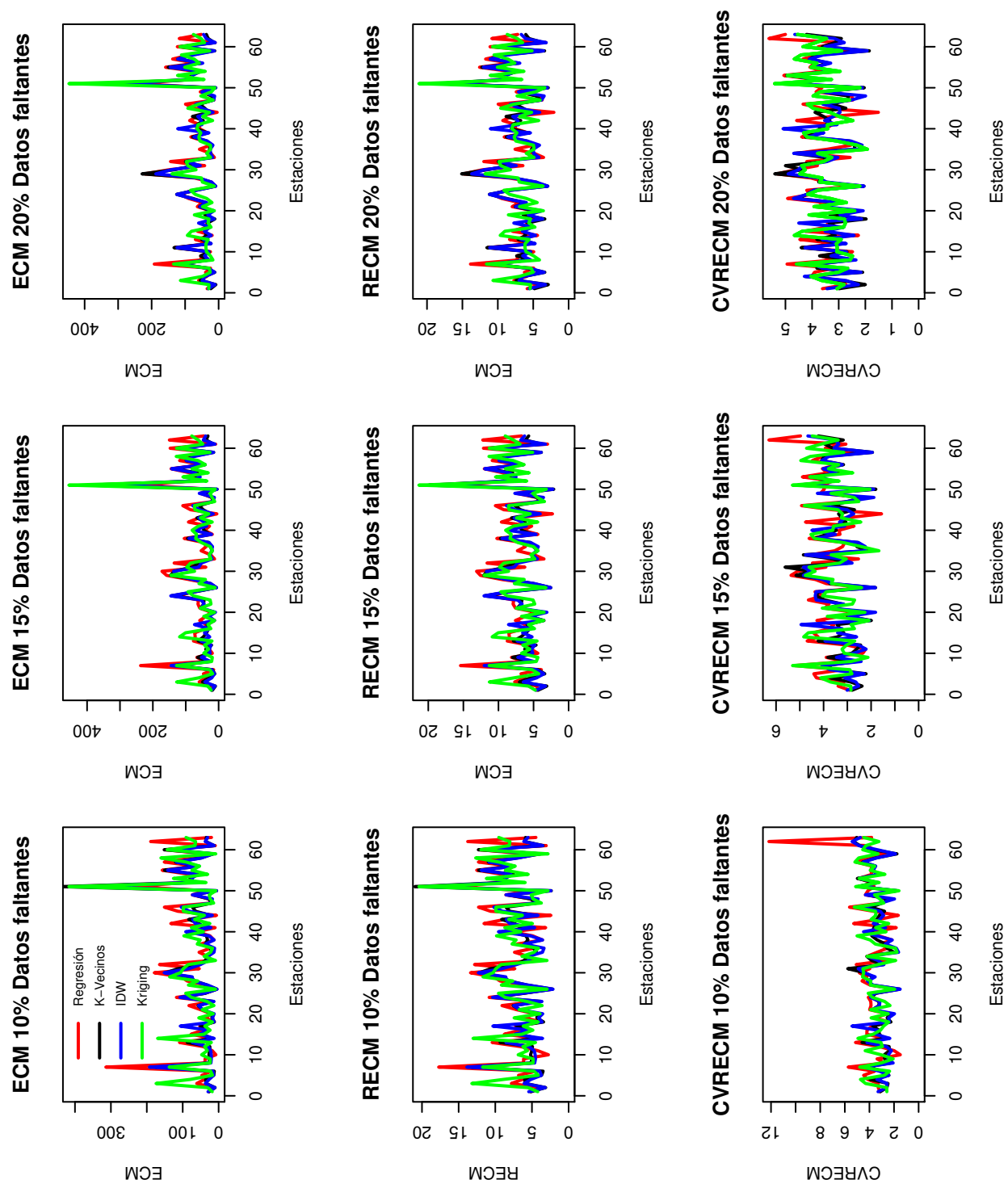


Figura 10.3: Comparación de métodos por zona Montaña

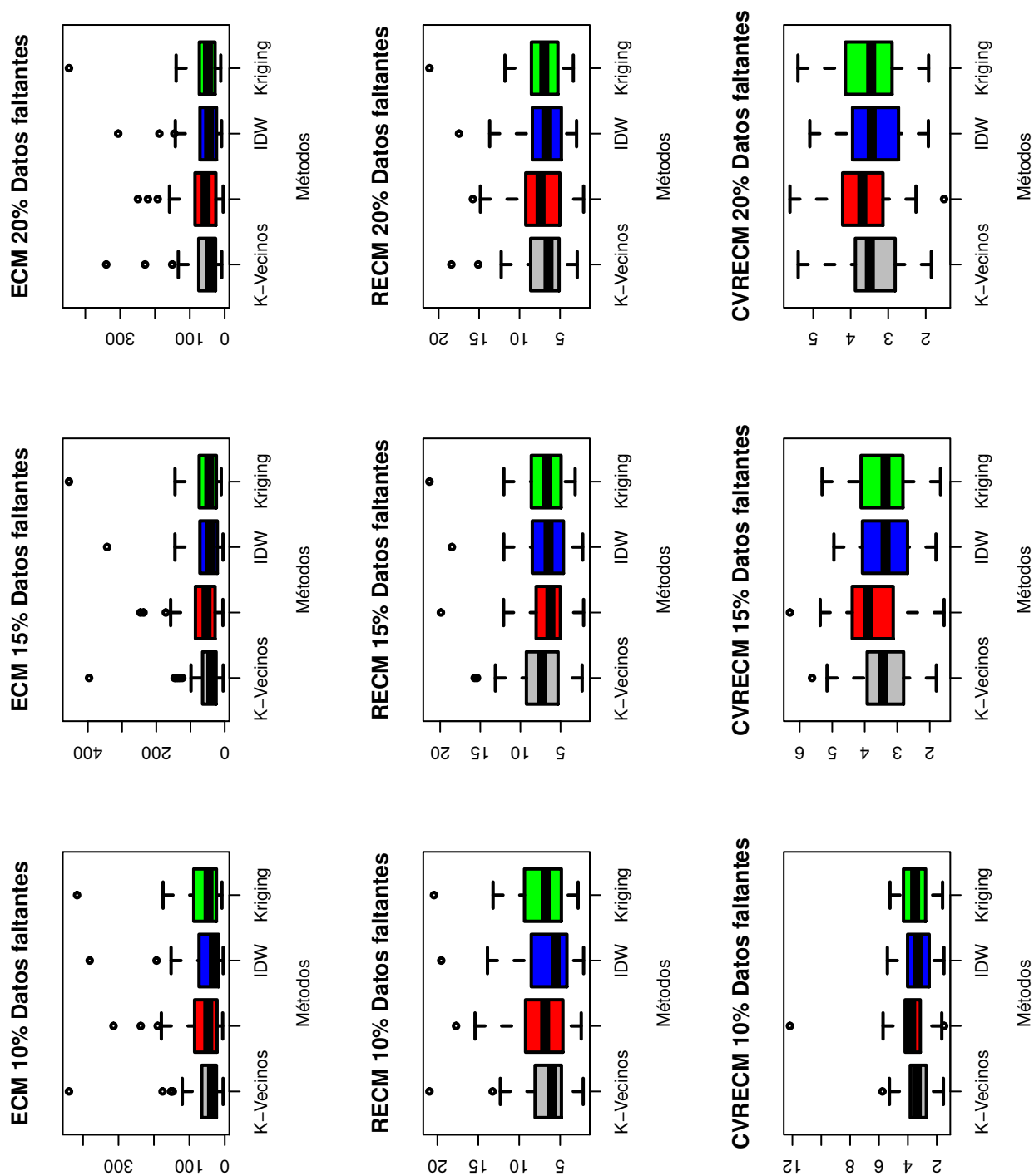


Figura 10.4: Comparación de métodos por medio de Boxplot Montaña

A continuación se presentarán los códigos de las metodologías implementadas en el software estadístico R, para la estimación de datos faltantes en bases de datos de precipitación:

