Identificación a través del análisis estadístico; las variables influyentes en la producción óptima de un tipo de vino y la prevención de su caducidad

Laboratorio de Modelación 2, MAT 289

Profesor: Pablo Aguirre	Mail: Pablo.Aguirre@usm.cl
Mandante: Eduardo Valenzuela	Mail: Eduardo.Valenzuela@usm.cl
Alumno: Andrés Sandoval	Mail: Andrés.Sandoval@usm.cl

DMAT
Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

EX UMBRA IN SOLEM
UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Descripción del problema

- Se tienen mediciones de diversas variables en el proceso de fermentación de un vino.
- Las variables medidas corresponden a la densidad, Brix, Acidez, Ácidos orgánicos, Azucares, Alcoholes, Aminoácidos y Ácidos grasos.
- Se tiene registro de procesos de fermentación en el que el vino ha superado los estándares de calidad y se ha considerado como un producto bueno. Y se tiene registro de procesos de fermentación en los que el vino no ha pasado los estándares de calidad y se ha considerado como malo.

Problemas del Data Set

Se tienen los siguientes problemas en el Data Set:

- Primero, los procesos de fermentación considerados buenos son menos que los procesos de fermentación considerados malos.
- Segundo, existe Data Missing parcial en algunas variables.
- Tercero, existe Data Missing total en algunas variables.

Etanol [% v/v]	Glicerol [ppm]	Ac. Aspartico	Serina
0.002	<11,5	44.56	44.10
0.754	1756.497		
4.185	7057.491	21.83	
7.010	10801.162		
8.572	9208.230	15.12	
9.266	9496.187		
9.768	9475.375	18.41	
10.475	9645.938		
10.365	9704.046	28.25	
10.753	9644.646		
11.107	9882.663	34.21	
10.726			
11.462		32.44	
11.553	10080.896		
11.434	9834.974	39.36	
11.025	9794.224		
11.429	9724.651	38.11	
11.485	9850.626		

Figura: Ilustración de la forma de 4 variables de los datos (para una trayectoria). Etanol contiene la totalidad de datos. Glicerol y Ac. Aspertico sufren de Data Missing parcial. La variable Serina posee un Data Missing total.

Objetivos

- Filtrar la base de datos para poder trabajarla
- Encontrar cuales son las variables que influyen en el resultado del proceso de fermentación, en el sentido de que si este es considerado bueno o malo.
- Encontrar un método de predicción para saber antes de las 72 horas si es que el vino resultará bueno o malo.

Análisis de Datos Funcionales (FDA)

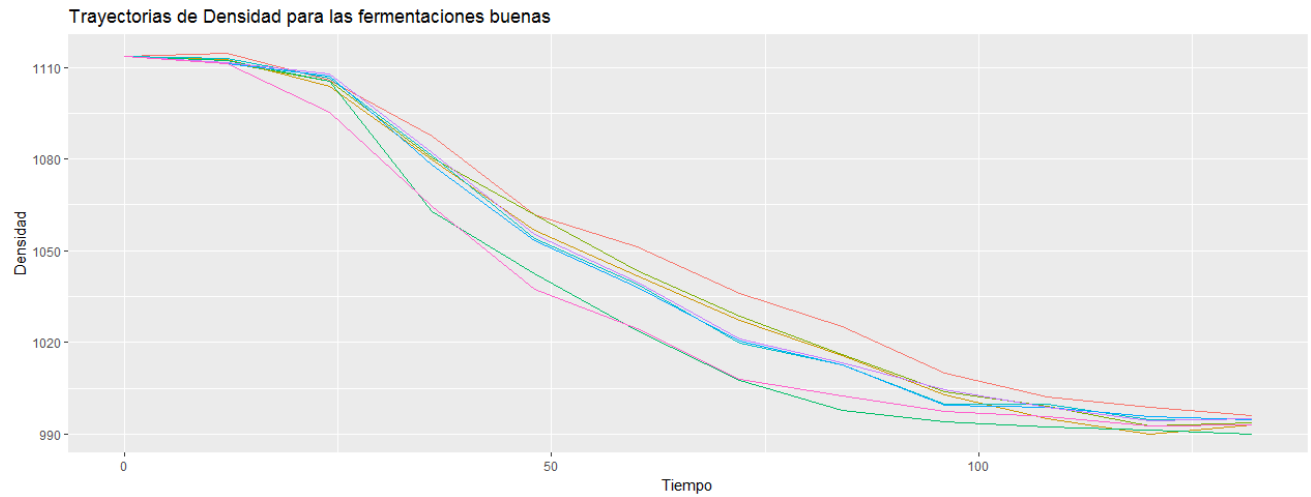
- En estadística se clásica se cuenta con observaciones de una variable de interés y se procede a su análisis:

$$X_1, X_2, ..., X_n \longmapsto \text{Análisis pertinente}$$

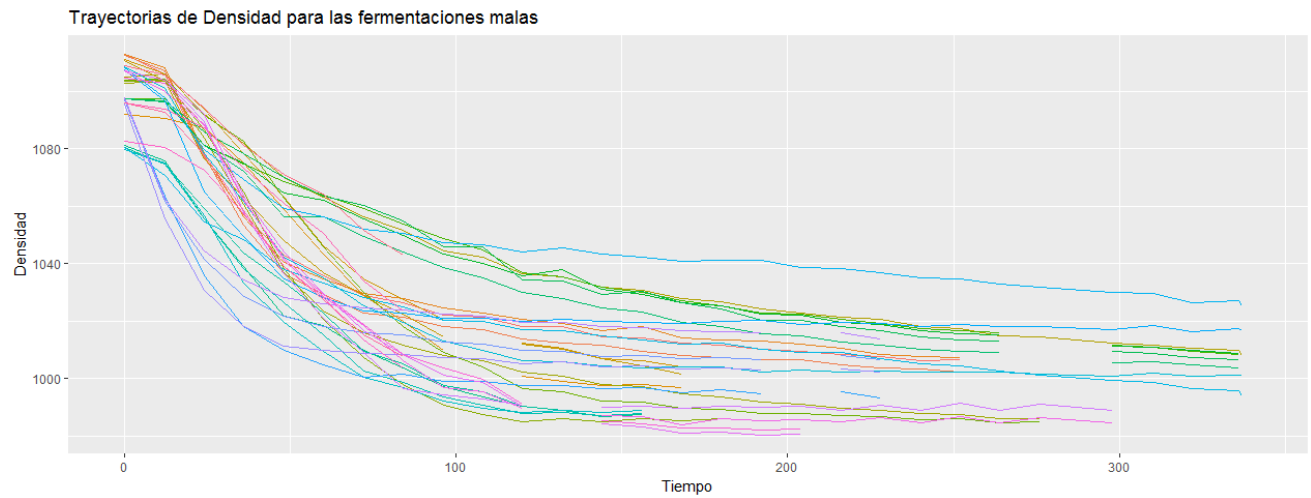
.

- En el FDA las observaciones son trayectorias:

$$X_1(t), X_2(t), ..., X_n(t) \longmapsto \text{Análisis pertienente usando FDA}$$



(a) Fermentaciones buenas



(b) Fermentaciones malas

Figura: Curvas de Densidad para las fermentaciones.

Filtrado del Data Set y generación de datos

- Las variables que contienen Data Missing total no son consideradas en el trabajo.

- Se trabajó con las variables que no contienen Data Missing y las que contienen Data Missing parcial. Se generaron datos de estas variables usando el método de B-spline (Explicado a continuación).

