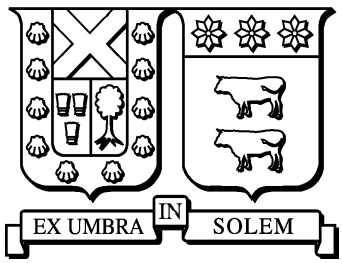
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"RESÚMENES LINGÜÍSTICOS PARA RIEGO DE CULTIVOS"

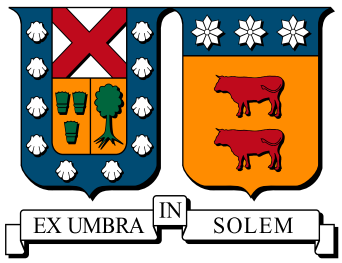
ÁLVARO RODRIGO ROJAS VALENZUELA

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: JOSÉ LUIS MARTÍ LARA

JUNIO 2018

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"RESÚMENES LINGÜÍSTICOS PARA RIEGO DE CULTIVOS"

ÁLVARO RODRIGO ROJAS VALENZUELA

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: JOSÉ LUIS MARTÍ LARA

PROFESOR CORREFERENTE: CECILIA REYES COVARRUBIAS

CORREFERENTE EXTERNO: CHRISTOPHER POPE SCHWARTZ

JUNIO 2018

Agradecimientos

En lo personal considero que es muy hipócrita agradecer cuando no es algo que nazca hacer de verdad, en mi caso tengo mucho que agradecer a muchas personas (y me nace hacerlo), gente genial que me acompañó, soportó, aceptó y enseñó a pesar de nuestras diferencias, a todos ellos les agradezco. Es difícil mencionar a todos y cada uno en forma particular, pero es bueno destacar algunos, como los “cabros del voley”, mis amigos de carrera, mis amigos del colegio y de la vida, mis familiares (quienes fueron un soporte fundamental) y los (buenos) profesores del colegio y la Universidad: personas que más allá de hacer un trabajo como cualquier otro se preocupan de los aspectos humanos de los estudiantes. Se podría decir que este trabajo más que mio es de todos UDs. (así que deberían sentirse orgullosos/avergonzados por el trabajo xD).

Resumen

Esta memoria aborda el tema de resúmenes lingüísticos, teniendo como objetivo principal construir resúmenes para apoyar la toma de decisiones en casos reales de riego de cultivos. Para esto se emplea lógica difusa en datos temporales, obteniendo frases y luego párrafos que resumen el estado hídrico del campo y las prácticas de riego. En particular, se utiliza un "calificador temporal" para la creación de algunos tipos de frases, y se usan reglas difusas *if-then* para inducir valores con los que no se cuentan. Finalmente, se cumple parcialmente con criterios mínimos para la validación de las frases mediante la verificación de las Máximas de Grice.

Abstract

The subject matter of this thesis is Linguistic Summaries, and its main objective is to build summaries to help decision-making in real cases of crop irrigation. To achieve this, fuzzy logic is applied over temporary data, obtaining phrases and subsequently paragraphs that summarize the field's hydric status and irrigation practices. In particular, a "temporal qualifier" is used for the creation of some types of phrases, and *if-then* fuzzy rules are applied to induce unavailable values. Finally, minimum criteria for phrase validation are met partially by verifying them according to Grice's Maxims.

Índice de Contenidos

Agradecimientos III

Resumen IV

Abstract IV

Índice de Contenidos V

Lista de Tablas VII

Lista de Figuras IX

Glosario XI

Introducción 1

1. Definición del Problema 3

1.1. Presentación del Caso 3

1.2. Objetivos 7

1.3. Alcance 7

2. Estado del Arte 8

2.1. Lógica difusa 8

2.2. Resúmenes Lingüísticos 14

2.3. Trabajos relacionados 23
2.4. Metodología para minería de datos: CRISP-DM 25

3. Propuesta de Solución **28**
3.1. Comprensión del negocio 28
3.2. Compresión de los datos 30
3.3. Preparación de los datos 34
3.4. Modelado 34

