UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Programa de Doctorado en Ingeniería Ambiental

Análisis composicional del acceso a servicios de agua, saneamiento e higiene y medida de sus desigualdades

Tesis doctoral realizada por:

Filimon Alejandro Quispe Coica

Dirigida por:

Agustí Pérez Foguet

Departamento de Ingeniería Civil y Ambiental

Barcelona, octubre 2021

ABSTRACT

Access to water, sanitation and hygiene (WASH) services is closely linked to public health. The benefits of having safe WASH are vast. The WHO/UNICEF Joint Monitoring Program (JMP) estimates of the population accessing different levels of WASH services are, by definition, constant sum; they are therefore compositional data. The JMP strives to generate accurate global and regional estimates. However, the methodological alternative is far from being the best for all possible situations at the global level, an issue that becomes more evident in the presence of data with non-linear patterns.

Therefore, the objective of this research thesis has been to incorporate new statistical alternatives to monitor population access to WASH services and to propose a new measure of inequality applied to the urban-rural comparison. In order to test and validate the new analysis techniques proposed, we have used case studies in multiple countries in both rural and urban contexts. It has also been validated on a broader spectrum by complementing the analysis with the sub-national level of a particular country.

In detail, in Chapter IV the methods of imputation of irregular data have been evaluated using two different options, in Chapter V a new measure of inequality in tripartite hygiene information is proposed; Chapter VI integrates the analysis algorithm for composite data that takes into account the pre-processing of the data, the robust fit of the model (linear and non-linear) and the uncertainty of the data; the performance of the algorithm is validated in the subnational context.

The results obtained justify differentiated analysis strategies both for minimum data (< 6) and for data greater than or equal to six. Fitting using robust models has better performance compared to the usual ones (that is, in non-robust techniques). The new measure of inequality has a unique value and is applicable for public policies and sector investments with a general vision. Finally, the compositional analysis on the information on access to WASH services is consolidated, which has been tested and validated at the subnational level in a wider range of possible situations than at the global level (carried out in Chapter IV). Therefore, it is expected that the application of the proposals made in this research will contribute to improving data analysis in the WASH sector and that future research will also go in that direction.

RESUMEN

El acceso a servicios de agua, saneamiento e higiene (ASH) está estrechamente relacionado con la salud pública. Lo beneficios de tener un ASH seguro son amplios. Las estimaciones del Programa Conjunto OMS/UNICEF de Monitoreo (PCM) de la población que accede a diferentes niveles de servicios ASH son, por definición, una suma constante; por tanto, son datos de composición. El PCM se esfuerza por generar estimaciones globales y regionales precisas. Sin embargo, la alternativa metodológica dista de ser la mejor para todas las situaciones posibles a nivel global, cuestión que se hace más evidente ante la presencia de datos con patrones no lineales.

Por lo tanto, el objetivo de esta tesis de investigación ha sido incorporar nuevas alternativas estadísticas para monitorear el acceso de la población a los servicios de ASH y proponer una nueva medida de desigualdad aplicada a la comparación urbano-rural. Con la finalidad de testear y validar las nuevas técnicas de análisis que se propone, hemos utilizado estudios de caso en múltiples países en contextos tanto rurales como urbanos. También se ha validado en un espectro más amplio al complementar el análisis con el nivel subnacional de un país en concreto.

En detalle, en el Capítulo IV se han evaluado los métodos de imputación de datos irregulares utilizando dos opciones diferentes, en el Capítulo V se propone una nueva medida de desigualdad en la información tripartita sobre higiene; El Capítulo VI integra el algoritmo de análisis de datos compositivos que toma en cuenta el preprocesamiento de los datos, el ajuste robusto del modelo (lineal y no lineal) y la incertidumbre de los datos; el desempeño del algoritmo es validado en el contexto subnacional.

Los resultados obtenidos justifican estrategias de análisis diferenciadas tantos para datos mínimo (< 6) como para datos mayores o iguales a seis. El ajuste mediante modelos robustos tiene un mejor rendimiento en comparación con los habituales (es decir, sobre las técnicas no robustas). La nueva medida de desigualdad es de valor único y es aplicable para políticas públicas e inversiones sectoriales con una visión general. Finalmente, se consolida el análisis composicional sobre la información de acceso a los servicios de ASH, lo cual ha sido probado y validado a nivel subnacional en una gama más amplia de situaciones posibles que en el nivel global (realizado en el Capítulo IV). Por lo tanto, se espera que la aplicación de las propuestas realizadas en esta investigación contribuyan a mejorar el análisis de datos en el sector ASH y que las investigaciones futuras también vayan en esa dirección.

AGRADECIMIENTOS

Un agradecimiento especial a mi esposa Yuly por su apoyo incondicional en todo el proceso y que a pesar de la distancia siempre estuvo para escucharme y motivarme. A mi madre Claudia por sus sabios consejos. A mi hermano Ángel y a mis hermanas por sus palabras de aliento.

También agradezco al Dr. Agustí por sus *enseñanzas*, *paciencia* y *motivación* en todo el proceso de doctorado. Al grupo de investigación "Sciences and Global Development (EScGD)" por acogerme en sus instalaciones. Por último, al Programa Nacional de Becas del Perú (PRONABEC; Beca Presidente de la República), por la financiación de mi estancia y mi estudio.

Sin todos estos insumos no hubiese sido posible la culminación de este trabajo.

LISTA DE PUBLICACIONES DERIVADAS DEL TRABAJO DE TESIS

Artículos publicados en revistas indexadas:

Quispe-Coica, A., Pérez-Foguet, A., 2020b. Preprocessing alternatives for compositional data related to water, sanitation and hygiene. Sci. Total Environ. 743, 140519. https://doi.org/10.1016/j.scitotenv.2020.140519

Quispe-Coica, A., Pérez-Foguet, A., 2021. A new measure of hygiene inequality applied to urban-rural comparison. Int. J. Hyg. Environ. Health. (En producción con N° IJHEH-113876)

Artículos publicados en otras revistas:

Quispe-Coica, A., Fernández, S., Acharte Lume, L., Pérez-Foguet, A., 2020. Status of Water Quality for Human Consumption in High-Andean Rural Communities: Discrepancies between Techniques for Identifying Trace Metals. J — Multidiscip. Sci. J. 3, 162–180. https://doi.org/10.3390/j3020014

Trabajos presentados en congresos nacionales e internacionales:

Presentación oral:

Quispe-Coica, A., Pérez-Foguet, A., 2020a. Desigualdades de acceso al agua y al saneamiento según quintiles de riqueza: un análisis de tendencias en datos compositivos, in: 2nd Latin American and Caribbean Young Water Professional Conferences. Colombia, pp. 1–3.

Fernández-Alba, S., Magre-Vinardell, L., Álvarez-Pujol, C., Acharte-Lume, L., Quispe-Coica, A., Pérez-Foguet, A., 2019. Calidad del agua, contexto social e higiene en comunidades rurales alto-andinas de Huancavelica (Perú), in: VIII Congreso Universidad y Cooperación Al Desarrollo". Compostela.

Quispe-Coica, A., Pérez-Foguet, A., 2019. Joint evolution of access to water of urban and rural populations in South America through Compositional Data Analysis, in: Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019): Terrassa. pp. 130–142.

Quispe-Coica, A., Pérez-Foguet, A., 2018. Evolución del Acceso al Agua y Saneamiento en América del Sur Mediante Técnicas Estadísticas Composicionales, in: XXXVI Congreso Interamericano de Ingeniería Sanitaria y Ambiental. AIDIS, Guayaquil-Ecuador, pp. 753–757.

Póster:

Navarrete, D., Gonzales, W., Poma, C., Quispe-Coica, A., Pérez-Foguet, A., Meseguer, R., 2019. “Implementación de comunicación mediante LoRa para seguimiento de servicios de agua y saneamiento,” in: VIII Congreso Universidad y Cooperación Al Desarrollo. Compostela.

Manuscrito preliminar (en borrador):

Quispe-Coica, A., Pérez-Foguet, A., 2021. From the global to the subnational scale: landing the monitoring of drinking water and sanitation services.

Participación en proyectos de cooperación:

Título del proyecto: Comunicación dedicada para la gestión y monitoreo del agua en zona altoandina de Perú (Fase 1)
Código: 2019-B012
Responsable: Alejandro Quispe Coica
Período: 18-7-2019 al 31-07-2020

Monto de financiamiento del CCD: 15,126.04 €

Título del proyecto: Comunicación dedicada para la gestión y monitoreo del agua en zona altoandina de Perú (Fase 2)
Código: 2020-B004
Responsable: Alejandro Quispe Coica
Período: 30-7-2020 al 1-10-2021
Monto de financiamiento del CCD: 8,000.0 €

Participación en proyectos de investigación:

Título del proyecto: "Red de sensores inalámbricos en entorno remoto LPWAN para el monitoreo de parámetros de calidad de agua (Ph, turbidez, temperatura, colimetría) en los sistemas de abastecimiento de las comunidades rurales del distrito de Huancavelica"

Coinvestigadores UPC: Agustí Pérez Foguet (coinvestigador N°3), Roc Meseguer Pallares (coinvestigador N°4), Alejandro Quispe Coica (coinvestigador N°5)

Financiamiento de la Universidad Nacional de Huancavelica (UNH), Perú, a través de FOCAM y ejecutada conjuntamente por investigadores de la UPC y la UNH

Monto de financiamiento del CCD: S/ 150,000.00 (moneda peruana)

Período: 2021 al 2022 (actualmente en ejecución)

Referencia: 2017 SGR-01496-GRPRE, Generalitat de Catalunya, research group on Engineering Sciences and Global Development (EScGD)

Investigador principal: Agustí Pérez Foguet

Monto de financiamiento: € 1,500 euros

Período: Mayo de 2019 a Septiembre del 2021

Organización de conferencia:

Temática: Tecnologías inteligentes de monitoreo del agua aplicadas a contextos rurales – "Ciencia y tecnología para el desarrollo local"

Formato: virtual

Organizador responsable: Alejandro Quispe Coica
Fecha: 10-06-2021 al 30-06-2021
Descripción: Ponencia de 9 temas por los diferentes especialistas de Perú y España
Link: https://www.upc.edu/ccd/ca/noticies/conferencia-virtual-201ctecnologias-inteligentes-de-monitoreo-del-agua-aplicadas-a-contextos-rurales201d

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

ACRONIMOS Y ABREVIATURAS

ASH	Agua, saneamiento e higiene
AyS	Agua y saneamiento
OMS	Organización Mundial de la Salud
UNICEF	Fondo de las Naciones Unidas para la Infancia
PCM	Programa Conjunto OMS/UNICEF de Monitoreo
INEI	Instituto Nacional de Estadística e Informática
ENDES	Encuesta Demográfica y de Salud Familiar
ENAHO	Encuesta Nacional de Hogares
ENAPRES	Encuesta Nacional de Programas Presupuestales
ODS	Objetivos de Desarrollo Sostenible
ALC	América Latina y el Caribe
ODM	Objetivo de desarrollo del milenio
CoDa	Datos composicionales (Acrónimo en ingles de Compositional Data)
OLS	Acrónimo en ingles de Ordinary Least Squares
GAM	Acrónimo en ingles de Generalized Additive Model
ilr	Acrónimo en ingles de isometric log-ratio

CAPITULO I. INTRODUCCION

El acceso de la población a servicios de agua potable y saneamiento son fundamentales para el desarrollo, la salud y el bienestar de las personas. En 2016, la diarrea fue responsable de más de 1.6 millones de muertes en personas de todas las edades (Naghavi et al., 2017). A nivel regional, la diarrea fue responsable de más de 24 mil muertes en América Latina y el caribe y de más de 600 mil muertes en África subsahariana (Troeger et al., 2018). Muchas de las enfermedades y muertes en los países de ingresos bajos y medianos se atribuyen al agua no potable (Prüss-Ustün et al., 2014). Asimismo, los principales factores de riesgo asociados con la diarrea en los niños menores de cinco años son el agua no potable, el saneamiento no seguro y la emaciación infantil (puntuación baja de peso para la altura) (Troeger et al., 2018). Por tanto, todo lo que implica al sector WASH siempre será materia de análisis y discución.

La Organización Mundial de la Salud (OMS) y el Fondo de las Naciones Unidas para la Infancia (UNICEF) realizan el Programa Conjunto de Monitoreo (PCM) de la población que accede a diferentes niveles de servicios de agua y saneamiento (AyS) desde 1990 (Bartram et al., 2014). Oficialmente el monitoreo global de servicios de AyS nace con la declaración del Milenio de las Naciones Unidas, firmada en septiembre de 2000 (Naciones Unidas, 2000); en el cual los países miembros se comprometen a “reducir a la mitad el porcentaje de personas que carezcan de acceso a agua potable o que no puedan costearlo” (Naciones Unidas, 2000) para el año 2015. A este período se le llamó Objetivos de desarrollo del Milenio (ODM).

En este período, el seguimiento del acceso a AyS estuvo dentro del objetivo 7, “garantizar la sostenibilidad del medio ambiente”. El seguimiento se basó en cuatro niveles de servicios, también llamadas escaleras de agua y saneamiento. En el caso del acceso a servicios de agua, las catagorías de seguimiento han sido: agua entubada en las instalaciones (hogares y locales), otras fuentes mejoradas, fuentes no mejoradas y acceso a agua superficial. En el caso del acceso a servicios de saneamiento, las categorías de seguimiento han sido: mejorada, compartidas, fuentes no mejoradas y defecación al aire libre.

Al finalizar 2015, el balance fue positivo para el progreso de los servicios de agua (que alcanzó una cobertura del 91%), ya que se superó la meta global del ODM del 88% de la población en acceso a una fuente mejorada de agua potable (i.e., agua entubada y por otras fuentes mejoradas). En términos de población, desde 1990 hasta 2015, 2,600 millones de personas han tenido acceso a una fuente de agua potable mejorada (OMS/UNICEF, 2015). Mientras que, en saneamiento, no se cumplió la meta global de acceso a instalaciones mejoradas del 77% y tan solo se llegó a la cobertura del 68%. Aun así, el seguimiento global de los ODM ha sido positivo porque los países se esforzaron por cumplir con las metas estipuladas en cada uno de ellos y, además, se generó considerable información de AyS. También se genera una discusión sobre las deficiencias en los indicadores de seguimiento de agua potable y saneamiento (Bain et al., 2012; Onda et al., 2012; Bartram et al., 2014; Weststrate et al., 2018).

Post 2015, la asamblea general de las Organización de las Naciones Unidas (ONU) también adopta la Agenda 2030. La Agenda 2030 es un ambicioso plan de acción promovido por las Naciones Unidas y tiene el espíritu de “no dejar a nadie atrás” (United Nations General Assembly, 2015). La inclusión de todas las personas para alcanzar el objetivo global en 2030 es el motor que impulsa a todos los países adherentes. Aborda 17 Objetivos de Desarrollo Sostenible (ODS), de los cuales nuestro interés en este estudio es el ODS 6.1 y 6.2. relacionado con el acceso a

servicios de agua y saneamiento/higiene, respectivamente. La relevancia de los ODS 6.1 y 6.2, radica porque transversalmente está vinculado con otros objetivos de la Agenda 2030 (UN-Water, 2016; Requejo-Castro et al., 2020).

Las metas y objetivos trazados en los ODS son ambiciosos y tienen como fecha límite 2030. No obstante, ya hay alertas que saltan de que el mundo no está en camino de alcanzar el objetivo trazado sobre agua limpia y saneamiento para el 2030 (Sadoff et al., 2020). África es un claro ejemplo de que 2030 es demasiado pronto y que no lograrán alcanzar las metas planteadas si no movilizan agresivamente recursos para lograr servicios universales de ASH (Nhamo et al., 2019).

El monitoreo ha evolucionado de los Objetivos de Desarrollo del Milenio (ODM) a los Objetivos de Desarrollo Sostenible (ODS), y se ha expandido de cuatro a cinco escaleras de seguimiento. Las nuevas escaleras de seguimiento para agua, son: servicio gestionado de manera segura, servicio básico, servicio limitado, instalación no mejorada y acceso a gua superficial. La categoría "gestionado de forma segura", mide el progreso de las poblaciones urbanas y rurales que acceden a una fuente mejorada ubicada en las instalaciones que está disponible cuando es necesario y libre de contaminación. El servicio básico está relacionado con el acceso a agua mejorada, pero con un tiempo de ida y vuelta para recoger agua de menos de 30 minutos; mientras que, si la recolección de agua de una fuente mejorada excede los 30 minutos, se clasifica como un servicio limitado. La instalación no mejorada, se refiere al acceso a fuentes de agua no mejoradas, como beber agua de un pozo excavado o de un manantial sin protección. Finalmente, el acceso a agua superficial, se refiere al acceso a beber agua directamente de un río , estanque, canal, etc.

En el caso de saneamiento, son igual cinco escaleras de servicio: servicio gestionado de manera segura, servicio básico, servicio limitado, instalación no mejorada y defecación al aire libre. Es también una novedad la incorporación de Higiene al monitoreo global, en específico dentro del ODS 6.2. El monitoreo de la higiene se hace sobre información tripartita: servicio básico (instalación disponible en el hogar para lavarse las manos con agua y jabón), servicio limitado (disponibilidad de una instalación para lavarse las manos sin jabón y/o agua) y sin instalaciones (no hay instalaciones para lavarse las manos en el lugar).

Por otro lado, el ODS 10 está relacionado con "reducir la desigualdad en y entre los países". El monitoreo de las desigualdades ha cobrado cada vez más relevancia, dado que es fundamental para lograr el acceso universal. En materia de agua y saneamiento, el seguimiento de la desigualdad está implícita en las metas del ODS 6.1 y 6.2. Para agua, la meta del ODS 6.1 es "*para 2030, lograr el acceso universal y equitativo al agua potable a un precio asequible para todos*". Mientras que para el saneamiento, la meta del ODS 6.2 es "*para 2030, lograr el acceso a servicios de saneamiento e higiene adecuados y equitativos para todos* …". También hay una extensiva literatura que pone en manifiesto las implicancias de la desigualdad en el progreso del monitoreo global (Bain et al., 2014; Cetrulo et al., 2020).

Desde el punto de vista del monitoreo, también es la PCM la que se encarga de monitorear la desigualdad en el acceso a ASH de manera rutinaria desde diferentes aspectos, siempre que se disponga de información. Las principales características del monitoreo están relacionadas con: desigualdad por área de residencia entre urbano y rural, por quintiles de riqueza, regiones subnacionales y según niveles de servicio de la escalera ASH. Si bien ha habido cambios en el seguimiento de ASH desde los ODM hasta los ODS, se sigue manteniendo el método de análisis estadístico sobre datos con suma constante del 100%, que detallamos en los siguientes ítems.

1.1. Monitoreo global del acceso a agua, saneamiento e higiene

El monitoreo internacional del acceso a los servicios ASH se realiza estimando el porcentaje de población con niveles de servicio definidos. Los niveles de servicio de ASH suman 100% —si son porcentajes o 1 si son proporciones— y, por tanto, son datos composicionales (John Aitchison, 1986; Egozcue y Pawlowsky-Glahn, 2005; Lloyd et al., 2012; Pérez-Foguet et al., 2017). Si bien las partes del total han variado históricamente (por ejemplo, las categorías utilizadas para monitorear el sector de agua y saneamiento se expandieron de cuatro a cinco), siempre se han enfocado en la evolución del porcentaje de población con acceso a los diferentes niveles de servicio.

Por otra parte, las estimaciones de los niveles de servicio se hacen mediante la regresión lineal de mínimos cuadrados ordinarios (OLS, acrónimo en inglés de ordinary least squares), acondicionado a una serie de reglas —mirar en detalle en WHO/UNICEF (2018a)—. La alternativa de regresión lineal para generar las estimaciones ha sido cuestionado por varios autores (Wolf et al., 2013; Yerg, 2013; Yerg et al., 2013; Bartram et al., 2014; Fuller et al., 2016; Pérez-Foguet et al., 2017). El principal problema de la regresión lineal es que, en presencia de datos no lineales, tiende a subestimar o sobreestimar las estimaciones (Fuller et al., 2016).

Otro problema es que en países que han progresado significativamente en dotar servicios ASH mejorados (que comúnmente ocurre en países de mediano a alto ingreso), la data tiende a valores extremos (extremo inferior con valor cero y extremo superior con valor de 100%), por lo que las reglas de extrapolación de la regresión lineal pueden llegar a superar los valores extremos del todo o nada. En este caso, el PCM fija las estimaciones en el valor de cero y 100% (WHO/UNICEF, 2018). Además, en el modelo lineal, es probable que haya un mayor sesgo cuando las tasas de cobertura se acercan al 0% o al 100% (Bartram et al., 2014).

Debido a los problemas precitados, en la literatura muchos autores discuten y plantean la necesidad de cambiar la regresión lineal por métodos alternativos para modelar mejor las tendencias temporales y reducir el sesgo inherente (Bartram et al., 2014; Fuller et al., 2016; Pérez-Foguet et al., 2017; Wolf et al., 2013). Yerg et al. (2013) propone la regresión logística y Wolf et al. (2013) el modelado multinivel. También a finales del 2014 el grupo de trabajo de la PCM celebró debates plenarios para discutir sobre el método OLS y otros desafíos relacionados con el método de monitoreo de la PCM como el error de muestreo y la incertidumbre de la data (WHO/UNICEF, 2014).

Fuller et al. (2016) propone modelar las tendencias temporales con modelos aditivos generalizados (GAM; acrónimo en inglés de Generalized additive model), pero con un mínimo de datos. La alternativa es más solida y funciona bien en presencia de datos no lineales. No obstante, i) las técnicas estadisticas aplicadas son las habituales y no para datos compositivos, y ii) los datos con tendencia a los valores extremos siguen siendo un problema, ya que hay países en el que las estimaciones (incluso haciendo modelos con GAM) llegan a superar el extremo superior del 100% y en el extremo inferior dar valores negativos.

Las características compositivas de los datos recién son tomados en cuenta en el estudio de Pérez-Foguet et al. (2017), quienes introducen técnicas estadísticas para datos compositivos como alternativa estadística para el monitoreo global de acceso a servicios de ASH. La nueva alternativa ofrece ventajas en datos con tendencias a los valores extremos y, como consecuencia, la interpolación y extrapolación de los modelos no supera los límites extremos.

Seguidamente se complementa con el estudio de Ezbakhe and Pérez-Foguet (2019) quienes introducen la incertidumbre de la data acoplado al modelo; todo esto para datos compositivos.

A pesar de los avances estadísticos, el PCM sigue usando el método de regresión lineal para estimar el acceso de la población a servicios de ASH. La principal justificación para esto es que, i) “para muchos países, y especialmente para los indicadores de nivel de servicio, no hay suficientes puntos de datos para justificar el uso de métodos no lineales” y ii) “las técnicas no lineales también son limitada en su capacidad de extrapolar incluso unos pocos años, lo que a menudo es necesario para el método de estimación del PCM ” (WHO/UNICEF, 2018).

Finalmente, la cantidad de información ha aumentado en algunos países, lo que se presenta como una oportunidad para aplicar técnicas estadisticas no lineales como GAM.

1.2. Monitoreo de la desigualdad en ASH

El monitoreo de ASH ha evolucionado sustancialmente en los últimos 20 años. Un punto clave es el paso del uso de indicadores únicos de desempeño (como la cobertura de agua y saneamiento por tecnologías mejoradas y no mejoradas) a marcos multidimensionales que entienden ASH en relación con conceptos como pobreza (Giné-Garriga y Pérez-Foguet, 2013a, 2019) y derechos humanos (Baquero et al., 2015; Giné-Garriga et al., 2017), o desde la perspectiva de grupos vulnerables y marginados (Redman-Maclaren et al., 2018; Ezbakhe et al., 2019; Anthonj et al., 2020a). La integración de estos conceptos conduce a una complejidad mucho mayor que la simple cobertura de una población mediante una solución técnica u otra.

Esta naturaleza multidimensional se midió primero a través de indicadores agregados como el índice de pobreza de ASH (Giné-Garriga y Pérez-Foguet, 2013a, 2013b) que extendió la propuesta seminal del Índice de Pobreza Hídrica (Sullivan, 2002; Sullivan et al., 2003; Giné-Garriga y Pérez-Foguet, 2010; Pérez-Foguet y Giné-Garriga, 2011). Asimismo, algunas limitaciones de los indicadores agregados, como la compensabilidad entre dimensiones y la falta de mecanismos para considerar influencias cruzadas entre dimensiones, se han abordado con diferentes técnicas (Ezbakhe y Pérez-Foguet, 2018; Giné-Garriga et al., 2018), principalmente dentro del enfoque de apoyo a procesos específicos de toma de decisiones.

Algunas de estas ideas están actualmente integradas en el monitoreo de la escalera de ASH impulsado por el PCM, que ha pasado de una perspectiva de cobertura a un enfoque de nivel de servicio. El grupo de trabajo sobre el monitoreo de las desigualdades, mediante un informe recomendó la necesidad de revisar varias formas de visualización de los datos para mostrar varios tipos de desigualdades, la necesidad de análisis desagregados y más (WHO/UNICEF, 2015).

Actualmente, el seguimiento de la desigualdad se basa en datos desagregados. En la medida de lo posible, las estimaciones de la PCM suelen desglosarse de forma rutinaria por zona de residencia (urbana y rural), por quintiles de riqueza, por regiones subnacionales y por niveles de servicio. El método de monitoreo de la desigualdad urbano-rural se hace mediante una simple diferencia entre las partes (WHO/UNICEF, 2019a, 2021). Esta forma de comparar es muy habitual también en la literatura del sector (Hasan y Alam, 2020).

Sin embargo, en cualquiera de las características de monitoreo precitadas: i) el marco básico para el monitoreo local e internacional aún sigue necesitando hacer tendencias temporales, ii) el ajuste del modelo se hace en indicadores primarios estipulados por el PCM (WHO/UNICEF, 2018), cuya característica particular es que describen las partes de un todo con suma constante

del 100% y iii) tanto en el monitoreo multidimensional como en el monitoreo de desigualdad realizado por el PCM, no se aborda la naturaleza compositiva de los datos, lo que puede conducir a correlaciones espurias entre las partes (Pérez-Foguet et al., 2017).

1.3. Datos composicionales

Los datos de composición son multivariantes por naturaleza (van den Boogaart y Tolosana-Delgado, 2013a) y se definen como partes de un todo que transmiten información relativa (John Aitchison, 1986; Egozcue y Pawlowsky-Glahn, 2011). El espacio muestral de los datos de composición es el simplex S^D representada en Ec. (1.1).

La composición $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, ..., \mathrm{X}_\mathrm{D})$ con *D* partes, son números reales que forman vectores estrictamente positivos y donde k también es una constante positiva. En el caso del monitoreo de acceso a ASH el valor de k es de 100% si son porcentajes o 1 si son proporciones.

$$S^D = \left\{ \mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, ..., \mathrm{X}_\mathrm{D}) : \forall \mathrm{X}_i > 0,\ i = 1, 2, 3, ..., D;\ \sum_{i=1}^{D} \mathrm{X}_i = k \right\} \tag{1.1}$$

Para aplicar cualquier técnica estadística habitual, es necesario primero hacer transformaciones log-cociente: de razón logarítmica aditiva (alr; acrónimo en inglés de additive log-ratio) representada en Ec. (1.2) o de razón logarítmica central (clr; acrónimo en inglés de centered log-ratio) representada en Ec. (1.3), ambos propuesto por Aitchison (1986); o la razón logarítmica isométrica (ilr; acrónimo en inglés de isometric log-ratio) de Egozcue et al. (2003) representada en Ec. (1.4).

De los precitados, la transformación ilr es la más sólida, ya que conserva todas las propiedades métricas de los datos compositivos como la isometría, matriz de covarianzas no singular, etc. (Egozcue et al., 2003); superando así los problemas que tienen las transformaciones alr y clr (más detalle ver Vera Pawlowsky-Glahn & Buccianti (2011)) para una aplicación generalizada. Sin embargo, para hacer un biplot compositivo, es recomendable utilizar coordenadas clr (Daunis-i-Estadella et al., 2011).

$$\mathrm{alr}(\mathrm{x}) = \left[\log\frac{\mathrm{x}_1}{\mathrm{x}_\mathrm{D}} \cdots \log\frac{\mathrm{x}_{\mathrm{D}-1}}{\mathrm{x}_\mathrm{D}} \right] \tag{1.2}$$

Donde D es el número de partes de X.

$$\mathrm{clr}(\mathrm{x}) = \left[\log\frac{\mathrm{x}_1}{\mathrm{g}(\mathrm{x})} \cdots \log\frac{\mathrm{x}_\mathrm{D}}{\mathrm{g}(\mathrm{x})} \right] \tag{1.3}$$

Donde g(x) la media geométrica de las partes de X.

$$\mathrm{ilr}(\mathrm{x}) = \sqrt{\frac{\mathrm{r_i} \times \mathrm{s_i}}{\mathrm{r_i} + \mathrm{s_i}}} \ln\frac{\mathrm{g_m}(\mathrm{x_r}+)}{\mathrm{g_m}(\mathrm{x_s}-)} \tag{1.4}$$

Donde "r" es el número de variables positivas en el balance V, "s" es el número de variables negativas en el balance V. gm () es la media geométrica de las variables X, los que están con signo positivo en la proporción superior (Xr +) y con un signo negativo en la proporción inferior (Xs −). Los balances V, son definidos siguiendo el procedimiento de partición binaria secuencial

(SBP, acrónimo en inglés de sequential binary partition) of Egozcue y Pawlowsky-Glahn (2005). Para definir el orden del SBP, hay una serie de alternativas, sin embargo, estudios anteriores muestran que es más apropiado por grupos afines y coherentes el análisis sectorial de ASH (Quispe-Coica y Pérez-Foguet, 2019).

Las alternativas descritas giran entorno al ratio entre sus partes (e.g., A/B), dónde el denominador de la razón (i.e, B) no puede ser un valor cero, tampoco puede ser un dato perdido; a este tipo de datos también se le conoce como datos irregulares (J. Aitchison, 1986; Egozcue et al., 2019). En situaciones como esta, es necesario hacer un tratamiento diferenciado de la data (Hron et al., 2010; Martín-Fernández et al., 2012; Quispe-Coica y Pérez-Foguet, 2020a) o sino excluir el vector que presentan estas irregularidades (Quispe-Coica y Pérez-Foguet, 2018).

CAPITULO II. OBJETIVOS Y METODOS

En esta sección, primero se define el problema de la investigación, para seguidamente concretar nuestros objetivos. Luego se detalla la metodología seguida y una breve descripción de los temas tratados en la tesis

2.1. Definición del problema de la investigación

Si bien las nuevas técnicas estadísticas para datos compositivos del sector ASH propuestos por Pérez-Foguet et al. (2017) y Ezbakhe y Pérez-Foguet, (2019) son más sólidas comparado con la regresión lineal del PCM, no pueden aplicarse a un amplio espectro de situaciones propias del sector, ya que las nuevas técnicas estadísticas giran en torno al log-cociente entre sus partes y cualquier valor en el denominador no puede ser de valor cero ni ser un dato faltante. Una alternativa sencilla para abordar esta situación es la exclusión del dato con estas irregularidades. No obstante, una exclusión *per se* puede llevar a la exclusión de ciertos países del análisis y ser perjudicial para el modelo, más aún si ya de por sí los países carecen de información (Quispe-Coica y Pérez-Foguet, 2018).

Por otro lado, la calidad de la información proveniente de diferentes fuentes de información entra en cuestión. Bain et al. (2018) indican que “el PCM debe seguir explorando métodos para evaluar la calidad de los datos de diversas fuentes”. Actualmente, el PCM hace una validación puntual de la información, excluye los que considera atípicos (ver pestaña “data summary” de cualquier base de datos del PCM; https://washdata.org/). No obstante, esta forma de validar es i) univariante y no sigue procedimientos estadísticos multivariantes (característica de los datos compositivos), lo cual aumenta el sesgo en la identificación de datos atípicos y ii) no se garantiza que los modelos no estén influenciados por los valores atípicos. En consecuencia, es probable que las estimaciones subestimen o sobrestimen los datos de un año en concreto. Un problema similar al manifestado por Fuller et al. (2016) en el caso de no aplicar GAM.

Tal como se dijo en el ítem 1.2 de la introducción, el grupo de trabajo sobre el monitoreo de las desigualdades recomendó la necesidad de revisar formas de visualización de los datos para mostrar distintos tipos de desigualdades (WHO/UNICEF, 2015). Actualmente, el monitoreo de la desigualdad se hace mediante una simple diferencia de categorías únicas entre urbano y rural (WHO/UNICEF, 2019a, 2021), lo cual no captura la característica multivariante de los datos.

La representación en mapas temáticos de las categorías de servicios y las medidas de desigualdad son también univariadas. Esta forma de analizar es muy común en la literatura del sector (WHO/UNICEF, 2019a, 2020). Esto implica que en un ranking de desigualdad un país puede ser el primero o el último según la categoría de análisis. También es probable que una interpretación del resultado de una sola categoría de servicio genere sesgo en la lectura, dado que los datos de ASH son composicionales e inherentemente multivariantes (van den Boogaart y Tolosana-Delgado, 2013a; Pérez-Foguet et al., 2017), donde un cambio en una de las categorías de servicio afecta una parte o el total restante.

En el espacio geométrico del simplex, existen pocas alternativas geométricas que nos permitan obtener una medida generalizada de desigualdad urbano-rural. Además, depende de la cantidad de información que exista. Cuando hay tres partes, los datos de composición se pueden representar en un diagrama ternario, cuando son cuatro partes en el tetraedro regular y para cinco o más partes, primero es necesario realizar transformaciones de razón logarítmica para aplicar cualquier técnica estadística común (Egozcue y Pawlowsky-Glahn, 2006;

Pawlowsky-Glahn y Egozcue, 2006). Una medida generalizada en cualquiera de las alternativas geométricas del simplex, son aún materia en estudio, más aún, en el sector ASH en el que hay poca literatura de aplicación de nuevas técnicas estadísticas para datos compositivos.

Finalmente, en la mayoría de países en desarrollo la descentralización de los servicios en la gestión de agua y saneamiento (WAS) urbano ha tendido hacia el nivel subnacional o compartían responsabilidades entre los gobiernos nacionales y subnacionales (Herrera y Post, 2014). Sin embargo, el monitoreo global rara vez considera la gobernanza a nivel subnacional o local (Herrera, 2019).

Muchos países de la región América Latina y el Caribe (ALC), por citar algunos, Perú (ver http://ods.inei.gob.pe/ods/), México (ver http://agenda2030.mx/index.html?lang=es#/home) y República Dominicana (ver http://ods.gob.do/Home/Inicio) están implementando plataformas de monitoreo de los ODS. No obstante, en sus plataformas de seguimiento se ha observado falta de información, falta de armonización de los indicadores globales con los subnacionales y más. Agregarle que las técnicas estadísticas aplicadas tienen que ser de acorde a la característica de los datos. En consecuencia, surge la oportunidad de desagregar la información global junto con sus indicadores de seguimiento para apoyar el seguimiento de los ODS a las instituciones encargadas en cada país y que los resultados sean útiles para los gestores subnacionales y locales del AyS.

De todo lo anterior se desprende que, para obtener resultados fiables en el seguimiento de los ODS, en concreto de los ODS 6.1-6.2, siguen siendo necesarias nuevas alternativas estadísticas que busquen resolver los problemas mencionados, y esto es lo que aporta esta investigación.

2.2. Objetivos

El objetivo general de la investigación presentada en esta tesis es:

- Incorporar y proponer nuevas alternativas estadísticas para monitorear el acceso de la población a los servicios de agua, saneamiento e higiene y al monitoreo de la desigualdad urbana-rural

Objetivos específicos:

- Cuantificar la presencia de datos irregulares del monitoreo global
- Evaluar alternativas de tratamiento de datos irregulares y generar modelos robustos
- Proponer nuevas alternativas de monitoreo de la desigualdad urbana-rural
- Validar los métodos propuestos en contextos subnacionales de un país en concreto

2.3. Método

El método seguido para cumplir con los objetivos propuestos consistió en cuatro pasos principales, que se describen brevemente a continuación:

El primer paso consistió en una revisión bibliográfica sobre cuatro temas principales, i) datos de composición, ii) monitoreo global del acceso a ASH, iii) seguimiento global de la desigualdad, iv) seguimiento de los ODS a nivel subnacional, con las siguientes palabras clave principales: CoDa, transformaciones log-ratio, correlación espuria, balance, imputación, estadística multivariante, métodos estadístico robustos, datos irregulares, valores atípicos, diagrama

ternario, biplot; datos composicionales en ASH, programa conjunto de monitoreo de la OMS y UNICEF, Agua potable, saneamiento e higiene, escalera de servicio, ODS, método de monitoreo del PCM, errores de muestreo, encuestas de hogares; monitoreo de la desigualdad, desigualdad urbana-rural, ODS 6.1 y 6.2, diagrama ternario, indicadores de desigualdad, lavado de manos, COVID-19, medidas multivariantes de desigualdad; ODS a nivel subnacional, ALC, etc.

Consultamos una amplia literatura de revistas científicas, resoluciones de acuerdos de organismos internacionales con los países suscritos, libros, actas de conferencias, reportes técnicos, informes globales publicados por el PCM, ONU, WHO, etc. También se ha participado en cursos relacionados con el análisis estadístico de datos composicionales (curso CoDa, CoDaWork 2019 - Terrassa), en foros académicos y en conferencias sobre la materia con el objetivo de consolidar el conocimiento de las nuevas técnicas estadísticas. Toda esta revisión y aprendizaje a aterrizado al marco teórico de esta tesis y en la introducción de los capítulos siguientes.

El segundo paso consistió en hacer una revisión integral de los datos de monitoreo global obtenidos de https://washdata.org/. La revisión consistió en cantidad, calidad y características de los datos. Esto nos permitió identificar los problemas comunes que surgen en los datos, que se abordan en el capítulo IV de esta tesis.

El tercer paso fue identificar y evaluar alternativas estadísticas para el preprocesamiento de datos irregulares y proponer un algoritmo del análisis estadístico robusto para el monitoreo de los servicios de ASH. Todo el cálculo computacional se realizó en R con sus diferentes paquetes, lo cual está a disposición del público en el repositorio de Zenodo (https://zenodo.org/communities/escgd). Se testeo en países con características que abarquen todas las situaciones posibles del sector ASH, con la finalidad de que el algoritmo sea una expresión general y no particular para un país en concreto.

El cuarto paso consistió en proponer una alternativa multivariante de desigualdad entre urbano y rural. Se construyó un marco geométrico de medida de desigualdad basado en el diagrama ternario. La información de Higiene se obtuvo de la base de datos PCM de 2017, con la que se validó la propuesta de la nueva medida de desigualdad en el conjunto de datos globales de Higiene. También se cruzó los resultados de la clasificación ternaria de los países con la cantidad de muertes o caso por COVID-19, con el fin de demostrar que nuestra propuesta tiene una potencial aplicación a contextos actuales y más amplios como es el seguimiento de los ODS con información tripartita.

El quinto y último paso consistió en la integración de la incertidumbre de Ezbakhe and Pérez-Foguet (2019) a nuestra propuesta de algoritmo descrito en el paso 3. El resultado es una técnica estadística que contempla todos los pasos de análisis de datos, pero en este caso para el monitoreo de los datos compositivos del sector ASH. Se seleccionó un país en concreto con la finalidad de probar y validar la nueva alternativa al nivel subnacional tanto en contextos rurales como urbanos. En total, fueron validados en 24 departamentos del Perú, tanto para agua como para saneamiento, generándose un total de 96 modelos de tendencia temporal.

Cabe señalar que cada capítulo posterior ha tenido su propia introducción y metodología a seguir para su desarrollo al igual que los cálculos computacionales realizados en R Core Team (2020).

2.4. Resumen de los temas tratados

Esta tesis incluye cuatro capítulos adicionales. La primera parte se centra en la revisión de las características de la data de ASH a nivel global (Capítulo III), la segunda está orientado a resolver problemas comunes que se presentan en la data, como: valores ceros, valores perdidos y ambos simultáneamente (Capítulo IV). La tercera, está orientado a proponer una alternativa de monitoreo de la desigualdad (Capítulo V) y finalmente el Capítulo VI, está orientado a aplicar los métodos propuesto al monitoreo de la escala subnacional.

Con más detalle:

Capítulo III. Característica de la data y correlación espuria

Se explora la cantidad de información que existe de ASH a nivel global y la característica de estos datos. Como resultado, se logra identificar la presencia de datos con valores ceros, valores perdidos y ambos simultáneamente, a estos datos también se les llaman datos irregulares. También revelamos la existencia de correlaciones espurias, bajo un supuesto predeterminado, que se generan al no aplicar métodos estadísticos adecuados para datos composicionales.

Capítulo IV. Preprocesamiento de datos irregulares y alternativas robustas de regresión

Se testean alternativas de tratamiento de datos irregulares. Con los datos tratados hacemos modelos robustos de tendencia temporal. En general, el resultado es un algoritmo que incorpora el preprocesamiento de datos irregulares y el ajuste de modelos robustos con un mínimo de seis datos y también con puntos de datos menores a esta cantidad.

Capítulo V. Método alternativo de monitoreo de la desigualdad urbano-rural post 2015

En este capítulo se propone una nueva medida de desigualdad urbana-rural en concordancia con las características compositivas de los datos. Se valida la nueva medida de desigualdad en información tripartita de higiene. Los resultados muestran la aplicabilidad al sector y con un potencial de réplica al monitoreo de comparativas entre urbano y rural en otros sectores de monitoreo.

Capítulo VI. Marco de monitoreo subnacional de los servicios de agua y saneamiento

En este capítulo, se testea el algoritmo propuesto en el capítulo IV. Se aplica al nivel subnacional de Perú, tanto para contextos urbanos como rurales. El resultado muestra que es posible aterrizar el monitoreo global del ODS 6 al nivel subnacional. Las nuevas técnicas estadísticas dan resultados coherentes y son más robustas que la del PCM.

CAPITULO III. CARACTERISTICA DE LA DATA Y CORRELACIÓN ESPURIA

Resumen

Poco se ha evaluado sobre las características de los datos en ASH y las dudosas inferencias que pueden generar al aplicarles un análisis de correlación. Por otro lado, la principal justificación del PCM para seguir utilizando la regresión lineal es la escasa cantidad de datos que existen en el sector del ASH. Hay evidencia de que este método no captura correctamente los datos cuya tendencia es no lineal.

Por tanto, en este capítulo, hemos analizado la data global, con el objetivo de corroborar el porcentaje de países en el cual se puede aplicar técnicas estadísticas no lineales y también para saber los desafíos que enfrenta la data para aplicar transformaciones log-ratio. Los resultados, muestran que la característica de la data es múltiple, hay datos con presencia de valor cero, valor perdido y ambos simultáneamente. Hemos encontrado que hay un alto porcentaje de datos ausentes en como mínimo una de las categorías, siendo estos de 34.2% en la data de agua-urbano, 32.1% en la data agua-rural, 47.1% en la data de saneamiento-urbano y 46.2% en la data de saneamiento rural. El tratamiento de los datos que faltan ayudará a aumentar la cantidad de información y, en consecuencia, a mejorar las estimaciones. También encontramos que es posible aplicar métodos no lineales en un gran porcentaje de países, principalmente en los países —con datos mínimos de seis— de las regiones de América Latina y el Caribe y del África subsahariana. Por último, los resultados muestran que, bajo los supuestos establecidos, existen correlaciones espurias al aplicar las técnicas estadísticas estándar. Por lo tanto, en presencia de datos de composición, es imperativo aplicar transformaciones log-ratio.

3.1. Introducción

El período de los ODM ha sido una iniciativa global en el que se generó gran cantidad de información sobre el acceso a agua y saneamiento por las diferentes opciones tecnológicas. Ha habido muchos aprendizajes y cuestionamientos a los indicadores de seguimiento (Craven et al., 2013; Bartram et al., 2014; Bain et al., 2018; Weststrate et al., 2018), los cuales en parte han sido mejorados para el ODS. Post-2015, hay nuevos indicadores, p. ej., la escalera de agua y saneamiento de cuatro categorías ahora tiene cinco categorías de seguimiento. No obstante, todos parten de la clasificación de servicios mejorados (gestionado de forma segura, básico y limitado) y no mejorados (otros no mejorados y el acceso agua superficial o si es saneamiento la defecación al aire libre).

Para realizar las estimaciones, el PCM parte con la identificación de fuentes de datos representativas a nivel nacional, que pueden proceder de múltiples fuentes de información, como censos, encuestas de hogares, datos administrativos, etc. Por lo general, esta información la recogen las propias instituciones estadísticas de cada país mediante entrevistas a los hogares o las instituciones sectoriales que supervisan el desempeño de los servicios de agua y saneamiento.

Al final del periodo de los ODM, se observó en la base de datos global que el número total de información de AyS —proveniente de múltiples fuentes representativos a nivel nacional— había aumentado considerablemente, pasando de 235 en el año 2000 a 3,607 en el año 2015 (WHO/UNICEF, 2018). Este aumento de información se presenta como una oportunidad para aplicar técnicas estadísticas no lineales, cuestión que ha sido propuesto en la literatura del sector por su ventajas estadísticas (Fuller et al., 2016).

No obstante, tal como se dijo en el ítem 1.1 de la presente tesis, actualmente el PCM sigue aplicando la regresión lineal. La principal justificación para continuar usando este método es que para muchos países no hay suficientes puntos de datos para justificar el uso de métodos no lineales (WHO/UNICEF, 2018).

Por tanto, en este capítulo queremos resolver dos preguntas básicas, i) ¿Hay la cantidad de datos suficientes para aplicar técnicas estadísticas no lineales?, y ii) ¿Como son las características de los datos?. La respuesta a la segunda pregunta nos ayudará a elegir entre todas las opciones las técnicas de preprocesamiento estadístico que existen las más adecuadas para el monitoreo del acceso a servicios de ASH.

3.2. El estado de los datos globales: número de puntos de datos

Se obtuvo información actualizada de la base de datos del PCM de https://washdata.org/. El análisis se ha basado en tres indicadores primarios, para agua potable son: servicio mejorado (M), agua entubada (X_1) y agua superficial (X_3). Para saneamiento son: servicio mejorado (M), alcantarillado (X_1) y defecación al aire libre (X_3). La información procesada solo contempla los estados que contienen información en al menos una de sus categorías en urbano y rural. La base de datos y scripts en R se encuentran publicados en Quispe-Coica (2021).

El resultado se muestra en la Tabla 3.1. En total se obtuvieron información de 200 países en agua y 198 países en saneamiento. La clasificación de los países según su número de puntos de datos ha sido en seis partes. Los países que no tienen puntos de datos en una de las categorías o en el total de urbano o rural están denotados con “A”. P. ej., Aruba no tiene

información en agua urbano de la categoría M y se clasifica en “A”, no obstante, en X_1 y X_3 tienen un punto de dato cada uno y, por lo tanto, están denotados con “B”. Otro caso, Anguila no tiene información en las tres categorías de saneamiento rural y están dentro de la clasificación de “A”, sin embargo, en saneamiento rural si tienen información y están clasificados en “C” la categoría M y en “B” la categoría X_1 y X_3. En general, bajo estos criterios, los países han sido clasificados según el número de puntos de datos que tienen.

En los servicios de agua-urbano, el 65% de los países tienen puntos de datos ≥ 6 en la categoría M, este valor es del 64.5% en la categoría X_1 y en la categoría X_3 este valor es tan solo del 45%. En rural, no hay variación significativa respecto al urbano en el porcentaje de países con puntos de datos ≥ 6 (M = 64.5%, X_1 = 64.0% y X_3 = 45.5%).

En los servicios de saneamiento-urbano, 137/198 (69.2%) de los países tienen puntos de datos ≥ 6 en la categoría M, 103/198 (52.0%) de los países en X_1 (datos ≥ 6) y 90/198 (45.5%) de los países en X_3 (datos ≥ 6). Mientras que, en rural, tienen puntos de datos ≥ 6: 135/198 (68.2%) países de la categoría M, 101/198 (51.0%) de los países en X_1 y 89/198 (44.9%) de los países en X_3.

La cantidad de países sin datos ha sido bajo en las categorías M y X_1 tanto en agua y saneamiento. No obstante, la proporción de países sin datos fue elevado en la categoría X_3, con valor mínimo de 26.3% y valor máximo de 28.5%.