Expansión en Bases

- Considerar una curva $X(t)$, ponemos:

$$X(t) = Z(t) + \epsilon(t)$$

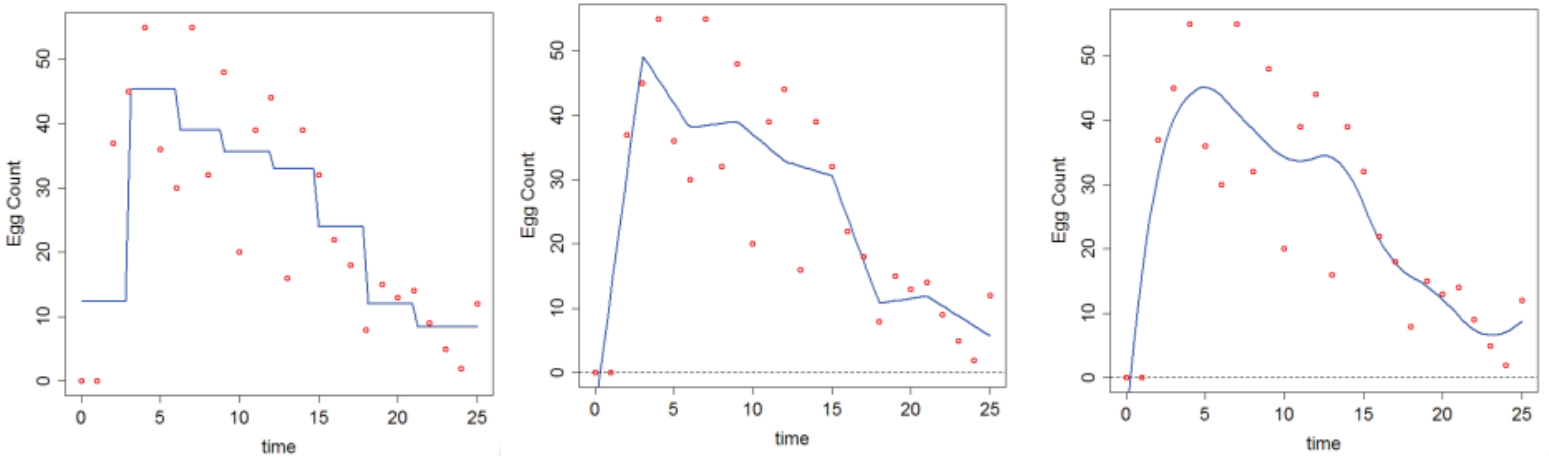
. donde:

$$Z(t) = \sum_{j=1}^{k} c_j \phi_j(t)$$

y $\epsilon(t)$ es el error de aproximación.

- Ejemplos: Bases de Fourier, Base de B-splines.

Ejemplo del Ajuste por B-splines



(a) B-splines Constantes (b) B-splines Lineales (c) B-splines Cuadráticos

Figura: Ajustes usando B-splines

Test de Hipótesis

- Supongamos que contamos con las siguientes observaciones (discretas):

$$U_{i,j} = X_i(t_j) + \epsilon_{i,j}, \qquad \text{para } j = 1, 2, ..., m$$

y

$$V_{i,j} = Y_i(t_j) + \epsilon_{i,j}, \qquad \text{para } j = 1, 2, ..., m$$

- $X_1, ..., X_n$ son trayectorias de una v.a X e $Y_1, ..., Y_n$ son trayectorias provenientes de una v.a Y
- Queremos contrastar la hipótesis $\mathcal{H}_0$ de que X e Y poseen la misma distribución (Ver [3])
- En aplicación a nuestro problema, se realizo el test de Hipótesis (que se detallará a continuación) a las variables del Data Set. La familia X corresponde a las fermentaciones buenas y la familia Y corresponde a las fermentaciones malas.
- Problema: Hay más mediciones para las fermentaciones malas.