10.1. KV

Código para estimar datos faltantes en una base de datos con el método K-Vecinos cercanos.

```
estimate_row=function(row,distan,k){

matrix_dist=as.matrix(distan)
pos=which(is.na(row))
pos_kcercanos=lapply(pos,function(x) names(sort(matrix_dist[-pos,x]))[1:k])
estimation=unlist(lapply(pos_kcercanos,function(x) mean(as.matrix(row[names(row)%in%x]))))
row[pos]=estimation
return(row)
}
```

10.2. Regresión

Código para estimar datos faltantes en una base de datos con el método de Regresión Lineal.

```
regresion=function(data){

y_estimation=list()
resul=matrix(dim(data)[2],3)
count=0

for(i in 1:dim(data)[2]){

Y=data[,i]
nas_Y=is.na(Y)
sta_reference = names(data)[i]
R_squar="NA"
sta_estimation="NA"

if(sum(nas_Y)!=0){

pos_na=which(nas_Y)
ind=1:dim(data)[2]
ind=ind[-which(ind==i)]
all_matrix=is.na(data[,ind])
all_pos=apply(all_matrix,2,function(x)if(sum(x)==0){return(0)}else{return(which(x))} )
station_all=lapply(all_pos,function(x) sum(pos_na %in% x))
station_model=data.frame(unlist(station_all))
station_model1=cbind(rownames(station_model)[station_model[,1]==0],station_model[station_model[,1]==0,])
pos_model=as.vector(sapply(station_model1[,1],function(x) grep(x,names(data))))
data_estimation=data[,pos_model]
bases_model=apply(data_estimation,2,function(x) return(na.omit(cbind.data.frame(Y,x))))
R_cuadrados=sapply(bases_model,function(x) summary(lm(x[,1]~x[,2]-1))$r.squared)
R_squar=max(R_cuadrados)
sta_estimation=names(which(R_cuadrados==R_squar))
Beta=coefficients(lm(Y~data[,names(data)==sta_estimation]-1))
data_x=data[pos_na,grep(sta_estimation,names(data))]
count=count+1
y_estimation[[count]]=Beta*data_x

}

resul[i,]=cbind(sta_reference,sta_estimation,R_squar)

}

colnames(resul)=c("Est_Y","Est_X","R")
all_output=list(resul,y_estimation)
return(all_output)
}
```

10.3. IDW

Código para estimar datos faltantes en una base de datos con el método IDW.

```

library(sp)
library(gstat)

idw_estimate=function(date){

  base=t(date)
  colnames(base)=c("Lon","Lat",paste0("Day",1:365))
  pos=which(is.na(base),arr.ind = TRUE)[,2]

  count=0
  estimates=0

  for(i in pos ){

    base_sin=as.data.frame(base[is.na(base[,i])==FALSE,c(1,2,i)])
    base_sin$x <-base_sin$Lon
    base_sin$y <-base_sin$Lat
    coordinates(base_sin) = ~x + y

    grd <- expand.grid(x = base[is.na(base)[,i]==TRUE,1], y = base[is.na(base)[,i]==TRUE,2])
    coordinates(grd) <- ~x + y
    count=count+1
    estimates[count] <- idw(formula = base[is.na(base)[,i]==FALSE,i]~1, locations = base_sin, newdata = grd)$var1.pred

  }

  return(estimates)
}