Tabla 3.1. Número de países con datos de AyS

Servicio	Sector	Indicador	Número de países con el siguiente número de puntos de datos						
			A (0)	B (1 – 2)	C (3 – 5)	D (6 – 10)	E (11 – 15)	F (≥ 16)	Total
Agua	Urbano	M	4 (2.0%)	34 (17.0%)	32 (16.0%)	49 (24.5%)	34 (17.0%)	47 (23.5%)	200
		X_1	2 (1.0%)	38 (19.0%)	31 (15.5%)	49 (24.5%)	33 (16.5%)	47 (23.5%)	200
		X_3	56 (28%)	28 (14%)	26 (13%)	37 (18.5%)	21 (10.5%)	32 (16%)	200
Agua	Rural	M	12 (6.0%)	29 (14.5%)	30 (15.0%)	52 (26.0%)	29 (14.5%)	48 (24.0%)	200
		X_1	11 (5.5%)	32 (16.0%)	29 (14.5%)	52 (26.0%)	28 (14.0%)	48 (24.0%)	200
		X_3	57 (28.5%)	29 (14.5%)	23 (11.5%)	38 (19.0%)	19 (9.5%)	34 (17.0%)	200
Saneamiento	Urbano	M	3 (1.5%)	29 (14.6%)	29 (14.6%)	55 (27.8%)	30 (15.2%)	52 (26.3%)	198
		X_1	8 (4.0%)	37 (18.7%)	50 (25.3%)	50 (25.3%)	21 (10.6%)	32 (16.2%)	198
		X_3	52 (26.3%)	24 (12.1%)	32 (16.2%)	34 (17.2%)	23 (11.6%)	33 (16.7%)	198
Saneamiento	Rural	M	12 (6.1%)	21 (10.6%)	30 (15.2%)	56 (28.3%)	26 (13.1%)	53 (26.8%)	198
		X_1	19 (9.6%)	29 (14.6%)	49 (24.7%)	48 (24.2%)	21 (10.6%)	32 (16.2%)	198
		X_3	53 (26.8%)	24 (12.1%)	32 (16.2%)	33 (16.7%)	22 (11.1%)	34 (17.2%)	198

Por otro lado, se ilustra la cantidad de datos por región en la Figura 3.1. Las regiones de África Sub-sahariana y América Latina y el Caribe tienen en general la mayor cantidad de puntos de datos disponibles en comparación con el resto de las regiones. Esto los hace idóneos para la aplicación de cualquier técnica estadística no lineal.

La región de Europa y América del Norte sigue en el siguiente orden. No obstante, la mayor cantidad de información solo se encuentra en las categorías de M y X_1 y, por lo tanto, es poco probable completar la escalera de AyS.

En las cinco regiones restantes (Australia y Nueva Zelanda, Asia central y meridional, Asia Oriental y Sudoriental, Oceanía, África del Norte y Asia Occidental), el número de puntos de datos es igual o inferior a 208 en las tres categorías de análisis. La región de Australia y Nueva Zelanda es el caso más desfavorable, ya que tienen la menor cantidad de datos disponibles.

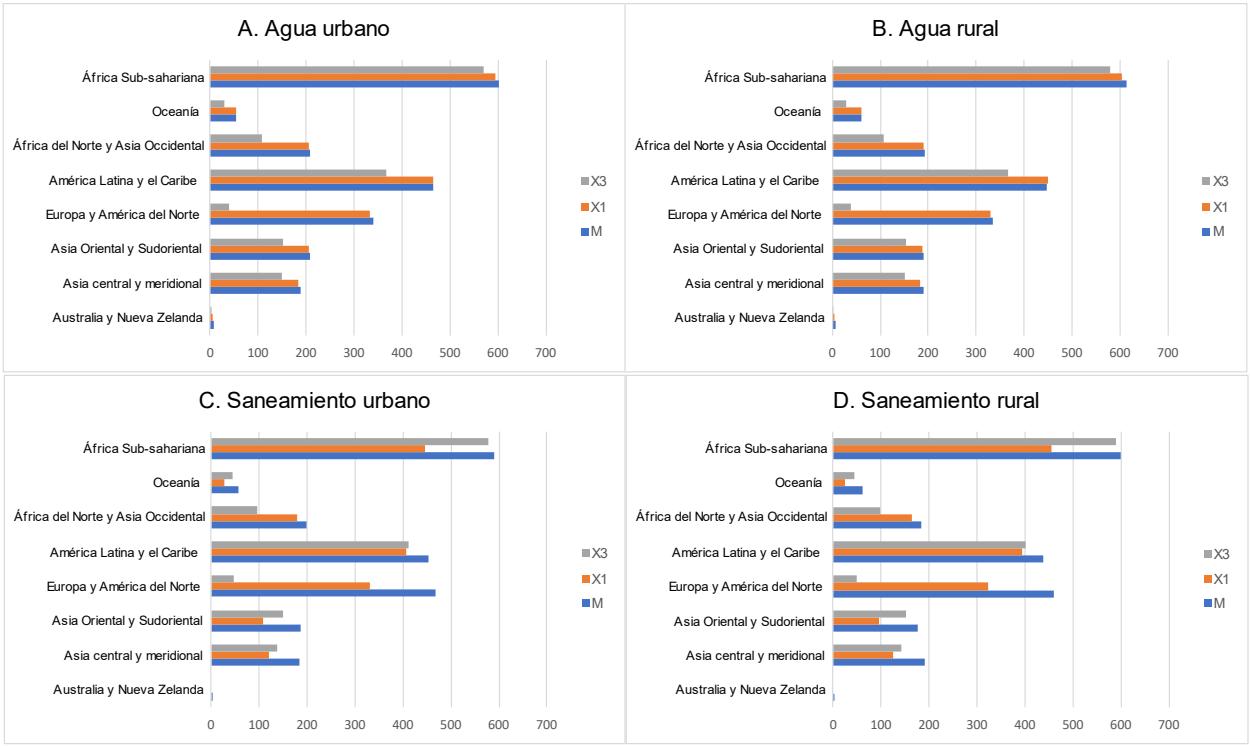


Figura 3.1. Número de datos disponibles por región.

3.3. Características de los datos

En esta sección mostramos el porcentaje de datos de AyS con valores cero, valores perdidos y ambos simultáneamente. El análisis se ha basado en las cuatro categorías de servicio que tienen valor de cierre del 100%. Es decir, la suma de $X_1 + X_2 + X_3 + X_4 = 100\%$. Dónde $X_2 = M - X_1$ y $X_4 = 100 - M - X_3$.

3.3.1. Datos con valor cero

La existencia de datos irregulares del tipo de valor cero a nivel global se muestra en la Figura 3.2. Hay diez patrones de comportamiento en agua urbano/rural y en saneamiento rural, mientras que en saneamiento urbano hay solo ocho patrones de comportamientos. En agua-urbano (Figura 3.2A), el 75.26% de los datos no tiene valor cero en las cuatro categorías de servicio. El análisis individual por categoría nos muestra que X_2 y X_3 son los que tienen la mayor cantidad

de datos con valor cero, con el 12.58% y 13.58% de los datos, respectivamente. La presencia de valor cero se presenta comúnmente cuando el acceso a agua entubada (X_1) tiene tendencia al límite superior del 100%. Es decir, la cobertura total de los hogares es mediante agua entubada. El patrón número 7 es un claro ejemplo. Además, según las estimaciones del PCM para 2020, el 99% de la población urbana tiene acceso a agua mejorada (WHO/UNICEF, 2021). Por lo que es esperable valores de X_1 cercanos al 100%.

En agua-rural (Figura 3.2B), el 84.8% de los datos no tiene valor cero en las cuatro categorías (patrón N° 1). De las cuatro categorías, X_2 tiene el mayor porcentaje de datos con valor cero, con un 10.88%. El análisis individual por categoría muestra que el acceso de los hogares al agua por otras formas no mejoradas (X_4) presentan valores elevados en los patrones N°5 y el N°7, con valores promedio de 23.5% y 49.3%, respectivamente.

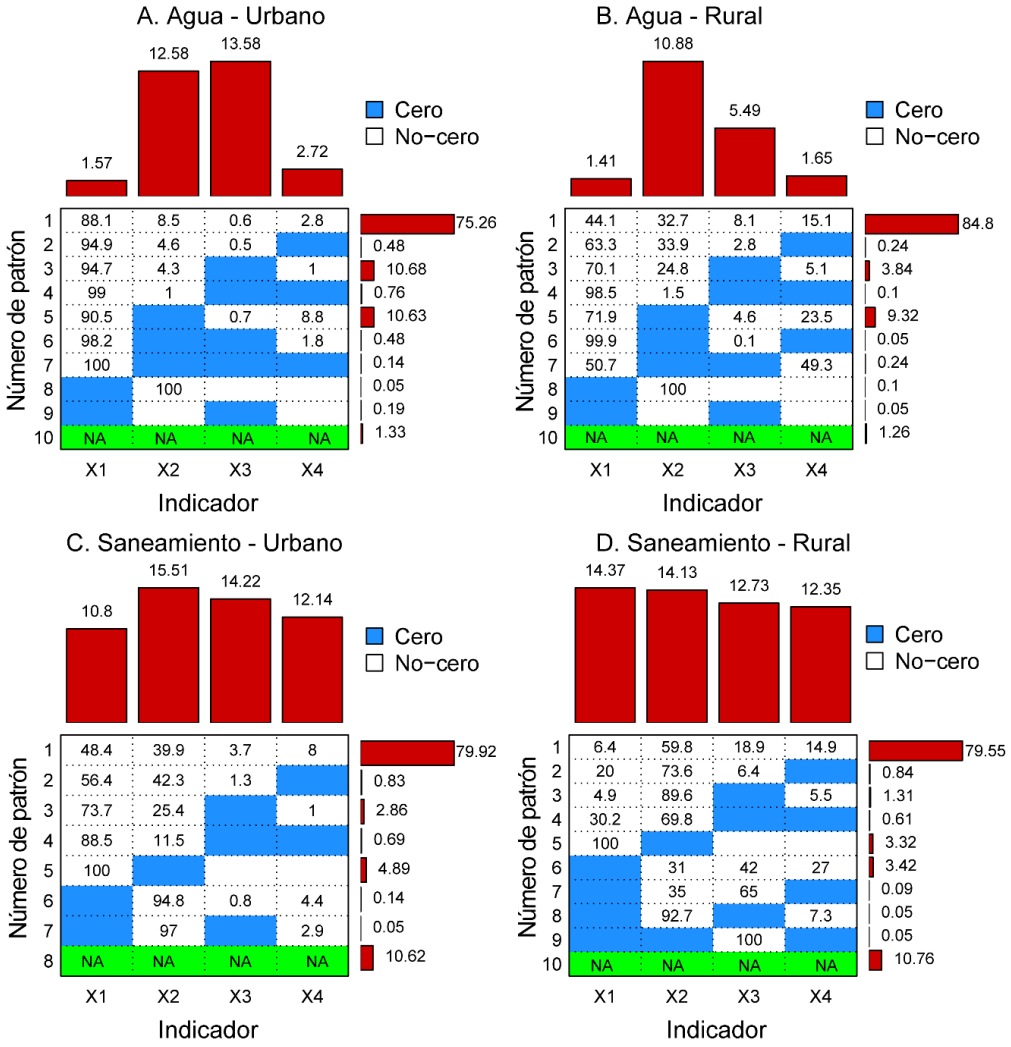


Figura 3.2. Datos con valore cero

Nota: los valores perdidos se denotan con NA

En saneamiento urbano y rural, en el 79.92% y 79.55%, respectivamente, no tiene valor cero en las cuatro categorías de servicio. En comparación con el servicio de agua, las cuatro categorías de saneamiento tanto en urbano como en rural tienen porcentaje de valores cero más altos. Un análisis exploratorio de los datos muestra que, para el saneamiento urbano (Figura

3.2C), los valores medios se distribuyen mayoritariamente en X_1 y X_2. Esto significa que los hogares urbanos acceden mayoritariamente a los servicios a través del alcantarillado y por otras formas mejoradas de saneamiento.

En saneamiento rural (Figura 3.2D), los valores medios también se distribuyen en X_1 y X_2, pero con mayor porcentaje en X_2. Esto significa que los hogares rurales acceden mayormente a servicios de saneamiento por otras formas mejoradas (X_2). También es en saneamiento rural donde el valor de X_3 es superior a las X_3 de saneamiento urbano. Esto significa que, históricamente, ha habido y sigue habiendo un número importante de hogares en los que la población rural sigue defecando al aire libre. Esta situación se ve reafirmada por los datos globales del último informe del PCM, donde el 13% de la población de los hogares rurales sigue defecando al aire libre, mientras que en urbano este valor es tan solo del 1% (WHO/UNICEF, 2021).

Por otro lado, tanto en agua como en saneamiento hay países que no disponen de datos en las cuatro categorías de análisis, estos se encuentran denotados con NA en la Figura 3.2. La razón es que la información que tiene el país es sólo para el indicador mejorado y no está desglosada en ninguna de las categorías. Esto representa el 1.33% para el agua urbana, el 1.26% para el agua rural, el 10.62% para el saneamiento urbano y el 10.78% para el saneamiento rural. Los países con estas características no son propicios para completar las escaleras de AyS, pero si para un análisis bivariado entre servicios mejorados y no mejorados.

3.3.2. Datos perdidos

Es muy común que la información del sector ASH no esté siempre disponible y completa. En muchos países, sólo se recoge una de las X categorías de intereses propios. Por tanto, es de esperarse que se vean reflejados estas situaciones en la base de datos global, lo cual se ilustra en la Figura 3.3.

En los servicios de agua, hay seis patrones de comportamiento comunes en la forma de presentar los datos, mientras que en el saneamiento hay siete patrones de comportamiento. En el acceso a agua, el 65.78% de urbano y el 67.9% de rural tienen las cuatro categorías de servicio con datos. Mientras que, en saneamiento tan solo el 52.91% en urbano y 53.81% en rural tienen las cuatro categorías con datos. Esto significa que, en materia de saneamiento, aproximadamente sólo la mitad de los datos están completos para la aplicación directa de las transformaciones log-ratio. El resto de los datos requerirá un tratamiento previo con métodos estadísticos para completar los datos.

También se aprecia que un poco más de la tercera parte son datos perdidos en las categorías X_3 y X_4 tanto en agua (urbano es de 32.7% y 34.03% y rural es de 30.69% y 31.86%, respectivamente) como en saneamiento (urbano es de 32.46% y 32.69% y rural es de 30.93% y 31.17%, respectivamente). En la categoría X_1 y X_2 de saneamiento urbano (es de 25.21% y 26.45%, respectivamente) y rural (es de 25.97% y 27%, respectivamente) son también elevados el porcentaje de datos perdidos; mientras que, en X_1 y X_2 de agua urbano (es de 2.53% y 2.86%, respectivamente) y rural (es de 2.48% y 2.67%, respectivamente) son bajos.

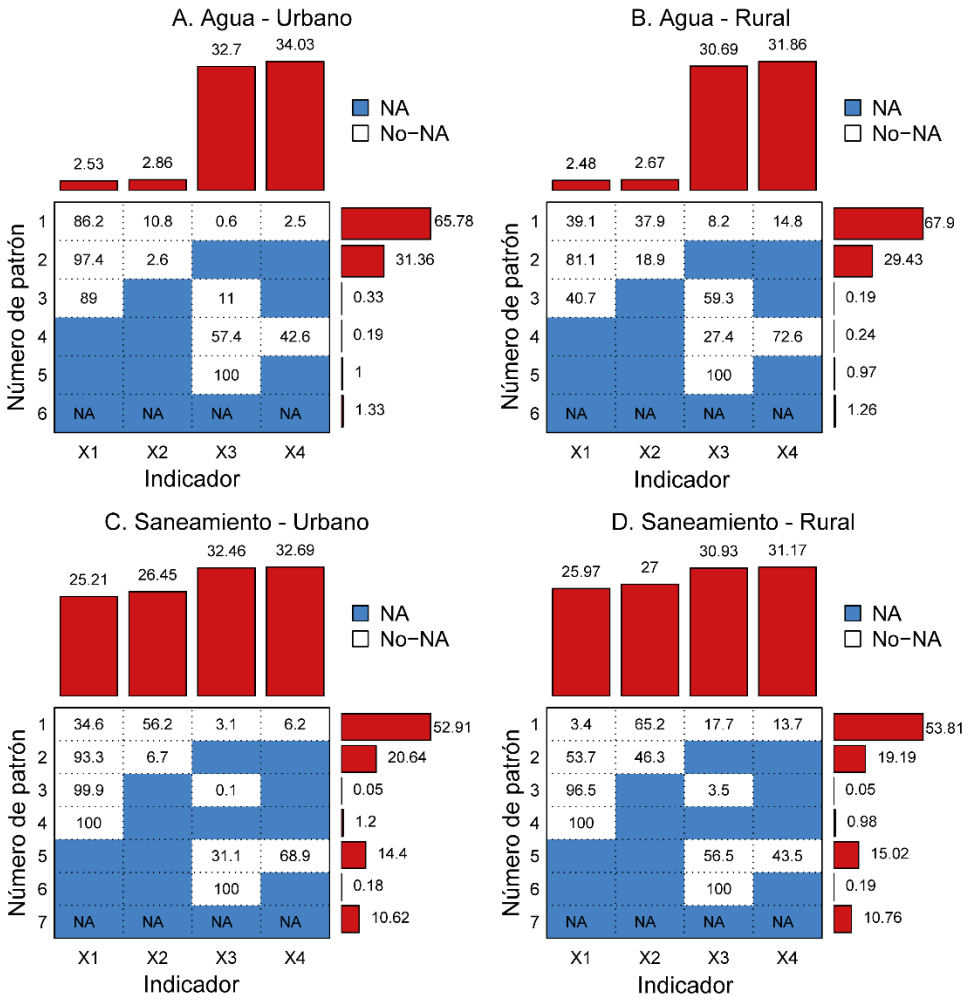


Figura 3.3. Datos perdidos.

Nota: los valores perdidos se denotan con NA

Estos resultados muestran que existe un alto porcentaje de datos ausentes en el sector del agua y el saneamiento, lo que requerirá métodos de imputación coherentes con la característica de composición de los datos para completar la información.

3.3.3. Datos con valor cero y valores perdidos simultáneamente

Se ha comprobado que también existe simultáneamente datos con valor cero y datos perdidos en el vector de composición. En total hay 667/8460 (7.9%) vectores con estas características. Desglosados por nivel de servicio y área de residencia, hay en agua urbano 244/2098 (11.6%) vectores, en agua rural 211/2059 (10.2%) vectores, en saneamiento urbano 128/2166 (5.9%) y en saneamiento rural 84/2137 (3.9%) vectores de composición con datos perdidos y de valor cero simultáneamente.

3.4. Correlación espuria

En la literatura de AyS, existe la hipótesis de que aplicar las técnicas estadísticas clásicas directamente sobre los datos con suma constante puede generar una correlación espuria (Pérez-Foguet et al., 2017; Ezbakhe y Pérez-Foguet, 2019). De hecho, es uno de los motivos principales por el que se propone el análisis estadístico sobre datos transformados. Sin embargo, en el

sector ASH, esta afirmación aún no se ha puesto a prueba. Debido a esto, se realiza un análisis de correlación de la población que accede a los diferentes niveles de servicios de agua y saneamiento y, para ser más explícitos, se realiza el siguiente ejemplo.

Dos organizaciones externas (denotados con “A” y “B”) quieren conocer el acceso de la población a servicios de agua y saneamiento urbano por diferentes niveles de servicio en Panamá y Serbia, respectivamente. Para lo cual, la organización "A" solicita datos de población de las cuatro categorías [X_1, X_2, X_3, X_4], mientras que la organización "B" sólo necesita conocer tres de las cuatro categorías [X_1*, X_2*, X_3*]. Con la información obtenida, la organización "A" procede a realizar proporciones para visualizar el porcentaje de la población que representa cada categoría en la composición. La organización "B" sigue el mismo procedimiento, pero solo para las tres categorías de servicios. El resultado se muestra en la Tabla 3.2.

Tabla 3.2. Proporción de acceso a agua y saneamiento

País	Fuente	Año	Organización A				Organización B		
			X_1	X_2	X_3	X_4	X_1*	X_2*	X_3*
Agua									
Panamá	ENASSER	2015	98.4	1.6	0	0	98.4	1.6	0.0
Panamá	MICS	2013	98	1.9	0	0.1	98.1	1.9	0.0
Panamá	EPM	2014	98.6	1.2	0	0.1	98.8	1.2	0.0
Panamá	EPM	2017	98.5	1.3	0	0.2	98.7	1.3	0.0
Panamá	EPM	2018	98.4	1.4	0	0.2	98.6	1.4	0.0
Panamá	EPM	2016	98.2	1.6	0	0.2	98.4	1.6	0.0
Panamá	EPM	2015	98.6	1.1	0	0.3	98.9	1.1	0.0
Panamá	ENASSER	2009	97.1	2.2	0.1	0.6	97.7	2.2	0.1
Panamá	ENSPA	2019	97.6	0.9	0.1	1.4	99.0	0.9	0.1
Panamá	LSMS	2008	97.2	0.6	0.1	2.1	99.3	0.6	0.1
Panamá	CEN	2010	96.3	1.4	0	2.2	98.6	1.4	0.0
Panamá	LSMS	2003	96.6	0.2	0.4	2.8	99.4	0.2	0.4
Panamá	CEN	2000	96	1.1	0.1	2.8	98.8	1.1	0.1
Saneamiento									
Serbia	MICS	2006	83.7	16.2	0.1	0	83.7	16.2	0.1
Serbia	MICS	2010	85.5	14.4	0	0.1	85.6	14.4	0.0
Serbia	MICS	2019	86.8	12.8	0.09	0.27	87.0	12.9	0.1
Serbia	MICS	2014	83.3	16.1	0	0.6	83.8	16.2	0.0
Serbia	LSMS	2007	38.9	8.3	0.04	52.7	82.4	17.6	0.1

Nota: Se muestran solo los valores redondeados a un decimal.

Posteriormente, cada organización realiza un análisis de correlación (de Pearson) de las categorías seleccionadas. El resultado se muestra en la Tabla 3.3.

Tabla 3.3. Matriz de correlación de los datos sobre el acceso a AyS.

1) Agua - Panamá

A					B			
	X_1	X_2	X_3	X_4		X_1*	X_2*	X_3*
X_1	1.00	0.24	–0.50	–0.91	X_1*	1.00	–0.98	0.47
X_2		1.00	–0.63	–0.62	X_2*		1.00	–0.63
X_3			1.00	0.62	X_3*			1.00
X_4				1.00				

2) Saneamiento - Serbia

A					B			
	X_1	X_2	X_3	X_4		X_1*	X_2*	X_3*

X_1	1.00	0.87	0.05	–0.99	X_1*	1.00	–0.99	–0.12
X_2		1.00	–0.06	–0.90	X_2*		1.00	0.09
X_3			1.00	–0.04	X_3*			1.00
X_4				1.00				

Notas: La matriz de correlación se realiza con los datos brutos de las series temporales del país. Los valores opuestos al caso A se muestran en gris.

Con los resultados obtenidos en la Tabla 3.3, la organización "A" deduce que el acceso al agua urbana entre las dos opciones tecnológicas de X_1 y X_2 tienen una correlación positiva y baja (0.24). Esto significa que existe una relación directa entre las dos opciones tecnológicas. Mientras que la organización "B" concluye que la relación entre las dos variables (X_1* y X_2*) es alta y al mismo tiempo inversa, con un valor de –0.98. Similar situación ocurre entre las categorías X_1 y X_3. En saneamiento urbano se obtienen resultados aún más discrepantes, ya que en las tres categorías de servicios los valores son opuestos.

Tabla 3.4. Ratio de composiciones

País	Fuente	Año	Proporción en composición completa "A" [X_1, X_2, X_3, X_4]		Proporción en Subcomposición "B" [X_1*, X_2*, X_3*]	
			X_1/X_2	X_3/X_2	X_1*/X_2*	X_3*/X_2*
Agua						
Panamá	ENASSER	2015	61.5	0.0	61.5	0.0
Panamá	MICS	2013	51.6	0.0	51.6	0.0
Panamá	EPM	2014	82.2	0.0	82.2	0.0
Panamá	EPM	2017	75.8	0.0	75.8	0.0
Panamá	EPM	2018	70.3	0.0	70.3	0.0
Panamá	EPM	2016	61.4	0.0	61.4	0.0
Panamá	EPM	2015	89.6	0.0	89.6	0.0
Panamá	ENASSER	2009	44.1	0.0	44.1	0.0
Panamá	ENSPA	2019	108.4	0.1	108.4	0.1
Panamá	LSMS	2008	162.0	0.2	162.0	0.2
Panamá	CEN	2010	68.8	0.0	68.8	0.0
Panamá	LSMS	2003	483.0	2.0	483.0	2.0
Panamá	CEN	2000	87.3	0.1	87.3	0.1
Saneamiento						
Serbia	MICS	2006	5.2	0.0	5.2	0.0
Serbia	MICS	2010	5.9	0.0	5.9	0.0
Serbia	MICS	2019	6.8	0.0	6.8	0.0
Serbia	MICS	2014	5.2	0.0	5.2	0.0
Serbia	LSMS	2007	4.7	0.0	4.7	0.0

Sin embargo, al hacer una relación entre las partes X_1/X_2 (se ha seleccionado las variables X_1 y X_2 para fines prácticos) tanto para la organización “A” como para la organización “B”, el resultado es igual para ambos. El resultado se muestra en la Tabla 3.4. Lo mismo ocurre al hacer el hacer la ratio entre las variables X_3 y X_2. Puede probar hacer ratio con las variables en diferentes formas, el resultado de la relación siempre será el mismo entre "A" y "B".

Por lo tanto, las alternativas estadísticas para datos compositivos (CoDa; es el acrónimo en inglés de compositional data) se basan en un enfoque log-cociente (John Aitchison, 1986; Egozcue et al., 2003), cada una con sus propias particularidades tal como se detalló en el ítem 1.3 de la presente tesis. Dado que el sector ASH no es ajeno con los supuestos planteados, la aplicación de una estadística adecuada para los datos de composición se hace obligatoria.

Por otro lado, para responder la pregunta ¿qué tanto se presentan las correlaciones espurias —bajo el supuesto planteado— en el monitoreo global de los servicios de AyS?, hemos realizado el mismo análisis en todos los países que tienen información. Como resultado de la exclusión de los vectores con datos ausentes, nos hemos quedamos con 119 y 122 países en saneamiento rural y urbano, respectivamente. En el servicio de agua tan solo nos hemos quedado con 121 países. El resultado de todos los países se muestra en Quispe-Coica (2021).

Tabla 3.5. Correlación espuria

Servicio	Rural			Urbano		
	f1	f2	f3	f1	f2	f3
Saneamiento	119	119	119	122	122	122
I	13 (10.9%)	13 (10.9%)	15 (12.6%)	10 (8.2%)	8 (6.60%)	11 (9.0%)
D	105 (88.2%)	105 (88.2%)	104 (87.4%)	112 (91.8%)	109 (89.3%)	106 (86.9%)
NA	1 (0.8%)	1 (0.8%)	0 (0%)	0 (0%)	5 (4.1%)	5 (4.1%)
Agua	121	121	121	121	121	121
I	8 (6.6%)	6 (5.0%)	9 (7.4%)	8 (6.6%)	15 (12.4%)	2 (1.7%)
D	113 (93.4%)	114 (94.2%)	111 (91.7%)	113 (93.4%)	102 (84.3%)	115 (95%)
NA	0 (0%)	1 (0.8%)	1 (0.8%)	0 (0%)	4 (3.3%)	4 (3.3%)
Total	240	240	240	243	243	243

Nota: f1 = comparativa de los resultados entre la correlación (X_1 y X_2) y la correlación (X_1* y X_2*). f2 = comparativa de los resultados entre la correlación (X_1 y X_3) y la correlación (X_1* y X_3*). f3 = comparativa de los resultados entre la correlación (X_2 y X_3) y la correlación (X_2* y X_3*). I = el número de países de la comparación con valores de correlación opuestos. D = el número de países de la comparación con valores de correlación que van en el mismo sentido. NA = No se pudo calcular la correlación de Pearson porque sólo se disponía de un vector como dato.

En la Tabla 3.5 resumimos los resultados obtenidos de la correlación (bajo el mismo supuesto inicial de este ítem). En saneamiento rural, el número de países con valores de correlación opuestos es 13/119 (10.9%) en f1 y f2, mientras que en f3 es 15/119 (12.6%). En saneamiento urbano, el número de países con valores opuestos expresado en porcentaje disminuye al 8.2% en f1, 6.6% en f2 y 9.0% en f3. En agua rural, el número de países con valores de correlación opuestos expresado en porcentaje es del 6.6% en f1, 5.0% en f2 y 7.4% en f3. Mientras que, en agua urbano es del 6.6% en f1, 12.4% en f2 y del 1.7% en f3.

Hemos analizado sólo las correlaciones con sentidos opuestos, ya que son los resultados más discrepantes. Y, aun así, nuestros resultados muestran que se trata de un porcentaje que consideramos elevado. Por lo tanto, es imprescindible aplicar las técnicas adecuadas a los datos de suma constante disponibles en el sector.

3.5. Discusión

Claramente la cantidad de información ha aumentado significativamente desde los ODM hasta el último informe de los ODS, por lo que la justificación principal del PCM para continuar usando la regresión lineal — es que para muchos países no hay suficientes puntos de datos para justificar el uso de métodos no lineales (WHO/UNICEF, 2018)— poco a poco va perdiendo

fuerza.

Si partimos del estudio de Fuller et al. (2016), que indica que para datos iguales o mayores a seis el modelo no lineal de GAM tuvo un buen desempeño. Y, tomando en cuenta que el menor porcentaje de los tres indicadores obtenidos (M, X_1 y X_3) es casi el mismo porcentaje de países que tienen las tres categorías con valores completos, GAM se podría aplicar —a nivel global— de forma univariada en el 65% de los países para hacer modelos en la categoría mejorado, en el 64.5% de los países para hacer modelos en la categoría X_1 y en tan solo el 45% de los países para hacer modelos en la categoría X_3; todo esto en la data de agua urbano. Dado que en agua rural no hay variación significativa respecto al urbano en el porcentaje de países con puntos de datos ≥ 6 (M = 64.5%, X_1 = 64.0% y X_3 = 45.5%), también se podría aplicar GAM en como mínimo el 45.5% de los países.

Por otro lado, las escaleras de AyS surgen de la desagregación de los servicios mejorados y no mejorados. Por ejemplo, en agua, las tres primeras escaleras de seguimiento (gestión segura, básico y limitado) son obtenidas al multiplicar el servicio mejorado por los indicadores segundarios, mientras que el servicio no mejorado se desagrega directamente en la categoría superficial (X_3) y en la categoría otras formas no mejoradas (X_4; dónde $X_4 = 100 – M – X_3$).

Tal como se puede ver, se puede asumir que el porcentaje mínimo de las tres categorías analizadas es también el porcentaje de países con las X completas (véase Ezbakhe y Pérez-Foguet (2019)). Bajo esta premisa, GAM en CoDa se pueden aplicar en el 45% de los países en agua urbano, en el 45.5% de los países en agua rural, en el 45.5% de los países en saneamiento urbano y en el 44.9% de los países en saneamiento rural.

Imputar los valores perdidos encontrados hará que este porcentaje de países aumente. Por lo tanto, aplicar técnicas estadísticas no lineales sobre datos compositivos ya es posible, más aún por las ventajas que ha demostrado tener (Fuller et al., 2016; Pérez-Foguet et al., 2017). En este estudio hemos identificado que los países de las regiones de África Sub-sahariana y América Latina y el Caribe son los que tienen más información, por lo que los países de estas regiones con datos mínimos (≥ 6) son una oportunidad para aplicar técnicas estadísticas no lineales como GAM. Con esto se logrará mejorar significativamente las estimaciones.

Finalmente, de los diferentes tipos de valores ceros que se presentan en la data —ceros por redondeo, ceros de recuento y ceros esenciales (Aitchison y Kay, 2003; Martín-Fernández et al., 2015; Templ et al., 2016)—, en el sector ASH es del tipo ceros de redondeo. Esto se debe a que asumimos que hay al menos una persona sin los servicios de una de las categorías de análisis. Por ejemplo, 1 persona en un millón expresada en proporciones es 0.000001; en fuentes de información como censos y encuestas es habitual redondear estos valores a dos o tres decimales para facilitar la lectura. A esto se suma las deficiencias que tienen las encuestas y censo de hogares (BBC, 2013; Neupert, 2017; PERU21, 2017).

CAPÍTULO IV. PREPROCESAMIENTO DE DATOS IRREGULARES Y ALTERNATIVAS ROBUSTAS DE REGRESIÓN

Resumen

Los niveles de servicio de ASH al que accede la población suman el 100%; por lo tanto, son datos de composición. A pesar de las evidencias de valor cero, datos perdidos y valores atípicos en las fuentes de información, aún no se ha analizado el tratamiento de estas irregularidades con diferentes técnicas estadísticas para CoDa en el sector ASH. Por lo tanto, los resultados pueden presentar estimaciones sesgadas, y las decisiones basadas en estos resultados no serán necesariamente adecuadas. Por lo tanto, en este artículo: i) evaluamos alternativas metodológicas de imputación que abordan el problema de tener valores cero o valores perdidos, o ambos simultáneamente; y ii) proponemos la necesidad de complementar la identificación punto a punto del Programa Conjunto de Monitoreo (PCM) de la OMS/UNICEF con otras alternativas robustas, para tratar los valores atípicos en función del número de puntos de datos. Estas sugerencias se han tenido en cuenta aquí utilizando estadísticas para CoDa con transformación isométrica log-ratio (ilr). Se presenta una selección de casos ilustrativos para comparar el rendimiento de las distintas alternativas.

Este capítulo está basado en:

- Quispe-Coica, A., Pérez-Foguet, A., 2020b. Preprocessing alternatives for compositional data related to water, sanitation and hygiene. Sci. Total Environ. 743, 140519. https://doi.org/10.1016/j.scitotenv.2020.140519

4.1. Introducción

El seguimiento internacional del acceso a los servicios de agua, saneamiento e higiene se lleva a cabo mediante la estimación del porcentaje de población de cada país con niveles de servicio definidos. Los datos son, por tanto, compositivos (John Aitchison, 1986; Egozcue y Pawlowsky-Glahn, 2005; Lloyd et al., 2012; Pérez-Foguet et al., 2017). Aunque las partes del total han variado históricamente (por ejemplo, las categorías utilizadas para supervisar el sector del agua y el saneamiento se ampliaron de cuatro a cinco), siempre se han centrado en la evolución del porcentaje de población con acceso a los distintos niveles de servicio.

Fuller et al. (2016) clasificaron la evolución temporal del acceso al agua y al saneamiento según la linealidad o no linealidad de las tendencias y propusieron el uso de Modelos Aditivos Generalizados (GAM; acrónimo en inglés de Generalized additive model) cuando los datos son mínimos. El carácter compositivo de los porcentajes de población se incluye en el análisis presentado por Pérez-Foguet et al.(2017), que concluyen que el uso de GAM para las transformaciones logarítmicas isométricas (ilr) de las variables de seguimiento habituales es adecuado. De este modo, se trata adecuadamente la no linealidad de las restricciones de suma igual a constante. Esto es especialmente relevante cuando partes del total tienden a valores cercanos a los extremos de todo o nada. Sin embargo, la propuesta no aborda situaciones comunes, como la presencia de valores reportados como cero, o datos faltantes en partes del total, lo que impide una aplicación directa del enfoque composicional.

Los datos con valor cero suelen presentarse en países que han realizado importantes avances en la prestación de servicios mejorados de agua y saneamiento; como consecuencia, las poblaciones con acceso a fuentes no mejoradas se han reducido drásticamente, siendo el número en muchos casos nulo o cercano a cero. Por lo tanto, las transformaciones de ilr en los datos no pueden llevarse a cabo si no se excluyen o imputan primero los valores cero. La exclusión es una alternativa fácil para abordar el problema, pero si la cantidad de datos del sector es baja, esto puede afectar a la capacidad de predicción de los modelos. Así, en la literatura se han propuesto alternativas para la imputación de los valores cero en cada situación según las propiedades de la CoDa, incluyendo ceros redondeados (Palarea-Albaladejo et al., 2007; Palarea-Albaladejo y Martín-Fernández, 2008; Martín-Fernández et al., 2012; Templ et al., 2016; Chen et al., 2018), ceros de recuento (Martín-Fernández et al., 2015)y ceros esenciales (Aitchison y Kay, 2003)).

Las técnicas relacionadas con los ceros redondeados son las alternativas de imputación más convenientes para el sector ASH, dado que, incluso en los países más desarrollados, es probable que haya al menos pequeños porcentajes de población que no tienen acceso a ningún tipo de servicio de agua. La sustitución simple y la sustitución multiplicativa ya han sido abordadas en estudios anteriores del sector (Pérez-Foguet et al., 2017; Ezbakhe y Pérez-Foguet, 2019). A pesar de su sencillez en la aplicación, estos métodos tienden a subestimar la variabilidad de los datos, por lo que es aconsejable que sólo se utilicen cuando la presencia de ceros es baja (Palarea-Albaladejo y Martín-Fernández, 2008). En presencia de grandes cantidades de valores cero, se recomiendan otras alternativas de imputación, de acuerdo con la variabilidad de los datos que existen en la serie temporal.

La falta de datos que definan la composición es también un tema de especial importancia en el sector, ya que afecta a algunas categorías de análisis. Por ejemplo, según la encuesta nacional

(PNAD17) en el sector rural de Brasil, el 88.4% de la población tiene acceso a fuentes mejoradas de agua potable (y el 82.7%, por tubería), pero no se da información sobre el acceso por fuentes superficiales (WHO/UNICEF, 2019b). La falta de uno o más puntos de datos para un año específico significa que la transformación ilr no puede aplicarse directamente, por lo que la información de ese año se pierde en el seguimiento de todas las partes (Quispe-Coica y Pérez-Foguet, 2018). Una primera alternativa es excluir del análisis los datos incompletos, pero esto puede dar lugar a sesgos (Strike et al., 2001), a graves pérdidas de información, a estimaciones inexactas que no ayudan a los gestores a tomar las mejores decisiones, etc. Existen diferentes alternativas basadas en completar los datos faltantes, incluyendo un reemplazo multiplicativo de Martín-Fernández et al. (2003), un algoritmo EM modificado de Palarea-Albaladejoy Martín-Fernández (2008) y un método clásico y robusto de imputación de Hron et al. (2010); sin embargo, aún no se han determinado las técnicas más adecuadas para los casos específicos del sector ASH.

Por último, la calidad de los datos disponibles puede clasificarse en muchos casos como baja o muy baja. El PCM valida los datos y los metadatos (información de la fuente de datos) uno por uno para determinar qué puede utilizarse. Las discrepancias entre los datos no son en sí un motivo de exclusión. Por citar un ejemplo, el porcentaje de la población con acceso a agua entubada en las zonas rurales de Indonesia fue reportado como 6.6% por la Encuesta Nacional Socioeconómica en 2016, pero otra fuente de información reportó que era 41.5% (Performance Monitoring y Accountability; PMA16) (WHO/UNICEF, 2019b). Esto se deriva del uso de múltiples fuentes de información y no es fácil de remediar automáticamente, sin embargo, influye directamente en las estimaciones obtenidas bajo cualquier modelo. Recientemente, Ezbakhe y Pérez-Foguet (2019) propusieron un método para tratar las incertidumbres que se originan en el muestreo estadístico, utilizando modelos de composición de tendencias aplicados a los datos de agua y saneamiento. Sin embargo, queda pendiente completar la validación puntual del PCM con técnicas y procedimientos para la detección de valores atípicos u otros errores en los datos distintos del muestreo (Bain et al., 2018). Por ello, es necesario evaluar alternativas de identificación para CoDa del sector ASH.

Cuando se trabaja con CoDa, los valores atípicos no pueden identificarse para una variable independientemente del resto; son necesarios métodos de análisis multivariante para facilitar la detección adecuada de los valores atípicos y permitir la identificación de los datos con errores evidentes, que pueden alterar las estimaciones (Filzmoser y Hron, 2008; Filzmoser et al., 2009, 2012). Filzmoser y Hron (2008) propusieron el uso de técnicas de identificación robusta basadas en la distancia de Mahalanobis (MD; acrónimo en inglés de Mahalanobis distance). La propuesta se aplica a modelos de regresión generales, como el GAM. Sin embargo, la escasa cantidad de datos que tienen algunos países puede limitar el uso de esta aplicación. Otras alternativas, como la regresión por mínimos cuadrados ordinarios (OLS; acrónimo en inglés de ordinary least squares), ofrecen una mejor opción en esos casos. Sin embargo, la aplicación directa de OLS no es conveniente, ya que puede verse influida negativamente por la presencia de valores atípicos.

Por lo tanto, es necesario aplicar estimadores robustos para los modelos de regresión lineal. En la literatura existen varios métodos para ello, como la estimación M y la estimación S (Rousseeuw y Yohai, 1984) la estimación MM (Yohai, 1987) y otros (para más detalle véase Maronna et al.,2019). En este estudio se aplican los estimadores de tipo MM, en base a los

buenos resultados obtenidos con ellos en otros estudios. Cabe añadir que las estimaciones robustas no excluyen necesariamente los valores atípicos, sino que modulan su influencia en el modelo calibrado, lo que le confiere una gran ventaja para su uso con datos limitados.

Este trabajo propone y analiza diferentes estrategias acopladas para el tratamiento de ceros, datos perdidos y valores atípicos en los modelos de tendencia composicional, aplicados al seguimiento internacional del sector ASH, completando el trabajo anterior en este sentido y facilitando su aplicación práctica a los datos disponibles. En concreto, aborda los siguientes objetivos:

- Evaluar alternativas para el tratamiento de los ceros o de los datos ausentes, o de ambos simultáneamente, utilizando métodos robustos;
- Identificar y tratar los valores atípicos mediante métodos robustos de forma diferenciada para contextos con pocos o muchos (más de seis) datos temporales diferentes, según la clasificación de Fuller et al. (2016).

Para ello, se ha seleccionado un conjunto de doce tendencias, con diferentes características, que son representativas a nivel internacional y dentro del conjunto de situaciones del sector, tanto para entornos urbanos como rurales de ASH.

4.2. Método

El algoritmo propuesto y mostrado en la Figura 4.1 sigue procedimientos y técnicas estadísticas para CoDa que pueden aplicarse y reproducirse fácilmente en cualquier sector o área de análisis. Para entenderlos, hay que conocer primero algunos conceptos básicos, como: i) los CoDa representan vectores, con D partes estrictamente positivos, y la suma es una constante "k", como se muestra en la Ec. (4.1); ii) su espacio muestral es el simplex S^D; para el análisis estadístico, es necesario pasar al espacio euclidiano utilizando transformadas ilr, lo que requiere que los componentes D se pasen a (D-1).

$$S^D = X = (X_i, ..., X_D) \quad \forall X_i > 0, \; i = 1,2,....,D \quad \sum_{X_i=1}^{D} X_i = k \tag{4.1}$$

k: puede ser 1, 100 o cualquier otra constante positiva.

Estos conceptos y términos, aunque parecen sencillos, no son habituales en el sector ASH. Por lo tanto, es necesario tenerlos claros, para entender el método de análisis en CoDa.

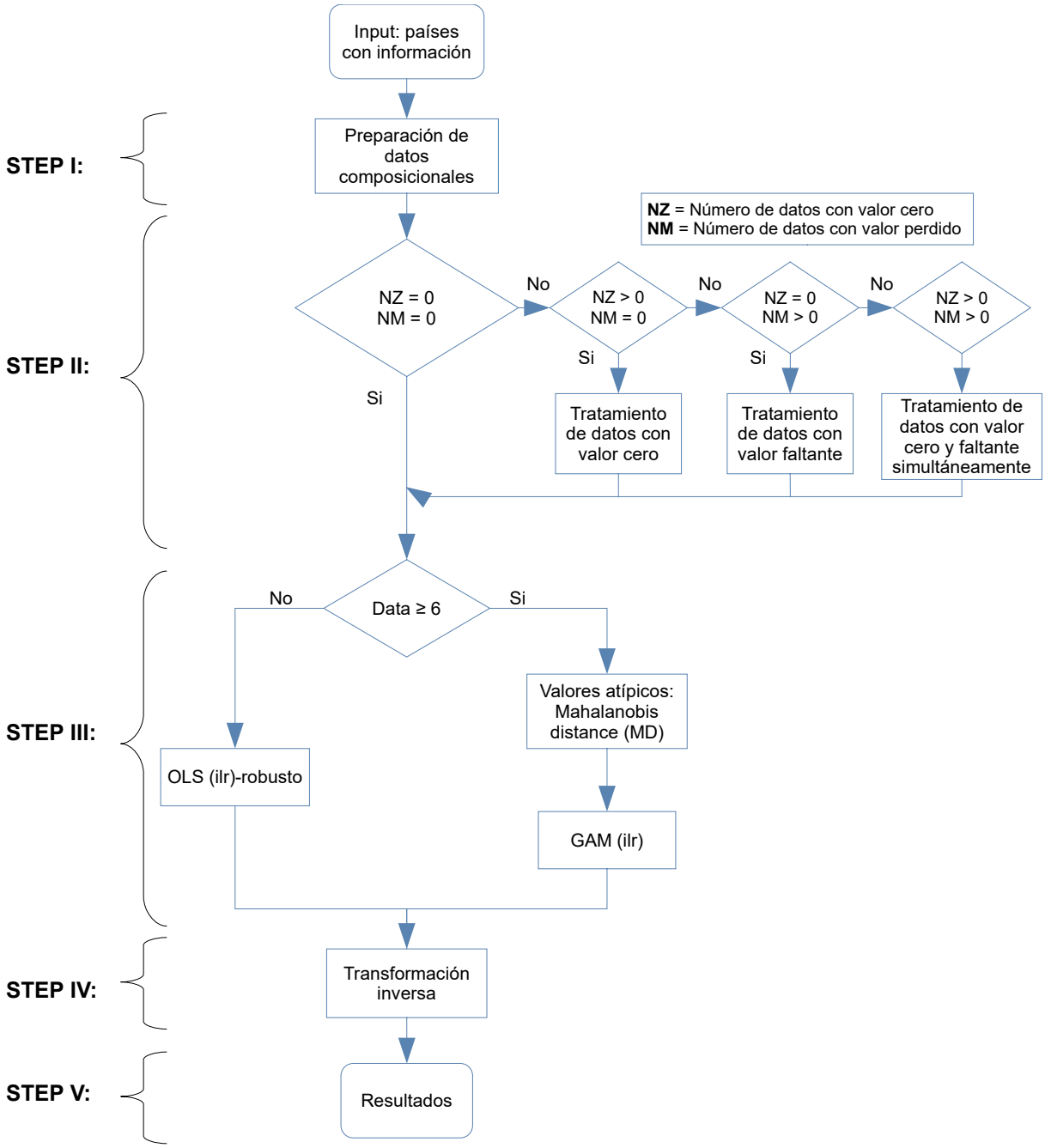


Figura 4.1. Análisis estadístico de CoDa en el sector ASH.

STEP I. Preparación de los datos de composición (CoDa)

Si la información está en unidades de población (P), las proporciones de las categorías de servicios se forman según la Ec. (4.2). Posteriormente, se construyen vectores con las partes, en los que la suma es una constante "k" (100% si se da en porcentaje, o uno si se da en proporciones). En los vectores que faltan datos, se utiliza la denotación "NA".

$$X_1^4(\%) = \frac{P_1^4}{\sum_{i=1}^{4} P_i} \times 100 \qquad (4.2)$$

$$X_1 + X_2 + X_3 + X_4 = 100 \quad (4.3)$$

$$X_{h1} + X_{h2} + X_{h3} = 100 \quad (4.4)$$

Los indicadores se forman de acuerdo con la Tabla 4.1. El agua y el saneamiento están representados en la Ec. (4.3) y cada uno de ellos consta de cuatro partes, mientras que los indicadores de higiene están representados en la Ec. (4.4) y constan de tres partes.