- Se construyen estimadores $\hat{X}_i$ e $\hat{Y}_i$ de X_i e Y_i. (Regresión no paramétrica)
- Se construyen $\hat{F_X}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\hat{X}_i \leq z)$ y $\hat{F_Y}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\hat{Y}_i \leq z)$
- Se constuye el estadístico de Cramér-von Mises:

$$\hat{T} = \int \left(\hat{F}_X(z) - \hat{F}_Y(z)\right)^2 \mu(\partial z)$$

Y se aproxima mediante la técnica de Monte Carlo:

$$\hat{T}_n = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{F}_X(M_i(t)) - \hat{F}_Y(M_i(t))\right)^2$$

- Para $M_1(t), M_2(t), ..., M_N(t)$ trayectorias de la variable aleatoria M, donde $M(t) = \sum_{k=1}^{\infty} \zeta_k \phi_k(t)$, donde ζ_k son variables aleatorias con media 0 y varianza unitaria (Se escogen de forma $N(0,1)$) y $\{\phi_k(t)\}$ corresponde a una base ortonormal de $L^2(\mathcal{I})$. (Se toma la base de funciones de la representación por componentes principales de las variables X e Y)

- Queremos una significancia de $1-\alpha$. Para finalizar se usará el método de Bootstrap (Ver[4]): ¿En que consiste?
- tenemos $\{\hat{X}_1, ..., \hat{X}_n\}$ y $\{\hat{Y}_1, ..., \hat{Y}_n\}$ nuestros estimadores previamente computados. Se realiza una muestra aleatoria simple con reposición digamos $\{X_1^*, X_2^*, ..., X_n^*\}$ y $\{Y_1^*, Y_2^*, ..., Y_n^*\}$. Se procede a calcular (via Monte Carlo)
$$T^* = \int \left(F_{X^*}(z) - \hat{F}_{Y^*}(z)\right)^2 \mu(\partial z)$$
- Se realiza el muestro aleatorio simple con reposición B veces. Obtenemos así una muestra $\{T_1^*, ..., T_B^*\}$
- Se construye el estadístico de orden $t_\alpha = T^*_{(B(1-\alpha)+1)}$
- Se rechaza $\mathcal{H}_0$ si $\hat{T} > t_\alpha$

Temperaturas en Australia (promedio)

- Se tiene un registro de temperaturas promedio en Australia medidas desde 1856 hasta 1936 (Ver [5]). Se consideran 4 periodos de tiempo: desde 1856 hasta 1876, desde 1877 hasta 1896, desde 1897 hasta 1916 y finalmente desde 1917 hasta 1936.