```

10.4. K-O

Código para estimar datos faltantes en una base de datos con el método Kriging Ordinario.

```

library(gstat)
library(geoR)
library(aqfig)

coord=t(base[1:2,-1:-3])
clim=t(aggregate(base,list(Quin=base$quincena),sum))
climatologia=cbind.data.frame(coord,clim[-1:-4,-1])
all_clim=lapply(3:26,function(x) as.geodata(climatologia,coords.col=1:2,data.col=x) )
all_vario=lapply(all_clim,function(x) variog(x, uvec = seq(0,1.4,1=17)))

model=list()

#Valle

model[[1]] <- variofit(all_vario[[1]], cov.model = "exponential", ini = c(10443.9,0.5188314), weights="equal")
model[[2]] <- variofit(all_vario[[2]], cov.model = "exponential", ini = c(15410.14,0.6162727), weights="equal")
model[[3]] <- variofit(all_vario[[3]], cov.model = "spherical", ini = c(1398.667,0.6126637), weights="equal")
model[[4]] <- variofit(all_vario[[4]], cov.model = "exponential", ini = c(2471.976,0.5260493), weights="equal")
model[[5]] <- variofit(all_vario[[5]], cov.model = "exponential", ini = c(7566.14,0.5332671), weights="equal")
model[[6]] <- variofit(all_vario[[6]], cov.model = "exponential", ini = c(9525.313,0.5188314), weights="equal")
model[[7]] <- variofit(all_vario[[7]], cov.model = "exponential", ini = c(25674.48,0.5296582), weights="equal")
model[[8]] <- variofit(all_vario[[8]], cov.model = "exponential", ini = c(12335.11,0.5296582), weights="equal")
model[[9]] <- variofit(all_vario[[9]], cov.model = "exponential", ini = c(7471.427,0.5332671), weights="equal")
model[[10]] <- variofit(all_vario[[10]], cov.model = "exponential", ini = c(9223.078,0.5260493), weights="equal")
model[[11]] <- variofit(all_vario[[11]], cov.model = "exponential", ini = c(19400.04,0.5260493), weights="equal")
model[[12]] <- variofit(all_vario[[12]], cov.model = "exponential", ini = c(7303.845,0.5260493), weights="equal")
model[[13]] <- variofit(all_vario[[13]], cov.model = "exponential", ini = c(1347.636,0.5188314), weights="equal")
model[[14]] <- variofit(all_vario[[14]], cov.model = "exponential", ini = c(2811.931,0.5260493), weights="equal")
model[[15]] <- variofit(all_vario[[15]], cov.model = "exponential", ini = c(8344.925,0.5224403), weights="equal")
model[[16]] <- variofit(all_vario[[16]], cov.model = "exponential", ini = c(1860.077,0.5296582), weights="equal")
model[[17]] <- variofit(all_vario[[17]], cov.model = "exponential", ini = c(9131.21,0.5332671), weights="equal")
model[[18]] <- variofit(all_vario[[18]], cov.model = "spherical", ini = c(3963.4134,0.5188314),nugget= 561.8301, weights="equal")
model[[19]] <- variofit(all_vario[[19]], cov.model = "spherical", ini = c(3951.643,0.5296582),nugget=550.0599, weights="equal")
model[[20]] <- variofit(all_vario[[20]], cov.model = "exponential", ini = c(24262.79,0.5296582), weights="equal")
model[[21]] <- variofit(all_vario[[21]], cov.model = "exponential", ini = c(14509.17,0.5332671), weights="equal")
model[[22]] <- variofit(all_vario[[22]], cov.model = "exponential", ini = c(22932.61,0.5260493), weights="equal")
model[[23]] <- variofit(all_vario[[23]], cov.model = "exponential", ini = c(18395.77,0.5260493), weights="equal")
model[[24]] <- variofit(all_vario[[24]], cov.model = "exponential", ini = c(26660.26,0.5260493), weights="equal")

##MONTAÑA

model[[1]] <- variofit(all_vario[[1]], cov.model = "exponential", ini = c(3009.414,0.352820317),nugget=770.426, weights="equal")
model[[2]] <- variofit(all_vario[[2]], cov.model = "spherical", ini = c(6693.749,0.3492113806),nugget=1829.602, weights="equal")
model[[3]] <- variofit(all_vario[[3]], cov.model = "gaussian",ini = c(2363.7647,0.356429253),nugget=638.1091, weights="equal")
model[[4]] <- variofit(all_vario[[4]], cov.model = "gaussian",ini = c(672.7936,0.262596907),nugget=410.7635, weights="equal")