4. Implementación y Validación **47**
4.1. Evaluación 47
4.2. Despliegue 60

Conclusiones **62**

Bibliografía **67**

Anexos **71**

A. Tabla de trabajos recientes sobre resúmenes lingüísticos con aplicaciones a datos reales **71**

B. Todas las posibles frases atemporales **73**

C. Elementos del archivo de configuración de párrafos **77**

D. IDs y etiquetas utilizadas en el código de los tipos de frases generadas **78**

E. Imágenes de resultados de encuestas **79**

Lista de Tablas

2.1. Ejemplos de frases para protoformas Tipo 1 y Tipo 2 de resúmenes lingüísticos. 16

2.2. Resumen de las correspondencias entre KDD, SEMMA y CRISP-DM . . 26

3.1. Diccionario de datos del archivo de la Figura 3.4 32

3.2. Elementos del archivo de entrada con los valores estadísticos y de tendencia de todos los riegos considerados que explica la imagen 3.5 33

3.3. Elementos del archivo de configuración con los valores de los *sets* difusos que se muestran en la figura 3.11a 43

4.1. Ejemplos de resúmenes temporales obtenidos con su valor de verdad, cobertura y PVM agrupados por temática del resumen 48

4.2. Ejemplos de resúmenes obtenidos con su valor de verdad. La cobertura es 1 para todas las frases 49

4.3. Ejemplos de párrafos de resúmenes obtenidos, con sus correspondientes valores de verdad 49

4.4. Resultados de la evaluación de frases por cada tema 53

4.5. Resultados de la evaluación de Máximas de Grice en el reporte 54

4.6. Valores de la Tabla 4.1 luego de la modificación a partir de los resultados de las encuestas 58

4.7. Valores de la Tabla 4.2 luego de la modificación a partir de los resultados de las encuestas 59

4.8. Valores de la Tabla 4.3 después de las encuestas 59

Lista de Figuras

1.1. Ejemplo de calicata a cielo abierto. 5

1.2. Ejemplo de sensor FDR para medir la humedad a diferentes profundidades de un cultivo 6

2.1. Gráficos de pertenencia para los conjuntos de “cargada” y “descargada” del ejemplo de la batería 10

2.2. Botellas del desierto del ejemplo de difusividad versus probabilidad . . . 11

2.3. Método de Mamdani para inferencia difusa. 13

2.4. Ejemplo *GLMP*, donde los rectángulos/cuadrados son *CPs* y las círculos/óvalos *PMs* 18

2.5. Ejemplo de función trapezoidal de pertenencia para un *set* difuso 22

2.6. Modelo de proceso CRISP-DM. 27

3.1. Gráfico de ejemplo de representación Líneas de Gestión en el Reporte . . 29

3.2. Gráfico de ejemplo de datos del sensor en un cultivo de paltos 30

3.3. Gráficos de datos de sensores del cultivo de paltos agrupados e identificación de riegos 31

3.4. Ejemplo del contenido del archivo de entrada con el detalle de los riegos . 32

3.5. Ejemplo del contenido del archivo de entrada con los valores estadísticos y de tendencia de todos los riegos considerados 33

3.6. GLMP para el riego de cultivos. Los círculos representan PMs y los rectángulos CPs 36

3.7. Definiciones de algunos *sets* difusos utilizados. 39

3.8. Matriz de reglas difusas para la deducción de los valores de las Líneas de Gestión 40

3.9. Matriz de reglas difusas para la deducción de los valores de la Calidad de Riego 41

3.10. Estructura del programa encargado de la generación de resúmenes en párrafos. 41

3.11. Ejemplo de archivo de configuración de *sets* difusos y representación gráfica del *set* difuso de duración de riegos. 42

3.12. Archivo de configuración reglas IF-THEN. 44

3.13. Archivo de configuración de párrafos. 45

4.1. Ejemplo cálculo e interpretación de palabra con el método del centroide. . 50

4.2. Ejemplo de pregunta sobre duración de riegos de la Encuesta 1. 51

4.3. Ejemplo de pregunta sobre tiempo entre riegos de la Encuesta 2. 52

4.5. Cambio de set difuso del promedio de la Distancia a la CC. 54

4.4. Ejemplo de set de preguntas sobre los datos de Paltos1b de la Encuesta 3. 55

4.6. Cambio de set difuso de eventos bajo el NR. 56

4.7. Cambio de set difuso de percolación profunda. 57

4.8. Ejemplo del cálculo e interpretación de etiqueta con el método del centroide luego del cambio en los *sets* difusos. 58