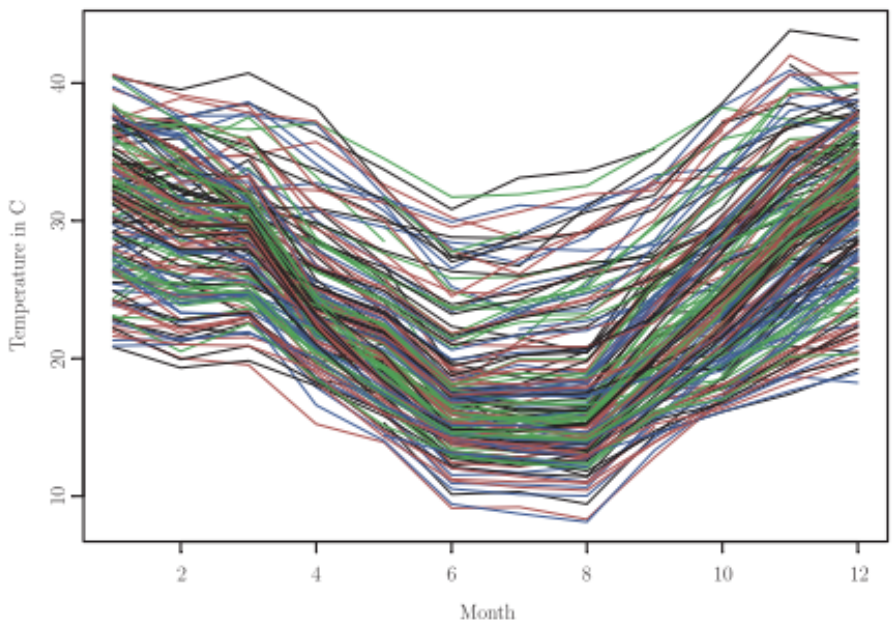


Figura: Temperaturas de Australia (promedio) medidas en los 80 años de estudio

Period	Period	P-value
1	2	0.001
1	3	0.081
1	4	0.024
2	3	0.001
2	4	0.000
3	4	0.053

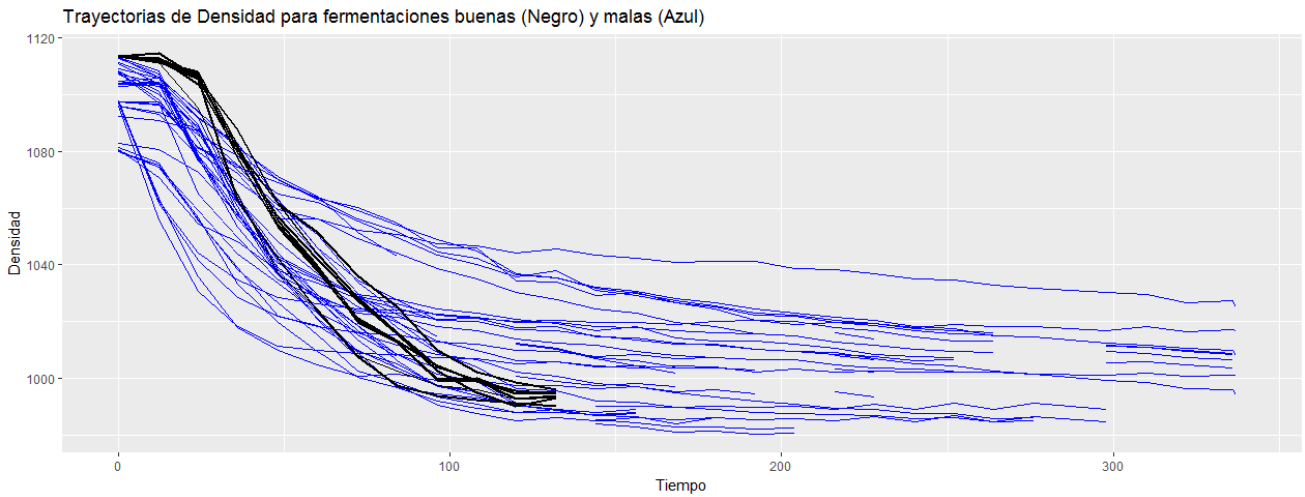
Figura: Valores p estimados por parejas usando el test descrito en la sección de Metodología

El autor concluyó que el clima promedio en Australia ha cambiado a lo largo del tiempo.