Tabla 4.1. Indicadores de composición del agua (A) y del saneamiento (S).
Agua (W): entubada, otros mejorados, superficial y otros no mejoradas.
Saneamiento (S): alcantarillado, otros mejorados, defecación al aire libre y otros no mejorados

Servicios		Indicador	
Agua y saneamiento	Mejorado (I)	X_1 (Agua entubada o alcantarillado)	X_1
		X_2 (otros mejorados; A o S)	I - X_1
	No mejorada (U)	X_3 (acceso a agua superficial o defecación al aire libre)	X_3
		X_4 (otros no mejorados; W o S)	U-X_3
Higiene	Instalación de lavado de manos en el local (H)	X_{h1} (servicios básicos)	X_{h1}
		X_{h2} (servicios limitados)	H - X_{h1}
	No hay instalaciones para lavarse las manos	X_{h3} (sin servicios)	100 - H

Los vectores de composición que presentan datos irregulares (por ejemplo, que son nulos, faltantes, o tanto nulos como faltantes simultáneamente) y los valores atípicos se tratan con funciones que implican transformaciones ilr según la Ec. (4.5) de Egozcue et al. (2003), cada una con particularidades en los balances *V*. Este procedimiento también se aplica para generar los modelos.

$$Y = ilr = \sqrt{\frac{r \times s}{r + s}} \ln \frac{g_m(X_r+)}{g_m(X_s-)} \quad (4.5)$$

r = número de variables positivas en el balance *V*
s = número de variables negativas en el balance *V*
g_m(-) = media geométrica de las variables

Sin embargo, para ilustrar el comportamiento de los modelos en los datos transformados, se realiza un tipo de balance, coherente con la forma habitual de análisis en el sector ASH. Por ejemplo, el seguimiento mundial se basa en la clasificación del acceso a los servicios de agua y saneamiento mejorados y no mejorados, que posteriormente se subdividen en categorías de servicios (WHO/UNICEF, 2017; Turman-Bryant et al., 2018); Asimismo, tanto las desigualdades en el acceso al agua como al saneamiento (Yang et al., 2013; Bain et al., 2014; WHO/UNICEF, 2019a; Anthonj et al., 2020b; Chitonge et al., 2020) como los estudios sobre el acceso a ASH y su relación con la salud (Prüss-Ustün et al., 2014; Freeman et al., 2017; Ashole Alto et al., 2020; Hasan y Alam, 2020; Patel et al., 2020) implican de un modo u otro la clasificación de servicios mejorados y no mejorados. Por lo tanto, el orden de los balances (Egozcue y Pawlowsky-Glahn, 2005) se define bajo este criterio (ver Figura 4.2), con el desglose de cada parte de la siguiente manera:

Los balances de agua y saneamiento constan de cuatro partes cada uno y siguen el mismo procedimiento (V_1) con el balance realizado entre la proporción de la población:

i. con acceso a servicios mejorados ($X_1 \times X_2$) y no mejorados ($X_3 \times X_4$);
ii. a continuación, con el acceso a los servicios de red (X_1) y otras formas de acceso mejoradas (X_2);
iii. por último, con el acceso a los servicios (X_3) y otras formas de acceso no mejoradas (X_4).

Los balances de higiene constan de tres partes y se realizan según el siguiente procedimiento (V_{h1}). El balance se realiza entre la proporción de la población:

i. con una instalación de lavado de manos en el local ($X_{h1} \times X_{h2}$) y sin instalación de lavado de manos (sin servicio) (X_{h3});
ii. a continuación, con acceso a servicios básicos (Xh1) y servicio limitado (X_{h2}).

$$V_1 = \begin{matrix} \\ ilr_1 \leftarrow \\ ilr_2 \leftarrow \\ ilr_3 \leftarrow \end{matrix} \begin{matrix} \begin{matrix} X_1 & X_2 & X_3 & X_4 \end{matrix} \\ \begin{pmatrix} +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{pmatrix} \end{matrix} \quad V_{h1} = \begin{matrix} \\ ilr_{h1} \leftarrow \\ ilr_{h2} \leftarrow \end{matrix} \begin{matrix} \begin{matrix} X_{h1} & X_{h2} & X_{h3} \end{matrix} \\ \begin{pmatrix} +1 & +1 & -1 \\ +1 & -1 & 0 \end{pmatrix} \end{matrix}$$

Figura 4.2. Balance en ASH.

El resultado de los balances se muestra en la Figura 4.2 y las transformaciones se muestran en Tabla 4.2.

Tabla 4.2. Transformaciones ilr

A. Agua y saneamiento	B. Higiene
$\mathrm{Y}_1^t = ilr_1 = \sqrt{\frac{2\times 2}{2+2}} \ln \frac{(\mathrm{X}_1 \times \mathrm{X}_2)^{1/2}}{(\mathrm{X}_3 \times \mathrm{X}_4)^{1/2}}$	$\mathrm{Y}_{h1}{}^t = ilr_{h1} = \sqrt{\frac{2\times 1}{2+1}} \ln \frac{(\mathrm{X}_{h1} \times \mathrm{X}_{h2})^{1/2}}{(\mathrm{X}_{h3})}$
$\mathrm{Y}_2^t = ilr_2 = \sqrt{\frac{1\times 1}{1+1}} \ln \frac{(\mathrm{X}_1)}{(\mathrm{X}_2)}$	$\mathrm{Y}_{h2}{}^t = ilr_{h2} = \sqrt{\frac{1\times 1}{1+1}} \ln \frac{(\mathrm{X}_{h1})}{(\mathrm{X}_{h2})}$
$\mathrm{Y}_3^t = ilr_3 = \sqrt{\frac{1\times 1}{1+1}} \ln \frac{(\mathrm{X}_3)}{(\mathrm{X}_4)}$	

STEP II. Tratamiento de los valores de cero, datos perdidos y cero más datos perdidos simultáneamente

Los países con datos que incluyen valores de cero, valores perdidos o valores de cero más valores perdidos simultáneamente se abordan de forma diferenciada con técnicas estadísticas robustas, ya que la baja calidad de los datos del sector puede influir en las imputaciones (Hron et al., 2010; Martín-Fernández et al., 2012; Maronna et al., 2019). Para los tres casos mencionados, se comparan dos alternativas de tratamiento. El número de ceros se denota por NZ, y el número de valores perdidos, por NM.

i) NZ = 0, NM = 0: no se realiza ningún tratamiento previo de los datos.
ii) NZ > 0, NM = 0: tratamiento de los valores nulos con dos variantes del algoritmo de maximización de expectativas (EM; acrónimo en inglés de Expectation–maximization) logarítmicas; función `lrEM` (Palarea-Albaladejo y Martín-Fernández, 2015) e `impRZilr` (Templ et al., 2019).
iii) NZ = 0, NM > 0: tratamiento de los valores perdidos mediante mínimos cuadrados recortados (LTS) (Hron et al., 2010), implementado en la función `impCoda`, o con el mismo algoritmo EM de logaritmo utilizado para el caso (*ii*) (NZ > 0 y NM = 0); función `lrEM`.

iv) NZ > 0, NM > 0: tratamiento de los valores nulos y ausentes simultáneamente, también con dos alternativas. Una es considerar los valores cero como un tipo especial de valores perdidos (Palarea-Albaladejo y Martín-Fernández, 2008; Martín-Fernández et al., 2011) y aplicar el mismo algoritmo LTS que antes (por ejemplo, la función `impCoda`). No se debe considerar lo contrario porque los valores perdidos no son necesariamente valores cero. La otra alternativa es utilizar la versión ampliada del algoritmo log-ratio EM, la función `lrEMplus`, presentada por Palarea-Albaladejo y Martin-Fernandez (2020).

STEP III. Modelos y estimaciones

Los países se clasifican en dos grupos según la cantidad de datos, siendo seis el límite de separación. Esta clasificación se describe en Fuller et al. (2016). Sin embargo, como la baja calidad de los datos también afecta a la capacidad de predicción de los modelos, optamos por realizar modelos robustos en ambos grupos, como se detalla a continuación:

Para los países con puntos de datos < 6: los modelos se construyen utilizando el método de regresión robusta OLS sobre los datos transformados de Tabla 4.2, para lo cual se utiliza la función `lmrob`, que calcula un estimador de regresión del tipo MM como se ha descrito anteriormente (Yohai, 1987; Yohai et al., 1991; Koller y Stahel, 2011). La evaluación de la influencia de los valores atípicos en los modelos de regresión lineal se lleva a cabo mediante pesos de robustez. Se añaden modelos de regresión lineal estándar sobre los datos transformados y no transformados para compararlos con la alternativa robusta.

Países con puntos de datos ≥ 6: el procedimiento de ajuste del modelo combina el método de identificación de valores atípicos como parte del preprocesamiento y luego excluye estos datos del análisis para generar modelos robustos, como se describe a continuación:

i. Los valores atípicos en los datos multivariantes se identifican mediante el cálculo de la distancia de Mahalanobis robusta (Ec. (4.6)) en coordenadas log-ratio isométricas de la Ecuación (4.5). Para el cálculo computacional, se aplica la función `outCoDa` (Templ et al., 2011).

$$\mathrm{MD}\left(Y_i^n\right) = \left[\left(Y_i - T\right)' C^{-1}\left(Y_i - T\right)\right]^{1/2} \; for\; i = 1,2,\ldots,n \qquad (4.6)$$

donde T y C son estimadores de la localización y la covarianza, respectivamente (Mahalanobis, 1936). La robustez se consigue cambiando T y C por el determinante de covarianza mínima (MCD; acrónimo en ingles de minimum covariance determinant), que son estimadores robustos (Filzmoser y Hron, 2008). Los valores atípicos potenciales son aquellos que tienen un MD robusto (cuadrado) mayor que el valor de corte, que es el cuantil 0.975 de la distribución χ^2_{D-1} con grados de libertad D-1 (Rousseeuw y van Zomeren, 1990). En el caso del agua y el saneamiento, el grado de libertad de la distribución chi-cuadrado es tres, y el valor de corte es 3.0575. Los puntos que están por encima de la distancia de corte no se tienen en cuenta en las estimaciones posteriores ($\mathrm{MD}\left(Y_i\right)^2 > \chi^2_{3,\,0.975}$).

ii. Después de identificar los valores atípicos, se construyen modelos de regresión con GAM, con cuatro grados de libertad (k = 4), sobre los datos transformados de la Tabla 4.2. El análisis se realiza para los datos con y sin presencia de valores atípicos. La capacidad predictiva del modelo entre ambos se compara con el coeficiente de determinación ajustado

(R-adj); los valores cercanos a uno la capacidad predictiva del modelo es mejor. El cálculo computacional para generar los modelos se realiza con la función `gam`.

STEP IV. Transformación inversa

Los valores de interpolación o extrapolación en los datos transformados se devuelven al espacio simplex, para lo cual se realiza la transformación inversa con la Ec. (4.7).

$$X = ilr^{-1}(\mathrm{Y}^t) \tag{4.7}$$

X = Vector de la Ec. (4.3) o de la Ec. (4.4).

Para el sector ASH, es importante ver las interpolaciones y extrapolaciones de los modelos en las diferentes categorías de acceso a ASH. Por lo tanto, es obligatorio realizar una transformación inversa.

STEP V. Resultados y prueba de calidad

Todo el proceso del algoritmo descrito hasta el STEP IV permite evaluar y comparar las interpolaciones y extrapolaciones de las diferentes alternativas en las categorías de acceso a ASH, utilizando métricas de calidad. Para ver el impacto de las alternativas en el STEP II en la escala de datos, se aplica la métrica del error cuadrático medio (RMSE; acrónimo en inglés de root mean square error) a los modelos expresados en términos de X. Por otro lado, la evaluación de la capacidad predictiva de los modelos en los datos se realiza mediante el indicador adimensional de bondad de ajuste de la eficiencia de Nash Sutcliffe (NSE) (Nash y Sutcliffe, 1970) aplicado a la X observada y estimada del modelo. Si NSE = 1, el ajuste del modelo es perfecto, mientras que NSE < 1 sugiere que la media observada es un mejor predictor que el modelo (Pérez-Foguet et al., 2017; Ezbakhe y Pérez-Foguet, 2019).

El cálculo estadístico de la Figura 4.1 se realiza mediante R Core Team, (2020)(v.3.6.3). El preproceso de datos y la integración de cada etapa de cálculo se presentan en Quispe-Coica y Pérez-Foguet (2020c). Se utilizan los siguientes paquetes: robCompositions (v2.2.1) de Templ et al. (2011) para `impRZilr`, `impCoda` y `outCoDa`; zCompositions (v1.3.4) de Palarea-Albaladejo y Martín-Fernández (2015)para lrEM y lrEMplus; robustbase (0.93-5) de Maechler et al. (2019) para `lmrob`; mgcv (v1.8-31) de Wood (2019) para `gam`; y compositions (v1.40-3) de (Boogaart et al., 2019).

4.3. Característica de los datos

Para probar el algoritmo propuesto en la Figura 4.1, seleccionamos diez países diferentes para los datos de acceso al agua y al saneamiento, y dos países para el caso de la higiene. Los datos anuales se extraen de la base de datos del PCM de 2000 a 2019, en la que tanto la cantidad de datos como la presencia o ausencia de irregularidades varían en diferentes proporciones, lo que permite abarcar las diversas situaciones que se dan en el sector de ASH (véase Tabla 4.3).

Los países que no presentan irregularidades en los datos están representados por Benin y Ghana para el acceso a la higiene, y por Indonesia para el acceso al agua rural. En el caso de la higiene, la escasa cantidad de datos se debe principalmente a la reciente incorporación de esta en los Objetivos de Desarrollo Sostenible (ODS 6.2) como parte de los indicadores de seguimiento (Craven et al., 2013); en cambio, el acceso al agua y al saneamiento se supervisa desde 1990 (Bartram et al., 2014). En este tipo de datos no se aplica el STEP II del algoritmo.

Tabla 4.3 . Acceso a agua, saneamiento e higiene (ASH)

Región	País	Sector	Servicio	Puntos de datos (X) [a]	Valor cero	Valor perdido
África subsahariana	Sudáfrica	Rural	Saneamiento	30 (×4)	0.00%	1.67%
América Latina y el Caribe	Brasil	Urbano	Agua	27 (×4)	0.00%	44.44%
Asia oriental y sudoriental	Indonesia	Rural	Agua	26 (×4)	0.00%	0.00%
África subsahariana	Nigeria	Rural	Agua	22 (×4)	1.14%	0.00%
América Latina y el Caribe	Paraguay	Urbano	Agua	21 (×4)	7.14%	0.00%
Asia central y meridional	Bangladesh	Rural	Saneamiento	20 (×4)	1.25%	30.00%
África subsahariana	Zambia	Rural	Saneamiento	16 (×4)	0.00%	6.25%
Norte de África y Asia Occidental	Egipto	Urbano	Agua	15 (×4)	10.00%	30.00%
América Latina y el Caribe	Uruguay	Urbano	Agua	15 (×4)	15.00%	3.33%
África subsahariana	Benin	Rural	Saneamiento	10 (×4)	0.00%	10.00%
África subsahariana	Benin	Rural	Higiene	5 (×3)	0.00%	0.00%
África subsahariana	Ghana	Rural	Higiene	4 (×3)	0.00%	0.00%

[a] Los puntos de datos del año están representados por tres o cuatro niveles de servicios ASH a los que la población tiene acceso.

Los datos con irregularidades se presentan de tres formas diferentes:

i) El primer caso está representado por Nigeria y Paraguay, que tienen valores nulos en los datos, de 1.14% y 7.14%, respectivamente. Las categorías de Paraguay revelan que esto ocurre cuando la provisión de servicios de agua por fuentes mejoradas es alta (Figura 4.3A); en consecuencia, los indicadores de acceso al agua no mejorada tienen tendencias nulas o valores cero. Otra particularidad que se observa en Paraguay es que el valor cero se presenta sólo en el indicador X_3, mientras que en Egipto se presenta en X_3 y X_4.

ii) El segundo caso se refiere a los países con valores ausentes en los datos y están representados por Sudáfrica, Zambia, Brasil y Benin. Brasil tiene el mayor porcentaje de valores perdidos (del 44.44%), que se distribuyen en las mismas proporciones en los indicadores X_3 y X_4 (Figura 4.3B).

iii) El tercer caso se refiere a los países que tienen tanto valores cero como valores ausentes en los datos y están representados por Bangladesh, Egipto y Uruguay. Egipto se muestra como ejemplo en Figura 4.3C para los datos con valores cero, y en la Figura 4.3D para los datos con valores perdidos. En ambos gráficos, los datos con valores cero y valores perdidos están en las categorías de X_3 y X_4.

Las irregularidades de los datos deben abordarse en STEP II, utilizando las funciones de imputación más adecuadas para cada caso.

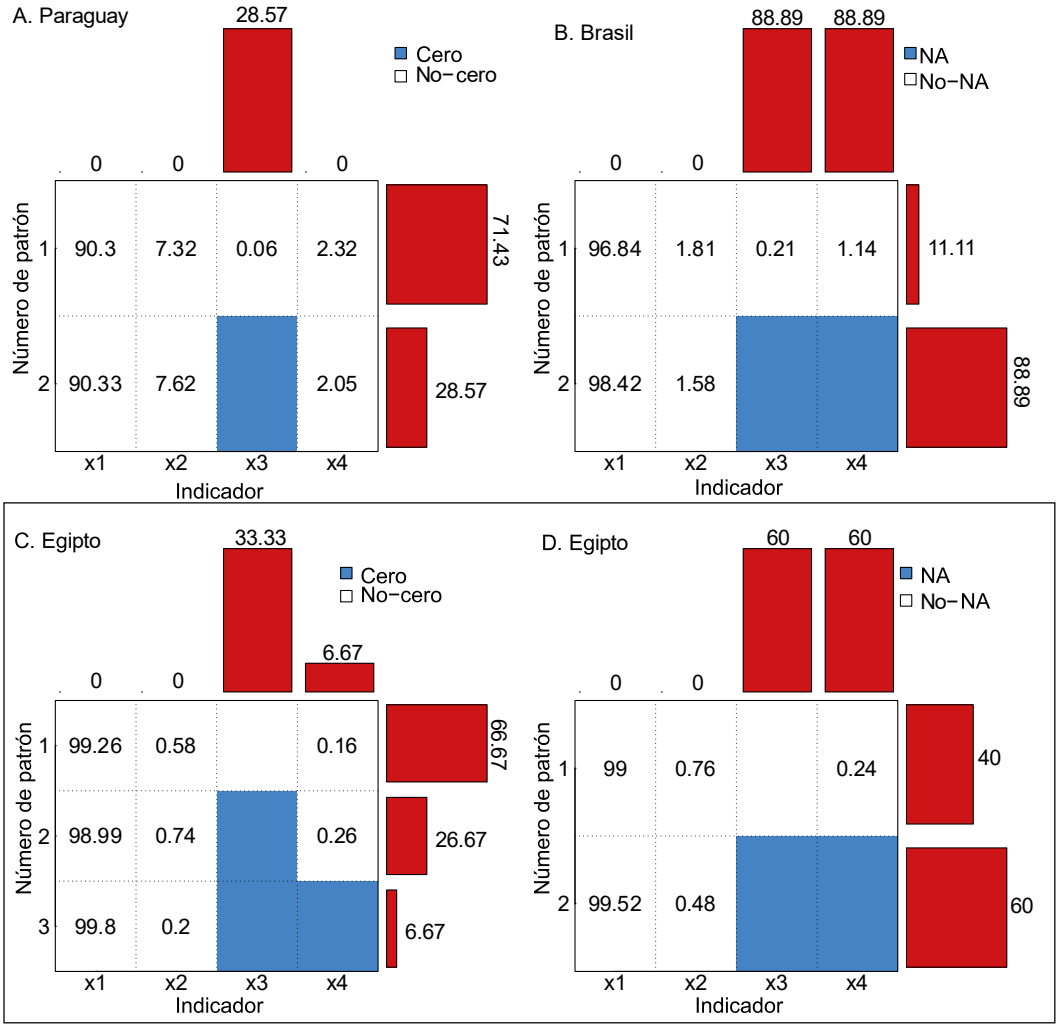


Figura 4.3. Países con datos irregulares.

Notas: A) Paraguay: patrones de valores nulos. B) Brasil: patrones de valores perdidos. (C y D) Egipto: valores cero y valores perdidos simultáneamente, con valores cero mostrados en C), y valores perdidos mostrados en D).

4.4. Resultados y discusión

En esta sección, discutimos y comparamos las alternativas de trabajo para tratar los valores de cero, los valores perdidos y ambos simultáneamente que suelen estar presentes en los datos. Posteriormente, analizamos la influencia de los valores atípicos en el modelo.

4.4.1. Países con datos con valor cero, valor perdido o simultáneamente ambos

De las diferentes características de los datos presentados en la sección 4.3, los países con datos irregulares han pasado por métodos de tratamiento diferenciados en STEP II. Por ejemplo, Paraguay y Nigeria tienen valores nulos en sus datos, del 7.14% y el 1.14%, respectivamente (Tabla 4.4A). El análisis bajo la métrica RMSE de las funciones de imputación `lrEM` e `impRZilr` que sustituyen los valores cero por valores pequeños no muestra diferencias significativas que permitan descartar por completo alguna de las dos alternativas; por tanto, ambas funciones pueden ser aplicables, ya que cualquiera de ellas ayuda a superar el problema de no poder realizar las transformaciones ilr de la Ec. (4.5).

Sin embargo, al tratar los valores perdidos en los datos (Tabla 4.4B), hay que tener en cuenta las diferencias en las métricas, lo que nos hace elegir la función `impCoDa` o la función `lrEM` según el país de análisis. Por ejemplo, las métricas de la función `impCoDa` son mejores para Benin y Sudáfrica, mientras que la función `lrEM` es mejor para Brasil; en cambio, para Zambia no hay diferencias significativas entre ninguna de las dos funciones (`impCoda` o `lrEM`).

Tabla 4.4. Métricas de calidad para seleccionar el método

País	Sector	Servicio	Valor cero	Valor perdido	Método - RMSE (%)				Método seleccionado
					impCoda	lrEM	lrEMplus	impRZilr	
A. Caso II: Datos con valores cero									
Paraguay	Urbano	Agua	7.14%	0.00%	-	0.0026	-	0.0027	lrEM
Nigeria	Rural	Agua	1.14%	0.00%	-	0.0060	-	0.0033	impRZilr
B. Caso III: Datos con valores perdidos									
Benin	Rural	Saneamiento	0.00%	10.00%	1.826	3.094	-	-	impCoda
Brasil	Urbano	Agua	0.00%	44.44%	0.648	0.321	-	-	lrEM
Sudáfrica	Rural	Saneamiento	0.00%	1.67%	0.015	0.026	-	-	impCoda
Zambia	Rural	Saneamiento	0.00%	6.25%	8.435	8.227	-	-	lrEM
C. Caso IV: Datos con valores cero y valores perdidos simultáneamente									
Bangladesh	Rural	Saneamiento	1.25%	30.00%	7.621 **(ª)**	-	8.690	-	impCoda
Egipto	Urbano	Agua	10.00%	30.00%	0.254 **(ª)**	-	0.269	-	impCoda
Uruguay	Urbano	Agua	15.00%	3.33%	0.052 **(ª)**	-	0.048	-	lrEMplus

Nota: [a] Los datos con valores de cero se consideran valores perdidos ("0" → "NA"); por tanto, se aplican métodos de imputación con la función impCoda.

Por otra parte, en los países con valores cero y valores perdidos simultáneamente (véase Tabla 4.4), la alternativa de sustituir los valores cero por "NA" y tratarlos como "valores perdidos" con la función `impCoda` da mejores resultados para Bangladesh y Egipto. Esto ocurre cuando hay un mayor porcentaje de datos con valores perdidos que con valores cero. Sin embargo, la situación contraria se da en el conjunto de datos de Uruguay, que tiene un 15% de valores cero y un 3.33% de valores perdidos, y para el que la función `lrEMplus` es una mejor alternativa.

Por último, si bien es cierto que cualquiera de los métodos evaluados es adecuado para al menos uno de los casos (dependiendo de cada caso), todos los métodos son ya mejores que las alternativas de imputación multiplicativa u otras alternativas simples, ya que permiten que exista variabilidad en los datos imputados. Esta ventaja es más significativa cuando los puntos de datos presentan un mayor porcentaje de estas irregularidades. Si no se aplica ninguna alternativa (ya sea simple o una de las que se muestran en este documento), muchos países del sector deben ser excluidos del análisis. Esto es especialmente importante si la pérdida de información es significativa (como ocurre en los países sudamericanos; Quispe-Coica y Pérez-Foguet, 2018). Por otro lado, una vez acordados los nuevos Objetivos de Desarrollo Sostenible (United Nations General Assembly, 2015; UN Water, 2016), cada país asumió la responsabilidad de reducir el acceso de la población a los servicios no mejorados de ASH. Para ello, muchos países están definiendo e implementando políticas públicas que cierren estas brechas, en cuyo caso los datos tenderán a irse a los valores extremos, haciendo aún más necesario el uso de alternativas de imputación para los valores cero.

4.4.2. Valores atípicos

4.4.2.1. Países con puntos de datos < 6

Esta sección aborda el caso de los países con pocos datos, donde la influencia de los valores atípicos se penaliza en el modelo acoplado. El acceso de las poblaciones rurales a los diferentes niveles de servicios de higiene en Benin y Ghana ilustra esta situación. En Figura 4.4A y E, presentamos el ajuste del modelo en los datos transformados mediante regresión lineal estándar y robusta. Las líneas de regresión de ambos métodos son similares en las transformaciones de ilr2, y difieren para ambos métodos en ilr1 (con cambios más drásticos en la subfigura E). La diferencia se debe principalmente a que, en el método robusto, los puntos 1 y 5 de Ghana y Benin, respectivamente, tienen un fuerte grado de influencia negativa en el modelo, por lo que asigna pesos de robustez de valor cero.

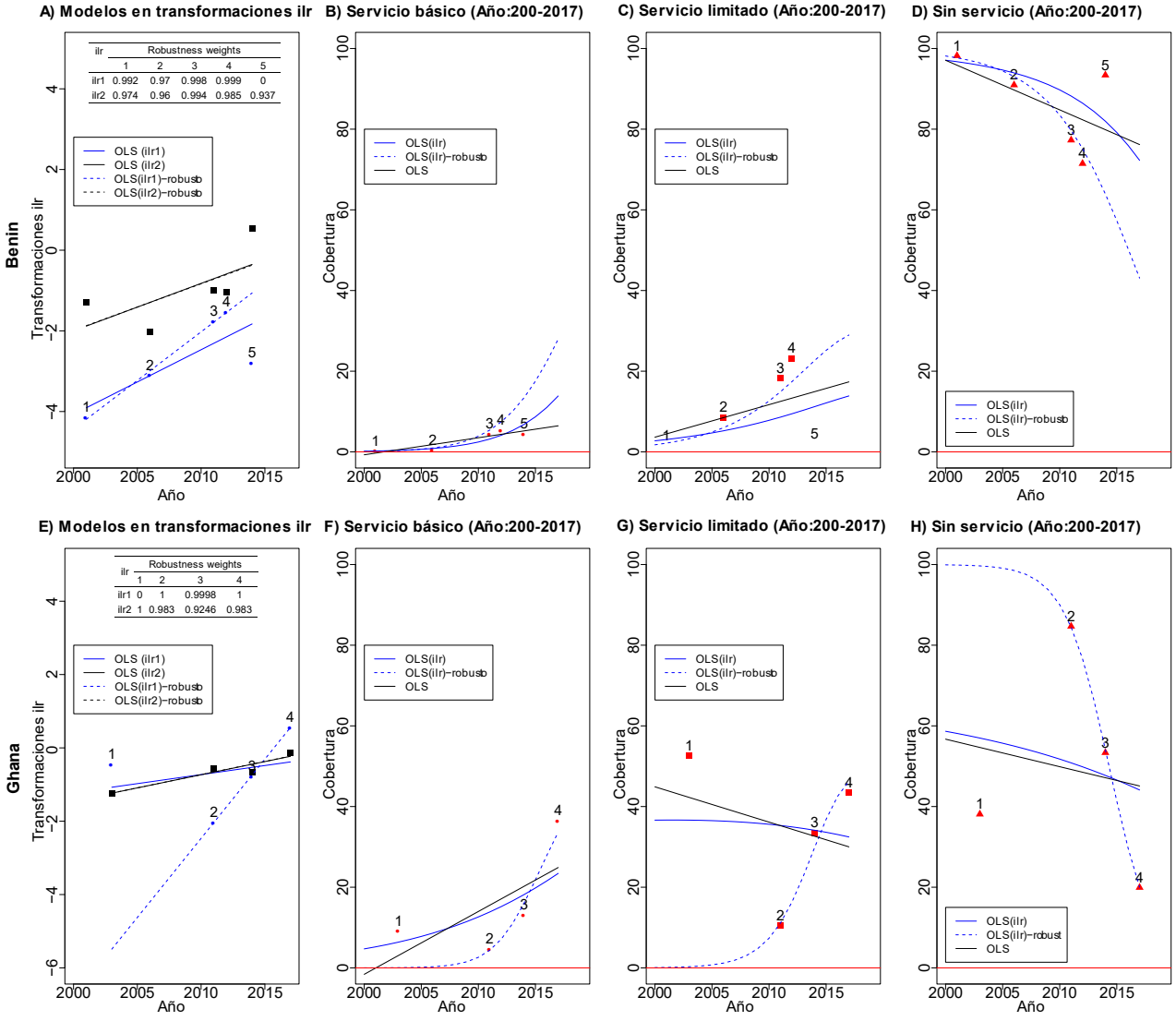


Figura 4.4. Modelos robustos con OLS

Notas: Modelo y estimaciones en CoDa de la higiene. (A, E) Se ajustan dos modelos diferentes en los datos transformados: i) OLS estándar (líneas sólidas azules y negras) y ii) OLS robusto (líneas discontinuas azules y negras). (B-D, F-H) Se ajustan tres modelos diferentes en CoDa: i) OLS estándar en datos originales (línea sólida negra), ii) inverso de OLS estándar de datos transformados (línea sólida azul) y iii) inverso de OLS robusto de datos transformados (línea discontinua azul).

La forma en que se modula la influencia de los datos crea diferencias significativas en las estimaciones de las categorías de servicios de higiene a las que accede la población. En el caso de Ghana, el efecto en cada categoría es aún mayor si lo comparamos con las otras alternativas (Figura 4.4F-H). Asimismo, tanto en Benin como en Ghana, la curva generada por OLS robusta (ilr) es la que mejor se ajusta a los datos. Por otra parte, si observamos los resultados cualitativamente, es más razonable excluir el punto 1 en Ghana y el punto 5 en Benin, lo que apoya la afirmación de que la alternativa de regresión lineal robusta es una excelente alternativa para los modelos de regresión en presencia de datos con valores atípicos.

Otra característica a tener en cuenta es que con OLS (ilr) u OLS robusto (ilr), las extrapolaciones de las categorías de servicios en 2000 y 2017 nunca superaron los límites extremos de 0 y 100% (Tabla 4.5). Esto ocurre porque la transformación inversa tiene un valor de cierre (Ec. (4.4)) que permite realizar estimaciones en la serie temporal sin restricciones. Lo contrario ocurre cuando la extrapolación se realiza con OLS estándar. Aquí, los valores negativos de Benin (-0.726) y Ghana (-1.594) en la categoría de servicios básicos del año 2000 ejemplifican esta situación; en estos casos, el PCM aplica restricciones de valor cero (WHO/UNICEF, 2018).

Tabla 4.5. Comparación de los valores estimados con diferentes métodos

Higiene País	Método	Año de estimación (2000)			Año de estimación (2017)		
		Servicio básico	Servicio limitado	No hay servicio	Servicio básico	Servicio limitado	No hay servicio
Benin	OLS (ilr)	0.162	2.740	97.098	13.862	13.868	72.270
	OLS (ilr)-robusto	0.105	1.752	98.142	27.959	28.968	43.073
	OLS [a]	–0.726	3.638	97.088	6.466	17.351	76.183
	Sitio web de PCM [b]	0.000	2.912	97.088	6.043	16.544	77.413
Ghana	OLS (ilr)	4.712	36.629	58.659	23.431	32.464	44.104
	OLS (ilr)-robusto	0.009	0.069	99.923	33.675	46.319	20.006
	OLS [a]	–1.594	44.893	56.701	24.890	30.010	45.100
	Sitio web de PCM [b]	NA	NA	NA	36.576	43.491	19.933

[a] Regresión OLS sobre datos no transformados. [b] Datos disponibles en el sitio web de PCM (Benin y Ghana, pestaña de Excel "Regresiones"). Los valores negativos están subrayados. NA: no disponible.

Por otro lado, los resultados de las estimaciones de 2017 con OLS robusta (ilr) difieren significativamente de las otras alternativas lineales en todas las categorías de Benin. En Ghana, solo los OLS robustos (ilr) y la regresión PCM dan resultados muy similares en las tres categorías. Aunque las alternativas de estimación difieren, existe un alto porcentaje de la población rural que no dispone de instalaciones para lavarse las manos (concretamente, el 43.07% en Benin, y el 20% en Ghana), si tenemos en cuenta los resultados de OLS robustos (ilr). En ambos países, se espera que esta tasa disminuya, dados los efectos positivos del lavado de manos con agua y jabón en la reducción y prevención de enfermedades, como la diarrea, la enfermedad por coronavirus 2019 (COVID-19), la infección respiratoria aguda y el impétigo, entre otras (Luby et al., 2005; Cairncross et al., 2010; Hirai et al., 2017; Prüss-Ustün et al., 2019; Brauer et al., 2020; Ma et al., 2020).

4.4.2.2. Países con puntos de datos ≥ 6

Las posibles razones de los valores atípicos en los datos pueden ser diversas. Sin embargo,

en los datos analizados aquí, es evidente que los valores atípicos se producen comúnmente cuando hay diferentes fuentes de información. Para ilustrar mejor este punto, presentamos el caso de la población rural de Sudáfrica, cuya información para las categorías de alcantarillado en 2011 proviene de tres fuentes diferentes: el Censo (CEN) reportó un 6.03% de acceso, la encuesta de Ingresos y Gastos de los Hogares (IES) reportó un 44.16% de acceso, y la encuesta General de Hogares (GHS) reportó un 5.07% de acceso. Teniendo en cuenta la diferencia significativa entre los datos de la IES y los de las otras dos fuentes de información (CEN y GHS), es normal suponer que se trata de un dato atípico sin necesidad de aplicar ningún método de validación. Por otra parte, dado que los datos del censo y la encuesta IES sólo difieren en un 0.96% puntos porcentuales (p.p.), es difícil saber si uno u otro valor es atípico o no.

Ante la duda que se genera, se puede aplicar la MD robusta a las series temporales del país. Los resultados obtenidos muestran que sólo el punto de datos del IES es un valor atípico (Figura 4.5 A.2), lo que confirma la hipótesis anterior. La validación puntual realizada por el JMP (2019)(véase la pestaña de Excel "Resumen de datos/saneamiento para 2011") identifica y excluye del modelo los puntos de datos del CEN y del IES. Estas diferencias de identificación que se manifiestan para un país y un año concretos pueden darse también para otros países cuando se analiza una serie temporal.

Tabla 4.6. Identificación de valores atípicos en ASH

País	Sector	Servicio	Puntos de datos (X)	Método			
				RMD [a]	PCM [b]		
					Mejorado	X_1	X_3
Sudáfrica	Rural	Saneamiento	30	7	3	7	3
Brasil	Urbano	Agua	27	1	0	0	0
Indonesia	Rural	Agua	26	9	3	4	3
Nigeria	Rural	Agua	22	3	1	0	1
Paraguay	Urbano	Agua	21	8	0	1	0
Bangladesh	Rural	Saneamiento	20	7	2	1	2
Zambia	Rural	Saneamiento	16	3	3	6	5
Egipto	Urbano	Agua	15	1	0	0	0
Uruguay	Urbano	Agua	15	4	1	1	1
Benin	Rural	Saneamiento	10	3	0	0	0

[a] La distancia de Mahalanobis robusta representa todas las partes en un solo punto, y las que superan el umbral se consideran valores atípicos.

[b] El PCM realiza la validación puntual de los datos de cada país. Datos disponibles en el sitio web del PCM (pestaña "Resumen de datos" del país/Excel).

La coherencia y las contradicciones en el número de valores atípicos identificados mediante los dos métodos, la MD robusta y el PCM, se muestran en Tabla 4.6. El número de valores atípicos identificados por la MD robusta es mayor que el identificado por el PCM en nueve de los diez países, siendo Paraguay el que presenta la mayor diferencia, mientras que lo contrario se observa para Zambia en las categorías X_1 y X_3. En cambio, tanto en Sudáfrica como en Zambia, el número de valores atípicos identificados es el mismo entre las dos alternativas (MD robusta y PCM) en las categorías X_1 y mejorada, respectivamente.

Estas diferencias sugieren que la identificación de valores atípicos con el método de análisis habitual del PCM es insuficiente y requiere herramientas adicionales. Por lo tanto, el método MD

robusto refuerza y complementa la forma de análisis habitual. Además, permite identificar metódicamente los valores actuales y otros valores atípicos, lo que reduce el sesgo de identificación. El inconveniente del método MD es que la distancia calculada representa las cuatro partes (véase Figura 4.5 A2, B.2 y C.2), por lo que la exclusión de los puntos que superan el umbral conduce a la pérdida de información para las cuatro categorías del año. Esto no ocurre ni con el método PCM ni con los métodos de identificación de estadísticas univariantes.

Siguiendo la secuencia del algoritmo (Figura 4.1), se puede aplicar STEP III (Figura 4.5). En Indonesia y Sudáfrica, la exclusión de los valores atípicos mejoró la calidad de los modelos de todos los datos transformados (Figura 4.5 A.3 y B.3). Las métricas de calidad R^2 ajustado confirman esta afirmación. Sin embargo, en Uruguay, las métricas de calidad sólo mejoraron en las transformaciones ilr3; esto demuestra la flexibilidad del GAM, que busca ajustarse a los datos, independientemente de si tiene o no valores atípicos. Por otro lado, si bien los modelos se generan en datos transformados, en el sector ASH es más importante ver la calidad de la capacidad predictiva en cada categoría de análisis. Por lo tanto, es necesario devolver las interpolaciones y extrapolaciones del modelo al espacio del simplex, sin descartar que todo lo que ocurra en los datos transformados influya en los resultados de los diferentes niveles de servicio.

Los resultados de la aplicación de la transformación inversa en STEP IV del algoritmo se muestran en Figura 4.6. La presencia de valores atípicos influyó en el ajuste de los modelos de forma diferenciada; esto afectó a las estimaciones. En Indonesia, la estimación del porcentaje de población rural que tiene acceso al agua entubada en 2020 es del 10.8% si el modelo se generó con datos que incluyen valores atípicos; sin embargo, este valor disminuyó al 5.7% si se excluyeron los valores atípicos del análisis, lo que dio lugar a una diferencia del 5.1% p.p. entre las dos estimaciones. En Sudáfrica, esta diferencia aumentó al 7.2% p.p. si analizamos la categoría de la población rural que tiene acceso al saneamiento a través de otras formas mejoradas.

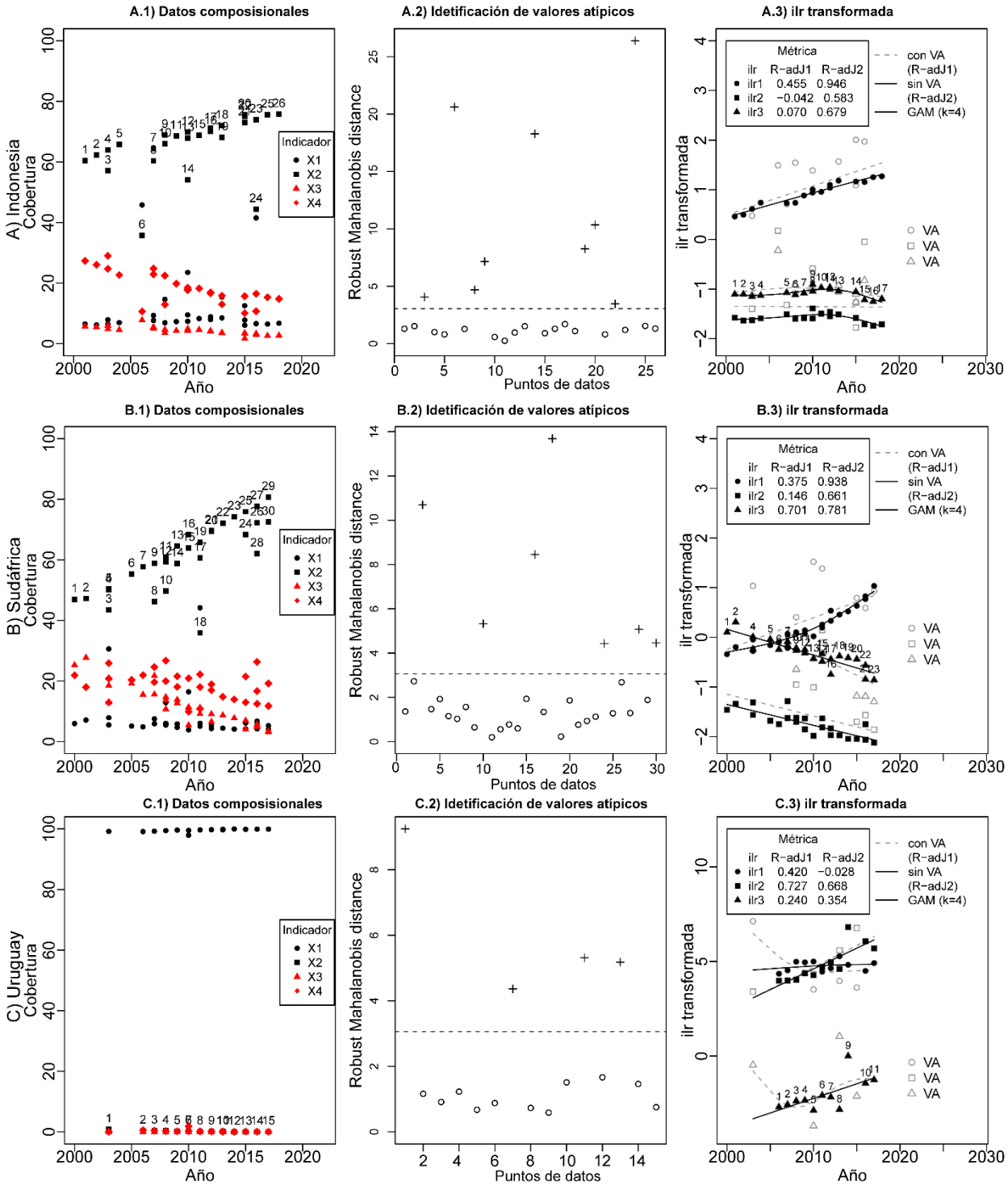


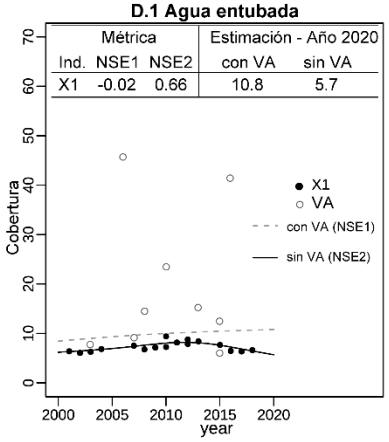
Figura 4.5. Modelos sobre datos transformados

Notas: (A.2, B.2 y C.2) Distancia de Mahalanobis robusta. Las distancias superiores al valor de corte (líneas discontinuas) se consideran valores atípicos (VA). (A.3, B.3 y C.3) Se ajustan dos modelos diferentes en los datos transformados: i) GAM con valores atípicos (líneas sólidas) y ii) GAM sin valores atípicos (líneas discontinuas).

D. Indonesia

D.1 Agua entubada

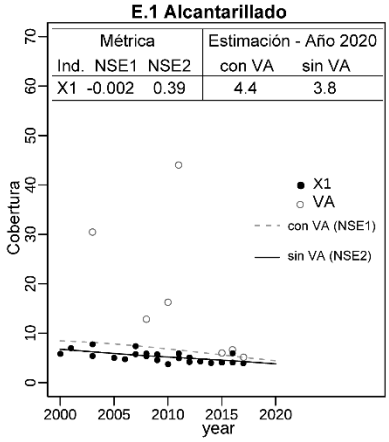
Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X1	-0.02	0.66	10.8	5.7



E. Sudáfrica

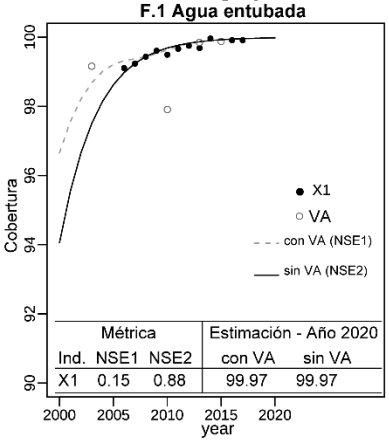
E.1 Alcantarillado

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X1	-0.002	0.39	4.4	3.8



F. Uruguay

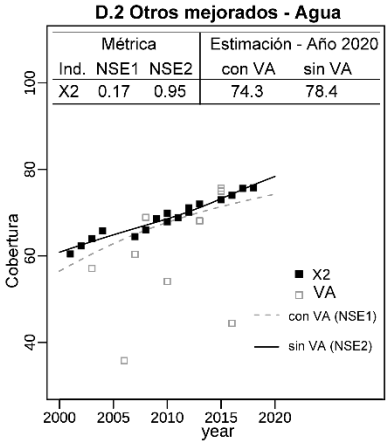
F.1 Agua entubada



Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X1	0.15	0.88	99.97	99.97

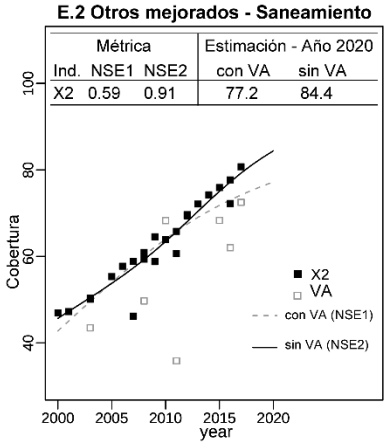
D.2 Otros mejorados - Agua

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X2	0.17	0.95	74.3	78.4



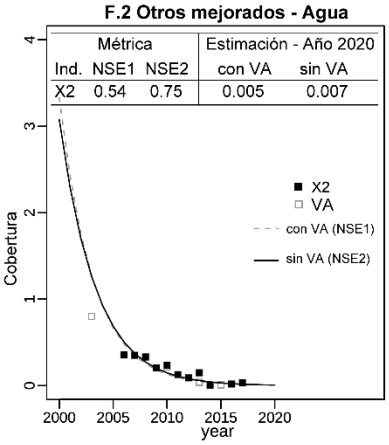
E.2 Otros mejorados - Saneamiento

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X2	0.59	0.91	77.2	84.4



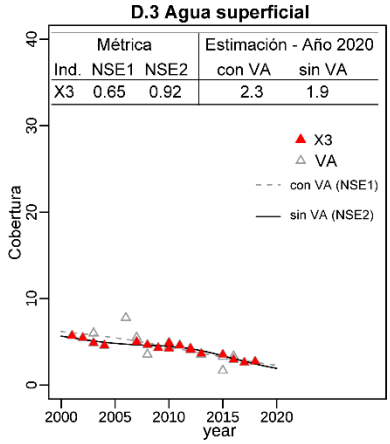
F.2 Otros mejorados - Agua

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X2	0.54	0.75	0.005	0.007



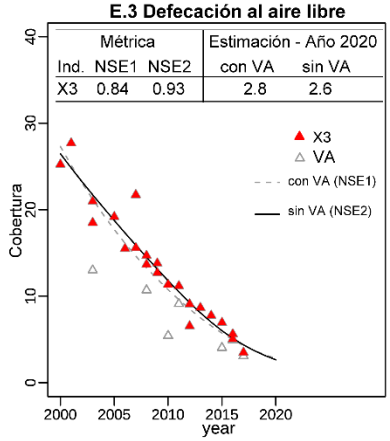
D.3 Agua superficial

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X3	0.65	0.92	2.3	1.9



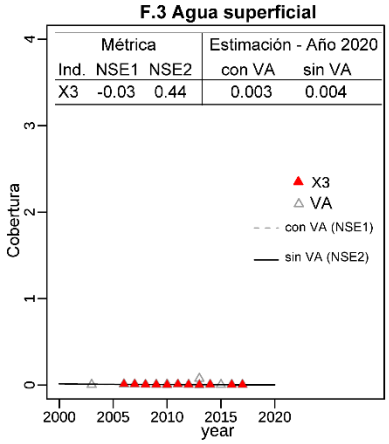
E.3 Defecación al aire libre

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X3	0.84	0.93	2.8	2.6



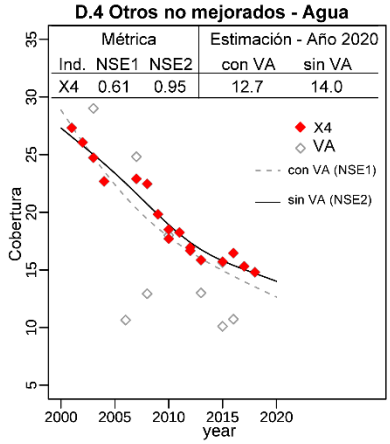
F.3 Agua superficial

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X3	-0.03	0.44	0.003	0.004



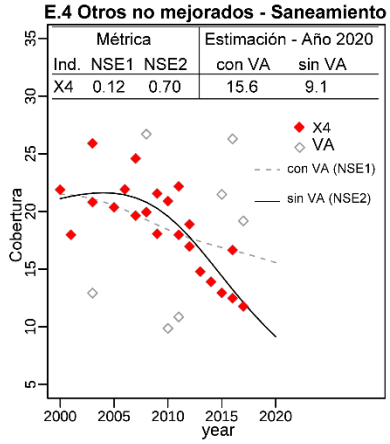
D.4 Otros no mejorados - Agua

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X4	0.61	0.95	12.7	14.0



E.4 Otros no mejorados - Saneamiento

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X4	0.12	0.70	15.6	9.1



F.4 Otros no mejorados - Agua

Métrica			Estimación - Año 2020	
Ind.	NSE1	NSE2	con VA	sin VA
X4	0.06	0.88	0.017	0.010

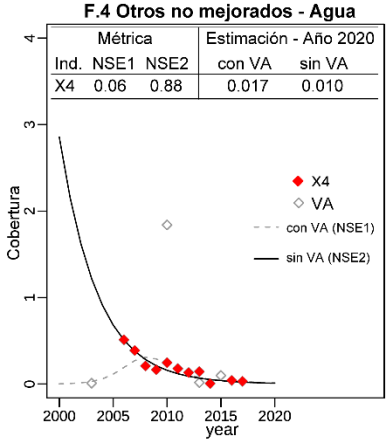


Figura 4.6. Resultados del ajuste de los modelos

Notas: (D, E y F) Se ajustan dos modelos diferentes en CoDa: i) Inverso de los datos transformados en GAM con valores atípicos (VA; líneas discontinuas) y ii) inversa de los datos transformados en GAM sin valores atípicos (VA; líneas sólidas).