```

```

model[[5]] <- variofit(all_vario[[5]], cov.model = "exponential", ini = c(3207.4478,0.349211381),nugget= 567.8781, weights="equal")
model[[6]] <- variofit(all_vario[[6]], cov.model = "spherical", ini = c(2257.7010,0.175982433),nugget=651.3019, weights="equal")
model[[7]] <- variofit(all_vario[[7]], cov.model = "spherical", ini = c(2257.7010,0.175982433),nugget=651.3019, weights="equal")
model[[8]] <- variofit(all_vario[[8]], cov.model = "gaussian", ini = c(3083.455,0.1759824327),nugget=1307.346, weights="equal")
model[[9]] <- variofit(all_vario[[9]], cov.model = "exponential",ini = c(2261.1537, 0.702887149),nugget=376.3808, weights="equal")
model[[10]] <- variofit(all_vario[[10]], cov.model = "gaussian",ini = c(3165.4567,0.793110559),nugget=632.7635, weights="equal")
model[[11]] <- variofit(all_vario[[11]], cov.model = "gaussian",ini = c(3165.4567,0.793110559),nugget=632.7635, weights="equal")
model[[12]] <- variofit(all_vario[[12]], cov.model = "gaussian",ini = c(839.8849,0.352968499),nugget=321.0523, weights="equal")
model[[13]] <- variofit(all_vario[[13]], cov.model = "exponential", ini = c(92.76812,0.262596907),nugget=11.03665, weights="equal")
model[[14]] <- variofit(all_vario[[14]], cov.model = "exponential", ini = c(66.51858,0.262596907),nugget=47.87722, weights="equal")
model[[15]] <- variofit(all_vario[[15]], cov.model = "exponential", ini = c(10.112762,0.269814779),nugget=2.271853, weights="equal")
model[[16]] <- variofit(all_vario[[16]], cov.model = "exponential", ini = c(80.69357,0.4394348), weights="equal")
model[[17]] <- variofit(all_vario[[17]], cov.model = "exponential", ini = c(80.69357,0.4394348), weights="equal")
model[[18]] <- variofit(all_vario[[18]], cov.model = "mat",ini = c(2674.415,0.183200306),nugget= 561.8301, weights="equal")
model[[19]] <- variofit(all_vario[[19]], cov.model = "gaussian",ini = c(2170.7220,0.352820317),nugget=578.9632, weights="equal")
model[[20]] <- variofit(all_vario[[20]], cov.model = "spherical", ini = c(1180.5288,0.1832003055),nugget= 576.1625, weights="equal")
model[[21]] <- variofit(all_vario[[21]], cov.model = "exponential", ini = c(13257.793, 0.175982433),nugget=2885.736, weights="equal")
model[[22]] <- variofit(all_vario[[22]], cov.model = "gaussian",ini = c(3218.613,0.360038190),nugget=1068.890, weights="equal")
model[[23]] <- variofit(all_vario[[23]], cov.model = "exponential", ini = c(951.4718, 0.172373496),nugget=199.8059, weights="equal")
model[[24]] <- variofit(all_vario[[24]], cov.model = "exponential", ini = c(319.8803,0.4286079817),nugget=123.0167, weights="equal")

all_missing=lapply(4:dim(base)[2],make_missing,base,por=0.2)

parameters1=function(ind,data,model){

  without_miss_day=as.geodata(t(data[c(1,2,ind[1]),c(-1:-3,-ind[2])]),coords.col = 1:2,data.col = 3)
  month=data[ind[1],1]
  coor_miss=as.matrix(t(data[1:2,ind[2]]))
  p<-krweights(without_miss_day$coords, coor_miss,krige = krige.control(obj.m =model[[month]]))
  p[p<0]=0
  pesos=p/sum(p)
  est=sum(without_miss_day$data*pesos)
  return(est)

}

fit=function(data_missing,model){

  index=which(is.na(data_missing),arr.ind = TRUE)
  index_l=split(index,1:dim(index)[1])
  r=lapply(index_l,parameters1,data_missing,model)
  return(r)

}

Res<-lapply(all_missing,fit,model)

```