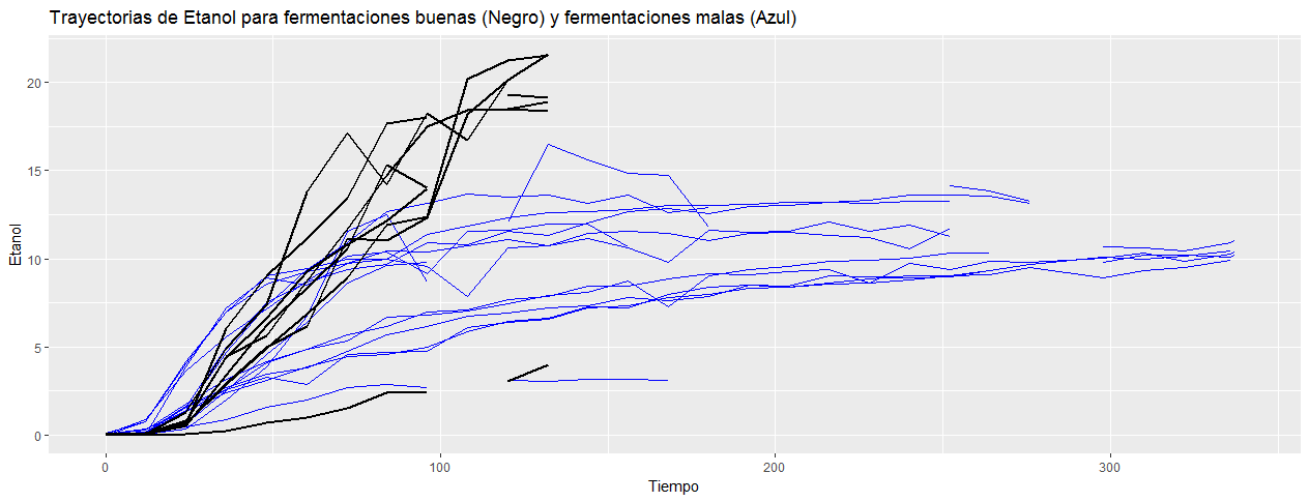
Resultados

Variable	p-value
Densidad	0.022
Brix	0.015
Acidez	0.068
Nitrógeno Asimilable	0.009
Ácido Cítrico	0.012
Ácido Tartárico	0.018
Ácido Málico	0.021
Ácido Succinico	0.032
Ácido Láctico	0.029
Ácido Acético	0.045
Fructosa	0.018
Etanol	0.017
Glicerol	0.098
Histidina	0.066
Glicina	0.071
Treonina	0.087

Cuadro: p-values para las distintas variables.

