**Medias Ponderadas para las bases de datos epidemiológicas**

**Contexto**

La técnica de ponderación se aplicará para corregir sesgos producidos por desequilibrios en la composición de las muestras tomadas en cada uno los monitoreos, que corresponden a Lotes sensores y a Brigadas fitosanitarias. Los sesgos que se identificaron en la visualización de los datos corresponden a desequilibrios en el tamaño de las muestras tomadas por cada año y por cada departamento.

Por lo anterior, se consultaron los datos históricos de área sembrada del cultivo de arroz mecanizado en Colombia de la Encuesta Nacional de Arroz Mecanizado (ENAM) para determinar los pesos del área sembrada por departamento y año para definir un parámetro de ponderación externo y adicionalmente en cada uno de los monitoreos se calculó el peso de la cantidad de muestreos por departamento y año para tener un parámetro de ponderación propio de cada base de datos.

El Departamento Administrativo Nacional de Estadística (DANE) y la Federación Nacional de Arroceros (Fedearroz) – Fondo Nacional del Arroz (FNA) realizan, como parte del convenio entre las dos entidades, la Encuesta Nacional de Arroz Mecanizado (ENAM) para generar información estadística sobre los principales indicadores del sector arrocero. Se realiza censo en la zona arrocera de los Llanos, se utilizan registros administrativos de los distritos de riego asociados al cultivo de arroz, y se ejecuta una encuesta a partir de una muestra probabilística en las demás zonas productoras de arroz.

Esta integración de metodologías genera la medición de las variables de área, producción y rendimiento del arroz mecanizado, asegurando una cobertura nacional. El área sembrada se presenta a nivel nacional, departamental, por zonas arroceras, según mes de siembra y por sistema de producción del cultivo. Las estimaciones de área cosechada y producción se presentan a nivel nacional y a nivel de los principales departamentos productores de arroz (Meta, Casanare, Tolima, Huila y Resto de departamentos). El rendimiento se presenta a nivel departamental y por sistema de producción.

**Metodología**

1. Supongamos que la población se compone de m subpoblaciones, correspondientes a cada municipio y año, monitoreados
2. Se calculó el factor peso de cada departamento productor de arroz por año tomando como referencia el área sembrada reportada en la ENAM y determinando la distribución porcentual de cada departamento en el área total sembrada.
3. Para la base de brigadas fitosanitarias (BFito) se determinó el factor peso calculando la participación porcentual del número de muestras por departamento y año en el total de muestras por año.
4. Para la base de lotes sensores (lsensores) se realizó el mismo procedimiento que en la base de brigadas para definir los factores de pesos.

Se cargó un archivo llamado pesos\_poderacion\_muestreos.xls, en la siguiente ruta <https://drive.google.com/drive/u/0/folders/1b1w5qffXU7osIduCF9-XzDFrIR3imljv> y el cual se encuentra dividido en las siguientes hojas:

* pesos\_area\_m\_bfito: en dicha hoja se definieron columnas de peso\_area (peso porcentual por área sembrada de arroz por año) y de peso\_mbf (distribución porcentual de las muestras de cada departamento y año en el monitoreo de brigadas fitosanitarias), ambos estan calculados para los años 2012 a 2021.
* pesos\_area\_m\_lsensores: en dicha hoja se definieron columnas de peso\_area (peso porcentual por área sembrada de arroz por año) y de peso\_mls (distribución porcentual de las muestras de cada departamento y año en el monitoreo de lotes sensores), ambos estan calculados para los años 2015 a 2019.

Por lo anterior, debemos calcular la media ponderada de las variables que conforman a cada tipo de monitoreo, teniendo en cuenta el peso de cada individuo en función de su subpoblación. Para ello aplicaremos la siguiente fórmula:

Donde,

W1i = corresponde al factor de peso\_area\_año

W2i = corresponde al factor peso\_mls ó peso\_mbf, dependiendo el monitoreo que se este analizando

Xi = Valor de la variable analizada en la ubicación i

Recordar que las bases de datos, se encuentran en el siguiente link <https://drive.google.com/drive/u/0/folders/1aTjLJUAdPFdDXliYyVGLE3tPWwuLEZf->

**VARIABLES**

**Las** variables a las cuales se les aplicará el cálculo de medias ponderadas son las siguientes, dependiendo del tipo de monitoreo**:**

Para el monitoreo de lotes sensores, son:

* Sev\_piricularia\_hoja listo
* Sev\_helmitosporium\_hoja listo
* Sev\_rhizoctonia listo
* Sev\_cercospora
* sev\_escaldado
* sev\_mancha\_cafe
* Sev\_gaeumanomices
* Sev\_manchado\_grano
* Porc\_vaneamiento\_ajustado
* inc\_entorchamiento
* inc\_hoja blanca
* inc\_piri\_hoja
* inc\_helmitosporium
* inc\_rhizoctonia
* inc\_cercospora
* inc\_escaldado
* inc\_mancha\_cafe
* inc\_gaeumanomices
* inc\_piricularia\_cuello
* inc\_sarna\_orejas
* inc\_complejo\_bacterial

Para el monitoreo de brigadas fitosanitaria, son:

* inc\_rhizoctonia
* inc\_piricularia
* inc\_gaeumanomices
* inc\_helmitosporium
* inc\_ barrenador
* inc\_mancha\_cafe
* sev\_mancha\_cafe
* Inc\_complejo\_bacteriano
* porc\_vaneamiento

Finalizado: validar en la reunión que sean todos los campos a calcular su media ponderada de peso

Sobre las medias ponderadas:

- Ya se calcularon, es importante validar principalmente 'brigadas fitosanitarias', al parecer no hay consistencia en la ponderación, revisé la metodología realizada en python y parece que esta bien; Sin embargo, me gustaría que validaras los 2 resultados {'media\_ponderada\_brigadas\_fitosanitarias.xlsx' y 'media\_ponderada\_lotes\_sensores.xlsx'} especialmente de 'porc\_vaneamiento', cuya media ponderada queda por encima de 3000 aprox.