4.9. Ejemplo de párrafos en un reporte. 60

Glosario

- **Calificador**. Componente de los resúmenes lingüísticos representado como una expresión lingüística que describe a las entidades, semánticamente representado como un conjunto difuso y determinando un subconjunto (difuso) de los datos. Por ejemplo: “bueno” para el atributo “rendimiento”.
- **Capacidad de campo (CC)**. Es la línea de gestión que indica la cantidad máxima de humedad que puede retener el campo donde se encuentra el cultivo.
- **CP (*computational percepction*)**. Es la percepción de parte de un fenómeno que se modela en un *GLMP*.
- **CRISP-DM (*cross industry standard process for data mining*)**. Es una metodología para proyectos de minería de datos.
- **Cuantificador**. Componente de los resúmenes lingüísticos que determina el tamaño relativo de la muestra de datos. Por ejemplo: “Muchos”, “Varios”.
- **Descriptor**. Componente de los resúmenes lingüísticos representado como una expresión lingüística que describe a las entidades, semánticamente representado como un conjunto difuso. Por ejemplo: “bajo” para el atributo “salario”.
- **GLMP (*granular linguistic model of a phenomenon*)**. Es una red de CP y PM que permite diseñar descripciones lingüísticas a partir de datos.
- **Línea de gestión (LG)**. Son valores asignados por un agrónomo y obtenidos dependiendo del tipo de cultivo y tierra. Estos valores permiten tener un control de los rangos en donde se mueve la humedad de la tierra.

- **Lógica Difusa o Borrosa**. Lógica basada en el concepto en que las cosas pueden ser y no ser. Esta lógica sienta el hecho de que entre si y no, o 1 y 0, puedes existir valores intermedios.

- **Mineria de Datos**. Análisis de grandes cantidades datos para la obtención de información relevante.

- **Nivel de riego (NR)**. Es la línea de gestión que indica el nivel mínimo de humedad deseado en el cultivo.

- **PM (*perception mapping*)**. Es un receptor de CPs y permite combinarlos para generar el GLMP.

- **Ponderador vectorial de métricas (PVM)**. Es la nueva métrica diseñada para escoger entre diferentes frases que hablen del mismo tema del negocio (con o sin temporalidad.

- **Protoforma**. Estructura básica de un resúmenes lingüísticos. Esta estructura determina la relación y restricciones entre los componentes del resumen.

- **Resumen Lingüístico**. Herramienta representada en frases en lenguaje de personas para la extracción de información desde una base de datos. Por ejemplo: “Muchas estudiantes tienen buenas calificaciones”.

- **Temporalizante**. Componente de los resúmenes lingüísticos inventado para este trabajo y representado como una expresión lingüística que describe temporalmente a las entidades, al igual que los descriptores y cuantificadores está representado como un conjunto difuso. Por ejemplo: “últimos días” o “los últimos riegos”.

Introducción

Las plantas requieren agua para desarrollarse. El agua cumple funciones importantes como mantener su temperatura similar a la del aire que la rodea, permitir la fotosíntesis, la mantención de un turgor celular adecuado y la incorporación de nutrientes desde el suelo a los órganos de crecimiento.

La optimización de recursos en la aplicación y control de riego es un punto de gran interés en la agronomía. Es por esto que se ha dedicado mucho tiempo y esfuerzo en la interpretación del comportamiento y adaptación de las plantas a los distintos cambios edafoclimáticos, y a la toma de decisiones sobre el riego informadas en tiempo real.

En este trabajo se plantea desarrollar un algoritmo generador de resúmenes lingüísticos para desplegar información relevante proporcionada por los sensores de humedad. Estos resúmenes funcionarán como un complemento a otros métodos para mostrar información como gráficos, tablas y estadísticas contenidos en un reporte general del cultivo.

El algoritmo utiliza lógica difusa para la confección de frases y párrafos, además de tener como entrada datos temporales de los sensores de humedad. Las técnicas utilizadas son variaciones de algunas propuestas en la literatura y explicadas en el estado del arte, como las series de tiempo de Janusz Kacprzyk y Sławomir Zadrozny, y modelar un *GLMP*.

Para la validación de las frases se utilizan métricas como la pertenencia a un conjunto difuso, valor de verdad y cobertura modificados, y encuestas donde se miden las máximas de Grice.