En las estimaciones de 2020 para Uruguay, no hay diferencias significativas entre las dos alternativas (modelos con valores atípicos incluidos y sin valores atípicos). Por ejemplo, para la categoría de acceso al servicio de agua entubada, la diferencia entre los dos modelos fue del 0.004% puntos porcentuales. Las tres categorías restantes tampoco difieren de estas afirmaciones. Parece que en los países que han cubierto casi toda la prestación de servicios de agua, la modelización y la comparación ya no son relevantes.

No obstante, no se puede descartar que la modelización sea necesaria para los datos de tendencia a valores extremos, ya que pequeñas proporciones pasadas a unidades de población pueden tener efectos significativos, como en China e India. Por otro lado, hay que destacar dos cosas: i) las estimaciones no pueden superar los valores extremos de 0 y 100% en ninguna categoría de servicio; y ii) es muy importante utilizar técnicas estadísticas adecuadas, como en STEP II, para tratar los valores de cero, según la variabilidad de los datos de las series temporales, ya que esto permite construir modelos sin excluir datos.

Tabla 4.7 . Métricas de calidad del modelo

País	Puntos de datos (X)	Métrica NSE							
		Con valores atípicos (NSE1)				Sin valores atípicos (NSE2)			
		X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
Sudáfrica-saneamiento rural	30 (×4)	–0.002	0.59	0.84	0.12	0.39	0.91	0.93	0.70
Brasil-agua urbana	27 (×4)	0.27	0.28	0.13	0.08	0.29	0.28	0.51	0.20
Indonesia-agua rural	26 (×4)	–0.02	0.17	0.65	0.61	0.66	0.95	0.92	0.95
Nigeria-agua rural	22 (×4)	0.17	0.29	–0.30	–0.09	0.24	0.61	0.79	0.08
Paraguay-agua urbana	21 (×4)	0.75	0.63	0.19	0.66	0.75	0.75	0.14	0.64
Bangladesh-saneamiento rural	20 (×4)	0.15	0.24	0.83	0.41	0.38	0.41	0.81	0.68
Zambia-saneamiento rural	16 (×4)	–0.04	0.03	0.05	0.16	0.07	0.07	0.01	–0.003
Egipto-agua urbana	15 (×4)	–0.06	–0.04	–9.78	–0.02	–0.07	–0.08	–1.61	–0.02
Uruguay-agua urbana	15 (×4)	0.15	0.54	–0.03	0.06	0.88	0.75	0.44	0.88
Benin-saneamiento rural	10 (×4)	0.22	0.54	0.68	0.45	0.22	0.74	0.70	0.52

Los valores de NSE2 inferiores a NSE1 se muestran en gris; los valores negativos están subrayados.

Las métricas de calidad de la Tabla 4.7 refuerzan la hipótesis de que los valores atípicos influyen en la calidad de los modelos. Las métricas de los cuatro indicadores son iguales o mejores cuando se excluyen los valores atípicos en seis de los diez países (a saber, Sudáfrica, Brasil, Indonesia, Nigeria, Uruguay y Benin). De estos países, Sudáfrica, Indonesia y Uruguay tienen métricas NSE2 cercanas a 1, lo que indica la alta capacidad de predicción de los modelos en estos países, según el indicador. Lo contrario se observa en Egipto, donde la media observada es un mejor predictor que el modelo en las cuatro categorías de análisis, tanto en los modelos con o sin valores atípicos. En Bangladesh, Paraguay y Zambia, la mejora sólo se produjo en algunas categorías.

Por otra parte, las tendencias temporales de las categorías de servicios muestran las desigualdades que existen en el acceso al agua y al saneamiento entre los sectores urbano y rural. En Indonesia y Sudáfrica, el acceso al agua y al saneamiento por otras formas mejoradas va en aumento (Figura 4.6 D.2 y E.2); sin embargo, en Uruguay, esta categoría tiende a valores

cero (Figura 4.6F.2). Si comparamos sólo Indonesia y Uruguay, la brecha rural-urbana en la categoría de acceso a agua entubada se incrementa aún más, reflejando la situación mundial reportada en la literatura con respecto a las disparidades que existen en el acceso a agua y saneamiento en ambos sectores (Bain et al., 2014; Chitonge et al., 2020). Dicho esto, y en el contexto de los ODS que buscan garantizar que nadie se quede atrás (Asamblea General de las Naciones Unidas, 2015), el sector rural, tanto en Indonesia como en Sudáfrica, se enfrenta a un mayor desafío en la prestación y gestión segura de los servicios de agua y saneamiento.

Por último, una vez identificados los datos atípicos, no se recomienda eliminarlos automáticamente, ya que esto puede hacer que se pierda información relevante que ayude a explicar la situación específica o la serie temporal del país. Además, existen otros factores que el analista no valora a la hora de excluir los datos (como el coste de obtener los datos a través de una encuesta, un censo u otras alternativas que sean representativas del país); por lo tanto, lo esencial antes de excluir los valores atípicos sería entender por qué son anómalos. Una alternativa que ayudaría a entender la presencia de estos datos podría ser consultar a las instituciones de origen de las fuentes de información. Sin embargo, la obtención de respuestas se complica cuando depende de instituciones de terceros (por ejemplo, para los informes a los ODS, los países asociados suelen tener instituciones estadísticas u otras especializadas que se encargan de recopilar, procesar y compartir la información con las partes interesadas). En estos casos, la exclusión es simplemente una necesidad por las mejoras que aporta a los modelos.

4.5. Mensajes clave

La existencia de valores nulos, de valores ausentes o de ambos simultáneamente hace necesario un tratamiento diferenciado de los datos, para lo que se dispone de distintas opciones de tratamiento. Aunque estas opciones no son equivalentes, no existe un criterio claro para elegir exactamente cuál utilizar, siendo todas las alternativas potencialmente igual de buenas. Además, estas opciones son adecuadas para analizar datos con variaciones en la evolución temporal, lo que no es posible si aplicamos el reemplazo multiplicativo (Martín-Fernández et al., 2003).

En países con poca cantidad de datos, concluimos que la regresión lineal robusta (OLS robusta (ilr)) es adecuada para el análisis de los datos del sector ASH, ya que limita la influencia de los valores atípicos en el modelo calibrado. Tanto cuantitativa como cualitativamente, la declaración de valores atípicos puede ser validada.

En los países con ≥ 6 puntos de datos, la identificación de valores atípicos con la distancia de Mahalanobis robusta tiende a darnos más que la clasificación cualitativa realizada con el PCM (y concretamente, para nueve de los diez países evaluados), lo que refuerza la clasificación habitual del PCM. Sin embargo, hay que tener en cuenta que, en el método de la distancia de Mahalanobis robusta, se excluyen todas las partes del año, mientras que en el PCM sólo es probable que se excluya una parte de la composición. Esta conclusión va de la mano de los ajustes del GAM a los datos, para los que la exclusión de los valores atípicos del análisis conduce generalmente a una mayor fiabilidad de los resultados de las interpolaciones y extrapolaciones.

Además, para todos los casos (por ejemplo, < 6 o ≥ 6 puntos de datos), la interpolación y la extrapolación de los modelos en las categorías de servicios nunca pueden superar el valor límite de 0 o 100%. Esta afirmación coincide y amplía la conclusión obtenida por Pérez-Foguet et al. (2017), ya que ahora hemos analizado una amplia gama de datos con diferentes irregularidades e incluimos el análisis del acceso a la higiene.

CAPÍTULO V. METODO ALTERNATIVO DE MONITOREO DE LA DESIGUALDAD URBANO-RURAL POST 2015

Resumen

El acceso a los servicios de higiene sigue siendo uno de los retos más urgentes a los que se enfrentan los países, especialmente los de bajos ingresos. Esto se ha vuelto mucho más crítico en el contexto actual de la pandemia de COVID-19. El PCM monitorea globalmente los niveles de acceso a los servicios de higiene. Como los datos están en tres partes con una suma constante y un valor positivo, son datos de composición. La desigualdad se supervisa en los datos desglosados; en el caso urbano-rural, esto se hace mediante una simple diferencia entre los niveles de servicio urbanos y rurales. Sin embargo, esta simple forma de cálculo no tiene en cuenta las características de los datos, lo que puede llevar a interpretaciones erróneas de los resultados. Por lo tanto, proponemos una medida alternativa de la desigualdad que utiliza un diagrama ternario y no infringe las propiedades de los datos.

Los resultados de la nueva medida de desigualdad urbano-rural muestra heterogeneidad espacial. La mayor desigualdad se da en Colombia, con un valor de 37.1 puntos porcentuales (p.p.), y la menor en Turkmenistán, con un valor de cero. Nuestros resultados también muestran que en 73 de los 76 países evaluados los servicios básicos de higiene son mayores en las zonas urbanas que en las rurales. Esto significa que los hogares urbanos tienen más disponibilidad de una instalación para lavarse las manos en el lugar con agua y jabón que los hogares rurales. Asimismo, al subdividir el diagrama ternario en parcelas ternarias, hemos agrupado y clasificado los países en función de las condiciones de los servicios de higiene en un orden jerárquico. Por último, nuestro estudio encuentra que una medida multivariante de desigualdad puede ser importante para las políticas públicas del sector con una visión general, lo que subraya el valor de tomar decisiones basadas en evidencia.

Este capítulo está basado en:

- Quispe-Coica, A., Pérez-Foguet, A., 2021. A new measure of hygiene inequality applied to urban-rural comparison. Int. J. Hyg. Environ. Health.
  (En producción con N° IJHEH-113876)

5.1. Introducción

La higiene de las manos con agua y jabón es una práctica de alto impacto y baja tecnología que se correlaciona con la buena calidad de la salud pública y es una forma sencilla y eficaz de reducir las enfermedades.

Los beneficios reportados del lavado de manos son amplios; por ejemplo, puede: reducir la transmisión de virus que causan enfermedades comunes; reducir el riesgo y la incidencia de enfermedades diarreicas (Shahid et al., 1996; Curtis y Cairncross, 2003); reducir el riesgo de infecciones respiratorias (Rabie y Curtis, 2006); y proporcionar beneficios económicos (Townsend et al., 2017). Adquiere mayor relevancia debido al nuevo coronavirus del síndrome respiratorio agudo severo-2 (SARS-CoV-2), que está causando la pandemia generada por la enfermedad del COVID-19 (Pal et al., 2020; Synowiec et al., 2021).

Actualmente, una de las principales recomendaciones para reducir el riesgo de infección por SARS-CoV-2 (además de la vacunación) es la práctica continua del lavado de manos con agua y jabón (Organización Mundial de la Salud [OMS], 2020a). Las incertidumbres que siguen existiendo sobre las vías de transmisión del SRAS-CoV-2 (transmisión aérea en aerosol, contacto superficial, transmisión fecal-oral (Heller et al., 2020; Pandey et al., 2021), etc.), y el hecho de que las personas asintomáticas pueden tener una elevada diseminación viral (por ejemplo, ser infecciosas), subrayan la importancia de utilizar el lavado de manos como medida esencial (Heller et al., 2020; Jones, 2020; Vuorinen et al., 2020; WHO, 2020b). Por ejemplo, un estudio en condiciones de laboratorio muestra que el SARS-CoV-2 puede permanecer viable e infeccioso en los aerosoles durante horas, y en las superficies hasta días, dependiendo del inóculo derramado (van Doremalen et al., 2020).

El seguimiento mundial de los servicios de higiene se ha incorporado al ODS 6.2 y se lleva a cabo desde 2015. La meta del Objetivo 6.2 es "para 2030, lograr el acceso a servicios de saneamiento e higiene adecuados y equitativos para todos...". La higiene adecuada está relacionada con que los hogares tengan instalaciones para lavarse las manos con agua y jabón. El Programa de Monitoreo Conjunto OMS/UNICEF (PCM) realiza un monitoreo global de los hogares con o sin instalaciones para el lavado de manos y los clasifica en las tres categorías de la llamada escalera del lavado de manos: básica, limitada y sin instalaciones. El número de casos de cada categoría se cuenta en un conjunto de hogares determinado (por ejemplo, un país), tras lo cual se calculan tres variables numéricas; una vez divididas por el número total de hogares, estas forman una proporción de suma constante de uno (o 100% si son porcentajes); se trata, por tanto, de datos de composición (John Aitchison, 1986; Egozcue y Pawlowsky-Glahn, 2011; van den Boogaart y Tolosana-Delgado, 2013b; Filzmoser et al., 2018).

El análisis por separado de cada porcentaje corresponde al análisis univariante y es muy habitual en el sector. El análisis conjunto de las tres variables es un análisis multivariante. Como las tres variables higiénicas tienen una restricción de suma constante y están implícitamente relacionadas con un total predefinido o con partes complementarias, ninguna variable puede interpretarse independientemente de las demás, sino que deben interpretarse como composiciones (van den Boogaart y Tolosana-Delgado, 2013a). En el monitoreo global, la aplicación de técnicas estadísticas para datos composicionales en agua, saneamiento e higiene (ASH) comenzó con Pérez-Foguet et al. (2017) y fue ampliada por Ezbakhe y Pérez-Foguet (2019) quienes incorporan la incertidumbre de los datos en el análisis; sin embargo, su aplicación

práctica al conjunto de datos globales no fue posible hasta que Quispe-Coica y Pérez-Foguet (2020) introdujeron el preprocesamiento de datos con valores cero, datos faltantes y valores atípicos (desarrollado en el Capítulo IV de esta tesis). Estos análisis también se han ampliado recientemente al ámbito sanitario relacionado con la mortalidad infantil (Ezbakhe y Pérez-Foguet, 2020).

La desigualdad de acceso a cualquier servicio es uno de los principales obstáculos para la cobertura universal. Por lo tanto, al igual que en el caso del agua y el saneamiento, el PCM también supervisa la desigualdad entre las zonas urbanas y rurales en materia de higiene. Esta información puede ayudar a los actores nacionales e internacionales del sector a orientar las intervenciones. En este sentido, los expertos han recomendado desagregar la información y medir las desigualdades desde diferentes aspectos (WHO/UNICEF, 2015; Economic y Council, 2016). Por lo tanto, el PCM actualmente hace un seguimiento de la desigualdad en materia de agua, saneamiento e higiene a partir de datos desglosados por quintiles de riqueza, residencia urbana y rural, regiones subnacionales y niveles de servicio de escalera (véase https://washdata.org/monitoring/inequalities). Los quintiles de riqueza se basan en un análisis de los activos de los hogares, y el resultado final expresado en proporciones también está representado por el lugar de residencia. Por lo tanto, en cualquiera de las alternativas de seguimiento descritas, se puede aplicar una medida de desigualdad urbano-rural.

La alternativa actual utilizada por el PCM para la presentación de informes globales es utilizar la simple diferencia entre la relación de los niveles de servicio urbanos y rurales. En el caso de la higiene, esto se lleva a cabo en las tres categorías de la escalera de servicio del lavado de manos (WHO/UNICEF, 2016), lo que proporciona tres medidas de desigualdad. Esto implica que, en una clasificación de la desigualdad, es probable que los países tengan diferentes posiciones según la categoría de análisis. Como los datos de composición tienen una suma constante, una parte o el resto también se verán afectados si una de las partes varía. Por consiguiente, al interpretar una categoría, también hay que tener en cuenta el resto de las categorías. De hecho, los datos composicionales son multivariados por naturaleza (van den Boogaart y Tolosana-Delgado, 2013a), lo que refuerza la idea de interpretar una categoría teniendo en cuenta el resto.

Conocer el espacio en el que operan estos datos es también una parte importante, ya que nos permite calcular la desigualdad seleccionando entre las diferentes alternativas multivariantes que existen, utilizando la que mejor se adapte a los datos. El espacio muestral en el que operan los datos composicionales es el simplex S^D, y la estructura del espacio vectorial se denomina geometría de Aitchison o simplex de Aitchison (que es un marco geométrico diferente del espacio vectorial euclidiano). Por lo tanto, antes de aplicar cualquier técnica estadística clásica, es necesario realizar primero transformaciones logarítmicas (Aitchison, 1982; John Aitchison, 1986; Egozcue et al., 2003). Sin embargo, cuando los datos están en tres o cuatro partes, es posible representarlos gráficamente en el mismo espacio simplex. Si los datos están en tres partes, la representación gráfica en simplex se hace mediante un diagrama ternario; si está en cuatro partes, mediante un tetraedro regular. Si los datos son mayores de cuatro partes, la representación gráfica en simplex no es posible; sin embargo, todo lo relacionado con las operaciones, definiciones e interpretaciones de los datos composicionales es válido para cualquier número de partes (Von Eynatten et al., 2002; Pawlowsky-Glahn y Egozcue, 2006).

Dicho esto, cuando la información es tripartita con una suma constante y es positiva, como

en el caso de los datos de higiene, hay pocas alternativas multivariantes que nos permitan calcular la desigualdad urbano-rural en el espacio muestral del simplex. Para nuestro propósito, la información se limita al diagrama ternario. La aplicación del diagrama ternario para representar gráficamente los datos es muy común en otras áreas de la ciencia, incluyendo las ciencias de la tierra, la geoquímica y la química (Miller, 2002; Graham et al., 2020; Verma, 2020). Últimamente, también se ha aplicado en epidemiología (Dumuid et al., 2020) y en la gestión de residuos, tanto para la visualización dinámica como estática (Bartl, 2014; Pomberger et al., 2017). Otros estudios presentan propuestas de representación gráfica ternaria en datos centrados, que permiten mejorar la visualización gráfica y la interpretación de la estructura de los datos (Von Eynatten et al., 2002).

Por otro lado, en la literatura del sector ASH es muy común el uso de mapas temáticos univariados de cualquiera de las categorías de servicios a los que accede la población, agrupando (por colores) aquellos que se encuentran dentro de un determinado rango (lo llamaremos "amplitud" en este estudio). Los informes globales realizados por la OMS/UNICEF son un claro ejemplo de que la lectura e interpretación de los resultados también son univariantes (WHO/UNICEF, 2019a, 2020). Sin embargo, si los puntos de datos de los países se representan en el diagrama ternario, cada punto representa una composición tripartita, y tendrán una lectura u otra, dependiendo de su ubicación en el diagrama ternario.

Uno de los principales objetivos de este estudio es proponer una medida multivariante de la desigualdad urbano-rural, teniendo en cuenta las características de composición de los datos. Para ello, primero discretizamos el diagrama ternario para representar en él los puntos de datos urbanos y rurales, y luego calculamos la distancia entre ambos puntos como medida global de desigualdad. Otro objetivo es representar la información tripartita en un mapa temático. Por último, lo aplicamos a un conjunto de datos globales sobre higiene en entornos urbanos y rurales.

5.2. Materiales y métodos

5.2.1. Análisis de datos: entrada

Para este estudio, la información se obtuvo de la plataforma PCM (www.washdata.org). Se filtraron los países con datos de 2017. Como resultado, el análisis se limita a los datos de 77 países en el área de residencia rural, y los datos de 76 países en el área de residencia urbana. La diferencia de una unidad entre lo urbano y lo rural se debe a Perú, que sólo presenta información para las instalaciones de higiene rurales y no urbanas. El desglose del número de países por región (de un total de 76) es el siguiente: 10 países de la región de Asia Central y Meridional (CSA), 9 países de la región de Asia Oriental y Sudoriental (ESEA), 12 países de la región de América Latina y el Caribe (ALC), 8 países de la región de África Septentrional y Asia Occidental (NAWA), 3 países de la región de Oceanía y 34 países de la región de África Subsahariana (SSA). Obsérvese que los países de varias regiones (Australia, Nueva Zelanda, Europa y América del Norte) no se incluyeron en el análisis, ya que no se encontró información sobre higiene para 2017 en el sitio web del PCM.

La fuente de información proporciona datos desglosados sobre la presencia o ausencia de una instalación de lavado de manos en los tres niveles de servicio: servicio básico (BS), servicio limitado (LS) y sin instalación (NF). El nivel de servicio básico se refiere a la disponibilidad de una instalación de lavado de manos en el local con agua y jabón; el nivel de servicio limitado se refiere a la disponibilidad de una instalación de lavado de manos en el local que carece de agua

y/o jabón; y NF se refiere a la ausencia de una instalación de lavado de manos en el local. Esta información se representa en un vector de composición de tres partes $S^{D=3} = \left\{ X = (X_1, X_2, X_3) : \forall X_i > 0,\ \sum_{i=1}^{3} X_i = 100 \right\}$, que posteriormente se representa en el diagrama ternario.

5.2.2. Conceptos básicos del diagrama ternario

El diagrama ternario es un diagrama que representa gráficamente las proporciones de las tres composiciones en un triángulo equilátero. La base matemática del triángulo equilátero es el conocido teorema de Viviani (Abboud, 2010), y tiene el potencial de expresar los datos de composición de tres partes como uno solo. Esto resulta ventajoso a la hora de realizar un análisis multivariante de la información tripartita.

La ubicación de los puntos de datos en el diagrama ternario puede determinarse de varias maneras. Aquí ilustramos dos formas. La primera consiste en trazar líneas paralelas a la base del diagrama ternario frente al vértice (Figura 5.1A). En el ejemplo representado, se da un valor aleatorio del 30% para la BS y del 20% para la NF. El primer paso es trazar el valor de BS con una línea recta cuyo valor es el 30% (línea azul sólida en la subfigura A); el valor de NF se traza entonces con una línea negra sólida cuyo valor es el 20%. La intersección de estas dos líneas será la ubicación del punto de datos "A" (BS = 30%, NF = 20%) en el diagrama ternario. Cabe señalar que no es necesario trazar la tercera línea (línea roja discontinua) para localizar el punto de datos de "A"; como hay un valor de cierre del 100%, el resultado de LS será simplemente una diferencia (es decir, LS = 100 - NF - BS). Por consiguiente, es posible trazar observaciones bidimensionales dentro de un diagrama ternario.

La segunda opción está relacionada con la conversión entre el diagrama ternario y las coordenadas XY (véase Pomberger et al. (2017)). Brevemente, la composición tripartita debe convertirse a coordenadas XY utilizando la Ec. (5.1) y la Ec. (5.2); esta transformación permite calcular la distancia entre dos puntos de forma clásica (Subfigura B, C).

$$x' = NF + BS/2 \qquad (5.1)$$

$$y' = BS \times \sqrt{3}/2 \qquad (5.2)$$

Se sigue el procedimiento de Pomberger et al. (2017) para construir el diagrama ternario a partir de las alternativas citadas. La principal justificación es facilitar la captura de los puntos de datos en el diagrama ternario mediante cálculos computacionales; además, esto permite calcular fácilmente la distancia entre dos puntos ($d_{\overline{UR}}$).

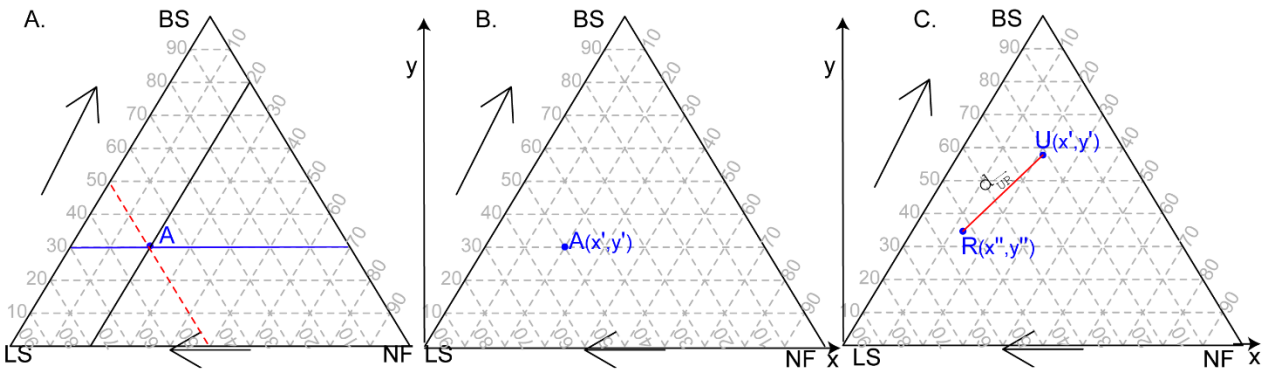


Figura 5.1. Diagrama ternario

Notas: Ubicación de los puntos de datos en el diagrama ternario (gráficos A y B) y medición de la distancia (gráfico B). Gráfico A: valor de BS en la línea azul continua, valor de NF en la línea negra continua y valor de LS en la línea roja discontinua.

5.2.3. Trazado y lectura de una parcela ternaria

La construcción de un mapa temático coloreado, al igual que los univariantes, requiere que los datos subyacentes estén ordenados. Esto se consigue discretizando primero el diagrama ternario en la parcela ternaria.

Las parcelas ternarias pueden tener diferentes amplitudes. Por ejemplo, si el diagrama ternario no está delimitado, la amplitud de la parcela ternaria será 100 (véase las Figura 5.1A) y, por tanto, la lectura ternaria sería BS ≤ 100, LS ≤ 100 y NF ≤ 100. Si limitamos la amplitud de la parcela ternaria a 50, el número de parcelas ternarias será cuatro, y la lectura ternaria de la parcela situada en el centro ternario sería BS ≥ 50, LS ≤ 50, y NF ≤ 50. Si limitamos la amplitud de la parcela ternaria a 20, el número de parcelas ternarias será de veinte y cinco (como se muestra en la Figura 5.2), y la lectura de una de las parcelas ternarias será BS > 80, LS < 20 y NF < 20 (véase la lectura de Q1, Figura 5.2A). Si la amplitud de la parcela ternaria es 10, el número de parcelas ternarias que se formarán será 100. La amplitud de la parcela ternaria está relacionada con la precisión de la lectura. En este estudio, seleccionamos una amplitud inicial de 20 para analizar los datos, que representaba un equilibrio entre la lectura y el ordenamiento de las parcelas. En los siguientes apartados se analiza la influencia de la amplitud y la relación con la precisión de la lectura.

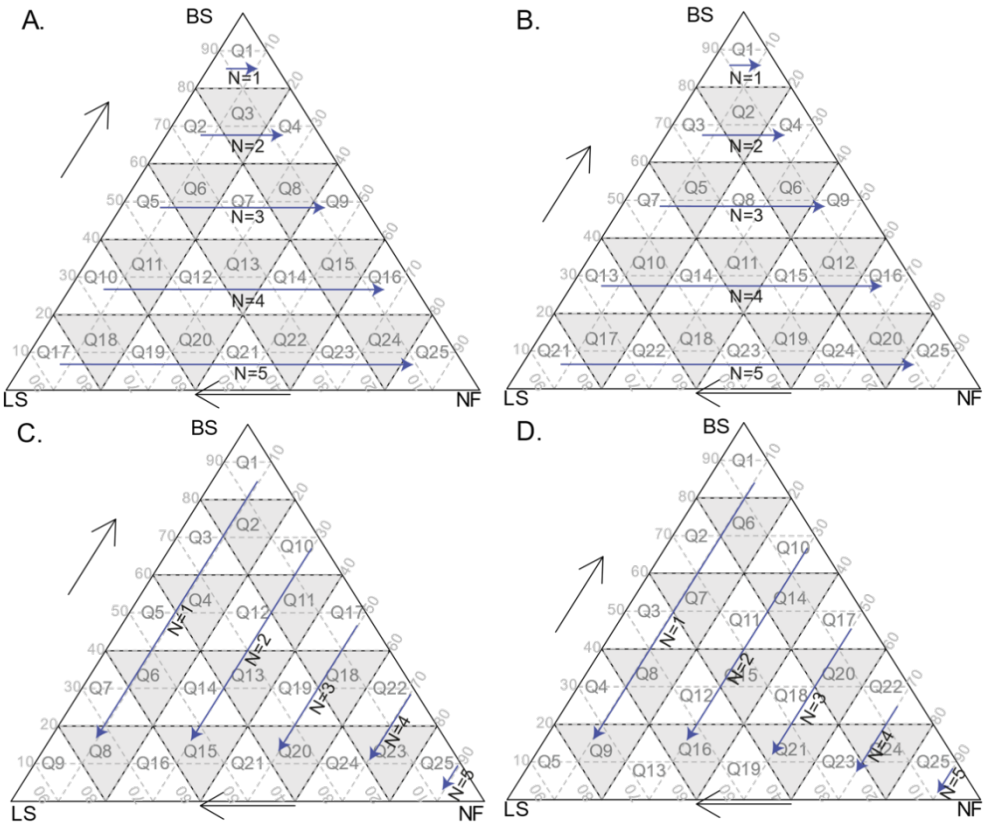


Figura 5.2. Parcelas ternarias

Notas: (A-C) Tres rutas alternativas para clasificar los países por grupos: horizontalmente de izquierda a derecha (A), verticalmente de arriba a abajo, con prioridad en las áreas sombreadas y luego en las áreas blancas, de izquierda a derecha (B), y diagonalmente de derecha a izquierda, y diagonalmente de arriba a abajo con prioridad en las áreas blancas y luego en las áreas sombreadas (C). N: número de nivel.

Por otro lado, para ordenar los grupos de países (parcelas), definimos los países que tienen valores cercanos al vértice BS como aquellos que tienen una alta cobertura de instalaciones para el lavado de manos con agua y jabón en el local como los que están en la mejor posición. Los países que tienen valores cercanos al vértice NF son los que tienen un alto porcentaje de hogares que no tienen instalaciones para lavarse las manos y se muestran cómo los últimos. Nótese que el orden puede tener cuatro alternativas de clasificación diferentes (mostradas en la Figura 5.2):

i) la primera alternativa de clasificación es la que se muestra en la subfigura A. Sigue la lógica de clasificar a los países con un alto grado de cobertura de instalaciones de servicios básicos de higiene, y luego de izquierda a derecha para priorizar el acceso a servicios de higiene limitados, y finalmente cerrar con el peor grupo de países en la Q25;

ii) en la segunda alternativa (subfigura B), los grupos se ordenan verticalmente de arriba a abajo, dando un mayor valor a los países que tienen una alta cobertura de servicios de higiene básica, y luego de izquierda a derecha para priorizar el acceso a los servicios de higiene limitada, y finalmente cerrando con el grupo de países más desfavorable en la Q25;

iii) en la tercera alternativa (subfigura C), los grupos se ordenan en diagonal de arriba a abajo y por niveles. El primer nivel es cuando NF ≤ 20, y el último, cuando 80 NF <. Este orden sigue el criterio de clasificar primero los países con un nivel bajo de cobertura NF (es decir, Q1 a Q9) y terminar la clasificación con el grupo de países con valores altos de NF (es decir, Q25).

Internamente, para N = 1, va en orden descendente, dando prioridad al nivel básico y terminando en el vértice LS;

iv) la cuarta alternativa (subfigura D) es una variante de la tercera alternativa; el único cambio es el orden interno de cada nivel (primero las parcelas ternarias blancas y luego las de color). El criterio utilizado para definir el orden en cada nivel se utiliza para dar un valor más alto al grupo de países con menor NF, y en diagonal de mayor BS a menor BS, tanto en la parcela ternaria blanca como en la parcela ternaria gris.

Para este análisis, elegimos la cuarta alternativa (Figura 5.2D), que garantiza que los cinco primeros órdenes de clasificación (es decir, Q1-Q5) incluyan el valor de cero para NF en comparación con las otras alternativas. En la primera y segunda alternativas, el grupo de países capturados por Q4 son los que no tienen un valor de NF de cero; para la tercera alternativa, los países que tienen un valor de cero tienen más probabilidades de estar en Q3 o Q5 que en Q2 o Q4. En resumen, al seleccionar la cuarta (Figura 5.2D), colocamos el grupo de países con el valor NF más bajo en los ocho primeros órdenes de clasificación (es decir, N = 1), y el grupo de países con el valor NF más alto (es decir, N = 5) en el último orden. Este criterio está relacionado con la visión de la Agenda 2030 de "no dejar a nadie atrás".

5.2.4. La desigualdad urbano-rural en un diagrama ternario

La desigualdad urbano-rural se mide a través de las distancias $d_{\overline{UR}}$ en el diagrama ternario mediante la Ecuación (5.3). La mayor desigualdad la presentarán los países con mayor distancia, y la menor desigualdad, los países con menor distancia. Un valor de cero indica que no hay desigualdad urbano-rural, mientras que un valor de 100 puntos porcentuales (p.p.) indica una desigualdad máxima entre urbano-rural.

$$d_{\overline{UR}} = \sqrt{(x' - x'')^2 + (y' - y'')^2} \qquad (5.3)$$

Donde x ', y' son los valores de la zona de residencia urbana, y x '', y '' son los valores de la zona de residencia rural.

La propuesta se compara con la medida absoluta de la desigualdad, es decir, la diferencia urbano-rural en términos de proporción; esta estrategia es utilizada actualmente por la OMS/UNICEF como parte de la comparación de la proporción de la población con acceso a los servicios ASH entre las zonas urbanas y rurales y se informa en los informes mundiales (WHO/UNICEF, 2019a).

Finalmente, se construyó una representación gráfica de mapas temáticos, diagramas ternarios y diagramas de caja en la plataforma R Core Team (2020) (v.4.0.3), para lo cual se utilizaron los siguientes paquetes de R: ggplot2 (v3.3.5; Wickham, 2016), tidyverse (v1.3.1; Wickham et al. 2019), y pgirmess (v1.7.0; Giraudoux, 2021) para construir el diagrama ternario; y tmap (v3.3-2; Tennekes, 2018) y sf (v1.0-1; Pebesma, 2018) para los mapas temáticos. La base de datos y los scripts de R se presentan en Quispe-Coica y Pérez-Foguet (2021).

5.3. Resultados

5.3.1. Clasificación ternaria de higiene en urbano y rural

La información de los países sobre el acceso a las instalaciones de higiene se presenta en las Figura 5.3 y Figura 5.4 para las residencias urbanas y rurales, respectivamente. Se utilizan

25 parcelas ternarias, sin información para seis parcelas ternarias en la zona urbana y cuatro en la zona rural. Los diagramas ternarios y su correspondiente clasificación en parcelas ternarias se presentan en la subfigura A y como mapa temático en la subfigura B. El mapa temático univariante de las tres categorías de higiene se muestra en las subfiguras C, D y E. La lista con la clasificación de todos los países se presenta en la Tabla A1, y el resumen del número de países encontrados en cada parcela ternaria se muestra en la Tabla 5.1.

Nuestro método de clasificación agrupó a 25 países por residencia urbana, y a 11 países por residencia rural, en el Q1. El desglose regional del Q1 muestra que 7 países de ALC, 6 de ESEA y CSA, 5 de NAWA y 1 de Oceanía correspondían a la residencia urbana, y 4 de CSA, 3 de NAWA, 2 de ALC y 2 de ESEA a la residencia rural. Ningún país de la región SSA aparecía en este grupo, pero sí a partir de Q2.

El grupo de países de la parcela ternaria Q1 (es decir, 80 < BS, LS < 20 y NF < 20) se caracteriza por tener las mejores condiciones para la práctica del lavado de manos in situ, ya que tienen una alta cobertura de servicios básicos (BS > 80), una baja cobertura de hogares con servicios limitados (LS < 20) y una baja cobertura de hogares sin instalaciones para el lavado de manos (NF < 20). Su representación en el mapa temático tiene el color más intenso, y la intensidad disminuye en el orden de Q1 a Q25. La diferencia entre lo urbano y lo rural en el número de países captados por Q1 muestra que hay una mayor probabilidad de tener mejores condiciones de lavado de manos en la residencia urbana que en la rural, tanto en el total como en el desglose regional.

El grupo de países de Q2 a Q9 se caracteriza por tener condiciones relativamente mejores para la práctica del lavado de manos, ya que tienen una baja cobertura de hogares sin instalaciones para el lavado de manos in situ (es decir, NF ≤ 20%), BS ≤ 80%, y LS de 0-100%. Obsérvese que la zona rural de Burundi es el único país que se encuentra en el Q5 y está muy cerca del vértice LS, lo que implica que tiene una alta cobertura de LS (> 80%), una baja cobertura de BS (< 20%) y una baja cobertura de hogares con NF (< 20%). También cabe destacar la aparición por primera vez de dos países de la región del SSA en el Q2 urbano (Tanzania y Namibia).

Las parcelas ternarias capturaron 9 países urbanos y 13 rurales en Q10 a Q16. Este grupo de países se caracteriza por tener valores de NF que están dentro del rango del 20% al 40%, mientras que los valores de BS y LS están entre cero y 80%. En comparación con Q1 a Q9, hay una mayor concentración de países rurales que urbanos en Q10-Q16, con 13 frente a 9 países, respectivamente; este patrón se observa también en los siguientes niveles de clasificación (de N3 a N5).

Siguiendo la secuencia, el grupo de países que fueron clasificados en Q17 a Q21 tienen la característica común de tener valores de NF del 40% al 60% y valores de BS y LS del 0 al 60%. En el desglose regional, hay un predominio de países de la región SSA sobre el resto. Para los casos urbanos, hay 7 países en SSA frente a 1 en NAWA. Para los casos rurales, hay 8 países en SSA frente a 3 en ALC y 1 en ESEA.

En el siguiente orden de clasificación, de Q22 a Q24, la situación es similar a la anterior: hay un predominio de países de la región del SSA sobre el resto. Esto es más drástico en la parcela ternaria Q24, que contenía únicamente países del SSA, tanto en la categoría urbana como en la rural.

Tabla 5.1. Cuantificación del número de países según su clasificación en la parcela ternaria.

N	Qn	Urbano							Rural						
		Global	Regional						Global	Regional					
			CSA	ESEA	ALC	NAWA	Oceanía	SSA		CSA	ESEA	ALC	NAWA	Oceanía	SSA
1	Q1	25	6	6	7	5	1		11	4	2	2	3		
	Q2	6	2	1		1		2	4	1		2	1		
	Q3	4	1	1				2	6	3	1	1			1
	Q4	1						1	3	1	1				1
	Q5								1						1
	Q6	4	1		1	1		1	8		2	4	1	1	
	Q7	5					2	3							
	Q8	2			1			1	3						3
	Q9	1						1	1						1
2	Q10	2		1	1										
	Q11	3						3							
	Q12	1						1	2						2
	Q13								1			1			
	Q14	1			1				4		2		1		1
	Q15	2						2	5	1			1	1	2
	Q16								1					1	
3	Q17								2		1	1			
	Q18	4						4	1						1
	Q19								2						2
	Q20	3				1		2	1			1			
	Q21	1						1	6			1			5
4	Q22	2			1			1							
	Q23								4				1		3
	Q24	4						4	3						3
5	Q25	5						5	8						8
	Total	76	10	9	12	8	3	34	77	10	9	13	8	3	34

Notas: N, número de nivel de la Figura 5.2D; Qn, parcela ternaria de orden n.

El último orden de clasificación corresponde a la Q25. Como en el caso de la Q24, todos los países incluidos en esta parcela ternaria pertenecen a la región del SSA. El grupo de países que se encuentra en esta parcela ternaria se caracteriza por una baja cobertura de hogares con BS (< 20), una baja cobertura de LS (< 20) y una alta cobertura de hogares con NF (> 80). Su representación en el mapa temático tiene el color más bajo en intensidad, lo que significa que los países de este grupo (del total de países analizados) tienen las condiciones más desfavorables para el lavado de manos.

Por otro lado, observamos que la interpretación del mapa temático univariante puede llevar a una interpretación errónea de la información. Utilizando a Burundi como ejemplo, si sólo interpretamos el valor de BS (de 4.1%) expresado en el mapa temático de la Figura 5.4C, observamos que está representado con el color más bajo en intensidad, dando la impresión de que Burundi tiene condiciones muy desfavorables para el lavado de manos. Sin embargo, en nuestra alternativa de clasificación, Burundi se encuentra en Q5 y está representado en el mapa temático de la Figura 5.4B con el color más intenso. Esto significa que tiene condiciones relativamente mejores para el lavado de manos, ya que tiene un valor alto de LS (94.5%) y un valor bajo de NF (1.4%). En consecuencia, un análisis univariante no muestra lo que ocurre en las otras partes, lo que pone de manifiesto el valor de utilizar una forma alternativa de explorar los datos cerrados con un mapa temático ternario.

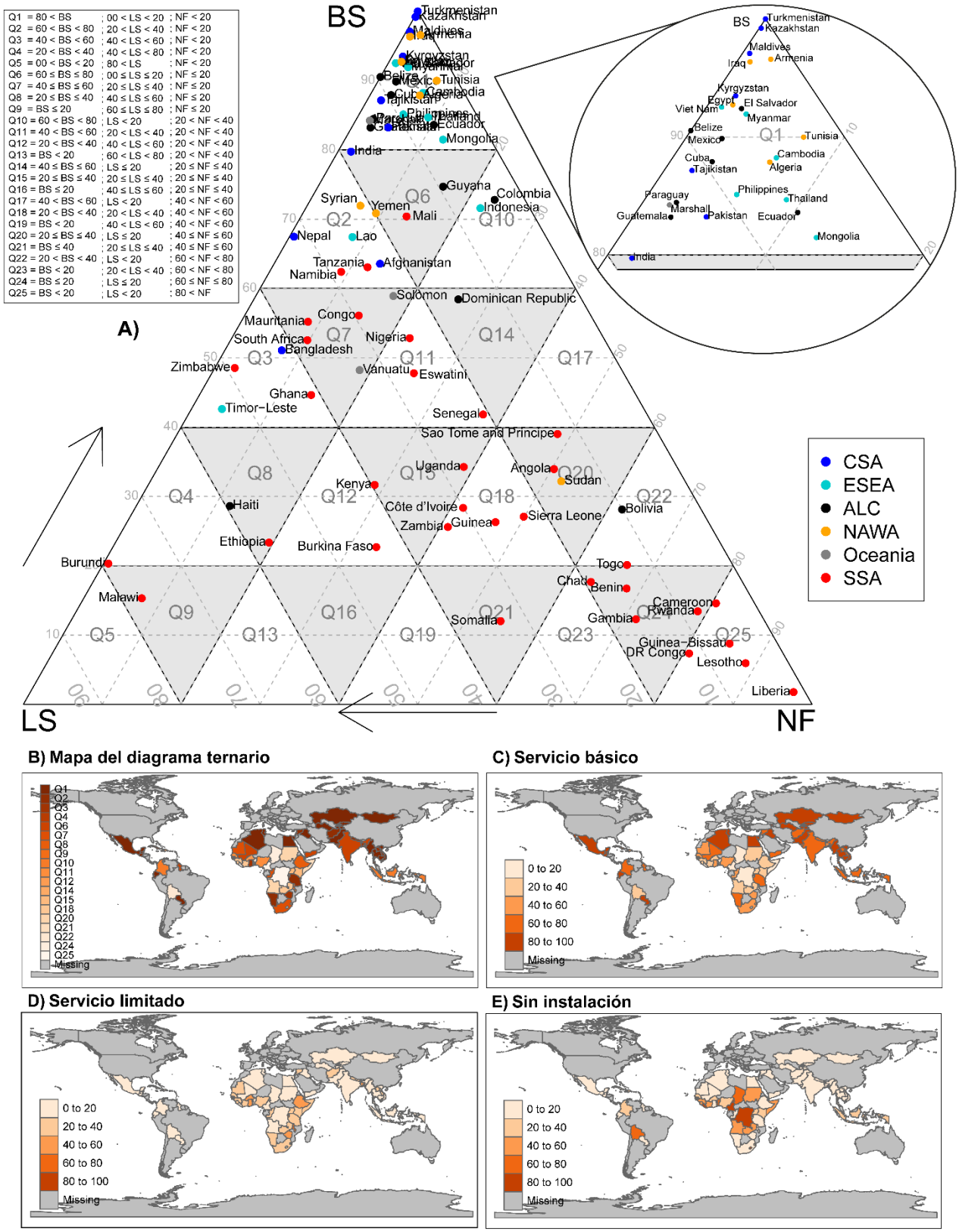


Figura 5.3. Mapas temáticos - Urbano

Notas: Acceso a los servicios de higiene urbana en 2017. A: diagrama ternario. B: mapa del diagrama ternario. C-E: diagrama univariante.

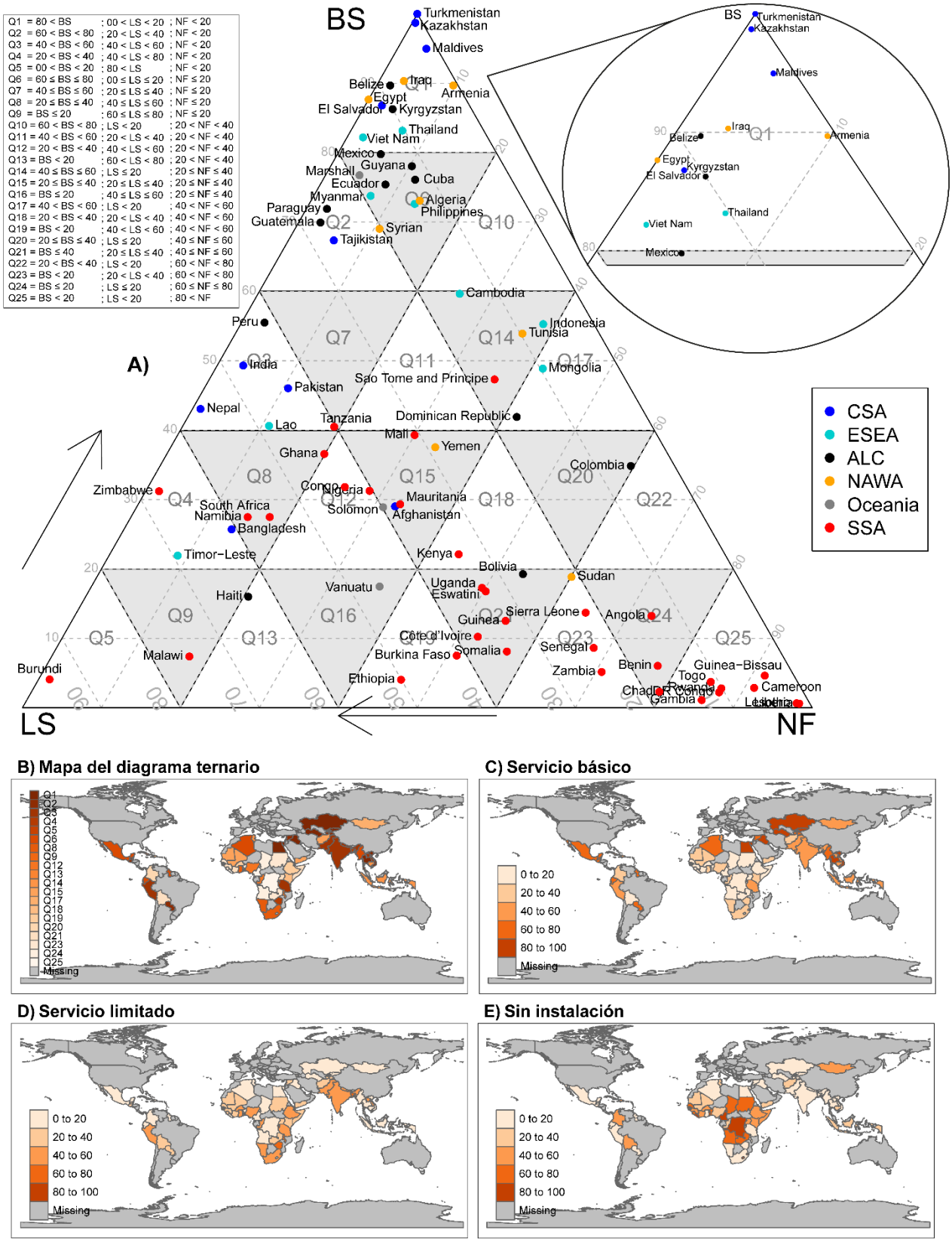


Figura 5.4. Mapas temáticos - Rural

Notas: Acceso a los servicios de higiene rural en 2017. A: diagrama ternario. B: mapa del diagrama ternario. C-E: diagrama univariante.