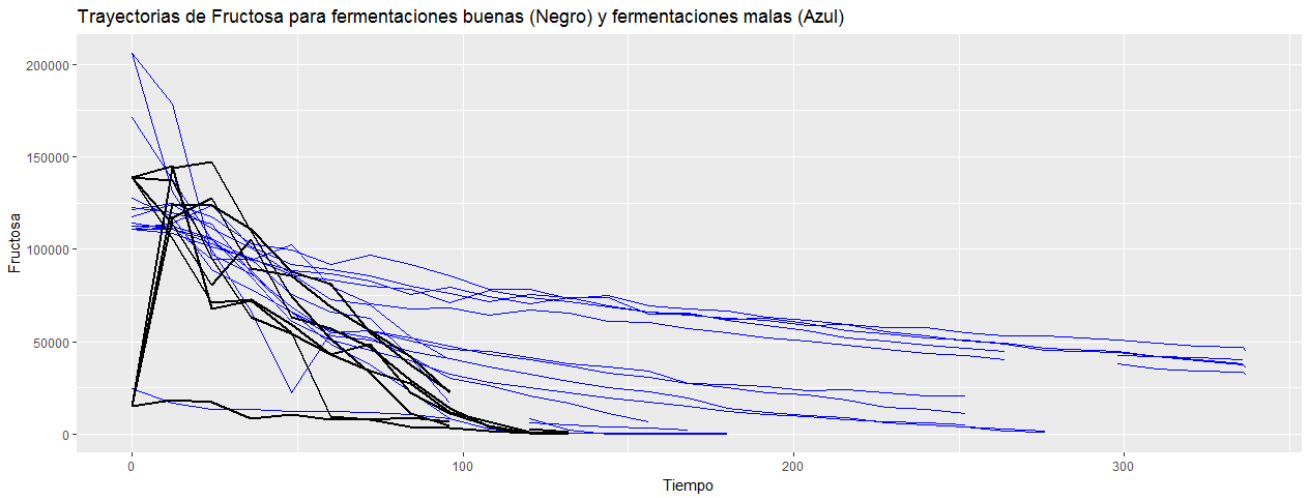


(a) Densidad

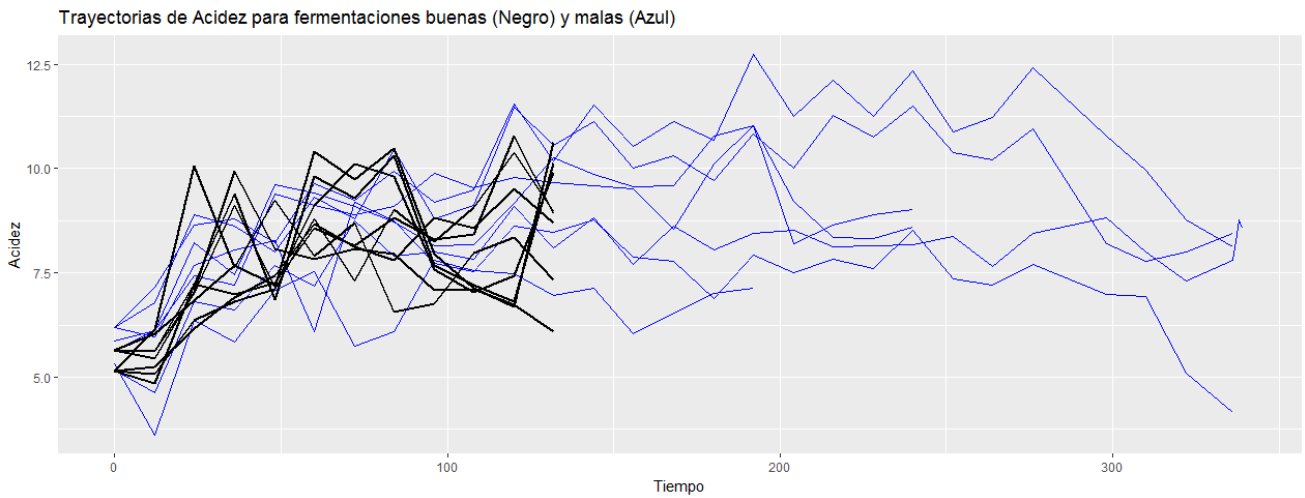


(b) Etanol

Figura: Trayectorias de Densidad y Etanol.



(a) Fructosa



(b) Acidez

Figura: Trayectorias de Fructosa y Ácidez.