El presente informe se estructura en cuatro grandes capítulos. En el primero se presenta el problema que se busca resolver, los objetivos y el alcance del trabajo realizado. Posteriormente, en el segundo capítulo se expone el estado del arte, donde se muestran técnicas y

trabajos similares, y que son la base teórica para lo desarrollado. En el tercer capítulo se da a conocer la propuesta diseñada, en donde además se explican brevemente el negocio y los datos a utilizar. El cuarto capítulo contiene la implementación de la propuesta, en donde se enseñan los resultados obtenidos y la forma de evaluar para obtenerlos. Por último se postulan algunas conclusiones del trabajo realizado, con algunas propuestas de trabajo futuro.

Capítulo 1

Definición del Problema

En este capítulo se explica a grandes rasgos la necesidad agronómica de agua en los cultivos, las formas con la que se mide el nivel de humedad, y la búsqueda de más precisión y tecnología para mejorar los procesos actuales.

1.1. Presentación del Caso

Como mencionan Callejas R. *et al* en su libro [11], las plantas requieren agua para desarrollarse. El agua cumple funciones importantes como mantener su temperatura similar a la del aire que la rodea, permitir la fotosíntesis, la mantención de un turgor celular adecuado y la incorporación de nutrientes desde el suelo a los órganos de crecimiento. El requerimiento de agua puede ser suplido por las precipitaciones y el ascenso capilar de agua desde napas freáticas superficiales.

La optimización de recursos en la aplicación y control de riego es un punto de gran interés en la agronomía. Es por esto que se ha dedicado mucho tiempo y esfuerzo en la interpretación del comportamiento y adaptación de las plantas, a los distintos cambios edafoclimáticos y a la toma de decisiones sobre el riego informadas en tiempo real.

La mayoría de los especialistas que estudian este recurso hídrico concuerdan en que lo único que ocurrirá en el futuro es que se hará paulatinamente más escaso, porque el aumento de la población implica un mayor consumo de agua dulce y de la demanda de alimentos,

lo que implica un incremento en la utilización del riego. Por lo que se hace esencial implementar sistemas de distribución de agua eficientes e incorporar el "riego inteligente". Con la aplicación de sistemas de riego eficaces y que proporcionen cantidades concordantes con los requerimientos de las plantas, se permite ahorrar agua y energía, evitar pérdidas de nutrientes por lixiviación, y aumentar los rendimientos y calidad de producción.

Una buena programación del riego consiste en estimar la cantidad de agua y momento adecuado requeridos por el cultivo. Consta de dos etapas, la programación propiamente tal (predicción) y el control de ésta, a través de cuantificación de la humedad del suelo y/o el estado hídrico de la planta.

En la etapa predictiva se busca establecer *a priori* los tiempos de duración y frecuencia de riegos que permitan obtener un desarrollo adecuado de los cultivos. Para esto se necesita considerar factores como condiciones del clima, características propias de los cultivos y otras propias del suelo.

La forma tradicional de obtener información, y utilizada aún, es mediante el análisis e interpretación cualitativo del aspecto de la tierra cercana a la planta realizando calicatas o barrenos (ver Figura 1.1). Este método consiste en cavar de forma ancha y profunda cerca de la planta para tomar diferentes muestras de suelo. Para medir bien es requerida la experiencia práctica del observador, ya que se debe interpretar al tacto y de forma visual la sensación de humedad según la textura y aspecto del suelo.

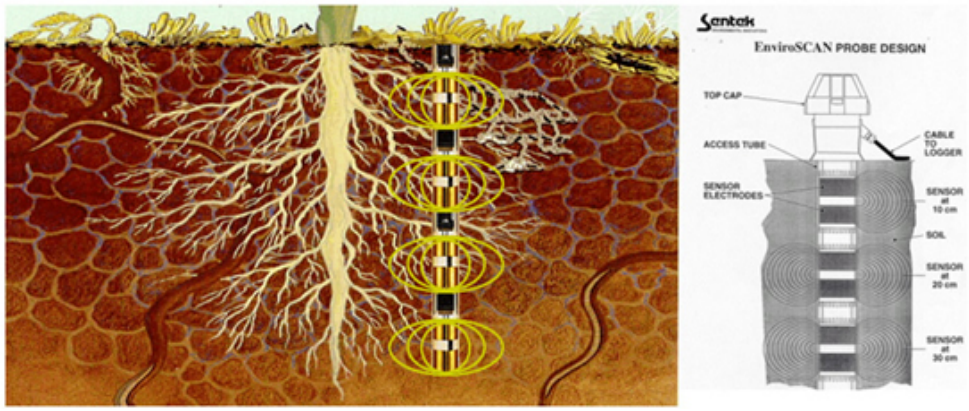
Dadas estas limitaciones y subjetividades, surge una de las maneras más recientes y tecnológicas de saber el estado hídrico de un cultivo: sensores de humedad (ver Figura 1.2), que mediante mediciones a diferentes profundidades proporcionan datos sobre la cantidad de agua aproximada en la tierra cercana de forma continua, rápida y casi sin alterar las condiciones del suelo circundante.