Por último, al cruzar la clasificación de los países con su nivel de renta (Figura A3), observamos que en el orden de clasificación de Q1 a Q9, hay una mayor concentración de países de renta media-alta que de países de renta baja y media-baja, tanto en la categoría urbana como en la rural. Esto disminuye en los niveles superiores; por ejemplo, en el nivel 2 (es decir, N2; de Q10 a Q16), sólo hay tres países de renta media-alta en las zonas de residencia urbana, y sólo uno en la rural; en el siguiente nivel (N3; de Q17 a Q21), sólo hay dos países de renta media-alta en las zonas rurales y ninguno en la urbana. En los niveles más altos, las parcelas ternarias no han captado ningún país de renta media-alta. En general, los resultados muestran que los países de renta media-alta se concentran en la parte superior del diagrama ternario, y los países de renta media-baja y baja, en la parte inferior.

5.3.2. Desigualdad urbana-rural en el acceso a las instalaciones de higiene

Los resultados de los vectores de desigualdad urbano-rural se muestran en el diagrama ternario de la Figura 5.5A y, espacialmente, en la Figura 5.5B. Los vectores utilizan la zona de residencia urbana como punto de partida y la zona de residencia rural como punto final. La distancia entre estos dos puntos, el módulo del vector, representa la desigualdad urbano-rural.

En el diagrama ternario, las direcciones de los vectores no tienen un orden definido, lo que indica un comportamiento diferenciado de la desigualdad entre los países; además, expresados como mapa temático, los resultados muestran heterogeneidad espacial (Figura 5.5B). Sin embargo, se repiten patrones comunes en 73/76 países, como la dirección del vector de arriba hacia abajo (en diagonal hacia la derecha o hacia la izquierda). Esto indica que hay un mayor nivel de BS en los entornos urbanos que en los rurales. Por el contrario, en Sao Tomé y Príncipe y Guyana, la dirección es de abajo a arriba (en diagonal hacia la izquierda), lo que significa que el nivel de BS es mayor en el entorno rural que en el urbano (véase también la Figura 5.6B).

En el caso de la LS, la diferencia urbano-rural tiene un valor negativo en 57/76 países, lo que indica que los hogares de la zona de residencia rural tienen una mayor cobertura de instalaciones de higiene que carecen de jabón o agua que los hogares de la zona de residencia urbana. Por último, en el caso de la NF, la diferencia urbano-rural tiene un valor negativo en 68/76 países, lo que significa que los hogares de la zona de residencia rural tienen una mayor cobertura de NF que los de la urbana.

El resultado de la desigualdad urbano-rural expresado en términos de distancia se muestra en la Figura 5.6A y los estadísticos resumidos por cuartiles en la Tabla 5.2 (boxplot en la Figura A1). Complementamos el valor de desigualdad ilustrado en la Figura 5.6A con los resultados obtenidos de la clasificación urbana (Qu) y rural (Qr) en el diagrama ternario. El valor delta en mayúsculas (Δ) indica un cambio de orden entre urbano y rural en la parcela ternaria (es decir, Δ = Qu - Qr). Un valor negativo de Δ significa que, en la clasificación, lo urbano está en un orden superior a lo rural; un valor positivo de Δ indica el proceso inverso. Un valor de Δ igual a cero significa que no hay ningún cambio y, por tanto, lo urbano y lo rural están en la misma parcela ternaria.

A nivel global, se identificaron 19/76 países pertenecientes al primer cuartil con distancias de desigualdad ≤ 8.7 p.p. (con un valor mínimo de cero en Turkmenistán), y con distancias de desigualdad > 8.7 p.p. en el 75% de los países restantes. En el tercer cuartil, las distancias de desigualdad son ≤ 23.1 p.p. para el 75% de los países, y > 23.1 p.p. para el 25% restante de los países, con un valor máximo de 37.1 p.p. en Colombia. El valor cero encontrado en Turkmenistán se basa en tener tanto NF como LS iguales a cero, con la categoría de servicio BS al 100%; por lo tanto, al aplicar la Ec. (5.3), se obtiene un vector con distancia nula. El significado del valor nulo es que hay igualdad entre lo urbano y lo rural en el acceso a las

instalaciones del servicio de higiene y, en el caso particular de Turkmenistán, las instalaciones de higiene también tienen jabón y agua.

Desde una perspectiva regional, las medidas de desigualdad muestran una gran heterogeneidad. En la región de ALC, Colombia es el país con mayor desigualdad urbano-rural (con un valor de 37.1 p.p.), y Belize es el país con menor desigualdad (con un valor de 1.5 p.p.) (Figura 5.6A). El valor de desigualdad de Colombia se desvía mucho del comportamiento normal de la región, donde el 75% de los países tienen distancias de desigualdad ≤ 13.5 p.p.; por lo tanto, se considera un comportamiento atípico. La región de ALC también tiene el menor rango intercuartil (IQR; 5.4 p.p.) con respecto al resto de las regiones, lo que se traduce en una menor dispersión de los datos.

En cambio, la región CSA tiene el valor más alto de IQR, de 24.6 p.p., lo que se traduce en una mayor dispersión de los datos. Asimismo, en CSA, el 75% de los países tienen una distancia de desigualdad de ≤ 28.3 p.p., con un valor de cero en Turkmenistán; el 25% restante las distancias tienen una distancia de desigualdad de > 28.3 p.p., con un valor máximo de 34.6 p.p. en Pakistán.

En la región ESEA, 7/9 países tienen distancias de desigualdad ≤ 25.2 p.p., con un valor mínimo de 3.5 p.p. en Tailandia; en los dos países restantes, las distancias de desigualdad son > 25.2 p.p., con un valor máximo de 31.0 p.p. en Mongolia. En la región de Oceanía, las distancias sólo pudieron calcularse para tres países; el valor mínimo es de 6.6% en Marshall, y el valor máximo es de 26.8 p.p. en Vanuatu. En la región del SSA, el 75% de los países tienen una distancia de desigualdad ≤ 21.6 p.p., con un valor mínimo de 1.2 p.p. en Liberia, y el 25% restante de los países tienen distancias > 21.6 p.p., con un valor máximo de 32.5 p.p. en Namibia.

Nótese que un valor bajo en la medida de la desigualdad no implica necesariamente que se den las mejores condiciones para el lavado de manos con agua y jabón. Por ejemplo, en las regiones CSA, ESEA, ALC y NAWA, un delta con un valor cero corresponde a los países con menos desigualdad, pero también está presente cuando lo urbano y lo rural están en la misma parcela ternaria de Q1, Q2 o Q6. Los países de Q1 tienen las mejores condiciones para lavarse las manos con agua y jabón. Sin embargo, en la región del SSA, los cuatro países con menor desigualdad se encuentran en Q21 o Q25, tienen un valor Δ igual a cero, pero presentan altos valores de NF (> 80% en Q25, y entre 40% y 60% en Q21) lo que se traduce en países con las condiciones más desfavorables para el lavado de manos con agua y jabón. En consecuencia, para evitar una interpretación errónea del valor de la desigualdad, es necesario que esta magnitud vaya acompañada de la clasificación ternaria.

Nuestros resultados también muestran que la medida multivariada de desigualdad tiene un comportamiento dual con respecto a la higiene de la BS: tiene una relación directa en la región del SSA (es decir, la magnitud de la desigualdad disminuye y el promedio de la BS urbana y rural también disminuye), y una relación inversa en las restantes regiones (es decir, la magnitud de la desigualdad disminuye, mientras que el promedio de la BS urbana y rural aumenta). Esta afirmación se ve respaldada por el resultado del ajuste lineal entre la medida de la desigualdad obtenida y el promedio de la BS urbana y rural, que tiene una pendiente negativa en las regiones CSA, ESEA, ALC, NAWA y Oceanía, mientras que la región SSA tiene una pendiente positiva (ver Figura A4).

Los gráficos univariados de la Figura 5.6B-D permiten una visualización muy intuitiva de la desigualdad urbano-rural en el acceso a los servicios de higiene, facilitando la lectura de cada categoría. Sin embargo, no es posible conocer el nivel de desigualdad en el país en su conjunto, ya que las tres categorías contienen información sobre la desigualdad y pueden puntuar mejor o peor según la categoría de análisis. Por ejemplo, si se ordena el valor absoluto

de la desigualdad en orden descendente (por ejemplo, con la mayor desigualdad en el primer orden y la menor en el último), la India ocupa el puesto 12 basado únicamente en la categoría BS, el puesto 3 basado únicamente en la categoría LS y el puesto 64 basado en la categoría NF. Sin embargo, la medida única propuesta en este estudio da a India un valor de 29.7 p.p., situándola en el 9º lugar.

Otro detalle a tener en cuenta es que la suma de los valores de desigualdad de las tres categorías da como resultado un valor cero (es decir, 30.4 p.p. (BS) – 28.8 p.p. (LS) – 1.6 p.p. (NF) = 0). Esto se debe a que, tanto para las zonas urbanas como para las rurales, la suma de las partes tiene un valor constante del 100% y, por lo tanto, la diferencia entre lo urbano y lo rural resulta cero. El valor constante de cierre es una peculiaridad de los datos de composición, y es también una de las razones por las que requiere un enfoque estadístico particular.

Por último, de los resultados obtenidos se deduce que el acceso a las instalaciones de servicios —como a las instalaciones de higiene con jabón y agua—, es un privilegio de los hogares situados principalmente en las zonas urbanas. Sao Tomé y Príncipe y Guyana son los únicos casos en los que el servicio básico era mayor en las zonas rurales que en las urbanas. Por su parte, los servicios limitados de higiene y NF son más elevados en las zonas rurales que en las urbanas.

Tabla 5.2. Resumen estadístico de la medida de desigualdad por cuartiles.

Región	Mínimo (p.p.)	Cuartil 1 (p.p.)	Cuartil 2 (p.p.)	Cuartil 3 (p.p.)	Máximo (p.p.)
Asia central y meridional (CSA)	0	3.7	20.5	28.3	34.6
Asia oriental y sudoriental (ESEA)	3.5	10.9	16.4	25.2	31.0
América Latina y el Caribe (ALC)	1.5	8.1	11.0	13.5	37.1
África del Norte y Asia Occidental (NAWA)	3.6	5.7	9.5	21.3	33.1
Oceanía	6.6	16.3	26.0	26.4	26.8
África subsahariana (SSA)	1.2	10.2	15.8	21.6	32.5
Global	0	8.7	14.4	23.1	37.1

Nota: Los valores se expresan en puntos porcentuales (p.p.) y se redondean a un decimal.

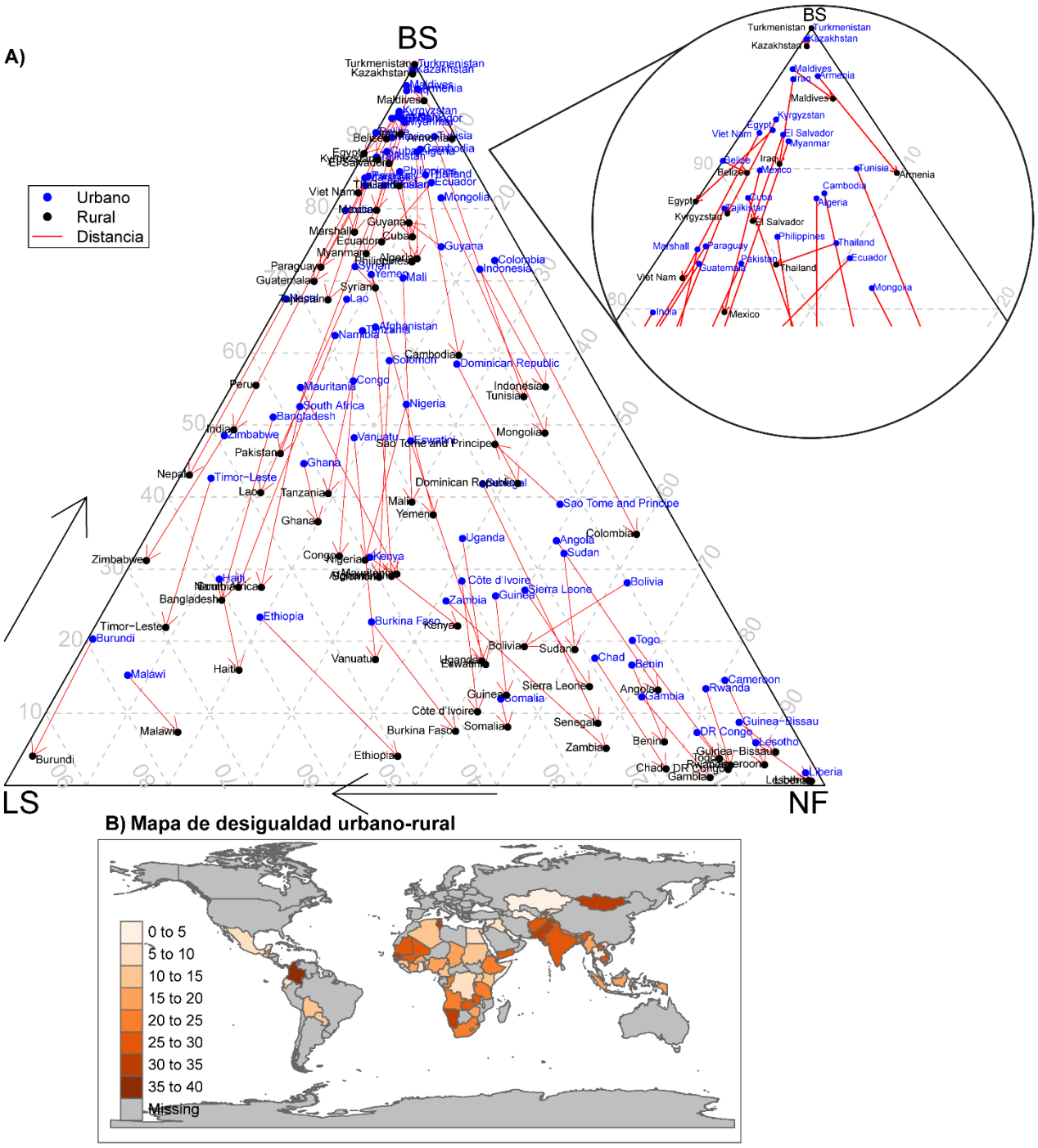


Figura 5.5. Mapa de desigualdad

Notas: Desigualdad urbano-rural del diagrama ternario, expresada como distancia (A) o mapa (B). El punto inicial del vector es la posición urbana, y el punto final del vector es la posición rural.

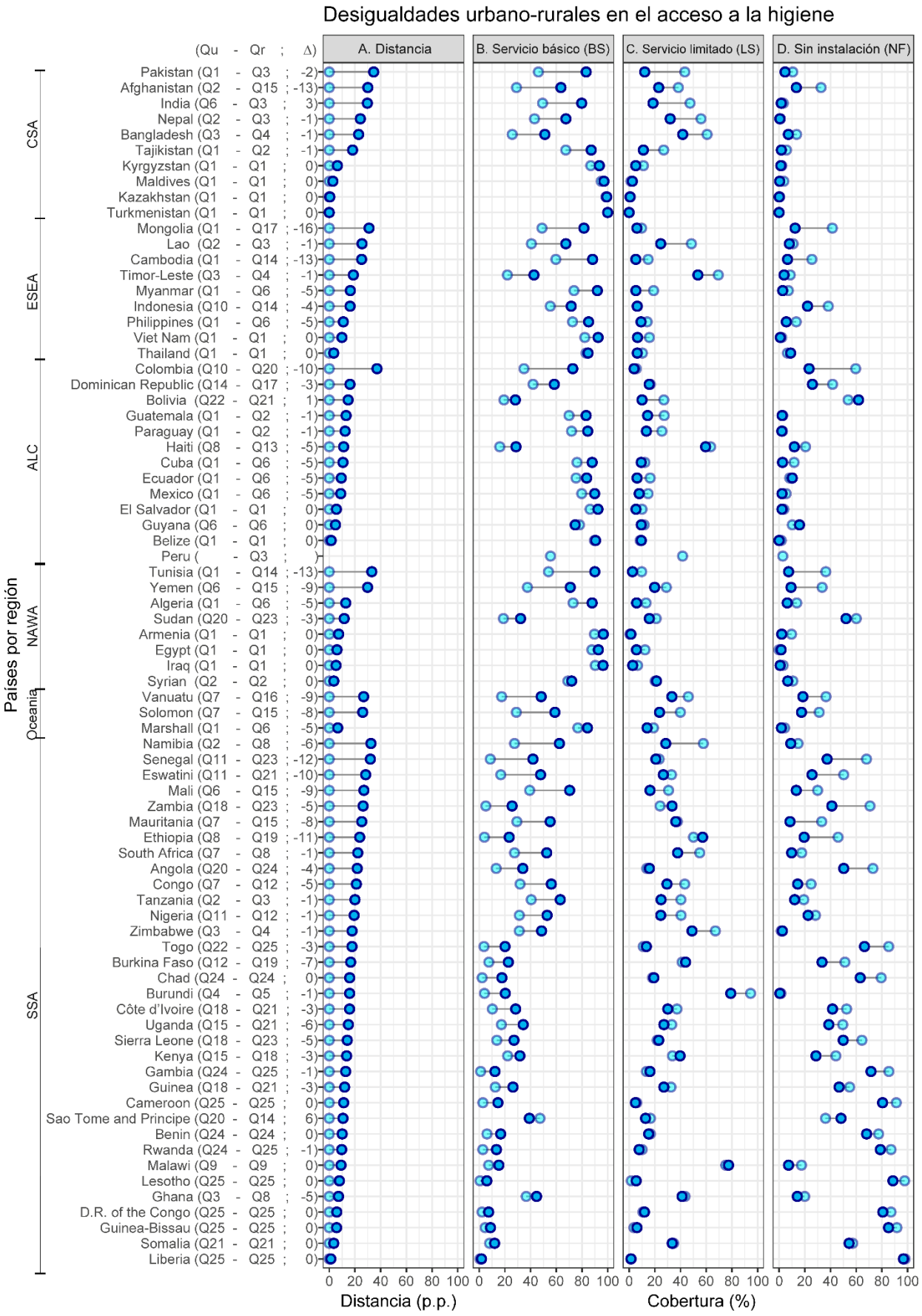


Figura 5.6. Desigualdad multivariante *y* univariante.

Notas: A) Una medida de la desigualdad expresada en términos de distancia. El círculo azul representa el límite superior de la distancia del diagrama ternario. B-D: Gráfico de desigualdad urbano-rural utilizado por WHO/UNICEF

(2019). Un círculo azul denota lo urbano, y un círculo azul claro, lo rural. ∆ indica el cambio de orden entre lo urbano (Qu) y lo rural (Qr) en la parcela ternaria (por ejemplo, en Pakistán $\Delta = Q_{u=1} - Q_{r=3} = -2$).

5.4. Discusión

5.4.1. La desigualdad urbano-rural en un diagrama ternario

Construimos veinticinco parcelas ternarias con una precisión de lectura de veinte, lo que nos permitió agrupar los países con comportamientos similares en orden desde la parcela ternaria Q1 hasta la parcela ternaria Q25. Obtuvimos resultados que muestran una mayor concentración de países con hogares urbanos (por ejemplo, 25/76 países en Q1) que rurales (por ejemplo, 11/77 países en Q1) en el vértice BS. Esto se traduce en una mayor disponibilidad de instalaciones para el lavado de manos con agua y jabón en los hogares urbanos que en los rurales. En cambio, los países de las parcelas ternarias Q24 y Q25, tanto urbanas como rurales, corresponden únicamente a países pertenecientes a la región del SSA.

Para algunos países en análisis, la amplitud de veinte no fue beneficiosa, ya que genera resultados engañosos en el orden de clasificación. Por ejemplo, en el área de residencia urbana (Figura 5.3A), Kenia (BS = 31.7%, LS = 39.6%, NF = 28.7%) tiene mejores indicadores que Burkina (BS = 22.7%, LS = 43.9%, NF = 33.4%) y, por lo tanto, es mejor; sin embargo, según nuestra clasificación, está en el Q15, mientras que Burkina está en el Q12. Sin embargo, si ajustamos la precisión de la lectura a una magnitud de diez, Kenia se situaría en el Q48, y Burkina en el Q56, en la nueva clasificación. Esto nos permite afirmar que una de las limitaciones del método tiene que ver con la selección adecuada de la precisión de lectura.

El método también está limitado cuando el punto de datos está muy cerca del límite de la parcela ternaria subdividida o de sus vértices; al no tener en cuenta la incertidumbre de los datos, es muy probable que estén en uno de ellos o en ambos. Tomando como ejemplo el mismo caso anterior, el punto de datos de Kenia está muy cerca entre el límite de Q12 y Q15 (el límite entre ambos tiene el valor de 40 en LS), por lo que, si añadimos la incertidumbre de los datos, es muy probable que Kenia esté doblemente clasificada. Otro caso más drástico es el de la Tanzania rural: el punto de datos está muy cerca de uno de los vértices de Q3, lo que implica que, si se añade su incertidumbre, también podría estar en cualquiera de ellos (Q7, Q8, Q11, Q12 o Q15). Esto pone de manifiesto la necesidad de incorporar la incertidumbre de los datos (Ezbakhe y Pérez-Foguet, 2019) para obtener una mayor precisión en el orden de clasificación.

Además, los valores de las escaleras de servicio tienen una tendencia a ir al límite superior de uno, lo que, expresado en el diagrama ternario, indica que los puntos de datos están muy cerca de los vértices. Esto es relevante ya que el ODS 6.2 busca "*...para 2030, lograr el acceso a un saneamiento e higiene adecuados y equitativos para todos y poner fin a la defecación al aire libre...*". Por lo tanto, a medida que aumenta la tasa de progreso del BS, los puntos de datos de los países tenderán al vértice del BS. Esto implica que será más relevante capturar los puntos de datos cuando estén en el Q1 bajo un comportamiento multivariado.

Nuestro método permite realizar ajustes a las nuevas condiciones, dando una lectura más precisa. Para ello, sólo será necesario ajustar la amplitud de la parcela ternaria a un valor inferior a veinte, lo que genera un aumento del número de parcelas ternarias subdivididas.

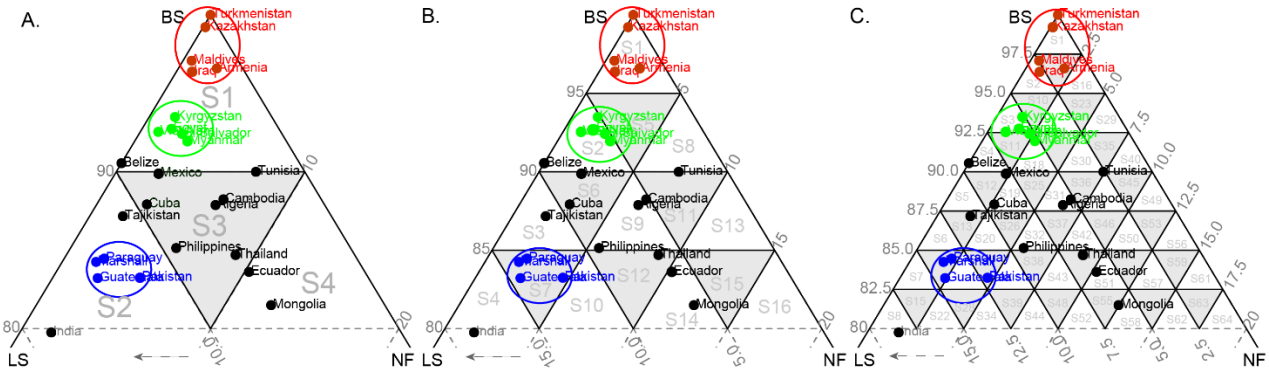


Figura 5.7. Zoom de parcelas ternarias

Notas: Ampliación de Q1 en la zona residencial urbana subdividida en parcelas ternarias con diferentes medidas de precisión de lectura. Izquierda: Cuando la precisión de lectura es 10, la subdivisión de Q1 da lugar a 4 parcelas ternarias. Centro: Cuando la precisión de lectura es de 5, la subdivisión de Q1 da lugar a 16 parcelas ternarias. Derecha: Cuando la precisión de lectura es de 2,5, la subdivisión de Q1 da lugar a 64 parcelas ternarias. (Obsérvese que, para reducir la confusión, las subdivisiones internas se indican con Sn).

Sin embargo, un aumento excesivo de la amplitud puede no ser beneficioso para la agrupación y el orden de clasificación de los países. Para ilustrar esta situación, realizamos un zoom de Q1 en la zona de residencia urbana (Figura 5.3A) y luego la subdividimos en parcelas ternarias con tres niveles de precisión de lectura (de 10, 5 y 2.5; Figura 5.7A-C). Aquí, la subfigura A se subdividió de una parcela ternaria con una amplitud de veinte, a cuatro parcelas ternarias con una amplitud de diez. El orden de la clasificación sigue la misma lógica esbozada en la Figura 5.2D y va de S1 como la mejor, a S4 como la menos favorable. También delimitamos intuitivamente tres posibles grupos similares (coloreados en rojo, verde y azul; Figura 5.7), que deberían recoger las siguientes subdivisiones a medida que aumenta la precisión de la lectura.

Al aumentar la precisión de lectura a cinco (Figura 5.7B), el grupo de países que se muestra en rojo y azul quedan perfectamente capturados por las parcelas ternarias S1 y S7, respectivamente, mientras que S2 captura el grupo de países que se muestra en verde y ha añadido Belize. Si seguimos aumentando la precisión de la lectura hasta una delimitación ternaria de 2.5, como se muestra en la Figura 5.7C, hay una mayor dispersión del grupo de puntos delimitados intuitivamente: el grupo de países en verde se distribuye ahora en S3, S11 y S18; los grupos de países en rojo y azul también tienen distribuciones múltiples. El orden de clasificación también se altera a medida que aumenta la amplitud de la lectura ternaria, como se observa en el caso de Myanmar (BS = 91.95%, LS = 5.25% y NF = 2.80%) y El Salvador (BS = 92.41%, LS = 5.31% y NF = 2.28%), mostrados en verde. Ambos países están en la misma parcela ternaria en la subfigura A y B; en la subfigura C, sin embargo, Myanmar está en S18 y El Salvador en S11, a pesar de tener puntos de datos muy cercanos.

Por lo tanto, un aumento excesivo de la precisión de la lectura tiene un efecto de dispersión en el grupo de países, que no es beneficioso para la agrupación u ordenación de la clasificación, además de la incertidumbre de orden mencionada anteriormente. Por lo tanto, hay que limitar la precisión de la lectura ternaria, teniendo en cuenta la incertidumbre de los datos y el nivel deseado de agrupación de los resultados finales.

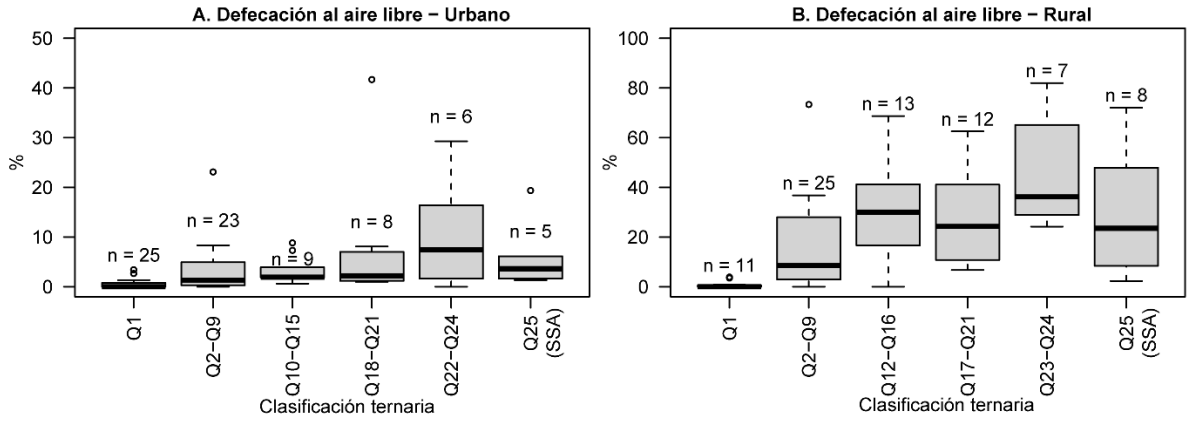


Figura 5.8. Clasificación ternaria vs defecación al aire libre

Notas: La defecación al aire libre (DO) se refiere a la eliminación de heces humanas en campos, bosques, arbustos, cuerpos de agua abiertos, playas, otros espacios abiertos o con residuos sólidos. En las regiones rurales, se excluyó del análisis a Siria, ya que no disponía de información sobre la defecación al aire libre.

En la misma línea que el ODS 6.2, que busca acabar con la defecación al aire libre y proporcionar una higiene adecuada, nuestro método nos permite explorar la relación entre los niveles de servicio de higiene y una categoría de servicio de saneamiento. Para ello, utilizamos un boxplot para cruzar la clasificación ternaria obtenida para cada país con la proporción de personas que siguen defecando al aire libre (véase la Figura 5.8). El boxplot muestra que la tasa de defecación al aire libre es más baja para los países que se encuentran en la Q1, y más alta para los que están en la Q25; esta tendencia es más drástica en las zonas rurales que en las urbanas. Esto demuestra que existe una relación en mayor o menor medida entre los que no tienen instalaciones para lavarse las manos con agua y jabón y la DO; es decir, los hogares con instalaciones de saneamiento que no practican la DO tienen más probabilidades de tener también instalaciones para lavarse las manos con agua y jabón.

Los países de las Q23, Q24 y Q25 pertenecen principalmente a la región del SSA; según la Figura 5.8, este grupo de países presenta valores elevados de DO y NF (NF > 60%). Estas condiciones en la región del SSA provocan enfermedades relacionadas con servicios ASH inadecuados, siendo la diarrea la principal. En el SSA, las mejoras en el saneamiento fueron responsables de una reducción de más del 10% en la tasa de mortalidad por diarrea en niños menores de 5 años (Troeger et al., 2018). En la misma región, Zerbo et al. (2021) descubrieron que el 7.75% (5.99% - 9.7%) de todas las muertes por enfermedades diarreicas se atribuyen a la falta de higiene. La revisión sistémica del efecto del lavado de manos con jabón sobre el riesgo de diarrea en la comunidad encontró que el lavado de manos con jabón puede reducir el riesgo de enfermedades diarreicas en un 42-47% (Curtis y Cairncross, 2003). Por lo tanto, una intervención destinada a reducir la DO y la NF en los hogares de la región del SSA tendrá un efecto positivo en su salud pública.

El ODS 6.2 también busca el acceso equitativo; en ese sentido, nuestro método da una medida integrada de una escalera de servicios de tres partes. La medida sirve para analizar la diferencia entre lo urbano y lo rural, lo que permite diferenciar las políticas públicas del sector con una visión general. Esto difiere de la simple diferencia entre la proporción de lo urbano y lo rural que realiza el PCM (WHO/UNICEF, 2016), que también se ha utilizado en otros estudios. Nuestra metodología también incorpora cálculos de distancias comunes, lo que facilita la comprensión de cada resultado obtenido, permite la comparación entre países (ya sea en un mapa temático o en medidas de distancia), y permite un ranking de desigualdad urbano-rural de los países.

Sin embargo, expresar la desigualdad en una medida única y multidimensional oculta

información a las partes, lo que no permite realizar intervenciones sectoriales específicas. Esta es una de las principales desventajas de una medida integral, como se ha señalado en otros estudios (Hsiao et al., 2005; Giné-Garriga y Pérez-Foguet, 2010; Giné-Garriga et al., 2017). Por tanto, la medida de desigualdad obtenida se refuerza aún más si se acompaña también de la medida habitual de la diferencia simple entre lo urbano y lo rural entre las partes o con la clasificación ternaria obtenida. Por ejemplo, no es lo mismo un valor de desigualdad de 3.5 p.p. en Tailandia que un valor de 3.5 p.p. en Somalia. Esto se puede visualizar mejor sumando la parcela ternaria a la que pertenecen: Tailandia está en Q1 (es decir, tiene 80 < BS, LS < 20, NF < 20), y Somalia está en Q21 (es decir, tiene BS ≤ 20, 20 ≤ LS ≤ 40, 40 ≤ NF ≤ 60). Esto ayudará a proporcionar una magnitud global para las políticas con una visión general y magnitudes desagregadas para las intervenciones específicas.

En este estudio, no hemos podido calcular la desigualdad urbano-rural para algunos países por falta de información. Sin embargo, para los países con información disponible, nuestros resultados muestran que las desigualdades urbano-rurales pueden diferir en magnitud de un país a otro, desde un valor mínimo de cero hasta un valor máximo de 37.1 p.p.. Con la variación geográfica regional, la desigualdad urbano-rural en CSA tiene un valor mínimo de cero y un valor máximo de 34.6 p.p.; en ESEA, el valor mínimo de 3.5 p.p. y un valor máximo de 31.0 p.p.; en ALC, un mínimo de 15 p.p. y un máximo de 37.1 p.p.; y en NAWA, un mínimo de 3.6 p.p. y un máximo de 33.1 p.p., en Oceanía, el mínimo es de 6.6 p.p. y el máximo de 26.8 p.p., y en SSA, el mínimo es de 1.2 p.p. y el máximo de 32.5 p.p.. Las medidas de desigualdad obtenidas justifican estrategias de gestión diferenciadas tanto para las zonas urbanas como para las rurales, con mayor énfasis en los hogares rurales, así como en los países que se encuentran en la parcela ternaria Q25, que son los países con mayor proporción de NF.

5.4.2. Instalaciones de higiene y COVID-19

Nuestros resultados se basan en la información de 2017. Sin embargo, consideramos que el análisis por parcelas ternarias con una amplitud de veinte ayuda a compensar los breves cambios que se han generado hasta la fecha en el análisis, lo que nos permite cruzar la información obtenida de la higiene con los datos de casos y muerte por la enfermedad COVID-19. La información cruzada de la higiene corresponde únicamente a los países que tienen lo urbano y lo rural en la misma parcela ternaria o en el rango de clasificación de las parcelas ternarias (por ejemplo, en la Figura 5.6, Ghana tiene lo urbano en la Q3 y lo rural en la Q8 y, por lo tanto, se considera en la clasificación de la Q2 a la Q9); los países que no cumplen esta condición fueron excluidos de este análisis. En cualquier caso, esta discusión es informativa para ayudar a los actores nacionales e internacionales a entender el vínculo entre la higiene y la salud pública.

No hemos podido calcular la desigualdad urbano-rural ni la realidad de algunos países en materia de higiene. La razón principal es la escasa o nula información. Es más drástico en las regiones de altos ingresos, como Europa, América del Norte, Australia y Nueva Zelanda, para las que no hay información ternaria sobre higiene para el año 2017. Un estudio reciente de Brauer et al. (2020) reafirma la hipótesis de que hay pocos datos globales sobre higiene. Aun así, con información complementaria, pudieron estimar la proporción de la población sin instalaciones para lavarse las manos con agua y jabón para el año 2019. Muestran que los países de altos ingresos tienen niveles bajos de la proporción de la población sin acceso a una estación de lavado de manos con agua y jabón. El hecho de tener casi todos los servicios básicos cubiertos es probablemente una de las razones por las que la recogida de información pasa a un segundo plano. Sin embargo, la mayoría de estos países también tienen el mayor número de casos confirmados y de muertes por COVID-19 según la OMS (véase la figura A2).

El caso de EE.UU. es el más ilustrativo, sobre todo porque las estimaciones de Brauer et al. (2020) sugerían que la proporción de la población sin acceso a una estación de lavado de manos con agua y jabón era de un valor del 0.4% [inferior = 0.3%, superior = 0.5%] en 2019. Sin embargo, a 15 de marzo de 2021, los EE.UU. ocupan el noveno lugar en el ranking de casos confirmados de COVID-19 por cada 100,000 habitantes y el decimotercer lugar en el ranking de muertes por COVID-19 por cada 100,000 habitantes.

En particular, Ahmad et al. (2020) mostraron que los condados con un mayor porcentaje de hogares con viviendas precarias tenían una mayor incidencia y mortalidad asociada al COVID-19. Los hogares que tienen cualquiera de los siguientes cuatro problemas se consideran condiciones precarias: hacinamiento, alta carga de costes de la vivienda, instalaciones de cocina incompletas e instalaciones de fontanería incompletas. Las instalaciones de fontanería incompletas están relacionadas con los hogares que carecen de agua entubada caliente y fría, de un inodoro con cisterna o de una bañera/ducha. La escasez de agua, la falta de agua o la inseguridad del agua en los hogares (Stoler et al., 2021) dificulta la higiene de las manos, lo que se traduce en un mayor riesgo de infección por el virus del SRAS-COV2. Dicho esto, es probable que, en los hogares sin las condiciones necesarias para practicar la higiene de las manos, haya un mayor aumento de la tasa de infección. Por lo tanto, la obtención de datos, aunque sean escasos, sigue siendo relevante para orientar las intervenciones.

Para los resultados obtenidos al cruzar nuestra clasificación ternaria con los casos confirmados y las muertes por COVID-19, la mediana de los países que están dentro del grupo Q1 es superior a la mediana del resto de los grupos tanto en los casos confirmados como en las muertes por COVID-19 (Figura 5.9A, B). Esto significa que algunos países que tienen niveles de servicio de higiene más altos (es decir, que pertenecen a 80 < BS, LS < 20, NF < 20) y tienen más casos confirmados y muertes por COVID-19 (total acumulado por cada 100,000 habitantes; al 15 de marzo de 2021). A su vez, los valores más bajos de la mediana de casos confirmados y muertes por COVID-19 se presentan en el rango de parcelas ternarias de Q17 a Q25 y, al mismo tiempo, los países que se encuentran en este rango de parcelas ternarias pertenecen todos a la región SSA.

Resulta paradójico que, con bajos niveles de higiene de BS, también haya una baja tasa de casos confirmados y de muertes por COVID-19 (por cada 100,000 habitantes). Sería de esperar que una menor proporción de BS condujera a una mayor tasa de contagio. Suponemos además que las cifras disponibles representan el número real de muertes de forma más o menos aproximada, pero con la misma exactitud independientemente del país; sin embargo, algunas noticias internacionales en el momento de escribir esta investigación arrojan dudas sobre esta hipótesis (ANDINA, 2021; GESTIÓN, 2021). No obstante, en la literatura están surgiendo posibles explicaciones sobre este fenómeno, que está ocurriendo en algunos países de la región del SSA. Una de las explicaciones es que los países del SSA tienen poblaciones más jóvenes, lo que podría actuar como un factor de protección contra el COVID-19. Otra posible explicación está relacionada con la baja proporción de adultos mayores en el SSA, dado que los adultos mayores tienen un mayor riesgo de hospitalización o muerte al contraer la enfermedad COVID-19 (CDC, 2019; Shahid et al., 2020). Según los Centros para el Control y la Prevención de Enfermedades (CDC), en comparación con las personas de 5 a 17 años, la tasa de hospitalización es 40 veces mayor en las personas de 65 a 74 años, 65 veces mayor en las de 75 a 84 años y 95 veces mayor en las de ≥85 años (CDC, 2019). Para contrastar estas afirmaciones, cruzamos la clasificación de los países en parcelas ternarias obtenidas con la edad media de la población y la proporción de la población de ≥65 años (Figura 5.9C, D). Los resultados muestran que la edad media en el SSA es inferior a la edad media de los

países que se encuentran en la parcela ternaria Q1; lo mismo ocurre con la proporción de la población de edad ≥65 años.

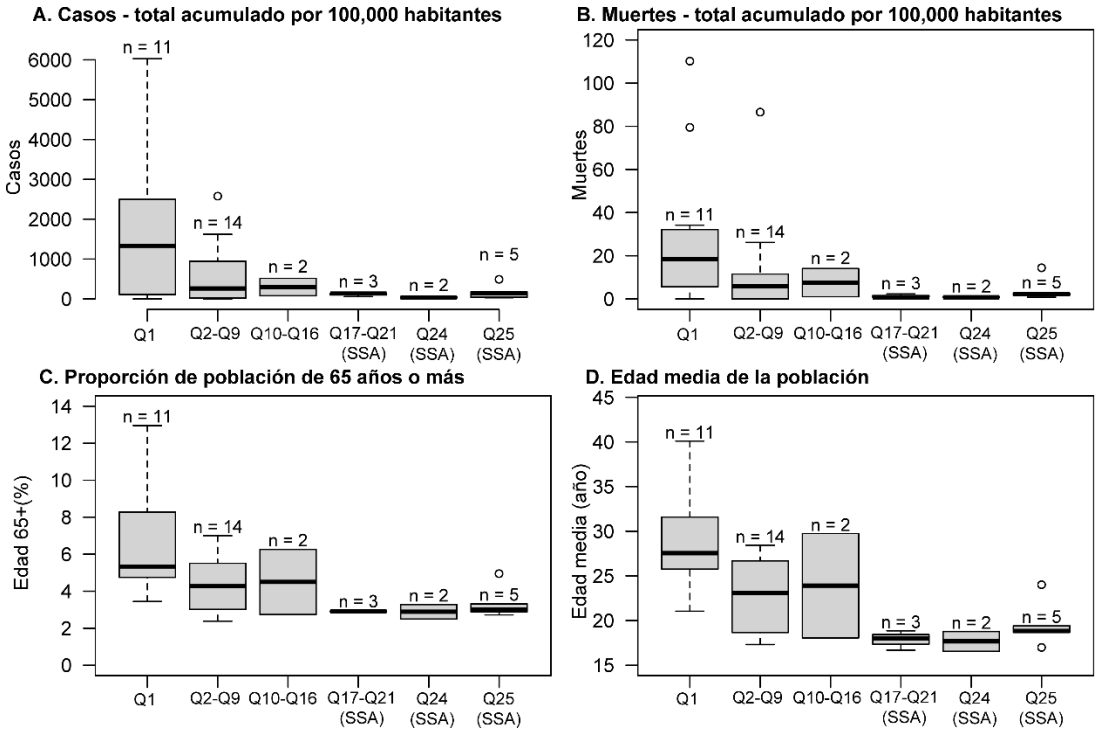


Figura 5.9. COVID-19 vs Higiene

Notas: A-D) Gráficos basados en el nivel de servicio, por ejemplo, los países de Q1 a Q9 pertenecen al nivel 1 (N1 a N5; véase también la Figura 5.2D). Casos (A) y muertes (B) por COVID-19 por cada 100,000 habitantes. Proporción de la población que tiene ≥ 65 años (C) y edad media de la población (D); información obtenida de United Nations Department of Economic and Social Affairs Population Division (2019). (United Nations Department of Economic and Social Affairs Population Division, 2019)

En la literatura se han sugerido otras explicaciones, como el papel del clima (Adedokun et al., 2020; Huang et al., 2020), la vacunación contra el Bacillus Calmette-Guerin (BCG) (que podría tener un papel protector contra el COVID-19) (Curtis et al., 2020; Miller et al., 2020), la densidad de población (Tcheutchoua et al., 2020), y otros (Lalaoui et al., 2020; Mbow et al., 2020; Tcheutchoua et al., 2020). Entre ellas, la más destacada es probablemente las lecciones aprendidas al hacer frente a enfermedades como el ébola, la malaria, el VIH y otras, dado que las infraestructuras construidas y los sistemas de gestión de la respuesta establecidos se han adaptado rápidamente para hacer frente a la pandemia actual (Lumu, 2020; Nachega et al., 2020; Payne, 2020).

Todo esto podría explicar en parte que, incluso estando en las parcelas ternarias de Q17 a Q25, algunos países de la región del SSA tengan valores bajos de casos confirmados y muertes por COVID-19. Lo más llamativo es que los países que están en Q25, como Camerún, Lesoto, RD Congo, Guinea-Bissau y Liberia, también tienen valores bajos de casos confirmados y muertes por COVID-19 y, al mismo tiempo, representan los países con las condiciones más desfavorables para la práctica de la higiene de manos, es decir, BS < 20, 80 < NF, LS < 20.

Por último, es muy posible que, a pesar de tener acceso a una instalación, se practique mal el lavado de manos con agua y jabón (Wolf et al., 2019), lo que aumenta aún más la probabilidad de contraer la enfermedad COVID-19.

5.5. Mensajes claves

Hasta donde sabemos, es la primera vez que se aplica una medida de desigualdad al sector de la higiene que tiene en cuenta las características multivariantes de los datos. Además, también es el primer intento de clasificar los países en parcelas ternarias que posteriormente pueden representarse en un mapa temático, con la potencial aplicación de un análisis espacial.

La construcción del mapa temático con cada parcela ternaria (Qn) ofrece una mejor visualización e interpretación de los resultados, ya sea agrupados por regiones o individualmente para cada país. Es necesario que cada parcela ternaria (Qn) se lea como delimitaciones ternarias, o (en el caso individual de un país representado por datos en el diagrama ternario) como datos con información tripartita. La omisión de alguna de las categorías en la lectura puede contribuir a la inexactitud en la interpretación de los resultados. En consecuencia, en este artículo, proporcionamos otra forma de explorar e interpretar los datos de higiene en su espacio, que es el simplex.

Asimismo, proponemos una medida alternativa de la desigualdad urbano-rural cuando la característica de los datos es compositiva y ternaria, sin infringir las propiedades de los datos (es decir, invariabilidad de escala, coherencia subcomposicional). Esta nueva medida de la desigualdad es diferente de la simple diferencia urbano-rural que se utiliza ampliamente en el seguimiento global del acceso de los hogares a los servicios de agua potable y saneamiento. Tiene el potencial de aplicarse al seguimiento de la desigualdad urbano-rural de las escaleras de servicio del PCM sobre estimaciones de agua, saneamiento, higiene, gestión de residuos y limpieza ambiental en los centros de salud (WHO/UNICEF, 2019c) que tienen información tripartita (servicio básico, servicio limitado y sin servicio). También puede utilizarse para supervisar la desigualdad entre zonas urbanas y rurales en las escaleras de servicios del PCM sobre agua potable, saneamiento e higiene en las escuelas (UNICEF and WHO, 2020) que también tiene información tripartita.

Los resultados que obtuvimos resaltan la necesidad de seguir haciendo esfuerzos en los hogares rurales, para reducir la brecha urbano-rural que se evidenció en 2017, ya que tienen menor disponibilidad de instalaciones básicas para el lavado de manos que los hogares urbanos. Esto se traduce en *no dejar a nadie atrás*. Por otro lado, Colombia es la prueba de que un país de ingresos medios-altos no es necesariamente una garantía de igualdad en el acceso a instalaciones para el lavado de manos en el lugar, dado que lo identificamos como el país con mayor desigualdad en 2017.

Lograr el acceso universal a instalaciones de higiene con agua y jabón para 2030 seguirá siendo un reto, principalmente en los países de la región del SSA que fueron clasificados en la parcela ternaria Q24 (urbana: Chad, Benín, Gambia, Ruanda; rural: Angola, Benín, Chad) y Q25 (urbana: Camerún, RD del Congo, Guinea-Bissau, Lesoto, Liberia; rural: Camerún, RD del Congo, Guinea-Bissau, Lesoto, Liberia, Togo, Ruanda, Gambia), ya que se encuentran dentro de las parcelas ternarias con peores condiciones. Además, la actual pandemia ha puesto de manifiesto la necesidad de que los hogares dispongan de instalaciones para lavarse las manos con agua y jabón; la higiene continua de las manos actuará como barrera protectora, reduciendo así el riesgo de contagio del virus SARS-CoV-2 u otras enfermedades comunes.

CAPÍTULO VI. MARCO DE MONITOREO SUBNACIONAL DE LOS SERVICIOS DE AGUA Y SANEAMIENTO

Resumen

El seguimiento de los servicios de agua y saneamiento (AyS) se realiza con una visión global, regional y nacional. Sin embargo, en la mayoría de los países, la descentralización de los servicios de AyS ha tendido al nivel subnacional o comparte responsabilidades entre los gobiernos nacionales y subnacionales. La gestión a nivel subnacional adquiere mayor importancia, ya que todo lo que se haga allí repercutirá en los objetivos y metas del país. Por lo tanto, en este estudio hemos propuesto una forma de desagregar la información y formar escaleras AyS a nivel subnacional.

Los resultados muestran que la desagregación de la información para hacer modelos de interpolación a nivel subnacional requiere superar tres puntos principales: la validación de los datos mediante métodos estadísticos, técnicas estadísticas que vayan de acuerdo a las características de composición de los datos y la incorporación de la incertidumbre acoplada al modelo. También muestra que el comportamiento subnacional es heterogéneo, lo que un análisis general no capta correctamente, es decir, hay un efecto de enmascaramiento de las tendencias subnacionales que la tendencia del país no representa. Sin embargo, estos han sido casos excepcionales en algunas categorías específicas. Por último, se contrasta la aplicabilidad de los modelos no lineales en un contexto más amplio que el realizado en el Capítulo IV de esta tesis.