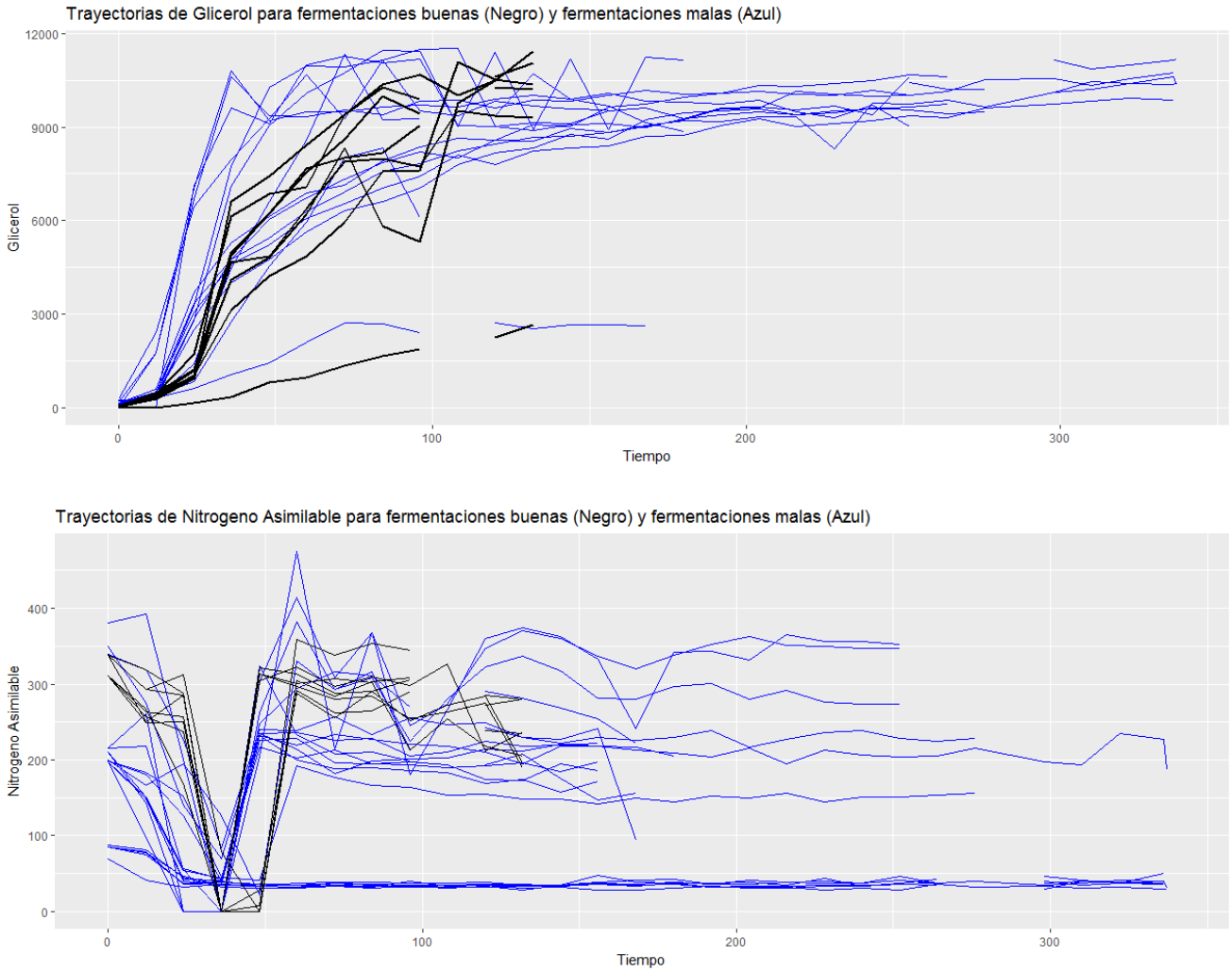


Figura: Trayectorias de Nitrógeno Asimilable

Tiempos de trabajo

Categoría	**Tiempo Esperado**	**Tiempo Efectivo**
Investigación	30	80
Código e Implementaciones	20	50
Reuniones	5	2
Redacción de informe	13	-
Total	68	132

Cuadro: Tiempos de trabajo.

Conclusiones

- Según el test descrito en la sección de Metodología; las variables que son influyentes en el resultado de fermentación del vino son : Densidad, Brix, Nitrogeno Asimilable, todos los tipos de Acidos la Fructosa y el Etanol

- La certeza de este tipo de test depende plenamente de la cantidad de datos observados y precisión en la aproximación computacional. Un estudio mas detallado o profesional debería incluir otro tipo de análisis como es el del error.

- Este trabajo me sirvió para profundizar mis conocimientos en estadística. El análisis funcional de datos es un tema reciente que desconocía. Pude aprender a usar mejor el software R entre otras cosas.

Bibliografía

[1] Rui. Castro *Lectures 2 and 3 - Goodness-of-Fit (GoF) Tests*

[2] Peter Hall *Two-sample test in functional data analysis from discrete data* . 2007.

[3]Geoffrey Jones *Application of the Bootstrap to Calibration Experiments*. 1996

[4] Han Lin Shang *Bootstrapping functional data: a study of distributional property of sample eigenvalues* . 2011

[5] Gina-Maria Pomann *Two Sample Hypothesis Testing for Functional Data* . 2011