Basándose en la dinámica de extracción de agua de las plantas y los parámetros que indican el balance hídrico del suelo, se establecen umbrales de referencia o líneas de gestión. Estos umbrales reflejan algunas necesidades o capacidades de las plantas y la tierra en donde se encuentran, tales como nivel de relleno, capacidad de campo y evapotranspiración, entre otros.

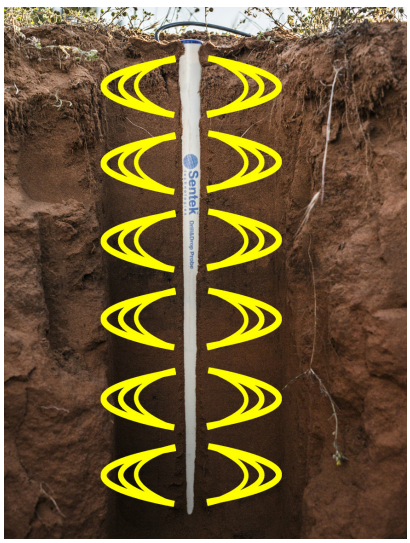
Figura 1.1: Ejemplo de calicata a cielo abierto.

Dado este escenario, se requiere implementar un algoritmo que genere conocimiento respecto al estado de hidratación de un campo de plantas, el cual utilice datos obtenidos de los sensores de humedad e/o información proporcionada por una estación meteorológica cercana. El programa que nace de este algoritmo debe informar acerca de los períodos, ubicaciones y cantidades de riegos, con los consumos y pérdidas de agua aproximados, apoyando así la planificación y permitiendo una correcta hidratación de las plantas y un uso óptimo del agua de riego.

Como mencionan van der Meulen, M. *et al* en [35] a diferencia de lo que se pensaba y asumía de estudios anteriores, la frase "una imagen vale más que mil palabras" no es siempre aplicable en todos los contextos, y en ciertos casos desplegar información en modo de texto es entendible e interpretable de manera más rápida y eficaz. Esto sugiere que una buena forma de expresar nuevo o mayor conocimiento a partir de los datos es mediante frases o párrafos en lenguaje natural.



(a) Esquema de utilización del sensor



(b) Sensor instalado en un campo

Figura 1.2: Ejemplo de sensor FDR para medir la humedad a diferentes profundidades de un cultivo

Tomando estos antecedentes en cuenta, se plantea desarrollar un algoritmo generador de resúmenes lingüísticos para desplegar información relevante proporcionada por los sensores de humedad. Estos resúmenes funcionarán como un complemento a otros métodos de mostrar la información como gráficos, tablas y estadísticas contenidos en un reporte general del cultivo.

1.2. Objetivos

Objetivo General

Apoyar la toma de decisiones en un caso real de riego de cultivos, mediante el diseño, implementación y evaluación de un algoritmo generador de resúmenes lingüísticos, logrando que la satisfacción de los usuarios finales y/o la empresa sea mayor al 70 % considerando las Máximas de Grice.

Objetivos Específicos

- Definir protoformas y fórmulas del valor de verdad adecuados para el tipo de resúmenes lingüísticos a generar, lo que permita obtener un valor de verdad mínimo de 75 % para cada resumen.
- Definir descriptores, calificadores y cuantificadores adecuados para la generación de resúmenes lingüísticos asociados al riego de cultivos, considerando elementos propios de la lógica difusa, el negocio y el entendimiento de los usuarios.
- Ayudar a planificar y optimizar el uso del agua y la calidad esperada de las plantas mediante la entrega de conocimiento obtenido de los datos arrojados por los sensores al proporcionar información relevante y no trivial.