6.1. Introducción

La relevancia de los ODS 6.1-2 radica porque transversalmente están vinculados a otros objetivos de la Agenda 2030 (UN-Water, 2016; Requejo-Castro et al., 2020). Asimismo, por los beneficios reportados en salud (prevención de enfermedades y muertes), economía, turismo, etc. Por lo tanto, todo lo que involucra al sector ASH siempre será objeto de análisis y discusión.

Como se ha indicado en los capítulos anteriores de esta tesis, el PCM realiza el seguimiento del acceso a los servicios ASH bajo una metodología preestablecida. La comparación se realiza desde una perspectiva global y regional. Varios autores han señalado que el método de análisis de datos tiene puntos clave a mejorar para obtener estimaciones más precisas (véase ítem 1.1). Por lo cual, existen propuestas en la literatura para mejorar la metodología de seguimiento global (véase tmbién ítem 1.1). De todas las alternativas existentes, hay dos que se refieren a datos composicionales, el de Pérez-Foguet et al. (2017) y la de Ezbakhe y Pérez-Foguet (2019). Sin embargo, su aplicación práctica al conjunto de datos globales no ha sido posible hasta que en el Capítulo IV de esta tesis se aborda el preprocesamiento de datos a problemas comunes que surgen en el sector, tales como: datos con valores cero, datos faltantes y valores atípicos. Las nuevas alternativas son robustas y mejoran significativamente las estimaciones del PCM.

A nivel regional, muchos países están implementando plataformas de seguimiento de los ODS a través de sus instituciones estadísticas y consideran la unidad geográfica subnacional dentro de ellas. Estas iniciativas intentan reproducir el seguimiento mundial de los ODS. Los ODS 6.1-2 también se incluyen en estas plataformas de seguimiento. Sin embargo, parece que cada país se enfrenta a múltiples obstáculos que no le permiten replicar las escaleras de seguimiento según las definiciones internacionales estipuladas por el PCM a la situación de cada país (es decir, a nivel subnacional).

Una breve revisión de algunos países que están implementando plataformas de monitoreo de los ODS muestra que las discrepancias no sólo son evidentes en un país en particular, sino también en otros países de la región de América Latina y el Caribe (ALC). Por ejemplo, la Oficina Nacional de Estadística (ONE) de República Dominicana ha implementado en su plataforma de monitoreo como indicador del ODS 6.1 el "Total de población con acceso a fuentes de agua mejoradas". Otras categorías de la escala de servicios según el marco internacional de monitoreo no han sido evidenciadas en la plataforma de la ONE. La falta de armonización entre los indicadores globales y subnacionales es evidente en este ejemplo. Esta discrepancia también se repite en otros países de la región de ALC.

Tanto en los países que ya tienen, como en los que tendrán, el establecimiento internacional de las metas de los ODS a nivel de país induce la necesidad (u oportunidad) de desagregar la información a escalas subestatales para apoyar las políticas sectoriales y alinearlas con las metas internacionales. El proceso no es sencillo, ya que hay que tener en cuenta que i) por un lado, los datos originales utilizados para la estimación del país no tienen por qué estar originalmente desagregados geográficamente; ii) y aunque lo estén, la información puede no estar disponible; iii) por otro lado, las estimaciones realizadas a nivel de país a través de regresiones temporales y una combinación de fuentes de información fijan los valores agregados para ser considerados como válidos, pero estas técnicas no tienen por qué ser adecuadas o útiles a escala subestatal, dado que la disponibilidad de datos será diferente.

Por ello, en este trabajo hemos seleccionado un país representativo del subconjunto de países con información disponible. Perú cuenta con datos generales y subnacionales y el acceso a la información pública es razonable. Por otro lado, Perú está dividido en 24

departamentos, otros países como Bolivia tienen 9 departamentos y Ecuador tiene 24 provincias. Los tres países mencionados son países andinos en sentido estricto.

El objetivo del artículo es presentar una metodología que permite aterrizar la escala del ODS 6.1-2 a nivel subnacional, concretamente en el caso de un país con un IDH medio y una renta media-alta, e ilustrar las posibilidades y dificultades encontradas en su aplicación práctica.

6.2. Antecedentes

6.2.1. Fuentes de datos globales y subnacionales

A nivel mundial, la cantidad de información documentada por el PCM entre 2000 y 2016 ha sido un total de 3,659, de los cuales 265 corresponden a censos, 1,323 a encuestas de hogares, 1,572 a conjuntos de datos administrativos y 499 a otros (WHO/UNICEF, 2018). Estos resultados muestran que tanto las encuestas de hogares como los datos administrativos son los principales contribuyentes de información para el PCM.

Un análisis desagregado a nivel de país, como es el caso de Perú, muestra que también cuenta con múltiples fuentes de información que sirven como insumos para las estimaciones realizadas por el PCM. Una revisión de la base de datos actualizada del PCM para Perú muestra que entre 2000 y 2018 existen 59 fuentes de información en diferentes años, de las cuales 2 corresponden a censos, 40 a encuestas de hogares, 14 a conjuntos de datos administrativos y 3 a otros (JMP, 2021). En este caso, es evidente que las encuestas de hogares (16 ENAHO, 15 ENDES y 9 ENAPRES) son las que aportan la mayor cantidad de información para las estimaciones de agua y saneamiento. Las encuestas de hogares que proporcionan la información son la Encuesta Nacional de Hogares (ENAHO), la Encuesta Demográfica y de Salud Familiar (ENDES) y la Encuesta Nacional de Programas Estratégicos (ENAPRES), cada una de ellas con microdatos que requieren de un preprocesamiento previo antes de conformar las escaleras de los servicios de agua y saneamiento. Por lo tanto, las encuestas de hogares son una fuente potencial de información que, tras el preprocesamiento de los datos, puede ser utilizada para formar las escaleras de los servicios de agua potable y saneamiento a nivel subnacional.

6.2.2. Iniciativas nacionales y subnacionales para aplicar las escaleras ASH

En la región de ALC existen iniciativas públicas y privadas para implementar plataformas de monitoreo de los ODS a nivel de país (ver Tabla 6.1). Lo más destacable es la participación de las instituciones estadísticas de cada país, que son las que lideran este proceso.

Sin embargo, cada país tiene peculiaridades, por lo que no se acaba de consolidar una plataforma de seguimiento por países según el marco internacional de seguimiento. En la revisión realizada, se muestra que en algunos países las definiciones de monitoreo no están de acuerdo con el marco internacional, en otros sólo hay indicadores generales (es decir, sólo se monitorea una categoría de monitoreo y no el total de las escaleras ASH). Otros países, como Argentina, Brasil y la República Dominicana, todavía están en proceso de implementación.

Es probable que la falta de información sea una de las principales causas y también el desconocimiento de cómo aterrizar la escala global junto con sus definiciones a la escala subnacional. La característica compositiva de los datos también influye en las estimaciones temporales, lo que agrega otro problema al conocimiento de la implementación de escaleras de servicio.

Tabla 6.1. Iniciativas de seguimiento del ODS 6.1-2 en la región de ALC

País	ODS 6.1	ODS 6.2	Institución	Plataforma de seguimiento	Descripción
Perú	Proporción de la población que dispone de agua a través de la red pública	Proporción de la población que utiliza los servicios de saneamiento gestionados sin riesgo	Instituto Nacional de Estadística e Informática (INEI[1])	http://ods.inei.gob.pe/ods/objetivos-de-desarrollo-sostenible/agua-limpia-y-saneamiento Fecha de revisión (23/08/2021)	Falta de armonización con los indicadores globales
México	Indicador general		Instituto Nacional de Estadística y Geografía (INEGI[1])	http://agenda2030.mx/index.html?lang=es#/home Fecha de revisión (23/08/2021)	Por el momento, sólo se observó una de las escaleras de seguimiento
Ecuador	Se han adaptado a las definiciones de los ODS (Molina-Vera et al., 2018)		Instituto Nacional de Estadística y Censos (INEC[1])	http://www.ecuadorencifras.gob.ec/objetivos-de-desarrollo-sostenible/ Fecha de revisión (23/08/2021)	Cumple con las definiciones de los ODS
Bolivia	En el INE no se observó la implementación de la plataforma de seguimiento de los ODS 6.1 y 6.2. Sólo se encontró información genérica en el enlace de Naciones Unidas Bolivia		Naciones Unidas - Bolivia	http://www.nu.org.bo/agenda-2030/13912-2/ods-6/ Fecha de revisión (23/08/2021)	Falta información desglosada a nivel subnacional.
	Hay índices generales de los ODS		Iniciativa privada	Fecha de revisión (23/08/2021)	
Argentina	En desarrollo	En desarrollo	Consejo Nacional de Coordinación de Políticas Sociales	http://www.odsargentina.gob.ar/VinculacionODS Fecha de revisión (23/08/2021)	En desarrollo (Consejo Nacional de Coordinación de Políticas Sociales, 2021)
Uruguay	Genérico. No sigue las escaleras de agua y saneamiento		-	Fecha de revisión (23/08/2021)	En el desarrollo de
Brasil	Indicador general		Instituto Brasileño de Geografía y Estadística	https://odsbrasil.gov.br/ Fecha de revisión (23/08/2021)	Sólo se observó una categoría de seguimiento
Paraguay	Indicador general		SDG Paraguay Comisión 2030	http://comisionods.mre.gov.py/agenda-2030 Fecha de revisión (23/08/2021)	No se encontró información sobre la parte metodológica, ni información desagregada del análisis subnacional (ODS-Paraguay, 2021)
República Dominicana	Población total con acceso a fuentes de agua mejoradas.	Indicador aún no disponible	Oficina Nacional de Estadística (ONE[1])	Fecha de revisión (23/08/2021)	Falta de armonización con los indicadores globales

[1] su acrónimo en español

6.3. Materiales y métodos

6.3.1. País en estudio: Perú

Según el World Bank (2020), Perú es considerado un país de renta media-alta. A nivel subnacional, se clasifica en 24 departamentos y una provincia constitucional (llamada Callao). La capital de Perú es Lima y, según el último censo peruano de 2017 (INEI, 2017), concentra casi un tercio de la población del país (Figura 6.1).

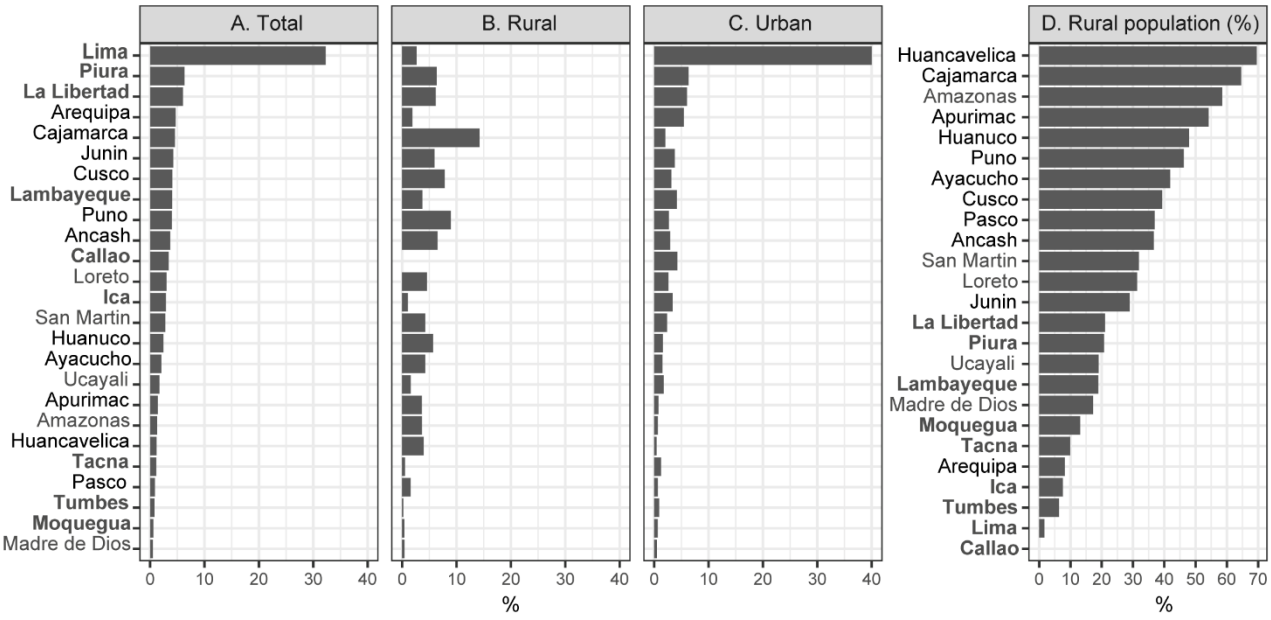


Figura 6.1. Población censada en 2017.

Notas: Los nombres de los departamentos están coloreados por región: Departamentos del Litoral (en gris negrita), de la Montaña (en negro), de la Selva (en gris). A: porcentaje de población por departamentos del total del país. B: población rural del departamento como porcentaje de la población rural total del país. C: la misma definición que en "B", pero para la urbana. D: tasa de población rural en cada departamento.

El país contiene tres regiones naturales claramente diferenciadas: la Costa, la Sierra y la Selva. La mayoría de los departamentos comparten al menos una de las regiones. Sin embargo, cada departamento tiene particularidades que permiten clasificarlos en una sola de las tres regiones —como en otros estudios (Hernández-Vásquez et al., 2016) —, así: costa (Tacna, Moquegua, Lima, Lambayeque, La Libertad, Ica, Piura y Tumbes), sierra (Cusco, Arequipa, Apurímac, Ayacucho, Ancash, Huancavelica, Cajamarca, Junín, Huánuco, Pasco y Puno) y selva (Amazonas, Madre de Dios, San Martín, Ucayali y Loreto).

Del total de la población rural, el departamento de Cajamarca es el que tiene el mayor porcentaje, seguido de Puno, y Cusco, a su vez, estos departamentos pertenecen a la región sierra (Figura 6.1B). En el caso urbano, del total de la población urbana, Lima es la que lidera seguido de Piura y La Libertad, a su vez, estos departamentos pertenecen a la región costa (Figura 6.1C).

En el análisis desagregado de la población departamental urbana y rural, Huancavelica, Cajamarca, Amazonas y Apurímac, son los que tienen una población rural mayor a la urbana (Figura 6.1D) y, a su vez, según el Instituto Nacional de Estadística e Informática (INEI) son los departamentos que históricamente han tenido una alta incidencia de pobreza (INEI, 2020).

6.3.2. Fuente de información

Para este estudio se extrajo información desagregada e histórica sobre agua y saneamiento de tres encuestas nacionales de hogares: ENDES (con 13 años de información, de 2000, 2009-2020), ENAHO (con 17 años de información, de 2004 a 2020) y ENAPRES (con 11 años de información, de 2010 a 2020). La información es de libre acceso y está disponible en la página web del INEI, http://iinei.inei.gob.pe/microdatos/.

6.3.3. Análisis de datos: compatibilidad metodológica entre lo global y lo subnacional

El procedimiento internacional para la construcción de las escaleras de monitoreo es establecido por el PCM (WHO/UNICEF, 2018). Cuenta con múltiples fuentes de información como insumos.

Tabla 6.2. Cálculo de escaleras de AyS a partir de múltiples fuentes de información

ENAPRES	**ENAHO**	**ENDES**
a) Agua - Fuente de información		
¿El suministro de agua en su casa proviene de?	¿El suministro de agua en su casa proviene de?	¿Cuál es la principal fuente de suministro de agua que utiliza en su casa para beber?
Red pública dentro de la vivienda (1), Red pública fuera de la vivienda, pero dentro de la edificación (2), Pilón o pileta de uso público (3), Camión cisterna u otro similar (4), Pozo (5), Río, acequia, manantial o similar (6), Otros (7)	Red dentro de la vivienda (1), Red fuera de la vivienda, pero dentro de la edificación (2), Pilón de uso público (3), Camión cisterna u otro similar (4), Pozo (5), Río, acequia, manantial o similar (6), Otros (7)	Red dentro de la vivienda (11), Red fuera de la vivienda, pero dentro de la edificación (12), Pilón y grifo público (13), Pozo dentro de la vivienda (21), Pozo público (22), Manantial (41), Río, presa, lago, estanque, arroyo, canal o canal de irrigación (43), Agua de lluvia (51), Camión cisterna (61), Agua embotellada (71) y Otros (96)
b) Escaleras de agua [$X_1 + X_2 + X_3 = 100$]		
Sp [(1)] = (6) x r; Sw [(2)] = (6) - Sp	Sp [(1)] = (6) x r; Sw [(2)] = (6) - Sp	
Mejorado (X_1): (1) + (2) + (3) + (4) + 0,5 [(3)] x [(5) + Sp]	Mejorado (X_1): (1) + (2) + (3) + (4) + 0,5 x [(5) + Sp]	Mejorado (X_1): (11) + (12) + (13) + (51) + (61) + (71) + 0.5 x [(21) + (22) + (41)]
Superficie (X_3): Sw	Superficie (X_3): Sw	Superficie (X_3): (43)
Otros no mejorados (X_2): X_1 - X_3	Otros no mejorados (X_2): X_1 - X_3	Otros no mejorados (X_2): X_1 - X_3
c) Saneamiento - Fuente de información		
El baño o el aseo que tiene su casa está conectado a:	El baño o el aseo que tiene su casa está conectado a:	Tipo de instalaciones sanitarias:
Red pública de alcantarillado dentro de la vivienda (1), Red pública de alcantarillado fuera de la vivienda, pero dentro del edificio (2), Letrina (3), Pozo séptico (4), Pozo ciego o negro (5), Río, acequia o canal (6), No tiene (7)	Red pública de alcantarillado dentro de la vivienda (1), Red pública de alcantarillado fuera de la vivienda, pero dentro del edificio (2), Letrina (3), Pozo séptico (4), Pozo ciego o negro (5), Río, acequia o canal (6), Otros (7), No tiene (8)	Conectado a red pública dentro de la vivienda (11), Conectado a red pública fuera de la vivienda (12), Letrina ventilada (21), Pozo séptico (22), Letrina —ciega o negra —(23), Letrina sobre río o lago (24), Río o canal (31), Sin servicio (32) y Otros (96)
d) Escaleras de saneamiento [$X_1 + X_2 + X_3 = 100$]		
Mejorado (X_1): (1) + (2) + (4) + 0.5 [(3)] x [(3) + (5)]	Mejorado (X_1): (1) + (2) + (4) + 0.5 x [(3) + (5)]	Mejorado (X_1): (11) + (12) + (21) + (22) + 0.5 x [(23)]
Defecación al aire libre (X_3): (7)	Defecación al aire libre (X_3): (8)	Defecación al aire libre (X_3): (32)
Otros no mejorados (X_2): X_1 - X_3	Otros no mejorados (X_2): X_1 - X_3	Otros no mejorados (X_2): X_1 - X_3

[(1)] Manantial en agua superficial corregida con información de ENDES, de 2004 a 2017. r = (41) / [(41) + (43)]. Para 2018 en adelante, ya no ha sido necesario corregir la información tanto de ENAHO como de ENAPRES.
[(2)] Agua superficial corregida por Sp. [(3)] El 50% se considera mejorado

Replicamos el procedimiento del PCM a nivel subnacional (en los 24 departamentos de

Perú) en tres indicadores de seguimiento: mejorados (X_1), otros no mejorados (X_2) y X_3 (defecación al aire libre en saneamiento o acceso a agua superficial). Los detalles de los cálculos se muestran en la Tabla 6.2. El análisis del indicador mejorado se justifica porque es el indicador del que surgen las tres primeras escaleras de agua y saneamiento (es decir, gestión segura, servicio básico y servicio limitado). Los indicadores X_2 y X_3 se justifican porque forman parte del cuarto y quinto orden de las escaleras de seguimiento.

6.3.4. Preprocesamiento de datos

El preprocesamiento de los datos consiste principalmente en el tratamiento de los datos de valor cero y en la identificación y validación de los valores atípicos (propuesta realizada en el Capítulo IV). Se aplica a la información tripartita de agua y saneamiento de la Tabla 6.2, que a su vez tiene una suma constante del 100% y forman vectores de composición como en la Ec. (6.1).

$$X_1 + X_2 + X_3 = 100 \tag{6.1}$$

Step i: Tratamiento de los valores cero

La Tabla 6.3 resume el número de departamentos que tienen datos con valores cero. Sin embargo, son mínimos y se dan principalmente en la categoría X_3 de aguas urbanas.

Tabla 6.3. Departamentos con valor cero según el indicador

Servicio	Sector	Indicador	Número de departamentos con el siguiente número de puntos de datos con valor cero					
			0	1 - 3	4 - 5	6 - 15	16 - 26	32
Agua	Urbano	X_1	24					
		X_2	22	2				
		X_3	1	1	3	8	10	1
Agua	Rural	X_1	24					
		X_2	24					
		X_3	18	6				
Saneamiento	Urbano	X_1	24					
		X_2	23	1				
		X_3	19	5				
Saneamiento	Rural	X_1	24					
		X_2	24					
		X_3	24					

Las tendencias temporales se realizan sobre transformaciones logarítmicas de razón de las variables X. Esto implica que el vector de valor cero se excluye del análisis o se somete a un tratamiento previo. La alternativa de excluir el vector con estas irregularidades conlleva la pérdida de información tripartita, lo que puede afectar a la calidad del modelo y, por tanto, la alternativa del tratamiento del valor cero es la más adecuada.

De las múltiples alternativas de tratamiento de valores cero que existen en la literatura, en este estudio utilizamos dos alternativas metodológicas: el reemplazo multiplicativo (Martín-Fernández et al., 2003) y el algoritmo de maximización de expectativas log-ratio (lrEM) (Palarea-Albaladejo y Martín-Fernández, 2015). La función `lrEM` se aplica cuando la información tiene una pequeña cantidad de datos de valor cero, mientras que, cuando la cantidad de datos de valor cero es muy significativa, aplicamos el reemplazo multiplicativo.

Step ii: Identificación de los valores atípicos y validación puntual de los datos

La identificación y exclusión de los valores atípicos es una práctica habitual en el análisis de datos; sin embargo, genera una pérdida de información que puede afectar a la capacidad predictiva del modelo. El PCM realiza una validación puntual de la información para cada

categoría de años y excluye los considerados atípicos. En el sector ASH, Quispe-Coica y Pérez-Foguet (2020) proponen utilizar la distancia de Mahalanobis (MD; acrónimo en inglés de Mahalanobis distance) como complemento a la validación puntual y con un mínimo de datos. La nueva alternativa es multivariante y cumple con las propiedades de los datos composicionales.

Por ello, en este estudio, la validación de los datos subnacionales sigue un doble procedimiento, primero se identifican los valores atípicos con el método MD que se encuentra en la función `outCoDa` (Templ et al., 2011), luego se complementa la validación puntual de los datos. Ambos procedimientos se realizan simultáneamente y ayudan a decidir si se excluye o no la información del año.

6.3.5. Estimación de los errores estándar

Ezbakhe y Pérez-Foguet (2019) incorporan la incertidumbre de los datos en los modelos de tendencia temporal como una aproximación. La aproximación se basa en curvas de error estándar relativo (RSE) generalizadas. La alternativa metodológica requiere como input el RSE mínimo y máximo de la fuente de información.

Este estudio utiliza el informe del INEI (INEI, 2021) como referencia para extraer el valor de la RSE del acceso al agua y al saneamiento a nivel subnacional. Los indicadores del informe difieren de lo propuesto en este estudio, sin embargo, sirven como referencia para establecer el valor de la RSE. En el caso de los servicios de agua, el valor mínimo es 2.8 y el máximo es 11.0 y, en el caso del saneamiento, el valor mínimo es 2.7 y el máximo es 15.3.

Por último, se calculan los intervalos de confianza del 95% a partir de los resultados de la regresión. Para más detalles sobre el procedimiento, véase el artículo de Ezbakhe y Pérez-Foguet (2019).

6.3.6. Análisis de tendencias

El análisis estadístico de las tendencias temporales se realiza sobre transformaciones log-ratio isométricas (ilr) de Egozcue et al. (2003) con un balance *V* definido (Egozcue y Pawlowsky-Glahn, 2005; Quispe-Coica y Pérez-Foguet, 2019). El procedimiento consiste en los siguientes pasos:

Paso i: definir el orden de los saldos

Se define un tipo de balance *V* consistente con la forma habitual de análisis en el sector ASH (Quispe-Coica y Pérez-Foguet, 2018, 2019, 2020a). El primero es entre la proporción de hogares con acceso a servicios mejorados (X_1) y no mejorados (X_2 × X_3), y el segundo es entre la proporción de hogares con acceso a otro servicio no mejorado (X_2) y X_3.

$$V = \begin{array}{c|ccc|cc} \text{Order} & X_1 & X_2 & X_3 & r & s \\ \hline 1 & +1 & -1 & -1 & 1 & 2 \\ 2 & 0 & +1 & -1 & 1 & 1 \end{array}$$

Paso ii: transformaciones logarítmicas isométricas (ilr)

La ecuación general de Egozcue et al. (2003) para las transformaciones ilr se muestra en la Ec. (6.2).

$$Y = \text{ilr} = \sqrt{\frac{r \times s}{r + s}} \ln \frac{g_m(X_r +)}{g_m(X_s -)} \tag{6.2}$$

Donde, *r* es el número de variables positivas en el balance *V*, *s* es el número de variables negativas en el balance *V* y $g_m(-)$ es la media geométrica de las variables.

Las coordenadas ilr donde se realiza el ajuste de los modelos son:

$$Y_1 = \sqrt{\frac{1\times 2}{1+2}} \ln \frac{(X_1)}{(X_2 \times X_3)^{1/2}} \quad (6.3)$$

$$Y_2 = \sqrt{\frac{1\times 1}{1+1}} \ln \frac{(X_2)}{(X_3)} \quad (6.4)$$

Paso iii: modelos y estimación

El ajuste del modelo se realiza para urbano y rural con OLS y GAM en las transformadas de la Ec. (6.3) y (6.4). Se utiliza el GAM debido a la presencia de no linealidad de los datos y porque cumple con los datos mínimos recomendados en la literatura (Fuller et al., 2016; Pérez-Foguet et al., 2017; Ezbakhe y Pérez-Foguet, 2019). El grado de libertad K del GAM es variable y se selecciona el que se ajusta a cada departamento. Este procedimiento se aplica a los servicios de agua y saneamiento.

Por último, los datos estimados se devuelven al espacio original mediante la transformación inversa ilr-1 expresada en la Ec. (6.5).

$$X = \text{ilr}^{-1}(Y^t) \quad (6.5)$$

Para la ponderación de los niveles de servicio con las unidades de población, se usa las estimaciones de población del INEI (2009). Los años sin datos se han completado extrapolando con los mismos datos existentes.

Step iv: validación del modelo con métricas de calidad

El ajuste del modelo se evalúa mediante la métrica de calidad NSE (Nash y Sutcliffe, 1970). Seguimos la clasificación de NSE realizada por otros autores de la siguiente manera: Un valor NSE = 1, significa que el modelo se ajusta perfectamente a los datos y, por lo tanto, se consideran muy buenos. NSE > 0.75 se consideran buenos, valores de NSE entre 0.36 y 0.75 se consideran satisfactorios. Un valor NSE < 0.36 se considera insatisfactorio, entre los cuales un valor de NSE = 0 significa que el modelo tiene la misma capacidad predictiva que la media observada, mientras que NSE < 0 indica que la media observada es mejor predictor que el modelo (Motovilov et al., 1999; Van Liew et al., 2003; Moriasi et al., 2007; Pérez-Foguet et al., 2017; Ezbakhe y Pérez-Foguet, 2019)

6.4. Resultados

En esta sección, mostramos los resultados de la aplicación de las técnicas de estadística composicional en datos subnacionales. En primer lugar, ilustramos las dificultades encontradas para una aplicación práctica del ajuste del modelo en datos desagregados por departamentos (abordamos el problema de los valores atípicos en el ajuste del modelo y la interpolación y extrapolación con métodos estadísticos estándar y sobre datos transformados). Posteriormente, mostramos el resultado de la aplicación de la metodología en forma de escaleras de agua y saneamiento a nivel subnacional, desagregadas en urbano y rural.

6.4.1. Ajuste del modelo e incertidumbre de los datos

El ajuste del modelo se realizó en múltiples fuentes de información en datos tripartitos. Hemos identificado tres situaciones comunes que deben superarse para mejorar el ajuste del modelo.

En primer lugar, es necesario imputar los datos con valores cero, ya que la alternativa estadística se realiza sobre datos transformados (véase la Ec. (6.2)). Se han identificado departamentos con presencia de valores cero en las categorías de servicios, principalmente

en X_3 (véase la Figura 6.2). En agua urbana, se identificaron datos de valor cero en 23/24 departamentos (Figura 6.2A); mientras que en agua rural (Figura 6.2B) y saneamiento urbano (Figura 6.2C) sólo estaban presentes en 6/24 departamentos. En contraste con el saneamiento rural, donde no hubo valores cero (en las categorías de servicio) en los departamentos analizados.

En el nivel de las categorías de servicio, cuando la tendencia de los niveles de servicio es a los extremos de todo (valor 100%) o nada (valor cero), es donde hay una mayor presencia de valores cero. La Figura 6.2A ilustra mejor esta afirmación. El valor cero de la categoría X_3 en la subfigura A significa que los hogares urbanos están en vías de cerrar el acceso al agua superficial. El departamento de Ica es el caso extremo, presentando 32/41 datos con valor cero en la categoría X_3; en departamentos con comportamientos similares, la alternativa de imputación mediante reemplazo multiplicativo es la que funcionó computacionalmente bien. La situación contraria se dio en el departamento de Loreto, donde no se encontró ningún dato con valor cero en ninguna de las categorías de servicios de AyS.

Todo lo anterior nos permite afirmar que es necesario utilizar uno u otro método para la imputación del valor cero en el análisis subnacional, teniendo en cuenta las limitaciones que cada método tiene en el cálculo computacional. Por ejemplo, la función `lrEM` tiene limitaciones en el cálculo computacional cuando hay un gran número de valores cero (caso Ica), lo que nos ha obligado a utilizar el método de imputación por sustitución multiplicativa. No tratarlos utilizando alguna de las alternativas de imputación de valores cero existentes para datos composicionales puede aumentar la pérdida de información con consecuencias para el ajuste del modelo.

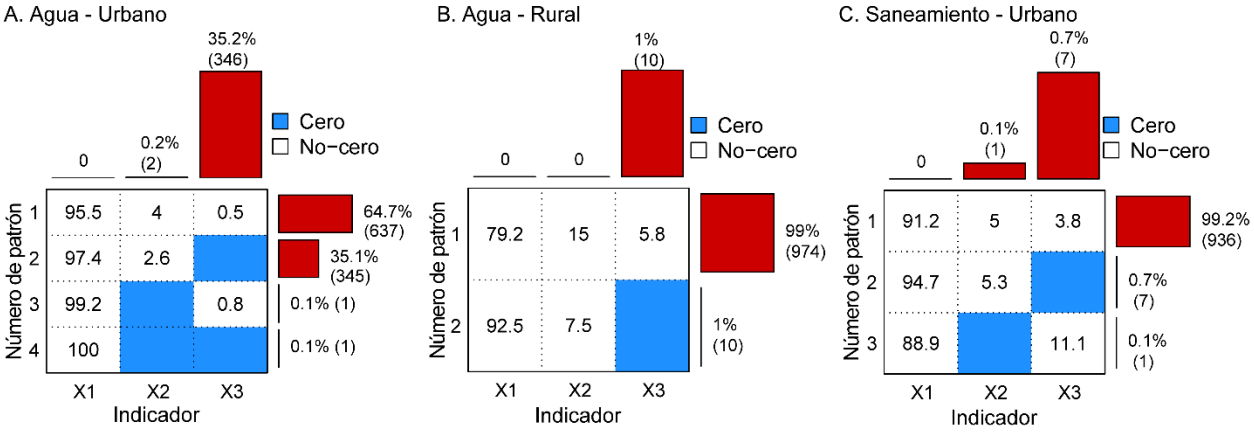


Figura 6.2. Número de patrones en datos irregulares.

En *segundo lugar*, se identificó la existencia de valores atípicos a nivel subnacional, que se validaron primero antes de ajustar los modelos. El procedimiento seguido en general ha sido, en primer lugar, aplicar el método robusto de identificación de valores atípicos de la MD en todos los departamentos, y después se aplicó la validación de puntos (el método actual utilizado por el PCM) sólo a los puntos de datos discrepantes. Esta forma de validar los puntos de datos funcionó correctamente en todos los casos. En la Figura 6.3 ilustramos la aplicación simultánea de estas dos alternativas para un departamento concreto.

En la serie temporal, la MD robusta ha identificado sólo seis puntos de datos como valores atípicos en el agua rural del departamento de San Martín. Estos puntos de datos han sido excluidos para el ajuste del modelo (el resultado se muestra en la Figura 6.3B-D en línea roja), ya que superan el umbral de corte de 2.716 (Figura 6.3A). Sin embargo, observando la Figura 6.3B se puede detectar que hay dos puntos de datos (que hemos pintado de negro a rojo para diferenciarlos del resto) que difieren significativamente del resto y que el método MD robusto

no ha identificado como valores atípicos. En nuestro estudio, también hemos excluido este tipo de datos del análisis. El nuevo modelo ajustado se muestra en azul. La principal justificación de este procedimiento se basa en el valor de la MD de estos dos puntos de datos, que están muy cerca del umbral de corte definido como atípicos con valores de 2.58 y 2.66 en 2011 y 2012, respectivamente. Es el procedimiento habitual que se realizó en todos los departamentos, tanto urbanos como rurales, para validar cada dato discrepante.

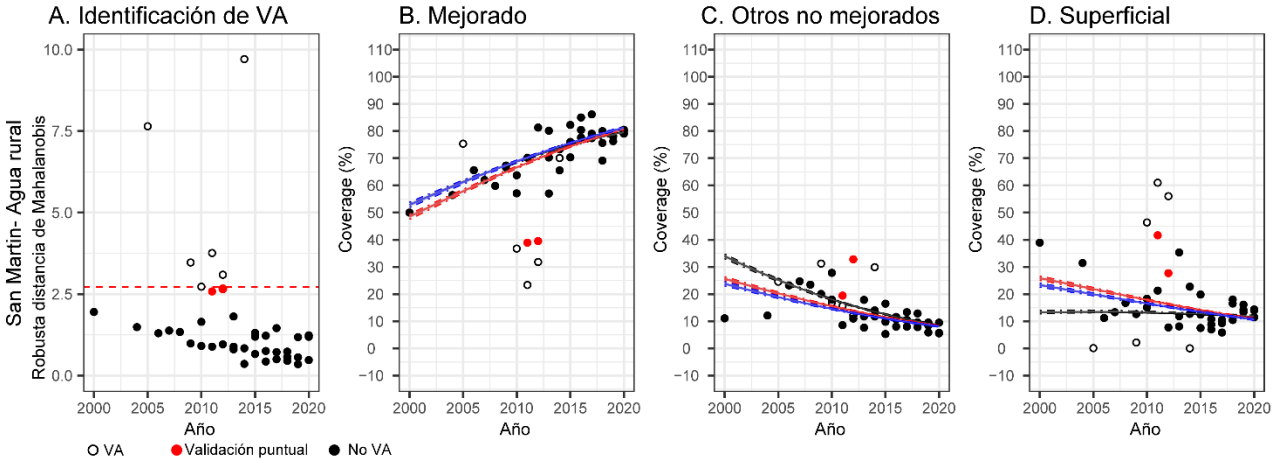


Figura 6.3. Ajuste de modelos robusto

Notas: A: Identificación de valores atípicos (VA) mediante la distancia de Mahalanobis robusta; la línea roja discontinua es el umbral de corte. B-D: Se generaron modelos (con OLS) con datos, i) que incluían valores atípicos (líneas negras), ii) sin valores atípicos (líneas rojas) y iii) sin valores atípicos (VA) y sin puntos de datos rojos (líneas azules). En la línea negra están los modelos que incluyen valores atípicos. En la línea azul están los modelos que incluyen valores atípicos.

Cabe señalar que esta forma de validar los puntos de datos se ha aplicado en menor medida a los indicadores de agua urbana, ya que el método de MD robusta ha sido más eficaz para identificar los valores atípicos que en el caso de los indicadores de agua y saneamiento rural. También hay que señalar que ha habido casos en los que el método MD robusto no ha identificado valores atípicos, en estos casos el ajuste del modelo ha sido directo, es decir, sobre todos los datos. En resumen, el ajuste del modelo calibrado se ha realizado excluyendo el 19.87% y el 18.65% de la información sobre los servicios de agua y saneamiento, respectivamente.

Por último, la identificación, validación y exclusión o no de los valores atípicos son procedimientos necesarios a nivel subnacional, ya que tendrán un impacto en el ajuste del modelo en las categorías de servicios. Por ejemplo, la Figura 6.3D muestra que el valor de interpolación es del 13.2% para el año 2000 cuando el modelo se ajusta a los datos tal cual, este valor aumenta al 23.3% si el modelo se ajusta a los datos validados (es decir, después de eliminar los valores atípicos y los datos rojos). Si comparamos los resultados de la interpolación con los datos obtenidos de la encuesta (del 38.9%) para el año 2000, observamos que el valor de la interpolación sobre los datos validados se acerca más a los datos de la encuesta que el valor obtenido del ajuste del modelo sobre los datos brutos. Esta situación también se ha comentado en el seguimiento global de ASH (Quispe-Coica y Pérez-Foguet, 2019, 2020a), por lo que el preprocesamiento previo al ajuste del modelo es también un procedimiento esencial en un análisis subnacional.

En *tercer lugar*, la incertidumbre de los datos se incluye en el ajuste del modelo de análisis (Ezbakhe y Pérez-Foguet, 2019). Hemos aplicado el algoritmo de simulación presentado por Ezbakhe y Pérez-Foguet (2019) a datos subnacionales. Los límites inferior y superior se calculan a partir de los resultados estimados para cada año. Esta forma de incorporar la

incertidumbre cuando el error de muestreo no está disponible tiene ventajas sobre no informar. Sin embargo, hemos encontrado que la alternativa metodológica de ajuste Ezbakhe y Pérez-Foguet (2019) tiene limitaciones en su aplicación a un contexto más amplio del sector; principalmente en el análisis del servicio de agua urbana, donde hemos encontrado que en 16/24 departamentos, en algunos años, los límites superior e inferior no incluyen el mejor ajuste dado por una sola tendencia temporal del modelo. Esto sucede principalmente con OLS, con algunas categorías convergiendo a cero y uno. Ver Figura 6.4.

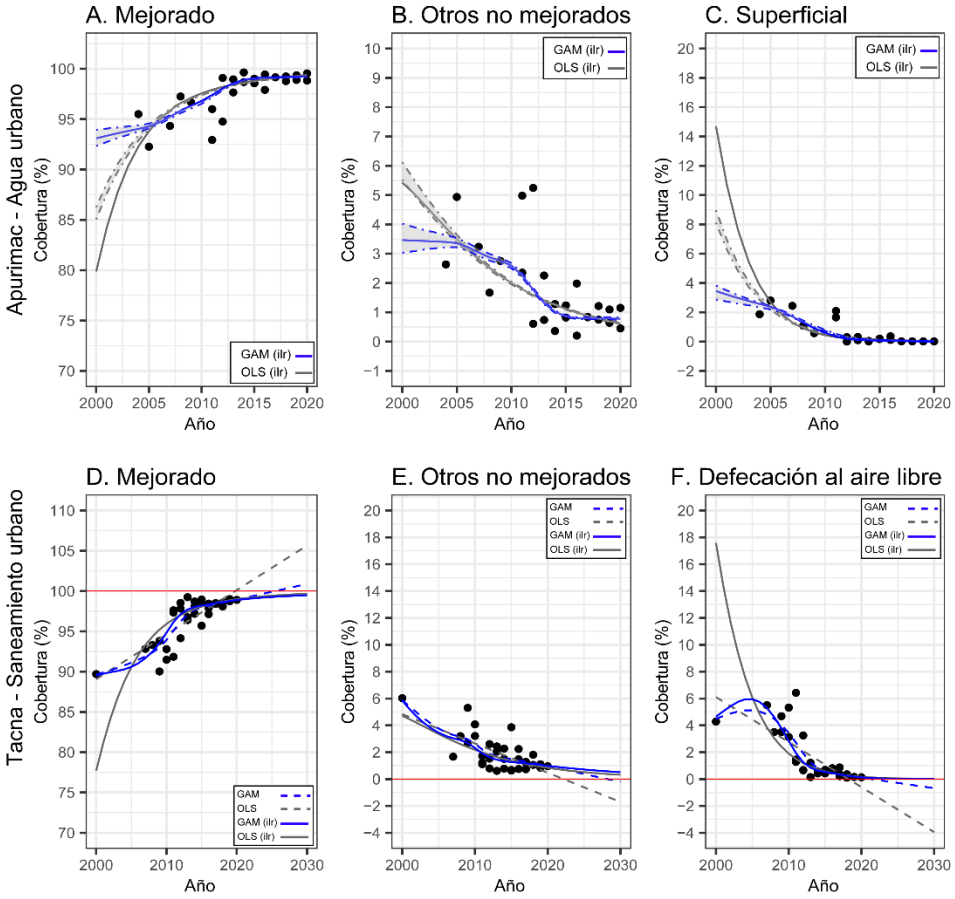


Figura 6.4. Modelo con incertidumbre

Notas: A-C: Ajuste del modelo con intervalos de confianza del 95% sobre datos transformados (ilr). D-F: Ajuste del modelo con GAM y OLS con estadísticas clásicas (líneas discontinuas) y sobre datos transformados (líneas sólidas).

Ilustramos estas limitaciones en la Figura 6.4A-C, donde hemos representado el ajuste del modelo con OLS (ilr) y GAM (ilr). El valor estimado de agua mejorada para el año 2000 es del 93.1% [92.3% - 93.9%] con GAM (ilr), mientras que con OLS (ilr) es del 79.9% (85.0% - 86.3%). El valor estimado de GAM (ilr) está dentro del límite inferior y superior, mientras que el límite inferior de OLS (ilr) está por encima de su valor estimado. Como puede observarse, en el caso de OLS (ilr) existe una discrepancia entre el valor estimado y su límite inferior, lo que no es coherente con las definiciones. Mientras que, para los servicios de saneamiento (urbano y rural), el ajuste del modelo sumado a sus incertidumbres tuvo comportamientos esperados con las definiciones. En agua rural, sólo el departamento de Tacna presentó esta discrepancia.

Por otra parte, el potencial del ajuste de los modelos en datos transformados se pone de manifiesto cuando los datos tienen una tendencia a los extremos de valor cero y uno (o 100 en porcentaje). Para ilustrar el potencial de los métodos estadísticos para los datos compuestos sobre los métodos estadísticos comunes en los modelos de tendencia temporal, hemos ampliado la extrapolación del modelo hasta 2030 (véase la Figura 6.4D-F). La Figura 6.4D muestra el caso en el que la extrapolación mediante GAM y OLS tiende a superar el valor

límite extremo del 100%, mientras que la extrapolación mediante los mismos métodos sobre datos transformados (es decir, GAM (ilr) y OLS (ilr)) no supera este límite extremo. La Figura 6.4E-F muestra el caso en que la extrapolación por los métodos estadísticos habituales tiene valores negativos, es decir, se obtienen valores por debajo de cero; esto no ocurre al extrapolar sobre los datos transformados. La comparación de los valores estimados para un año concreto, 2030, se muestra en la Tabla 6.4.

Tabla 6.4. Estimación con GAM y OLS para 2030

Urbano - Saneamiento	Método	Año de estimación (2030)		
		Mejorado	Otros no mejorados	Defecación al aire libre
Tacna	GAM (ilr)	99.482% (99.318% - 99.599%) [a]	0.513% (0.396% – 0.677%)	0.005% (0.003% - 0.008%)
	GAM	100.883%	-0.206	-0.677
	OLS (ilr)	99.670% (99.648 – 99.691)	0.315% (0.295% - 0.337)	0.015% (0.013 – 0.017)
	OLS	105.656	-1.706	-3.950

[a] Límite inferior y superior del intervalo de confianza del 95%. Se muestra con tres decimales, para ver el intervalo de confianza de la defecación al aire libre. Comparación del valor estimado utilizando estadísticas clásicas y sobre datos transformados (ilr).

Finalmente, todo el proceso anterior ha servido para tener los mejores ajustes posibles del modelo que han sido validados por las métricas de calidad. El resultado se muestra en la Tabla 6.5. En agua rural, el 58.4% de los modelos en X tienen valores de NSE clasificados como cualificados y buenos, valor que aumenta al 59.7% en agua urbana en las mismas clasificaciones. Mientras que, en el saneamiento rural, el 63.9% de los modelos en X tienen valores de NSE clasificados como cualificados y buenos, valor que aumenta al 66.6% en el saneamiento urbano. Por otro lado, también se obtuvieron valores de NSE negativos; en el servicio de agua urbano es donde se obtuvo el mayor número de modelos, que representan el 11.1% (8/72), y esto ocurrió en la categoría X_3. El valor negativo significa que la media observada es un mejor estimador que el modelo.

Tabla 6.5. Métrica de calidad de ajuste del modelo de los departamentos.

Sector	Métrica		Agua				Saneamiento			
			X_1	X_2	X_3	T [b]	X_1	X_2	X_3	T
Rural										
	Bien	(NSE > 0,75)	2	2		4/72 (5.6%)	5		11	16/72 (22.2%)
	Calificado	(0,36 < NSE < 0,75)	21	9	8	38/72 (52.8%)	15	3	12	30/72 (41.7%)
	No calificado	(NSE < 0,36)	1	13	16	30/72 (41.6%)	4	19	1	24/72 (33.3%)
	Negativo	(NSE < 0)				0		2		2/72 (2.8%)
	Total [a]		24	24	24	72/72 (100%)	24	24	24	72/72 (100%)
Urbano										
	Bien	(NSE > 0,75)	6	5	2	13/72 (18.0%)	3	0	4	7/72 (9.7%)
	Calificado	(0,36 < NSE < 0,75)	13	14	3	30/72 (41.7%)	16	8	17	41/72 (56.9%)
	No calificado	(NSE < 0,36)	5	5	11	21/72 (29.2%)	4	13	2	19/72 (26.4%)
	Negativo	(NSE < 0)			8	8/72 (11.1%)	1	3	1	5/72 (6.9%)

Total [a]	24	24	24	72/72 (100%)	24	24	24	72/72 (100%)

[a] Se refiere al número total de departamentos analizados. [b] Se refiere al número total de modelos de las categorías de servicios.

En resumen, el modelo se ajustó con GAM y OLS. En el caso del servicio de saneamiento, el GAM funcionó mejor en 21 de los 24 departamentos; mientras que, en el caso del servicio de agua, en el ámbito urbano, el GAM funcionó mejor en 22 de los 24 departamentos y, en el ámbito rural, en 17 de los 24 departamentos.

6.4.2. Escaleras de servicios de agua potable y saneamiento

Se ha podido estimar la cobertura del servicio a nivel subnacional tanto en el ámbito urbano como en el rural, mientras que el agregado nacional es el resultado de ponderar los valores estimados en las categorías de servicio a nivel subnacional con la población de cada una de ellas. La evolución temporal de 2000 a 2020 se muestra en la Figura 6.5 para el agregado nacional, mientras que para el nivel subnacional se ilustra en forma de escalera en la Figura 6.6.

Las estimaciones del PCM se generan a partir de datos nacionales agregados con OLS (WHO/UNICEF, 2018). Al compararlas con los resultados que hemos obtenido, encontramos que no hay cambios significativos en cinco categorías de servicios entre las dos formas de obtener el resultado del país. No ocurre así en la categoría "otros no mejorados" de saneamiento rural, donde hay un punto de ruptura en 2017 para las estimaciones del país al ponderar el nivel subnacional, es decir, desde el 2017 hasta el 2020, la proporción de la categoría "otros no mejorados" tiene una tendencia decreciente del 22.6% al 22.4%, respectivamente, mientras que en la estimación del PCM tiene una tendencia positiva del 21.1% al 21.6%.

Esta diferencia de tendencia tiene dos posibles explicaciones. En primer lugar, el método utilizado por el PCM es la regresión lineal, que no recoge correctamente los datos no lineales (Fuller et al., 2016; Pérez-Foguet et al., 2017). En segundo lugar, las tendencias de los modelos a nivel subnacional no son homogéneas, por lo que al ponderar por unidades de población los resultados agregados también pueden ser no lineales.