1.3. Alcance

Dado que los resúmenes lingüísticos apuntan a ser una síntesis de los datos, se espera que estos sean un complemento de los gráficos y estadísticas desplegadas en un reporte, es decir, los resúmenes no debiesen explicar la totalidad de la información de todos los datos por sí mismos.

Además, dado que la información a utilizar es, en su mayoría, sobre la humedad circundante a las plantas, se espera entregar frases que apunten a esta característica, evitando mencionar sobre otros procesos importantes de las plantas los cuales requieren de conocimiento mucho más experto y especializado para ser analizados y comprendidos a cabalidad.

Capítulo 2

Estado del Arte

En este capítulo se explican las bases teóricas para el desarrollo del trabajo las cuales son resúmenes lingüísticos y lógica difusa. Además se mencionan trabajos relacionados con los temas a tratar y su aporte al conocimiento o implementación de la teoría a presentar.

2.1. Lógica difusa

La Lógica Difusa, o *fuzzy logic* en inglés, es una lógica multivaluada que permite definir valores intermedios entre evaluaciones convencionales como verdadero/falso, sí/no, alto/bajo, etc. Nociones como "bastante altas" o "muy rápidas" pueden ser formuladas matemáticamente y procesadas por computadores, con el fin de aplicar una manera más humana de pensar y representar la incertidumbre y la vaguedad [15].

Si se supone que se tiene una batería completamente cargada, es decir tiene un 100 % de carga. Ahora, se conecta la batería a un artefacto eléctrico, de modo que consume parte de la energía disminuyendo la carga a un 90 %. ¿La batería sigue cargada? Si, aún lo está, pero ya no se puede decir que está completamente cargada, ahora está un poco más cerca de poder decirse que está descargada. Si continuamos usando la batería la carga recorrerá todos los valores porcentuales perteneciendo cada vez menos al estado de "cargada", hasta que alcance el estado de completamente "descargada".

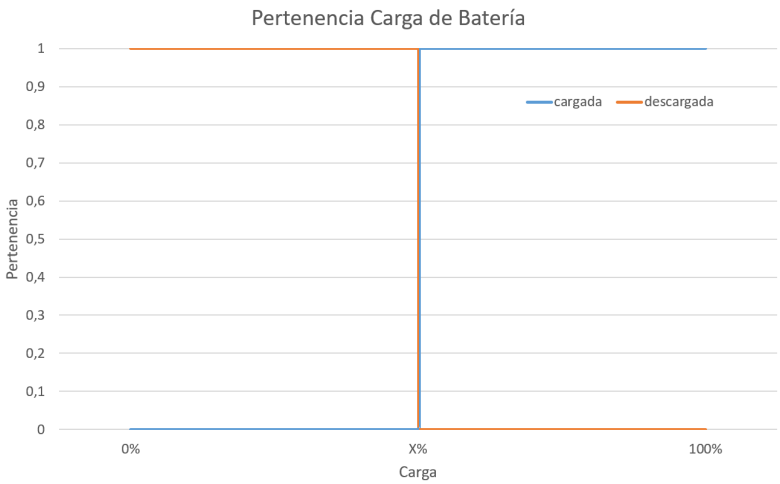
Entonces, ¿en qué momento la batería pasó de estar completamente cargada a estar completamente descargada? Según la lógica binaria tradicional, debiese existir un valor de inflección que refleje el paso de estado cargado a descargado que se representa como el valor X % de la Figura 2.1a, pero para la lógica difusa no existe una respuesta determinista.

Considerando el escenario planteado, ¿Entre 1 % y 99 % la batería está cargada o descargada? La respuesta difusa es que la batería está cargada y descargada a la vez como se observa en la Figura 2.1b. A diferencia de la lógica aristotélica un predicado lógico puede ser verdadero y falso al mismo tiempo, lo que se logra con un grado de pertenencia. Para este ejemplo, si se considera la fase inicial (en donde se consumió energía dejando la batería en 90 %), el predicado "la batería está cargada" es verdadero en un 90 % y falso en un 10 %, debido a que la carga es parte del conjunto "cargada" en un 90 % y "descargada" en un 10 % a la vez.