En general, Perú ha logrado avances significativos desde 2000 hasta 2020 en la prestación de servicios mejorados. La tasa media anual de progreso en el acceso al agua es del 0.34% para las zonas urbanas y del 1.36% para las zonas rurales, mientras que para el saneamiento la tasa media anual de progreso es del 0.53% para las zonas urbanas y del 1.70% para las zonas rurales. Sin embargo, aunque la tasa de progreso ha sido mayor para los hogares rurales que para los urbanos, los hogares urbanos son los que están más cerca del acceso universal a los servicios mejorados de AyS. La desigualdad entre las zonas urbanas y rurales sigue siendo importante.

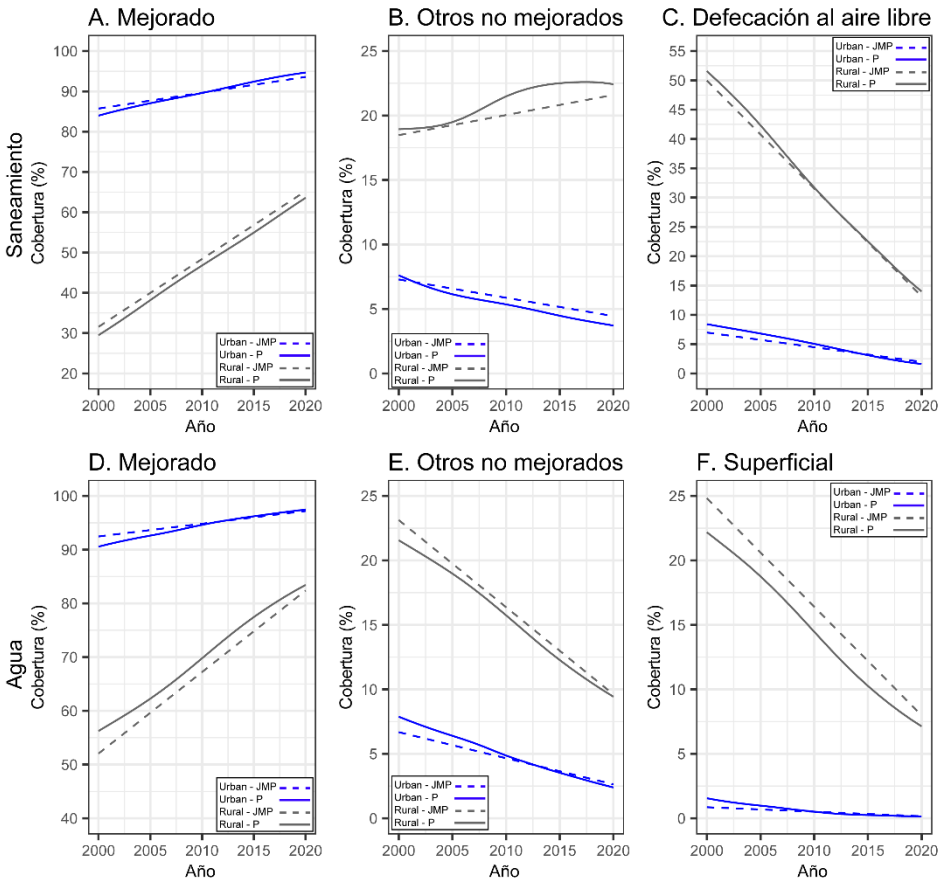


Figura 6.5. Cobertura nacional de servicios de AyS

Notas: P: cobertura del servicio obtenida ponderando la cobertura de los departamentos por su población respectiva (líneas sólidas). Estimación del PCM obtenida de washdata.org (líneas discontinuas). Urbano en líneas azules y rural en líneas grises.

Los resultados del acceso a los servicios de AyS a nivel subnacional muestran una heterogeneidad espacial y temporal. Con una mayor provisión de servicios mejorados de agua potable y saneamiento en la zona de residencia urbana que en la zona rural.

En el área de residencia urbana (Figura 6.6A), el avance de los departamentos a nivel subnacional ha hecho que, en el 2020, 19/24 departamentos tengan el valor de la categoría de servicio de agua mejorada en más del 95%. De los cuales 5 departamentos (Tacna, Cusco, Moquegua, Arequipa, Apurímac y Ayacucho) están a la cabeza con una cobertura superior al 99%. Por otro lado, los 5/24 departamentos restantes (Pasco, Tumbes, Puno, Ucayali y Loreto) tienen valores que oscilan entre el 87% y el 95%. Ucayali y Loreto tienen la menor cobertura de acceso a agua mejorada, con 89.8% (89.5 – 90.1) y 87.8% (87.5 – 88.1), respectivamente, y también tienen los valores más altos en la categoría de otros servicios de agua no mejorados, con 10.1% (9.8 – 10.4) y 10.1% (9.8 – 10.3), respectivamente.

En la sierra urbana, en cuanto al acceso a agua mejorada, los departamentos de Pasco y Huánuco son los que han tenido la mayor tasa de avance promedio por año del 2000 al 2020 con valores de 1.51% y 1.27%, respectivamente; aun así, solo han alcanzado una cobertura de 94.3% (94.0 – 94.6) y 97.2% (97.0 – 97.3), respectivamente, en el 2020. Puno es el que se ha mantenido casi estático, desde el 2000 al 2020, sólo pasando de 88.6% (87.8 – 89.3) a 90.0% (89.4 – 90.3), respectivamente. En el caso de la costa urbana, Tumbes —90.9% (90.6 – 91.1)— y Piura —95.3% (95.2 – 95.4)— tienen la menor dotación respecto al resto.

En cuanto al acceso al agua rural (Figura 6.6B), Loreto y Ucayali fueron nuevamente los

departamentos con los valores más bajos en la categoría de agua mejorada. También son los departamentos con mayor proporción de hogares con acceso a agua superficial en 2020, con 52.8% (52.2 – 53.4) y 30.2% (29.7 – 30.6), respectivamente. El resto de departamentos (22/24) también han incrementado su cobertura de agua mejorada, superando el porcentaje de 70 en 2020, estando a la cabeza los departamentos de Apurímac, Ayacucho, Moquegua y Cusco (con > 93%). Esto significa que se ha incrementado la provisión del servicio de agua entubada y/o de otras formas mejoradas de servicios de agua. La reducción de hogares con acceso a agua superficial sigue siendo un reto, principalmente en Loreto, Ucayali, Piura —17.2% (16.9 – 17.6)—, Madre de Dios —11.7% (11.5 – 11.9)— y San Martín —10.6% (10.4 – 10.8)—.

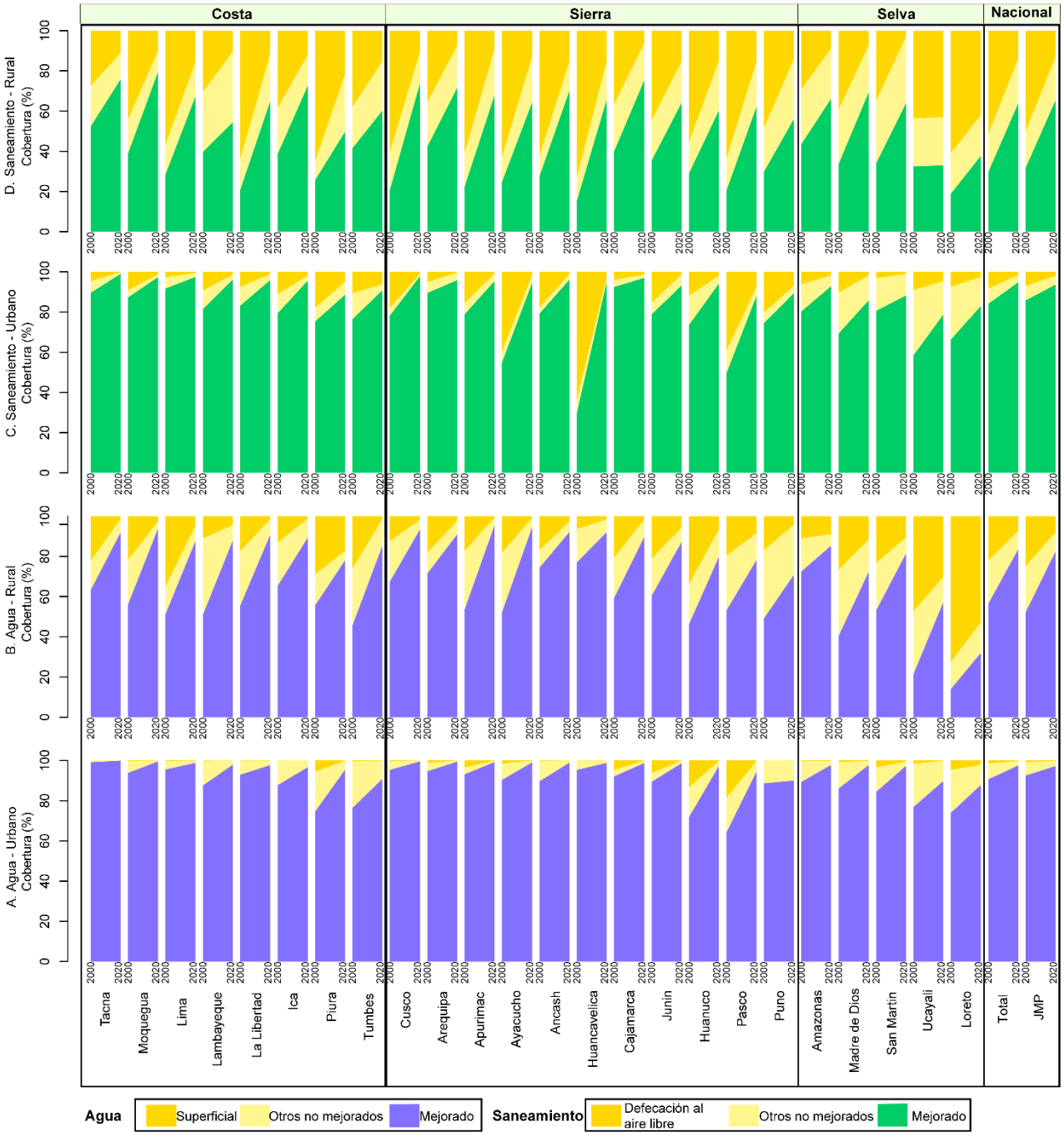


Figura 6.6. Escaleras de servicios AyS en el nivel subnacional.

Notas: El ajuste del modelo en cada departamento ha sido con GAM (ilr) u OLS (ilr) sobre datos preprocesados.

Respecto a los servicios de agua, nuestros resultados nos permiten afirmar que los hogares

urbanos han avanzado más que los rurales en la prestación de servicios mejorados de agua potable y, en consecuencia, el acceso a las aguas superficiales se ha reducido drásticamente, llegando al umbral más bajo de valor cero. Esta diferencia se acentúa más al comparar la calidad del agua en la dotación de servicios (Miranda et al., 2010; Instituto Nacional de Estadística e Informática, 2019; Quispe-Coica et al., 2020).

En cuanto al saneamiento urbano (Figura 6.6C), en general, la proporción de hogares con acceso a un saneamiento mejorado aumentó del 84.0% en 2000 al 94.7% en 2020. Este aumento hizo que tanto la defecación al aire libre como otros tipos de saneamiento no mejorados disminuyeran de 2000 a 2020. La defecación al aire libre se redujo del 8.4% al 1.6%, y los otros servicios de saneamiento no mejorados del 7.6% al 3.7%.

A nivel subnacional, el departamento de Huancavelica es el que experimentó el mayor incremento en saneamiento mejorado, pasando de 29.6% (28.7 – 30.6) en el 2000 a 95.1% (94.9–95.2) en el 2020; lo que representa una tasa de avance promedio anual de 3.3%. En consecuencia, la defecación al aire libre tuvo una drástica reducción del 63.6% (62.5 – 64.7) en 2000 al 2.6% (2.5 – 2.6) en 2020 y el acceso al saneamiento por otras formas no mejoradas tuvo una ligera reducción del 6.8% (6.5 – 7.1) en 2000 al 2.4% (2.3 – 2.5) en 2020. Por otro lado, 4 de los 24 departamentos - a saber: Ucayali, Loreto, Madre de Dios y San Martín; que a su vez pertenecen a la región de la selva- tienen la mayor cobertura de otro saneamiento no mejorado, con valores que oscilan en el rango de 10% a 17%. Esto significa que aún existen numerosos hogares que cuentan con tecnologías de saneamiento como letrinas y otras alternativas de saneamiento no mejorado.

En cuanto al saneamiento rural (Figura 6.6D), en general, el saneamiento mejorado ha pasado del 29.5% en 2000 al 63.6% en 2020. Sin embargo, este ritmo de avance no es suficiente para la cobertura universal de instalaciones de saneamiento mejoradas en los próximos años. El valor obtenido en 2020 sigue siendo bajo en comparación con el resto de los servicios evaluados (agua potable y saneamiento urbano). Por lo tanto, el país está obligado a realizar un mayor esfuerzo en saneamiento rural, más aún si la meta es cerrar la brecha de saneamiento al 2030 (Decreto supremo N° 007-2017-VIVIENDA, 2017).

A nivel subnacional, en el año 2020, se ha encontrado que en 6 de los 24 departamentos (Moquegua, Tacna, Cajamarca, Cusco, Ica y Arequipa), el acceso a saneamiento mejorado ha alcanzado valores que oscilan entre el 70.8% y el 80%. De los cuales Moquegua y Tacna son los que tienen mayor cobertura, siendo ésta de 79.3% (78.9 – 79.8) y 75.7% (75.0 – 76.4), respectivamente. Para el mismo año y el mismo indicador, el valor de los 13 departamentos restantes oscila entre 59% y 70.3%. Y finalmente, el acceso a saneamiento mejorado en los últimos cinco departamentos (Puno, Lambayeque, Piura, Loreto y Ucayali) oscila entre 32% y 57%. De los cuales, Loreto y Ucayali tienen la tasa más baja, siendo ésta de 37.6% (36.7 – 38.5) y 33.0% (32.4 – 33.7), respectivamente. También en ambos departamentos, la defecación al aire libre es la más alta, siendo ésta de 42.0% (41.1 – 42.9) y 43.0% (42.3 – 43.7), respectivamente. Una situación particular es la de Ucayali, donde las tres categorías de servicio se han mantenido casi constantes.

A nivel general, los departamentos que requieren especial atención son Loreto, Ucayali, ya que tienen y seguirán teniendo valores altos en la categoría de defecación al aire libre y acceso a agua superficial. Curiosamente, en estos departamentos, otros estudios mostraron altos índices de enfermedades diarreicas (Yori et al., 2014), específicamente en tres comunidades rurales de Loreto, Santa Clara de Nanay, Santo Tomás y La Unión. En la literatura ya se ha evidenciado la relación positiva entre las enfermedades diarreicas y el acceso a aguas superficiales no mejoradas, por lo que es muy probable que las altas enfermedades diarreicas

en Loreto estén relacionadas con los altos valores de defecación al aire libre y el alto valor de acceso a aguas superficiales.

6.5. Discusión

El desglose de la información armonizada con el marco de seguimiento global no es una novedad. La base de datos de desigualdad del PCM también incluye estimaciones de cobertura a nivel subnacional (véase https://washdata.org/). Sin embargo, las estimaciones se siguen realizando mediante OLS y no se subdividen en urbano y rural. La principal justificación para seguir utilizando el método OLS es que, i) "para muchos países, y especialmente para los indicadores de nivel de servicios, no hay suficientes puntos de datos que justifiquen el uso de métodos no lineales" y ii) "las técnicas no lineales también están limitadas en su capacidad de extrapolar incluso unos pocos años, lo que suele ser necesario para el método de estimación del PCM" (WHO/UNICEF, 2018).

En este sentido, en este estudio, mostramos que hay países como Perú que tienen suficiente información (que proviene de múltiples fuentes de información) a nivel general y subnacional donde el número de puntos de datos no es limitante para aplicar métodos no lineales como el GAM. Esto se justifica aún más, dado que a nivel subnacional existen patrones de curvatura muy diversos. P. ej., la curva en forma de campana ocurrió en 5/24 departamentos de saneamiento rural en la categoría X_3, lo que se ha reflejado en la agregación global de X_3 (véase la Figura 6.5B). El aumento de la curva en forma de campana se debe a la instalación inicial de otras opciones de saneamiento no mejoradas (letrinas no mejoradas, pozos negros, etc.) donde se realizaba la defecación al aire libre, que con el tiempo han sido sustituidas por infraestructura de saneamiento mejorada y, como consecuencia, la curva desciende y adquiere forma de campana. Estos patrones de comportamiento encontrados a nivel subnacional también se han manifestado en un análisis global de la literatura del sector (Bartram et al., 2014; Fuller et al., 2016).

Por otro lado, si se sigue con la misma metodología de estimación del PCM es probable que el modelo subestime o sobreestime la información del año; con probables repercusiones en la planificación de la inversión sectorial, la asignación eficiente de recursos y las políticas públicas subestatales acordes con las realidades subnacionales. Por tanto, es necesario migrar hacia técnicas estadísticas como las que planteamos en este estudio con tres enfoques principales, i) que tomen en cuenta las características compositivas de los datos (Pérez-Foguet et al., 2017), ii) que aborden la no linealidad (Fuller et al., 2016) e iii) incorporen la incertidumbre de los datos en los modelos (Ezbakhe y Pérez-Foguet, 2019). El preprocesamiento de la data (Capítulo IV) es un proceso inherente que tiene que ser implementado en cada análisis estadístico, cuestión que hasta el momento no se está ejecutando. Hemos detallado todo este procedimiento en este estudio con el fin de hacerlo replicable en un contexto más amplio.

La principal desventaja de la alternativa estadística presentada en este estudio para su aplicación al seguimiento subnacional es su complejidad para un público general. Sin embargo, a nivel de la región de ALC, se ha visto que son las instituciones estadísticas de cada país las que están implementando el monitoreo de los ODS del nivel subnacional. Por lo tanto, es previsible que cuenten con profesionales de la estadística, lo que ayudará a que la metodología propuesta sea replicada, en cada país, en contextos más amplios.

La calidad de la información es también otro aspecto a tener en cuenta para un buen ajuste del modelo. Para comprobarlo, aplicamos la métrica de calidad RMSE (acrónimo en inglés de root mean square error) a los datos brutos de las tres encuestas. Para la comparación entre el estimador y el valor observado, se toma como estimador el valor medio de las tres encuestas.

Un valor de RMSE cero significa que los tres puntos de datos del año son iguales. Los valores cercanos a cero significan que la dispersión del valor medio es baja y, por tanto, las tres fuentes de datos tienen valores casi iguales. Un valor de RMSE alto significa que para ese año hay una dispersión significativa de los datos con respecto al valor medio de las tres fuentes de información.

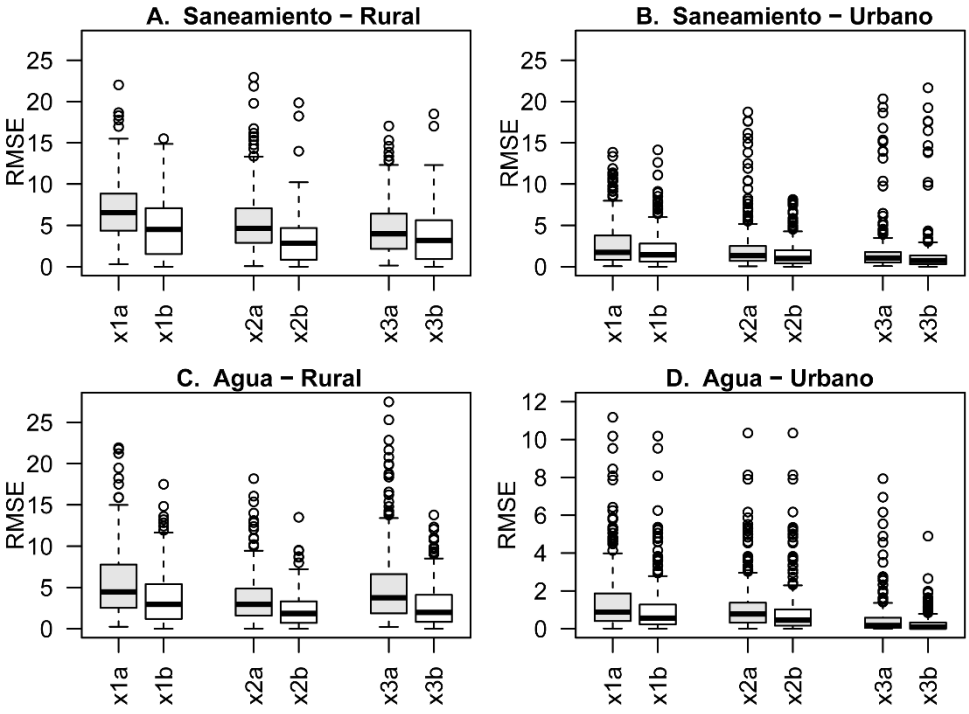


Figura 6.7. Dispersión de los datos

Notas: El gráfico de cajas en gris representa el valor de RMSE en los datos brutos (X1a, X2a y X3a). El gráfico de cajas en blanco representa el valor de RMSE en los datos validados (X1b, X2b y X3b).

El resultado del RMSE para cada año y para todos los departamentos se muestra en la Figura 6.7. El valor de la mediana de rural es mayor que el de urbano, lo que significa que existe una mayor dispersión de datos en el sector rural que en el urbano en cualquiera de las categorías de servicios. Tras la validación y exclusión de los puntos de datos atípicos mediante el método de MD robusto, fue posible reducir el valor de la mediana de los datos rurales y, aun así, siguen presentando una mayor dispersión que los datos urbanos.

Integrar parte de la información del DATASS (Diagnóstico de Abastecimiento de Agua y Saneamiento en Áreas Rurales; https://datass.vivienda.gob.pe/) en el sistema de monitoreo ayudaría a mejorar la calidad de la información rural, ya que tiene un tamaño de muestra que refleja casi el total del país rural. Por ejemplo, en Amazonas, el DATASS cuenta con información de 83,097 hogares, mientras que el tamaño de muestra predefinido para la encuesta ENDES es de sólo 940 hogares en 2020. La principal desventaja es que sólo se puede utilizar una parte de la información del DATASS, ya que no está diseñada para monitorear bajo el marco metodológico de las escaleras de agua y saneamiento.

Finalmente, aterrizar el marco de monitoreo global al nivel subnacional será cada vez más relevante, principalmente en países donde la ejecución de las inversiones en infraestructura de agua y saneamiento es compartida entre el gobierno nacional, los gobiernos regionales y locales, como es el caso de Perú. Un aspecto importante que el monitoreo global rara vez considera (Herrera, 2019). Por ejemplo, en el 2019 la ejecución de la inversión en agua y saneamiento ha sido de 22.3% (208,570,586 USD) en los gobiernos nacionales, 6.2% (57,974,932 USD) en los gobiernos regionales y 71.5% (669,470,499 USD) en los gobiernos

locales (MEF, 2021). Estos resultados nos muestran que todo lo que se haga en los gobiernos subnacionales y principalmente en los gobiernos locales impactará en los objetivos y metas del país. En consecuencia, nuestra propuesta abre el camino para ser replicada también a nivel local.

6.6. Mensajes claves

En este estudio hemos demostrado que la aplicación de una metodología de monitoreo subnacional según las características de los datos es relevante y puede servir de referencia para los países de ALC y las organizaciones regionales que están vinculados con el monitoreo del acceso a los servicios de AyS. Entre ellos, el SIASAR (Sistema de Información de Agua y Saneamiento Rural) y el observatorio para América Latina y el Caribe de Agua y Saneamiento (OLAS).

La propuesta, que considera métodos de interpolación lineal y no lineal sobre datos transformados, ofrece resultados válidos y proporciona una mejor manera de realizar un seguimiento subnacional (de abajo hacia arriba) en línea con el marco de seguimiento de los ODS. Se detectan casos de variables departamentales con valores muy cercanos a cero o a uno, principalmente en las tendencias temporales del agua urbana. En estas circunstancias, la interpolación y extrapolación de datos mediante técnicas estadísticas para datos de composición funciona bien y presenta ventajas sobre las técnicas estadísticas estándar.

También se ha demostrado que se pueden formar escaleras de AyS con información desglosada a nivel subnacional tanto para las zonas urbanas como para las rurales. Esto permitirá a los gobiernos subestatales disponer de información de primera mano que les ayude a tomar las mejores decisiones en el sector y también permitirá realizar intervenciones específicas para cada departamento.

En cifras generales, nuestros modelos estimaron para 2020 en agua urbana que el 97.5% de los hogares en Perú tienen acceso a servicios de agua mejorados (X_1), el 2.4% acceden a otros servicios de agua no mejorados (X_2) y sólo el 0.1% de los hogares siguen teniendo acceso a aguas superficiales (X_3). Mientras que, en el agua rural, el 83.4% de los hogares tiene acceso a servicios de agua mejorados, el 9.4% accede a X_2 y el 7.1% de los hogares sigue teniendo acceso a aguas superficiales. En saneamiento urbano y rural, el 94.7% de los hogares urbanos y el 63.6% de rural tienen acceso a servicios de saneamiento mejorados, el 3.7% de urbano y 22.4% de rural accede a otros servicios de saneamiento no mejorados y el 1.6% de urbano y el 14% de rural sigue defecando al aire libre. Estas cifras coinciden con las cifras globales estimadas por el PCM.

En el nivel subnacional, los departamentos de Loreto y Ucayali son los que han presentado las menores tasas de cobertura de servicios mejorados de agua y saneamiento tanto en los hogares urbanos como rurales. En cuanto al acceso al agua, los hogares rurales y urbanos tienen valores de 31.9% y 87.8% para Loreto y de 57.2% y 89.8% para Ucayali, respectivamente. Mientras que, en saneamiento, los hogares rurales y urbanos han alcanzado valores de 37.6% y 82.9% para Loreto y de 33.0% y 78.7% para Ucayali, respectivamente. El resto de departamentos han tenido comportamientos heterogéneos.

Asimismo, existe una distribución clara y uniforme de la desigualdad entre el acceso a los servicios mejorados de agua y saneamiento entre los hogares urbanos y los rurales. La población de los hogares urbanos es la que disfruta de los mejores servicios a través de la red pública y otras formas mejoradas de acceso al agua. La desigualdad es más significativa en el servicio de saneamiento, en el que los hogares rurales tienen la menor cobertura de saneamiento mejorado (es decir, por red pública, fosa séptica, letrinas mejoradas y otras

tecnologías de saneamiento mejorado). Por lo tanto, es necesario tener en cuenta los valores obtenidos en este estudio para las inversiones en infraestructura de agua y saneamiento enfocadas al cierre de brechas.

Finalmente, la información desagregada ha demostrado que existen patrones de curvatura heterogéneos (lineal, no lineal y de campana) que las tendencias temporales en los datos agregados no pueden capturar de manera eficiente. Por tanto, la propuesta de este estudio es útil para modelos en datos desagregados armonizados con los indicadores globales y permite mejorar su desempeño en la obtención de resultados más precisos.

CAPÍTULO VII. CONCLUSIONES

El contenido de esta tesis se ha centrado en demostrar cómo las nuevas herramientas estadísticas, aplicadas correctamente, pueden mejorar considerablemente los resultados del monitoreo de la población que accede a los diferentes niveles de servicio ASH y del monitoreo de la desigualdad; una mayor precisión en los resultados implica inversiones focalizadas, asignación eficiente de recursos, planes y políticas sectoriales más efectivas y acordes con la realidad del país.

Por lo tanto, esta tesis se ha dividido en cuatro partes: Característica de la data y correlación espuria (Capítulo III); preprocesamiento de datos irregulares y alternativas robustas de regresión (Capítulo IV); método alternativo de monitoreo de la desigualdad urbano-rural post 2015 (Capítulo V); y marco de monitoreo subnacional de los servicios de agua y saneamiento (Capítulo VI).

En general, esta tesis ofrece la consolidación de métodos — que partió con la propuesta de Pérez-Foguet et al. (2017) quienes introducen las técnicas estadísticas de CoDa al sector ASH y continuada por Ezbakhe y Pérez-Foguet (2019) quienes incorporan la incertidumbre de la data en el ajuste de modelos— para análisis de datos compositivos del sector ASH. También ofrece una nueva medida de desigualdad en Higiene aplicada a la comparación de acceso a niveles de servicios entre la zona de residencia urbana y rural. También proporcionamos un conjunto de casos de estudio y sus cálculos computacionales en R en cada capítulo de la tesis, con el fin de hacerla replicable en contextos más amplios. Por último, recomendamos algunas investigaciones que esta tesis no ha abordado.

7.1. Conclusiones principales

La principal conclusión de esta tesis es **que es necesario aplicar técnicas estadísticas acorde con la característica de los datos para el preprocesamiento, el ajuste del modelo y la medición de la desigualdad**. Se ha comprobado en esta tesis que la característica de los datos es múltiple, de entre ellos, los datos con valor cero, los datos con valor perdido y ambos simultáneamente (Capítulo III).

Hemos encontrado que **la cantidad de información con el que cuentan los países ha aumentado significativamente desde inicios del monitoreo de los ODM, principalmente en los países de las regiones del África subsahariana y de América Latina y el Caribe**. A nivel global, en agua, de las tres categorías analizadas tanto en urbano como en rural, el mínimo porcentaje de países con puntos de datos ≥ 6 es del 45% y esto se presenta en la categoría X_3 de agua urbano. En saneamiento es de 45.5% y esto se presenta en la categoría X_3 de saneamiento urbano (Capítulo III). Por lo tanto, **la hipótesis del PCM, sobre la no aplicación de métodos no lineales debido a la poca cantidad de datos, va perdiendo cada vez más consistencia.**

Se ha comprobado, bajo un supuesto preestablecido, **que existe una correlación espuria al aplicar la estadística habitual a datos compositivos** (Capítulo III), por lo que es imprescindible aplicar técnicas estadísticas adecuadas para el preprocesamiento de los datos (Capítulo IV), para el ajuste de los modelos a nivel global y subnacional (Capítulo IV y VI) y para el monitoreo de la desigualdad en el acceso a los servicios de ASH (capítulo V).

Es preocupante el alto porcentaje de valores perdidos que existen, en la información global, en al menos una de las categorías de análisis (Capítulo I). Es aún más preocupante en Higiene donde hay poca información, por lo que no hemos podido calcular la desigualdad entre el nivel de servicio urbano y rural para todos los países (Capítulo V). Para estos escenarios, en el Capítulo IV hemos testeado y validado alternativas de tratamiento de datos irregulares, de entre ellos los perdidos. Con esto se espera que los valores perdidos y otros datos irregulares del sector ya no sean un problema para aplicar cualquier transformación log-cociente (principio elemental del análisis estadístico para datos compositivos).

También en el Capítulo IV se ha **demostrado que es posible aplicar técnicas estadísticas robustas para todos los casos (puntos de datos < 6 o ≥ 6)**. Hacerlo o no puede tener implicaciones en el ajuste del modelo, principalmente debido a la influencia de los valores atípicos. Por lo tanto, se recomienda pasar a métodos de estimación robustos.

La novedad del Capítulo V es que se obtiene una medida integrada de desigualdad urbana-rural en las escaleras de higiene teniendo en cuenta la característica compositiva de los datos. Hasta donde sabemos, o al menos en el sector ASH, esta forma de calcular la desigualdad aún no se ha estudiado. Por lo tanto, la propuesta metodológica que realizamos es la más novedosa de toda la tesis doctoral, lo cual validamos con toda la información existente en el monitoreo global de Higiene.

La nueva medida de la desigualdad se complementa con la discretización del diagrama ternario en parcelas, lo que añade valor agregado a la novedad. Los resultados que obtuvimos nos permiten agrupar países con información tripartita, ya sea para hacer un ranking de la desigualdad o un ranking de acceso a servicios ASH o para expresar en un mapa temático el grupo de países capturados en las parcelas ternarias. Estos también son un valor agregado en la originalidad de la investigación, ya que en un ranking de acceso a los servicios ASH suele hacerse de forma univariante sin tomar en consideración que —en los datos de composición— un cambio en una parte también afectará al resto y también es común ver mapas temáticos univariados.

La propuesta realizada puede extenderse a contextos más amplios como, por ejemplo, el monitoreo de la desigualdad en: los niveles de servicio de AyS, ASH en instalaciones sanitarias, ASH en las escuelas, etc. También puede utilizarse para comparar el progreso de los países intra urbano o intra rural a través de medidas de distancias.

Finalmente, en el capítulo 6 se integra dos propuestas complementarias de análisis de datos para el sector ASH, el algoritmo propuesto en el capítulo 4 de esta tesis y el cálculo de la incertidumbre de Ezbakhe y Pérez-Foguet (2019). Se testea en nivel subnacional de Perú, en específico en la evolución temporal de los servicios de AyS en los 24 departamentos tanto en contextos urbanos como rurales generándose así 96 modelos (48 en agua y 48 en saneamiento). **Se demuestra la aplicabilidad del algoritmo integrado de análisis de datos para el monitoreo de la población al acceso de los servicios de AyS**. Los resultados muestran (cuando se tenga información) que es necesario un análisis en el nivel desagregado, ya que existe un efecto de enmascaramiento de la tendencia general sobre la tendencia del nivel subnacional. Como resultado, la evolución temporal del servicio de algunos departamentos no se refleja en la realidad global del país; con probables consecuencias en una mala planificación en las inversiones del sector, políticas públicas deficientes (dado que las políticas públicas comúnmente tienen una visión general de arriba abajo) y más.

Otra conclusión importante es que **se ha logrado armonizar y aterrizar los indicadores globales al nivel subnacional**, un problema general que aún persiste en los países de ALC. Es aún más relevante en países como Perú que cuentan con datos significativos y que comparten las responsabilidades de la gestión de AyS entre el gobierno nacional y los gobiernos subnacionales (formados por gobiernos regionales, provinciales y distritales).

Por último, se espera que la i) propuesta del algoritmo que integra modelos para una amplia gama de datos lineales y no lineales, con los valores atípicos incluidos y ii) la nueva medida de desigualdad que se propone para la información tripartita, contribuya a mejorar el análisis de datos en el sector y especialmente en aquellos que cuentan con múltiples fuentes de información. Este trabajo se complementa perfectamente con la propuesta realizada por Pérez-Foguet et al. (2017) y continuada por Ezbakhe y Pérez-Foguet (2019), sobre el análisis estadístico para CoDa en el sector ASH.

7.2. Limitaciones

Basándonos en nuestros resultados, hemos identificado dos limitaciones principales de nuestro estudio:

Por un lado, reconocemos el reducido número de contextos estudiados para testear los métodos de imputación de datos irregulares. Creemos que esto puede ampliarse a más países con la finalidad de tener afirmaciones más concluyentes.

En segundo lugar, existe información limitada de los datos de Higiene, lo que no nos ha permitido concluir sobre la situación global. Sin embargo, a medida que aumente la cantidad de información, esto ya será posible. Asimismo, nuestra propuesta de clasificación ternaria no tiene en cuenta las incertidumbres en los datos. Por lo tanto, es probable que, tras añadir la incertidumbre de los datos, la clasificación de un país cambie con respecto a su clasificación actual.

7.3. Investigaciones futuras

Sobre la base de nuestros resultados, recomendamos las siguientes investigaciones futuras que deberían llevarse a cabo:

- En este estudio hemos encontrado que hay una falta de información, en una de las partes del total, en los datos de ASH. Esta situación representa una oportunidad para testear y validar otras alternativas estadísticas para completar los datos faltantes además de los ya realizados en esta investigación. Esto será relevante en países con datos limitados, ya que aumentará la disponibilidad de puntos de datos para el ajuste del modelo calibrado.
- También hemos introducido el ajuste lineal robusto con modulación de valores atípicos acoplados al modelo (este tipo de análisis no es muy común en el sector ASH) y un modelo robusto no lineal (robustez basada en la identificación y exclusión de valores atípicos). Se ha comprobado que las estimaciones del modelo son resultados más fiables. Por tanto, dada las ventajas que ofrecen los modelos robustos, es necesario evaluar y validar en los datos de ASH otras alternativas estadísticas robustas. Se debe poner mayor énfasis en modelos no lineales como GAM con modulación de valores atípicos acoplados (Wong et al., 2014), más detalles en Maronna et al. (2019).
- La caracterización de la incertidumbre es también otro tema que debe ser abordado, ya que se ha visto que en casos particulares conduce a caracterizaciones inapropiadas (ver Figura 6.4 del Capítulo VI).
- Actualmente se realiza el análisis de la población urbano y rural que acceden a los diferentes niveles de servicio de manera independiente, a pesar de que estos son composicionales. Es decir, suman la población total del país. En este estudio también nos hemos centrado en el análisis desagregado. Por tanto, futuras investigaciones también deberán evaluar el juste de modelo acoplado entre urbano y rural.
- Por último, es posible explorar y medir la desigualdad urbano-rural de los datos de composición en el diagrama ternario cuando son datos tripartitos. Cuando son de cuatro partes, también se puede calcular en el tetraedro regular, tema que no se aborda en este estudio. Cuando se trata de cinco o más partes, es necesario realizar primero transformaciones log-cociente para aplicar cualquier técnica estadística habitual, tema que tampoco se aborda en esta investigación. Por lo tanto, las investigaciones futuras deberían abordar nuevas metodologías para medir la desigualdad urbano-rural cuando los datos de composición tienen cuatro o más partes. Recomendamos que se haga mayor énfasis en las medidas de desigualdad de cinco partes, ya que los ODS 6.1 y 6.2 tienen cinco niveles de servicio en las escaleras de agua y saneamiento.

REFERENCIAS BIBLIOGRÁFICAS

Abboud, E., 2010. Viviani's theorem and its extension. Coll. Math. J. 41, 203-211. https://doi.org/10.4169/074683410X488683

Adedokun, K.A., Olarinmoye, A.O., Olarinmoye, A.O., Mustapha, J.O., Kamorudeen, R.T., 2020. A close look at the biology of SARS-CoV-2, and the potential influence of weather conditions and seasons on COVID-19 case spread. Infect. Dis. Poverty 9, 1-5. https://doi.org/10.1186/s40249-020-00688-1

Ahmad, K., Erqou, S., Shah, N., Nazir, U., Morrison, A.R., Choudhary, G., Wu, W.-C., 2020. Association of poor housing conditions with COVID-19 incidence and mortality across US counties. PLoS One 15, e0241327. https://doi.org/10.1371/journal.pone.0241327

Aitchison, John, 1986. The statistical analysis of compositional data (Monographs on Statistics and Applied Probability), 1st ed. Chapman & Hall, Ltd., London, United Kingdom.

Aitchison, J., 1986. Irregular compositional data, en: The Statistical Analysis of Compositional Data. Springer Netherlands, Dordrecht, pp. 256-280. https://doi.org/10.1007/978-94-009-4109-0_11

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B 44, 139-160.

Aitchison, J., Kay, J.W., 2003. Possible solutions of some essential zero problems in compositional data analysis, en: In Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop. University of Girona, Girona (Spain).

ANDINA, 2021. Perú es el primer país en el mundo en sincerar sus cifras de fallecidos por covid-19. URL https://andina.pe/agencia/noticia-peru-es-primer-pais-el-mundo-sincerar-sus-cifras-fallecidos-covid19-847490.aspx (accedido 6.2.21).

Anthonj, C., Setty, K.E., Ezbakhe, F., Manga, M., Hoeser, C., 2020a. A systematic review of water, sanitation and hygiene among Roma communities in Europe: Situation analysis, cultural context, and obstacles to improvement. Int. J. Hyg. Environ. Health. https://doi.org/10.1016/j.ijheh.2020.113506

Anthonj, C., Tracy, J.W., Fleming, L., Shields, K.F., Tikoisuva, W.M., Kelly, E., Thakkar, M.B., Cronk, R., Overmars, M., Bartram, J., 2020b. Geographical inequalities in drinking water in the Solomon Islands. Sci. Total Environ. 712, 135241. https://doi.org/10.1016/j.scitotenv.2019.135241

Ashole Alto, A., Godana, W., Gedamu, G., 2020. Impact of Community-Led Total Sanitation and Hygiene on Prevalence of Diarrheal Disease and Associated Factors among Under-Five Children: A Comparative Cross-Sectional Study in Selected Woredas of Gamo Gofa Zone, Southern Ethiopia. Adv. Public Heal. 2020.

Bain, R., Johnston, R., Mitis, F., Chatterley, C., Slaymaker, T., 2018. Establishing Sustainable Development Goal Baselines for Household Drinking Water, Sanitation and Hygiene Services. Water 10, 1711. https://doi.org/10.3390/w10121711

Bain, R., Wright, J.A., Christenson, E., Bartram, J.K., 2014. Rural:urban inequalities in post 2015 targets and indicators for drinking-water. Sci. Total Environ. 490, 509-513. https://doi.org/10.1016/j.scitotenv.2014.05.007

Bain, R.E., Gundry, S.W., Wright, J.A., Yang, H., Pedley, S., Bartram, J.K., 2012. Accounting for water quality in monitoring access to safe drinking-water as part of the Millennium Development Goals: lessons from five countries. Bull World Heal. Organ 90, 228-235. https://doi.org/10.2471/BLT.11.094284

Baquero, Ó.F., Jiménez Fdez. de Palencia, A., Pérez-Foguet, A., 2015. Reporting progress on

the human right to water and sanitation through JMP and GLAAS. J. Water Sanit. Hyg. Dev. 5, 310-321. https://doi.org/10.2166/washdev.2015.151

Bartl, A., 2014. Moving from recycling to waste prevention: A review of barriers and enables. Waste Manag. Res. https://doi.org/10.1177/0734242X14541986

Bartram, J., Brocklehurst, C., Fisher, M., Luyendijk, R., Hossain, R., Wardlaw, T., Gordon, B., 2014. Global Monitoring of Water Supply and Sanitation: History, Methods and Future Challenges. Int. J. Environ. Res. Public Health 11, 8137-8165. https://doi.org/10.3390/ijerph110808137

BBC, 2013. El fracaso del «mejor censo en la historia de Chile» - BBC News Mundo. URL https://www.bbc.com/mundo/noticias/2013/08/130809_chile_problemas_del_censo_2012_ng (accedido 9.30.21).

Boogaart, K.G. van den, Tolosana-Delgado, R., Bren, M., 2019. compositions: Compositional Data Analysis.

Brauer, M., Zhao, J.T., Bennitt, F.B., Stanaway, J.D., 2020. Global Access to Handwashing: Implications for COVID-19 Control in Low-Income Countries. Environ. Health Perspect. 128, 057005. https://doi.org/10.1289/EHP7200

Cairncross, S., Hunt, C., Boisson, S., Bostoen, K., Curtis, V., Fung, I.C.C.H., Schmidt, W.-P., 2010. Water, sanitation and hygiene for the prevention of diarrhoea. Int. J. Epidemiol. 39, i193-i205. https://doi.org/10.1093/ije/dyq035

CDC, 2019. Older Adults and COVID-19. URL https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html (accedido 4.17.21).

Cetrulo, T.B., Marques, R.C., Malheiros, T.F., Cetrulo, N.M., 2020. Monitoring inequality in water access: Challenges for the 2030 Agenda for Sustainable Development. Sci. Total Environ. 727, 138746. https://doi.org/10.1016/j.scitotenv.2020.138746

Chen, J., Zhang, X., Hron, K., Templ, M., Li, S., 2018. Regression imputation with Q-mode clustering for rounded zero replacement in high-dimensional compositional data. J. Appl. Stat. 45, 2067-2080. https://doi.org/10.1080/02664763.2017.1410524

Chitonge, H., Mokoena, A., Kongo, M., 2020. Water and Sanitation Inequality in Africa: Challenges for SDG 6. Springer, Cham, pp. 207-218. https://doi.org/10.1007/978-3-030-14857-7_20

Consejo Nacional de Coordinación de Políticas Sociales, 2021. Argentina: Objetivos de Desarrollo Sostenible, Metas priorizadas e Indicadores de seguimiento. Ciudad Autónoma de Buenos Aires.

Craven, J., Giné-Garriga, R., Jiménez Fdez. de Palencia, A., Pérez-Foguet, A., 2013. Introducing hygiene elements into sanitation monitoring. Loughborough University.

Curtis, N., Sparrow, A., Ghebreyesus, T.A., Netea, M.G., 2020. Considering BCG vaccination to reduce the impact of COVID-19. Lancet. https://doi.org/10.1016/S0140-6736(20)31025-4

Curtis, V., Cairncross, S., 2003. Effect of washing hands with soap on diarrhoea risk in the community: A systematic review. Lancet Infect. Dis. https://doi.org/10.1016/S1473-3099(03)00606-6

Daunis-i-Estadella, J., Thió-Henestrosa, S., Mateu-Figueras, G., 2011. Including supplementary elements in a compositional biplot. Comput. Geosci. 37, 696-701. https://doi.org/10.1016/j.cageo.2010.11.003

Decreto supremo N° 007-2017-VIVIENDA, 2017. Decreto Supremo que aprueba la Política Nacional de Saneamiento. D. Of. del Bicenten. "El Peru. URL https://busquedas.elperuano.pe/normaslegales/decreto-supremo-que-aprueba-la-politica-

nacional-de-saneamie-decreto-supremo-n-007-2017-vivienda-1503314-7/ (accedido 6.11.20).

Dumuid, D., Pedišić, Ž., Palarea-Albaladejo, J., Martín-Fernández, J.A., Hron, K., Olds, T., 2020. Compositional Data Analysis in Time-Use Epidemiology: What, Why, How. Int. J. Environ. Res. Public Health 17, 2220. https://doi.org/10.3390/ijerph17072220

Economic, U.N., Council, S., 2016. Report of the inter-agency and expert group on sustainable development goal indicators. E/CN.3/2016/2/Rev.1. Stat. Comm 13.

Egozcue, J.J., Martín-Fernández, J.A., Palarea-Albaladejo, J., Pawlowsky-Glahn, V., 2019. The statistical analysis of compositional data: irregular data. CoDaCourse, CoDaWork2019.

Egozcue, J.J., Pawlowsky-Glahn, V., 2011. Basic Concepts and Procedures, en: Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, Chichester, UK, pp. 12-28. https://doi.org/10.1002/9781119976462.ch2

Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. Geol. Soc. London, Spec. Publ. 264, 145-159. https://doi.org/10.1144/GSL.SP.2006.264.01.11

Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37, 795-828. https://doi.org/10.1007/s11004-005-7381-9

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric Logratio Transformations for Compositional Data Analysis. Math. Geol. 35, 279-300. https://doi.org/10.1023/A:1023818214614

Ezbakhe, F., Giné-Garriga, R., Pérez-Foguet, A., 2019. Leaving no one behind: Evaluating access to water, sanitation and hygiene for vulnerable and marginalized groups. Sci. Total Environ. 683, 537-546. https://doi.org/10.1016/j.scitotenv.2019.05.207

Ezbakhe, F., Pérez-Foguet, A., 2020. Child mortality levels and trends: A new compositional approach. Demogr. Res. 43, 1263-1296. https://doi.org/10.4054/DEMRES.2020.43.43

Ezbakhe, F., Pérez-Foguet, A., 2019. Estimating access to drinking water and sanitation: The need to account for uncertainty in trend analysis. Sci. Total Environ. 696, 133830. https://doi.org/10.1016/j.scitotenv.2019.133830

Ezbakhe, F., Pérez-Foguet, A., 2018. Multi-Criteria Decision Analysis Under Uncertainty: Two Approaches to Incorporating Data Uncertainty into Water, Sanitation and Hygiene Planning. Water Resour. Manag. 32, 5169-5182. https://doi.org/10.1007/s11269-018-2152-9

Filzmoser, P., Hron, K., 2008. Outlier Detection for Compositional Data Using Robust Methods. Math. Geosci. 40, 233-248. https://doi.org/10.1007/s11004-007-9141-5

Filzmoser, P., Hron, K., Reimann, C., 2012. Interpretation of multivariate outliers for compositional data. Comput. Geosci. 39, 77-85. https://doi.org/10.1016/J.CAGEO.2011.06.014

Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. Sci. Total Environ. 407, 6100-6108. https://doi.org/10.1016/J.SCITOTENV.2009.08.008

Filzmoser, P., Hron, K., Templ, M., 2018. Compositional Data as a Methodological Concept, en: Applied Compositional Data Analysis. Springer Nature, Switzerland, pp. 1-16. https://doi.org/10.1007/978-3-319-96422-5_1

Freeman, M.C., Garn, J. V., Sclar, G.D., Boisson, S., Medlicott, K., Alexander, K.T., Penakalapati, G., Anderson, D., Mahtani, A.G., Grimes, J.E.T., Rehfuess, E.A., Clasen, T.F., 2017. The impact of sanitation on infectious disease and nutritional status: A systematic review and meta-analysis. Int. J. Hyg. Environ. Health 220, 928-949. https://doi.org/10.1016/j.ijheh.2017.05.007

Fuller, J.A., Goldstick, J., Bartram, J., Eisenberg, J.N.S., 2016. Tracking progress towards global drinking water and sanitation targets: A within and among country analysis. Sci. Total Environ. 541, 857-864. https://doi.org/10.1016/j.scitotenv.2015.09.130

GESTIÓN, 2021. COVID-19: tasa de letalidad se dispara de 3.5% a 9.4% tras sinceramiento de número de muertos. URL https://gestion.pe/peru/covid-19-en-peru-tasa-de-letalidad-se-dispara-de-35-a-94-tras-sinceramiento-de-cifra-de-muertos-nndc-noticia/ (accedido 6.2.21).