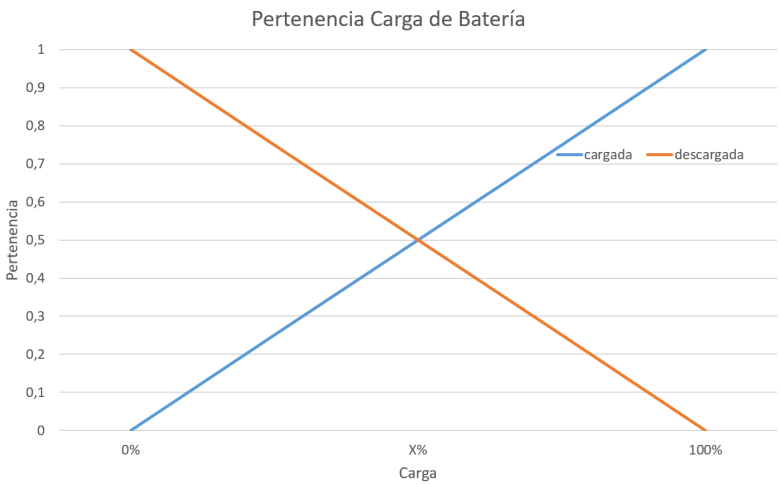
Con ayuda de la lógica difusa se puede explicar la lógica desde una perspectiva más humana, en donde las palabras del lenguaje natural pueden participar sin ser necesariamente tajantes o absolutas, es decir tienen un cierto grado de participación. Esto último no excluye los casos en donde las palabras son absolutas o participan completamente de un conjunto. En el ejemplo de la batería, en un comienzo estaba en un 100 % de carga, por lo que era parte del estado "cargado" por completo y pertenencia en un 0 % del estado "descargado".

Los conceptos empleados en lógica difusa y probabilidades están relacionados en cierto modo, pero son totalmente diferentes. La probabilidad representa información sobre frecuencia de ocurrencias relativas de eventos bien definidos sobre el total de eventos posibles. En cambio, el grado de pertenencia difuso representa las similitudes entre evento, donde las propiedades de estos eventos no están definidas de forma precisa.

La diferencia queda más clara con un ejemplo que entrega Carlos González Morcillo en [15]: un superviviente de un accidente se encuentra perdido y deshidratado en el desierto. En su camino encuentra dos botellas llenas de líquido, etiquetadas con "D" (difusa) y "P" (probabilista), además de un valor de 0,8 en ambas como se muestra en la Figura 2.2. La botella D difusa está etiquetada indicando que contiene líquido potable con un grado de pertenencia 0,8, mientras que la botella P probabilista está etiquetada señalando que contiene fluido potable con una probabilidad 0,8 de ser un líquido potable. ¿Cuál debería elegir el superviviente?.



(a) binario



(b) difuso

Figura 2.1: Gráficos de pertenencia para los conjuntos de “cargada” y “descargada” del ejemplo de la batería

La respuesta es que debe escoger la botella D, ya que indica que el líquido que contiene es bastante similar a otros que son potables. Este valor numérico depende de la función de pertenencia asociada al concepto de “líquido potable”. Supongamos que la función de pertenencia asocia 1,0 al agua pura, por lo que un valor de 0,8 indicaría que la botella D contiene agua no totalmente pura, pero todavía potable (o al menos no es un líquido perjudicial para el organismo). En cambio, la probabilidad asociada a la botella P indica que, tras realizar un alto número de experimentos, el contenido de la botella P es potable

el 80 % de las veces pero el otro 20 % el líquido no es potable.

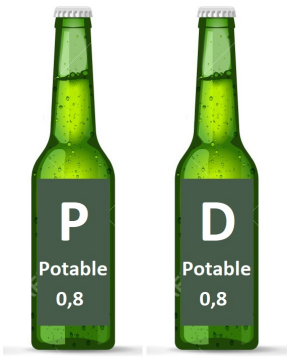


Figura 2.2: Botellas del desierto del ejemplo de difusividad versus probabilidad

Los conjuntos difusos se parecen mucho a los conjuntos finitos, salvo por la diferencia clave que es la capacidad de semi-pertenencia. Es por esto que comparten la mayoría (no todas) de sus propiedades y características.