Giné-Garriga, R., Flores-Baquero, Ó., Jiménez Fdez. de Palencia, A., Pérez-Foguet, A., 2017. Monitoring sanitation and hygiene in the 2030 Agenda for Sustainable Development: A review through the lens of human rights. Sci. Total Environ. 580, 1108-1119. https://doi.org/10.1016/j.scitotenv.2016.12.066

Giné-Garriga, R., Pérez-Foguet, A., 2019. Monitoring and targeting the sanitation poor: A multidimensional approach. Nat. Resour. Forum 43, 82-94. https://doi.org/10.1111/1477-8947.12171

Giné-Garriga, R., Pérez-Foguet, A., 2013a. Water, sanitation, hygiene and rural poverty: issues of sector monitoring and the role of aggregated indicators. Water Policy 15, 1018-1045. https://doi.org/10.2166/wp.2013.037

Giné-Garriga, R., Pérez-Foguet, A., 2013b. Unravelling the Linkages Between Water, Sanitation, Hygiene and Rural Poverty: The WASH Poverty Index. Water Resour. Manag. 27, 1501-1515. https://doi.org/10.1007/s11269-012-0251-6

Giné-Garriga, R., Pérez-Foguet, A., 2010. Improved Method to Calculate a Water Poverty Index at Local Scale. J. Environ. Eng. 136, 1287-1298. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000255

Giné-Garriga, R., Requejo, D., Molina, J.L., Pérez-Foguet, A., 2018. A novel planning approach for the water, sanitation and hygiene (WaSH) sector: The use of object-oriented bayesian networks. Environ. Model. Softw. 103, 1-15. https://doi.org/10.1016/j.envsoft.2018.01.021

Giraudoux, P., 2021. pgirmess: Spatial Analysis and Data Mining for Field Ecologists. URL https://cran.r-project.org/package=pgirmess (accedido 6.13.21).

Graham, T.R., Gorniak, R., Dembowski, M., Zhang, X., Clark, S.B., Pearce, C.I., Clark, A.E., Rosso, K.M., 2020. Solid-State Recrystallization Pathways of Sodium Aluminate Hydroxy Hydrates. Inorg. Chem. 59, 6857-6865. https://doi.org/10.1021/acs.inorgchem.0c00258

Hasan, M.M., Alam, K., 2020. Inequality in access to improved drinking water sources and childhood diarrhoea in low- and middle-income countries. Int. J. Hyg. Environ. Health 226, 113493. https://doi.org/10.1016/j.ijheh.2020.113493

Heller, L., Mota, C.R., Greco, D.B., 2020. COVID-19 faecal-oral transmission: Are we asking the right questions? Sci. Total Environ. 729, 138919. https://doi.org/10.1016/j.scitotenv.2020.138919

Hernández-Vásquez, A., Bendezú-Quispe, G., Díaz-Seijas, D., Santero, M., Minckas, N., Azañedo, D., Antiporta, D.A., 2016. Spatial analysis of childhood obesity and overweight in Peru, 2014. Rev. Peru. Med. Exp. Salud Publica 33, 489-97. https://doi.org/10.17843/RPMESP.2016.333.2298

Herrera, V., 2019. Reconciling global aspirations and local realities: Challenges facing the Sustainable Development Goals for water and sanitation. World Dev. 118, 106-117. https://doi.org/10.1016/j.worlddev.2019.02.009

Herrera, V., Post, A.E., 2014. Can developing countries both decentralize and depoliticize urban water services? Evaluating the legacy of the 1990s reform wave. World Dev. 64, 621-641. https://doi.org/10.1016/j.worlddev.2014.06.026

Hirai, M., Roess, A., Huang, C., Graham, J.P., 2017. Exploring the link between handwashing proxy measures and child diarrhea in 25 countries in sub-Saharan Africa: A cross-sectional study. J. Water Sanit. Hyg. Dev. 7, 312-322. https://doi.org/10.2166/washdev.2017.126

Howard, G., Bartram, J., Brocklehurst, C., Colford, J.M., Costa, F., Cunliffe, D., Dreibelbis, R., Eisenberg, J.N.S., Evans, B., Girones, R., Hrudey, S., Willetts, J., Wright, C.Y., 2020. COVID-19: urgent actions, critical reflections and future relevance of 'WaSH': lessons for the current and future pandemics. J. Water, Sanit. Hyg. Dev. 10, 379-396. https://doi.org/10.2166/washdev.2020.218

Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional data using classical and robust methods. Comput. Stat. Data Anal. 54, 3095-3107. https://doi.org/10.1016/J.CSDA.2009.11.023

Hsiao, C., Shen, Y., Fujiki, H., 2005. Aggregate vs. disaggregate data analysis—a paradox in the estimation of a money demand function of Japan under the low interest rate policy. J. Appl. Econom. 20, 579-601. https://doi.org/10.1002/jae.806

Huang, Z., Huang, J., Gu, Q., Du, P., Liang, H., Dong, Q., 2020. Optimal temperature zone for the dispersal of COVID-19. Sci. Total Environ. 736, 139487. https://doi.org/10.1016/j.scitotenv.2020.139487

INEI, 2021. Informe Perú: Indicadores de Resultados de los Programas Presupuestales, 2020. Lima, Perú.

INEI, 2020. Evolución de la pobreza monetaria 2008-2019. Lima.

INEI, 2017. National Censuses: XII of Population, VII of Housing and III of Indigenous Communities, 2017. INEI. URL http://censo2017.inei.gob.pe/ (accedido 7.17.21).

INEI, 2009. Perú: Estimaciones y Proyecciones de Población Urbana y Rural por Sexo y Grupos Quinquenales de Edad, Según Departamentos, 2000- 2015. URL https://proyectos.inei.gob.pe/web/biblioineipub/bancopub/Est/Lib0842/libro.pdf (accedido 10.20.20).

Instituto Nacional de Estadística e Informática, 2019. Perú: Formas de acceso al agua y saneamiento básico. URL https://www.inei.gob.pe/media/MenuRecursivo/boletines/boletin_agua_nov2019.pdf (accedido 3.10.20).

JMP, 2021. Estimates on the use of water, sanitation and hygiene in Peru. URL https://washdata.org/ (accedido 7.13.21).

JMP, 2019. Joint Monitoring Programme for Water Supply, Sanitation, and Hygiene: Estimates on the use of water, sanitation and hygiene in South Africa. WHO/UNICEF. URL https://washdata.org/data (accedido 11.11.19).

Jones, R.M., 2020. Relative contributions of transmission routes for COVID-19 among healthcare personnel providing patient care. J. Occup. Environ. Hyg. 17, 1-8. https://doi.org/10.1080/15459624.2020.1784427

Koller, M., Stahel, W.A., 2011. Sharpening Wald-type inference in robust regression for small samples. Comput. Stat. Data Anal. 55, 2504-2515. https://doi.org/10.1016/j.csda.2011.02.014

Lalaoui, R., Bakour, S., Raoult, D., Verger, P., Sokhna, C., Devaux, C., Pradines, B., Rolain, J.M., 2020. What could explain the late emergence of COVID-19 in Africa? New Microbes New Infect. https://doi.org/10.1016/j.nmni.2020.100760

Lloyd, C.D., Pawlowsky-Glahn, V., Egozcue, J.J., 2012. Compositional Data Analysis in Population Studies. Ann. Assoc. Am. Geogr. 102, 1251-1266.

https://doi.org/10.1080/00045608.2011.652855

Luby, S.P., Agboatwalla, M., Feikin, D.R., Painter, J., Billhimer, W., Altaf, A., Hoekstra, R.M., 2005. Effect of handwashing on child health: A randomised controlled trial. Lancet 366, 225-233. https://doi.org/10.1016/S0140-6736(05)66912-7

Lumu, I., 2020. COVID-19 Response in Sub-Saharan Africa: Lessons from Uganda. Disaster Med. Public Health Prep. 14, e46-e48. https://doi.org/10.1017/dmp.2020.248

Ma, Q.-X., Shan, H., Zhang, H.-L., Li, G.-M., Yang, R.-M., Chen, J.-M., 2020. Potential utilities of mask-wearing and instant hand hygiene for fighting SARS-CoV-2. J. Med. Virol. https://doi.org/10.1002/jmv.25805

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E.L.T., Anna di Palma, M., 2019. robustbase: Basic Robust Statistics.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India 2, 49-55.

Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M., 2019. Robust statistics: theory and methods (with R). John Wiley & Sons.

Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J., 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat. Model. An Int. J. 15, 134-158. https://doi.org/10.1177/1471082X14535524

Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2003. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Math. Geol. 35, 253-278. https://doi.org/10.1023/A:1023866030544

Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J., 2012. Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. Comput. Stat. Data Anal. 56, 2688-2704. https://doi.org/10.1016/J.CSDA.2012.02.012

Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A., 2011. Dealing with Zeros, en: Compositional Data Analysis. John Wiley & Sons, Ltd, Chichester, UK, pp. 43-58. https://doi.org/10.1002/9781119976462.ch4

Mbow, M., Lell, B., Jochems, S.P., Cisse, B., Mboup, S., Dewals, B.G., Jaye, A., Dieye, A., Yazdanbakhsh, M., 2020. COVID-19 in Africa: Dampening the storm? Science (80-.). 369, 624-626.

MEF, 2021. Seguimiento de la ejecución presupuestal - Consulta Amigable | Gobierno del Perú. URL https://www.gob.pe/802-seguimiento-de-la-ejecucion-presupuestal-consulta-amigable (accedido 8.27.21).

Miller, A., Reandelar, M.J., Fasciglione, K., Roumenova, V., Li, Y., Otazu, G.H., 2020. Correlation between universal BCG vaccination policy and reduced mortality for COVID-19. medRxiv. https://doi.org/10.1101/2020.03.24.20042937

Miller, W.E., 2002. Revisiting the geometry of a ternary diagram with the half-taxi metric. Math. Geol. 34, 275-290. https://doi.org/10.1023/A:1014842906442

Miranda, M., Aramburú, A., Junco, J., Campos, M., 2010. State of the quality of drinking water in households in children under five years in Peru, 2007-2010. Rev. Peru. Med. Exp. Salud Publica 27, 506—511. https://doi.org/10.1590/s1726-46342010000400003

Molina-Vera, A., Pozo, M., Serrano, J., 2018. Agua, saneamiento e higiene: medición de los ODS en Ecuador. Quito-Ecuador.

Moriasi, D.N., Arnold, J.G., Liew, M.W. Van, Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Trans. ASABE 50, 885-900. https://doi.org/10.13031/2013.23153

Motovilov, Y.G., Gottschalk, L., Engeland, K., Rodhe, A., 1999. Validation of a distributed hydrological model against spatial observations. Agric. For. Meteorol. 98-99, 257-277. https://doi.org/10.1016/S0168-1923(99)00102-1

Nachega, J.B., Mbala-Kingebeni, P., Otshudiema, J., Mobula, L.M., Preiser, W., Kallay, O., Michaels-Strasser, S., Breman, J.G., Rimoin, A.W., Nsio, J., Ahuka-Mundeke, S., Zumla, A., Tam-Fum, J.J.M., 2020. Responding to the challenge of the dual Covid-19 and ebola epidemics in the democratic republic of congo'priorities for achieving control. Am. J. Trop. Med. Hyg. https://doi.org/10.4269/ajtmh.20-0642

Naciones Unidas, 2000. Resolución aprobada por la Asamblea General, A/RES/55/2. Declaración del milenio. URL https://www.un.org/spanish/milenio/ares552.pdf (accedido 9.22.21).

Naghavi, M., Abajobir, A.A., Abbafati, C., Abbas, K.M., et al., 2017. Global, regional, and national age-sex specifc mortality for 264 causes of death, 1980-2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet 390, 1151-1210. https://doi.org/10.1016/S0140-6736(17)32152-9

Nash, J.E.E., Sutcliffe, J.V. V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10, 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Neupert, R., 2017. Los censos y la falacia de la planificación: el caso de Chile. Rev. Latinoam. Población 11, 105-116. https://doi.org/10.31406/RELAP2017.V11.I1.N20.5

Nhamo, G., Nhemachena, C., Nhamo, S., 2019. Is 2030 too soon for Africa to achieve the water and sanitation sustainable development goal? Sci. Total Environ. 669, 129-139. https://doi.org/10.1016/j.scitotenv.2019.03.109

ODS-Paraguay, 2021. Segundo informe voluntario Paraguay 2021: Caminando juntos hacia un Paraguay más inclusivo, participativo y resilente. URL https://www.mre.gov.py/ods/wp-content/uploads/2021/07/Segundo-Informe-Nacional-Voluntario-Paraguay-2021.pdf (accedido 8.23.21).

OMS/UNICEF, 2015. Informe de actualización 2015 y evaluación del ODM. Progresos en Mater. Saneam. y agua potable 90.

Onda, K., LoBuglio, J., Bartram, J., 2012. Global Access to Safe Water: Accounting for Water Quality and the Resulting Impact on MDG Progress. Int. J. Environ. Res. Public Heal. 2012, Vol. 9, Pages 880-894 9, 880-894. https://doi.org/10.3390/IJERPH9030880

Pal, M., Berhanu, G., Desalegn, C., Kandi, V., 2020. Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): An Update. Cureus 12. https://doi.org/10.7759/cureus.7423

Palarea-Albaladejo, J., Martin-Fernandez, J.A., 2020. Treatment of Zeros, Left-Censored and Missing Values in Compositional Data Sets.

Palarea-Albaladejo, J., Martín-Fernández, J.A., 2015. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemom. Intell. Lab. Syst. 143, 85-96. https://doi.org/10.1016/J.CHEMOLAB.2015.02.019

Palarea-Albaladejo, J., Martín-Fernández, J.A., 2008. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Comput. Geosci. 34, 902-917. https://doi.org/10.1016/J.CAGEO.2007.09.015

Palarea-Albaladejo, J., Martín-Fernández, J.A., Gómez-García, J., 2007. A Parametric Approach

for Dealing with Compositional Rounded Zeros. Math. Geol. 39, 625-645. https://doi.org/10.1007/s11004-007-9100-1

Pandey, D., Verma, S., Verma, P., Mahanty, B., Dutta, K., Daverey, A., Arunachalam, K., 2021. SARS-CoV-2 in wastewater: Challenges for developing countries. Int. J. Hyg. Environ. Health. https://doi.org/10.1016/j.ijheh.2020.113634

Patel, S.K., Pradhan, M.R., Patel, S., 2020. Water, Sanitation, and Hygiene (WASH) Conditions and Their Association with Selected Diseases in Urban India. J. Popul. Soc. Stud. 28, 103-115. https://doi.org/10.25133/JPSSv28n2.007

Pawlowsky-Glahn, V., Buccianti, A., 2011. Compositional Data Analysis, Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9781119976462

Pawlowsky-Glahn, V., Egozcue, J.J., 2006. Compositional data and their analysis: an introduction. Geol. Soc. London, Spec. Publ. 264, 1-10. https://doi.org/10.1144/GSL.SP.2006.264.01.01

Payne, C., 2020. COVID-19 in Africa. Nat. Hum. Behav. 4, 436-437. https://doi.org/10.1038/s41562-020-0870-5

Pebesma, E., 2018. Simple features for R: Standardized support for spatial vector data. R J. 10, 439-446. https://doi.org/10.32614/RJ-2018-009

Pérez-Foguet, A., Giné-Garriga, R., 2011. Analyzing Water Poverty in Basins. Water Resour. Manag. 25, 3595-3612. https://doi.org/10.1007/s11269-011-9872-4

Pérez-Foguet, A., Giné-Garriga, R., Ortego, M.I.I., 2017. Compositional data for global monitoring: The case of drinking water and sanitation. Sci. Total Environ. 590-591, 554-565. https://doi.org/10.1016/j.scitotenv.2017.02.220

PERU21, 2017. ¿Qué consecuencias tiene un censo mal planificado?. URL https://peru21.pe/peru/censo-2017-dia-peru-repitio-censo-2005-rarezas-381290-noticia/ (accedido 9.30.21).

Pomberger, R., Sarc, R., Lorber, K.E., 2017. Dynamic visualisation of municipal waste management performance in the EU using Ternary Diagram method. Waste Manag. 61, 558-571. https://doi.org/10.1016/j.wasman.2017.01.018

Prüss-Ustün, A., Bartram, J., Clasen, T., Colford, J.M., Cumming, O., Curtis, V., Bonjour, S., Dangour, A.D., De France, J., Fewtrell, L., Freeman, M.C., Gordon, B., Hunter, P.R., Johnston, R.B., Mathers, C., Mäusezahl, D., Medlicott, K., Neira, M., Stocks, M., Wolf, J., Cairncross, S., 2014. Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: a retrospective analysis of data from 145 countries. Trop. Med. Int. Heal. 19, 894-905. https://doi.org/10.1111/tmi.12329

Prüss-Ustün, A., Wolf, J., Bartram, J., Clasen, T., Cumming, O., Freeman, M.C., Gordon, B., Hunter, P.R., Medlicott, K., Johnston, R., 2019. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on low- and middle-income countries. Int. J. Hyg. Environ. Health 222, 765-777. https://doi.org/10.1016/j.ijheh.2019.05.004

Quispe-Coica, A., 2021. Data characteristic and spurious correlation. Zenodo. https://doi.org/10.5281/ZENODO.5579239

Quispe-Coica, A., Fernández, S., Acharte Lume, L., Pérez-Foguet, A., 2020. Status of Water Quality for Human Consumption in High-Andean Rural Communities: Discrepancies between Techniques for Identifying Trace Metals. J — Multidiscip. Sci. J. 3, 162-180. https://doi.org/10.3390/j3020014

Quispe-Coica, A., Pérez-Foguet, A., 2021. Multivariate measure of urban-rural inequality of hygiene facilities (R code). Zenodo. https://doi.org/10.5281/zenodo.5593837

Quispe-Coica, A., Pérez-Foguet, A., 2020a. Preprocessing alternatives for compositional data related to water, sanitation and hygiene. Sci. Total Environ. 743, 140519. https://doi.org/10.1016/j.scitotenv.2020.140519

Quispe-Coica, A., Pérez-Foguet, A., 2020b. Preprocessing alternatives for WASH estimates (R code). Zenodo. https://doi.org/10.5281/zenodo.3909303

Quispe-Coica, A., Pérez-Foguet, A., 2019. Joint evolution of access to water of urban and rural populations in South America through Compositional Data Analysis, en: Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019): Terrassa. pp. 130-142.

Quispe-Coica, A., Pérez-Foguet, A., 2018. Evolución del Acceso al Agua y Saneamiento en América del Sur Mediante Técnicas Estadísticas Composicionales, en: XXXVI Congreso Interamericano de Ingeniería Sanitaria y Ambiental. AIDIS, Guayaquil-Ecuador, pp. 753-757.

R Core Team, 2020. R: A Language and Environment for Statistical Computing.

Rabie, T., Curtis, V., 2006. Handwashing and risk of respiratory infections: A quantitative systematic review. Trop. Med. Int. Heal. https://doi.org/10.1111/j.1365-3156.2006.01568.x

Redman-Maclaren, M., Barrington, D.J., Harrington, H., Cram, D., Selep, J., Maclaren, D., 2018. Water, sanitation and hygiene systems in pacific island schools to promote the health and education of girls and children with disability: A systematic scoping review. J. Water Sanit. Hyg. Dev. https://doi.org/10.2166/washdev.2018.274

Requejo-Castro, D., Giné-Garriga, R., Pérez-Foguet, A., 2020. Data-driven Bayesian network modelling to explore the relationships between SDG 6 and the 2030 Agenda. Sci. Total Environ. 710, 136014. https://doi.org/10.1016/j.scitotenv.2019.136014

Rousseeuw, P., Yohai, V., 1984. Robust Regression by Means of S-Estimators, en: Robust and nonlinear time series analysis. Springer, New York, NY, pp. 256-272. https://doi.org/10.1007/978-1-4615-7821-5_15

Rousseeuw, P.J., van Zomeren, B.C., 1990. Unmasking Multivariate Outliers and Leverage Points. J. Am. Stat. Assoc. 85, 633-639. https://doi.org/10.2307/2289995

Sadoff, C.W., Borgomeo, E., Uhlenbrook, S., 2020. Rethinking water for SDG 6. Nat. Sustain. 3, 346-347. https://doi.org/10.1038/s41893-020-0530-9

Shahid, N.S., Greenough III, W.B., Samadi, A.R., Huq, M.I., Rahman, N., 1996. Hand washing with soap reduces diarrhoea and spread of bacterial pathogens in a Bangladesh village. J. Diarrhoeal Dis. Res. 85-89.

Shahid, Z., Kalayanamitra, R., McClafferty, B., Kepko, D., Ramgobin, D., Patel, R., Aggarwal, C.S., Vunnam, R., Sahu, N., Bhatt, D., Jones, K., Golamari, R., Jain, R., 2020. COVID-19 and Older Adults: What We Know. J. Am. Geriatr. Soc. 68, 926-929. https://doi.org/10.1111/jgs.16472

Stoler, J., Miller, J.D., Brewis, A., Freeman, M.C., Harris, L.M., Jepson, W., Pearson, A.L., Rosinger, A.Y., Shah, S.H., Staddon, C., Workman, C., Wutich, A., Young, S.L., Adams, E., Ahmed, F., Alexander, M., Asiki, G., Balogun, M., Boivin, M.J., Carrillo, G., Chapman, K., Cole, S., Collins, S.M., Eini-Zinab, H., Escobar-Vargas, J., Ghattas, H., Ghorbani, M., Hagaman, A., Hawley, N., Jamaluddine, Z., Krishnakumar, D., Maes, K., Mathad, J., Maupin, J., Owuor, P.M., Melgar-Quiñonez, H., Morales, M.M., Moran, J., Omidvar, N., Rasheed, S., Samayoa-Figueroa, L., Sánchez-Rodriguez, E.C., Santoso, M. V., Schuster, R.C., Sheikhi, M., Srivastava, S., Sullivan, A., Tesfaye, Y., Triviño, N., Trowell, A., Tshala-Katumbay, D.,

Tutu, R., 2021. Household water insecurity will complicate the ongoing COVID-19 response: Evidence from 29 sites in 23 low- and middle-income countries. Int. J. Hyg. Environ. Health 234, 113715. https://doi.org/10.1016/j.ijheh.2021.113715

Strike, K., El Emam, K., Madhavji, N., 2001. Software cost estimation with incomplete data. IEEE Trans. Softw. Eng. 27, 890-908. https://doi.org/10.1109/32.962560

Sullivan, C.A., 2002. Calculating a Water Poverty Index. World Dev. 30, 1195-1210. https://doi.org/10.1016/S0305-750X(02)00035-9

Sullivan, C.A., Meigh, J.R., Giacomello, A.M., Fediw, T., Lawrence, P., Samad, M., Mlote, S., Hutton, C., Allan, J.A., Schulze, R.E., Dlamini, D.J.M., Cosgrove, W., Delli Priscoli, J., Gleick, P., Smout, I., Cobbing, J., Calow, R., Hunt, C., Hussain, A., Acreman, M.C., King, J., Malomo, S., Tate, E.L., O'Regan, D., Milner, S., Steyl, I., 2003. The water poverty index: Development and application at the community scale. Nat. Resour. Forum 27, 189-199. https://doi.org/10.1111/1477-8947.00054

Synowiec, A., Szczepański, A., Barreto-Duran, E., Lie, L.K., Pyrc, K., 2021. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): a Systemic Infection. Clin. Microbiol. Rev. 34. https://doi.org/10.1128/cmr.00133-20

Tcheutchoua, D.N., Tankeu, A.T., Wouna Angong, D.L., Agoons, B.B., Yanwou Nguemnang, N.Y., Nana Djeunga, H.C., Kamgno, J., 2020. Unexpected low burden of coronavirus disease 2019 (COVID-19) in sub-Saharan Africa region despite disastrous predictions: reasons and perspectives. Pan Afr. Med. J. 37. https://doi.org/10.11604/pamj.2020.37.352.25254

Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: An R-package for Robust Statistical Analysis of Compositional Data, en: Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, Chichester, UK, pp. 341-355. https://doi.org/10.1002/9781119976462.ch25

Templ, M., Hron, K., Filzmoser, P., Facevicova, K., Kynclova, P., Walach, J., Pintar, V., Chen, J., Miksova, D., Meindl, B., Menafoglio, A., Di Blasi, A., Pavone, F., Zeni, G., 2019. Package «robCompositions».

Templ, M., Hron, K., Filzmoser, P., Gardlo, A., 2016. Imputation of rounded zeros for high-dimensional compositional data. Chemom. Intell. Lab. Syst. 155, 183-190. https://doi.org/10.1016/J.CHEMOLAB.2016.04.011

Tennekes, M., 2018. tmap: Thematic Maps in R. J. Stat. Softw. 84, 1-39. https://doi.org/10.18637/jss.v084.i06

Townsend, J., Greenland, K., Curtis, V., 2017. Costs of diarrhoea and acute respiratory infection attributable to not handwashing: the cases of India and China. Trop. Med. Int. Heal. 22, 74-81. https://doi.org/10.1111/tmi.12808

Troeger, C., Blacker, B.F., Khalil, I.A., Rao, P.C., Cao, S., Zimsen, S.R., Albertson, S.B., Stanaway, J.D., Deshpande, A., Abebe, Z., Alvis-Guzman, N., Amare, A.T., Asgedom, S.W., Anteneh, Z.A., Antonio, C.A.T., Aremu, O., Asfaw, E.T., Atey, T.M., Atique, S., Avokpaho, E.F.G.A., Awasthi, A., Ayele, H.T., Barac, A., Barreto, M.L., Bassat, Q., Belay, S.A., Bensenor, I.M., Bhutta, Z.A., Bijani, A., Bizuneh, H., Castañeda-Orjuela, C.A., Dadi, A.F., Dandona, L., Dandona, R., Do, H.P., Dubey, M., Dubljanin, E., Edessa, D., Endries, A.Y., Eshrati, B., Farag, T., Feyissa, G.T., Foreman, K.J., Forouzanfar, M.H., Fullman, N., Gething, P.W., Gishu, M.D., Godwin, W.W., Gugnani, H.C., Gupta, R., Hailu, G.B., Hassen, H.Y., Hibstu, D.T., Ilesanmi, O.S., Jonas, J.B., Kahsay, A., Kang, G., Kasaeian, A., Khader, Y.S., Khan, E.A., Khan, M.A., Khang, Y.H., Kissoon, N., Kochhar, S., Kotloff, K.L., Koyanagi, A., Kumar, G.A., Magdy Abd El Razek, H., Malekzadeh, R., Malta, D.C., Mehata, S., Mendoza, W., Mengistu, D.T., Menota, B.G., Mezgebe, H.B., Mlashu, F.W., Murthy, S., Naik, G.A., Nguyen, C.T., Nguyen, T.H., Ningrum, D.N.A., Ogbo, F.A., Olagunju, A.T., Paudel, D.,

Platts-Mills, J.A., Qorbani, M., Rafay, A., Rai, R.K., Rana, S.M., Ranabhat, C.L., Rasella, D., Ray, S.E., Reis, C., Renzaho, A.M., Rezai, M.S., Ruhago, G.M., Safiri, S., Salomon, J.A., Sanabria, J.R., Sartorius, B., Sawhney, M., Sepanlou, S.G., Shigematsu, M., Sisay, M., Somayaji, R., Sreeramareddy, C.T., Sykes, B.L., Taffere, G.R., Topor-Madry, R., Tran, B.X., Tuem, K.B., Ukwaja, K.N., Vollset, S.E., Walson, J.L., Weaver, M.R., Weldegwergs, K.G., Werdecker, A., Workicho, A., Yenesew, M., Yirsaw, B.D., Yonemoto, N., El Sayed Zaki, M., Vos, T., Lim, S.S., Naghavi, M., Murray, C.J., Mokdad, A.H., Hay, S.I., Reiner, R.C., 2018. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Infect. Dis. 18, 1211-1228. https://doi.org/10.1016/S1473-3099(18)30362-1

Turman-Bryant, N., Clasen, T.F., Fankhauser, K., Thomas, E.A., 2018. Measuring progress towards sanitation and hygiene targets: a critical review of monitoring methodologies and technologies. Waterlines 37, 229-247. https://doi.org/10.3362/1756-3488.18-00008

UN-Water, 2016. Water and Sanitation Interlinkages across the 2030 Agenda for Sustainable Development. Geneva.

UN Water, 2016. Monitoring Water and Sanitation in the 2030 Agenda for Sustainable Development. An Introd. Geneva, Switz.

UNICEF and WHO, 2020. Progress on drinking water, sanitation and hygiene in school: special focus on COVID-19. UNICEF, New York.

United Nations Department of Economic and Social Affairs Population Division, 2019. World Population Prospects, Online Edition. Rev. 1.

United Nations General Assembly, 2015. General Assembly Resolution A/RES/70/1. Transforming our world: the 2030 Agenda for Sustainable Development.

van den Boogaart, K.G., Tolosana-Delgado, R., 2013a. Introduction, en: Analyzing Compositional Data with R. Springer, Berlin Heidelberg, pp. 1-12. https://doi.org/10.1007/978-3-642-36809-7_1

van den Boogaart, K.G., Tolosana-Delgado, R., 2013b. Fundamental concepts of compositional data analysis, en: Analyzing Compositional Data with R. Springer, Berlin Heidelberg, pp. 13-50. https://doi.org/10.1007/978-3-642-36809-7_2

van Doremalen, N., Bushmaker, T., Morris, D.H., Holbrook, M.G., Gamble, A., Williamson, B.N., Tamin, A., Harcourt, J.L., Thornburg, N.J., Gerber, S.I., Lloyd-Smith, J.O., de Wit, E., Munster, V.J., 2020. Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. N. Engl. J. Med. 382, 1564-1567. https://doi.org/10.1056/nejmc2004973

Van Liew, M.W., Arnold, J.G., Garbrecht, J.D., 2003. Hydrologic simulation on agricultural watersheds: Choosing between two models. Trans. ASAE 46, 1539-. https://doi.org/10.13031/2013.15643

Verma, S.P., 2020. Multidimensional Techniques for Compositional Data Analysis, en: Road from Geochemistry to Geochemometrics. Springer Singapore, pp. 441-479. https://doi.org/10.1007/978-981-13-9278-8_11

Von Eynatten, H., Pawlowsky-Glahn, V., Egozcue, J.J., 2002. Understanding perturbation on the simplex: A simple method to better visualize and interpret compositional data in ternary diagrams. Math. Geol. 34, 249-257. https://doi.org/10.1023/A:1014826205533

Vuorinen, V., Aarnio, M., Alava, M., Alopaeus, V., Atanasova, N., Auvinen, M., Balasubramanian, N., Bordbar, H., Erästö, P., Grande, R., Hayward, N., Hellsten, A., Hostikka, S., Hokkanen, J., Kaario, O., Karvinen, A., Kivistö, I., Korhonen, M., Kosonen, R., Kuusela, J., Lestinen, S., Laurila, E., Nieminen, H.J., Peltonen, P., Pokki, J., Puisto, A., Råback, P., Salmenjoki, H., Sironen, T., Österberg, M., 2020. Modelling aerosol transport and virus exposure with numerical simulations in relation to SARS-CoV-2 transmission by inhalation indoors. Saf.

Sci. 130, 104866. https://doi.org/10.1016/j.ssci.2020.104866

Weststrate, J., Dijkstra, G., Eshuis, J., Gianoli, A., Rusca, M., 2018. The Sustainable Development Goal on Water and Sanitation: Learning from the Millennium Development Goals. Soc. Indic. Res. 2018 1432 143, 795-810. https://doi.org/10.1007/S11205-018-1965-5

WHO/UNICEF, 2021. Progress on household drinking water, sanitation and hygiene 2000-2020: Five years into the SDGs. UNICEF J.

WHO/UNICEF, 2020. Hygiene Baselines pre-COVID-19, JMP.

WHO/UNICEF, 2019a. Progress on household drinking water, sanitation and hygiene 2000-2017: special focus on inequalities, WHO. United Nations Children's Fund (UNICEF) and World Health Organization, New York.

WHO/UNICEF, 2019b. Joint Monitoring Programme for Water Supply, Sanitation, and Hygiene: Estimates on the use of water, sanitation and hygiene in Indonesia. JMP. URL https://washdata.org/data (accedido 5.11.19).

WHO/UNICEF, 2019c. WASH in health care facilities: global baseline report 2019. World Health Organization, Geneva.

WHO/UNICEF, 2018. JMP methodology 2017 update & sdg baselines.

WHO/UNICEF, 2017. Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines, World Health Organization and UNICEF. World Health Organization, Geneva.

WHO/UNICEF, 2016. Inequalities in sanitation and drinking water in Latin America and the Caribbean, en: Inequalities in sanitation and drinking water in Latin America and the Caribbean. p. 12.

WHO/UNICEF, 2015. Task Force on Monitoring Inequalities for the 2030 Sustainable Development Agenda Meeting Report. WHO/UNICEF, New York, NY, USA.

WHO/UNICEF, 2014. World Health Organization and UNICEF, Report. WHO/UNICEF JMP Task Force on Methods 2014.

WHO, 2020a. Recommendations to Member States to improve hand hygiene practices to help prevent the transmission of the COVID-19 virus: interim guidance, 1 April 2020. World Health Organization, Geneva PP - Geneva.

WHO, 2020b. Transmission of SARS-CoV-2: implications for infection prevention precautions: scientific brief, 09 July 2020. World Health Organization, Geneva PP - Geneva.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis, 2.ª ed, Use R! Springer Nature, New York. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the Tidyverse. J. Open Source Softw. 4, 1686. https://doi.org/10.21105/joss.01686

Wolf, J., Bonjour, S., Prüss-Ustün, A., 2013. An exploration of multilevel modeling for estimating access to drinking-water and sanitation. J. Water Health 11, 64-77. https://doi.org/10.2166/wh.2012.107

Wolf, J., Johnston, R., Freeman, M.C., Ram, P.K., Slaymaker, T., Laurenz, E., Prüss-Ustün, A., 2019. Handwashing with soap after potential faecal contact: global, regional and country estimates. Int. J. Epidemiol. 48, 1204-1218. https://doi.org/10.1093/ije/dyy253

Wong, R.K.W., Yao, F., Lee, T.C.M., 2014. Robust Estimation for Generalized Additive Models.

http://dx.doi.org/10.1080/10618600.2012.756816 23, 270-289. https://doi.org/10.1080/10618600.2012.756816

Wood, S., 2019. «mgcv»: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.

World Bank, 2021. GDP per capita (current US$) | Data. URL https://data.worldbank.org/indicator/NY.GDP.PCAP.CD (accedido 4.9.18).

World Bank, 2020. GDP per capita (current US$) - Latin America & Caribbean. URL https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=ZJ (accedido 11.23.20).

Yang, H., Bain, R., Bartram, J., Gundry, S., Pedley, S., Wright, J., 2013. Water Safety and Inequality in Access to Drinking-water between Rich and Poor Households. Environ. Sci. Technol. 47, 1222-1230. https://doi.org/10.1021/es303345p

Yerg, A., 2013. Modeling and Forecasting Drinking-water and Sanitation Access: A New Approach. https://doi.org/10.17615/6nf5-nf23

Yerg, A., Bain, R., Bartram, J., 2013. Estimating Progress of Millennium Development Goal Target 7C with Logistic Regression. Working Paper. Chapel Hill, NC: The Water Institute at UNC.

Yohai, V.J., 1987. High Breakdown-Point and High Efficiency Robust Estimates for Regression. Ann. Stat. 15, 642-656. https://doi.org/10.1214/aos/1176350366

Yohai, V.J., Stahel, W.A., Zamar, R.H., 1991. A Procedure for Robust Estimation and Inference in Linear Regression, en: Directions in Robust Statistics and Diagnostics. Springer New York, New York, NY, pp. 365-374. https://doi.org/10.1007/978-1-4612-4444-8_20

Yori, P.P., Lee, G., Olórtegui, M.P., Chávez, C.B., Flores, J.T., Vasquez, A.O., Burga, R., Pinedo, S.R., Asayag, C.R., Black, R.E., Caulfield, L.E., Kosek, M., 2014. Santa Clara de Nanay: The MAL-ED cohort in Peru. Clin. Infect. Dis. 59, S310-S316. https://doi.org/10.1093/cid/ciu460

Zerbo, A., Castro Delgado, R., Arcos González, P., 2021. Water sanitation and hygiene in Sub-Saharan Africa: Coverage, risks of diarrheal diseases, and urbanization. J. Biosaf. Biosecurity 3, 41-45. https://doi.org/10.1016/J.JOBB.2021.03.004

ANEXOS

Figura A1. Diagrama de caja de la desigualdad urbano-rural por región

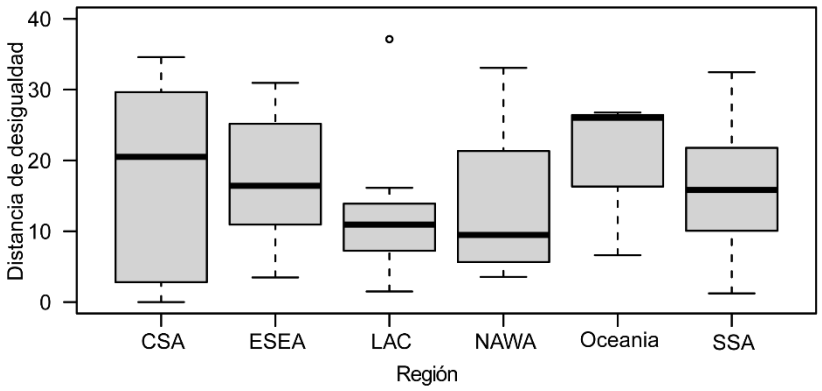


Figura A2. Servicio de higiene urbano (A) y rural (B) de 2017. Mapa temático de casos confirmados (C) y muertes (D) del COVID-19, obtenido el 15 de marzo de 2021 de la plataforma de la OMS (véase https://covid19.who.int/).

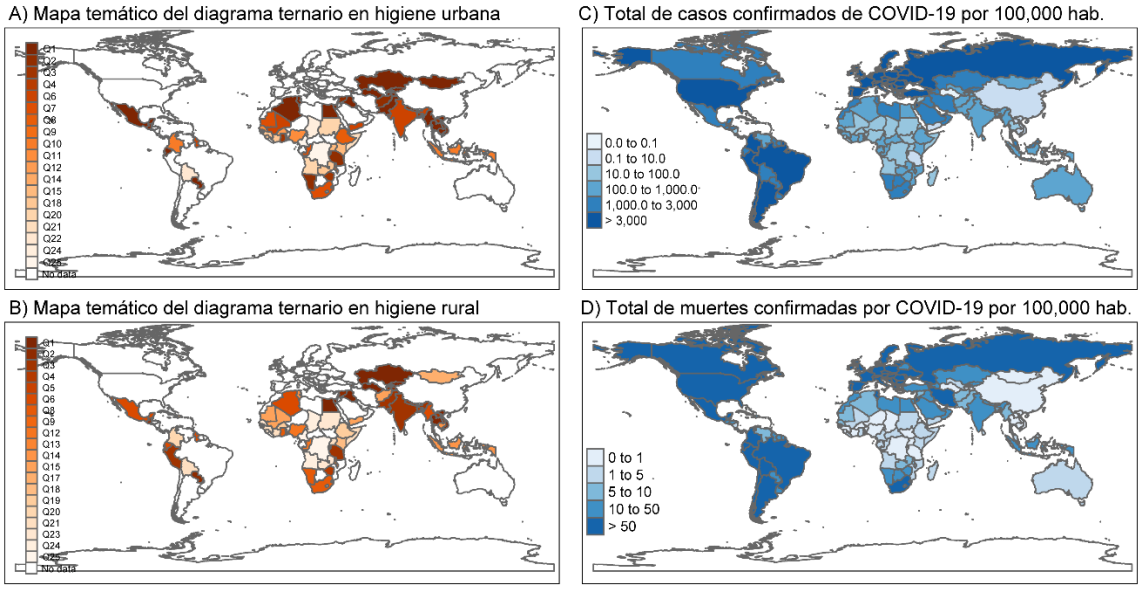


Figura A3. Representación gráfica urbana (izquierda) y rural (derecha) de los países en el diagrama ternario por ingresos. La información sobre la clasificación de los países según su nivel de riqueza se obtuvo del Banco Mundial(2021).

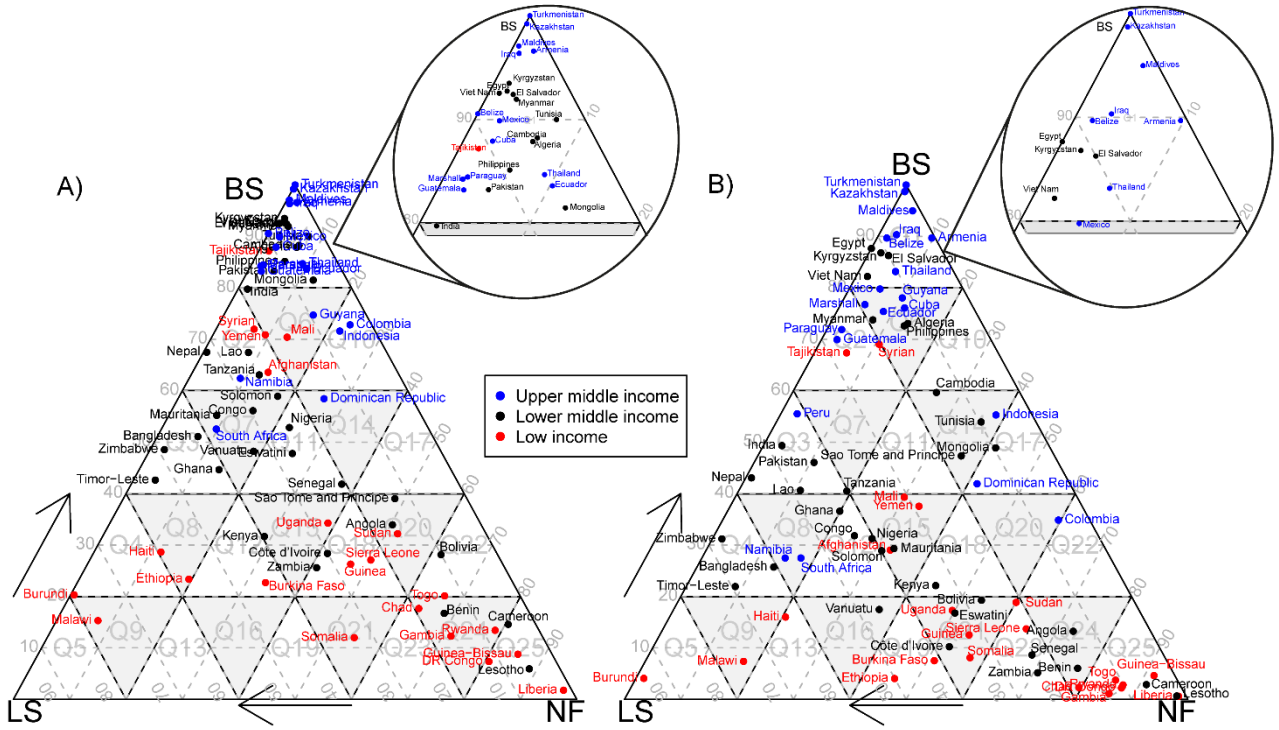


Figura A4. Ajuste lineal, en las seis regiones analizadas, entre la medida de desigualdad obtenida y el promedio de la BS urbana y rural. Las regiones CSA, ESEA, ALC, NAWA y Oceanía tienen una pendiente negativa, mientras que la región SSA tiene una pendiente positiva.

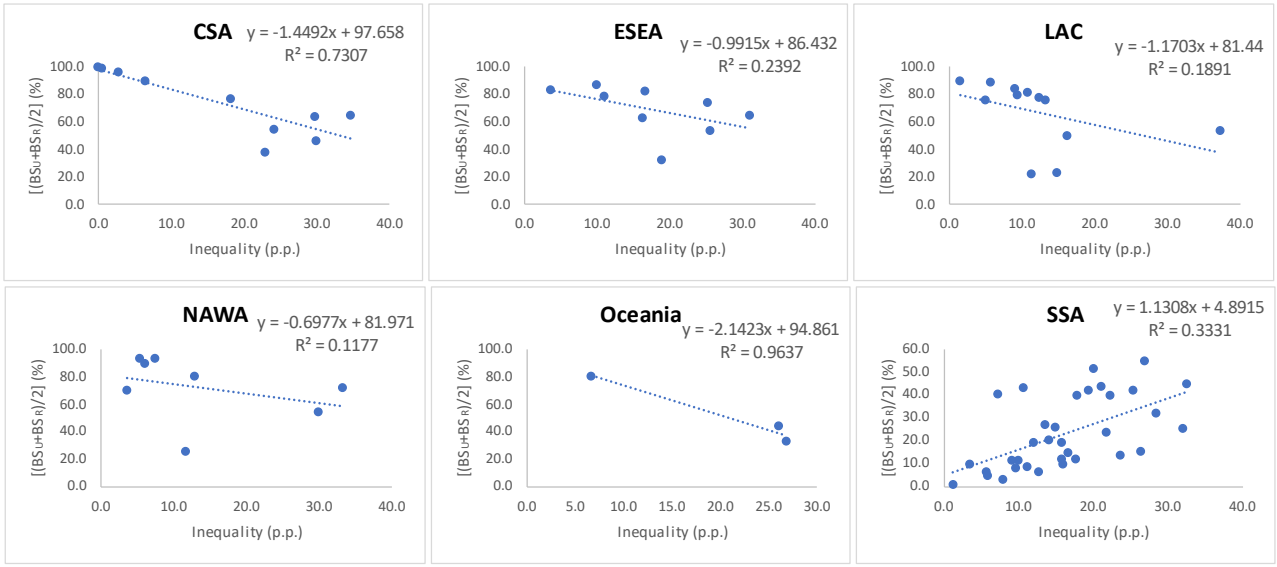


Tabla A1. Clasificación urbana y rural de los países en el diagrama ternario

ID	Urban	Rural
Q1	Algeria, Armenia, Belize, Cambodia, Cuba, Ecuador, Egypt, El Salvador, Guatemala, Iraq, Kazakhstan, Kyrgyzstan, Maldives, Marshall, Mexico, Mongolia, Myanmar, Pakistan, Paraguay, Philippines, Tajikistan, Thailand, Tunisia, Turkmenistan and Vietnam	Armenia, Belize, Egypt, El Salvador, Iraq, Kazakhstan, Kyrgyzstan, Maldives, Thailand, Turkmenistan and Vietnam
Q2	Afghanistan, Laos, Namibia, Nepal, Syria and Tanzania	Guatemala, Paraguay, Syria and Tajikistan
Q3	Bangladesh, Ghana, Timor-Leste, Zimbabwe	India, Laos, Nepal, Pakistan and Tanzania
Q4	Burundi	Bangladesh, Timor-Leste and Zimbabwe
Q5		Burundi
Q6	Guyana, India, Mali and Yemen	Algeria, Cuba, Ecuador, Guyana, Marshall, Mexico, Myanmar and Philippines
Q7	Congo, Mauritania, Solomon, South Africa and Vanuatu	
Q8	Ethiopia and Haiti	Ghana, Namibia and South Africa
Q9	Malawi	Malawi
Q10	Colombia and Indonesia	
Q11	Eswatini, Nigeria and Senegal	
Q12	Burkina Faso	Congo and Nigeria
Q13		Haiti
Q14	Dominican Republic	Cambodia, Indonesia, Sao Tome and Principe and Tunisia
Q15	Kenya and Uganda	Afghanistan, Mali, Mauritania, Solomon and Yemen
Q16		Vanuatu
Q17		Dominican Republic and Mongolia
Q18	Côte d'Ivoire, Guinea, Sierra Leone and Zambia	Kenya
Q19		Burkina Faso and Ethiopia
Q20	Angola, Sao Tome and Principe and Sudan	Colombia
Q21	Somalia	Bolivia, Côte d'Ivoire, Eswatini, Guinea, Somalia and Uganda
Q22	Bolivia and Togo	
Q23		Senegal, Sierra Leone, Sudan and Zambia
Q24	Benin, Chad, Gambia and Rwanda	Angola, Benin and Chad
Q25	Cameroon, DR Congo, Guinea-Bissau, Lesotho and Liberia	Cameroon, DR Congo, Gambia, Guinea-Bissau, Lesotho, Liberia, Rwanda and Togo