Sobre esas características se pueden definir operaciones básicas para el valor de verdad como:

- $T(\neg a) = 1 - T(a)$
- $T(a \wedge b) = min(T(a), T(b))$
- $T(a \vee b) = max(T(a), T(b))$
- $T(a \Rightarrow b) = min(1, 1 + T(b) + T(a))$
- $T(a \Leftrightarrow b) = 1 - |T(a) - T(b)|$

A los conjuntos difusos que cumplen con las propiedades de asociatividad, conmutatividad, elemento neutro y monotonicidad se les llaman normas triangulares o t-normas (*t-norms*). Las más utilizadas son:

- **Mínimo**: $T(x, y) = Min(a, b)$
- **Producto (algebraico)**: $Prod(x, y) = x \cdot y$
- **Operación de Lukasiewicz o Producto limitado**: $W(x, y) = Max(0, x + y - 1$

- **Producto drástico**: $Z(x,y) = \begin{cases} x & \text{si } y = 1 \\ y & \text{si } x = 1 \\ 0 & \text{en otro caso} \end{cases}$

Estas t-normas se evalúan punto a punto en todo el universo y se relacionan por medio de las desigualdades de la condiciones 2.1:

$$Z(x,y) \leq W(x,y) \leq Prod(x,y) \leq Min(x,y) \qquad \forall x, y \in [0,1] \tag{2.1}$$

También existe la inferencia difusa, en donde se pueden emplean reglas difusas *if-then*, además de las operaciones de conjuntos anteriormente mencionadas, para obtener un resultado en base a premisas.

Una regla difusa (regla de producción difusa *if-then*) es expresada simbólicamente como:

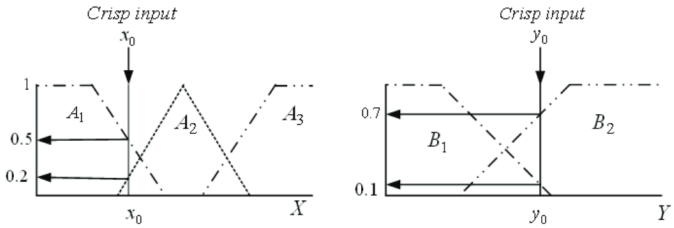
IF < proposición difusa (**antecedente**)> THEN < proposición difusa (**consecuente**) >

donde <proposición difusa> puede ser una proposición difusa atómica o compuesta.

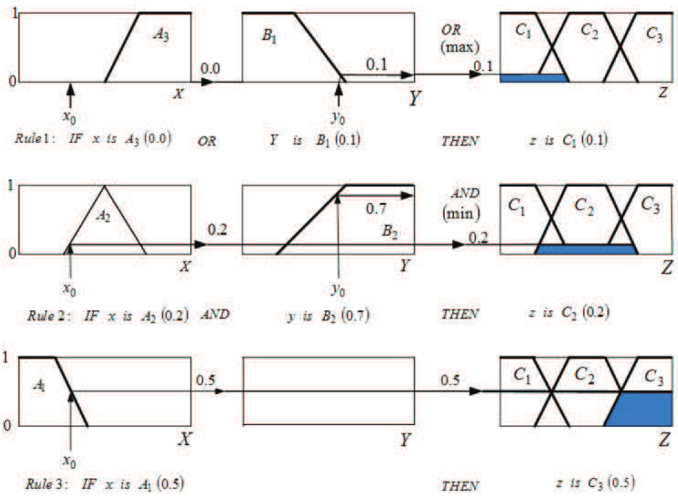
Basado en la interpretación de ($A \rightarrow B$) "A junto con B" o "A y B ambos están" se pueden utilizar las cuatro funciones t-normas para resolver la relación. Uno de los métodos más utilizado es el propuesto por Ebrahim Mamdani en 1975 [25] que consiste en cuatro pasos:

1. ***Fuzificación***: de las variables de entrada, lo que implica considerar los valores iniciales a usar y determinar el grado de pertenencia de estas entradas a los conjuntos difusos asociados.

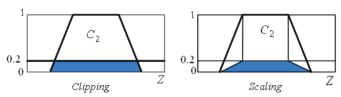
2. **Evaluación de Reglas**: se consideran las entradas anteriores y se aplican a los antecedentes de las reglas difusas. Si una regla tiene múltiples antecedentes, se utiliza el operador OR (max) o AND (min) para obtener un único número que represente el resultado de la evaluación. Este número (el valor de verdad) se aplica al consecuente mediante un recorte (*clipping*) o escalado según el valor de verdad del antecedente. El método más comúnmente utilizado es el recorte que como su nombre lo dice "corta" el consecuente con el valor de verdad del antecedente; el escalado proporciona



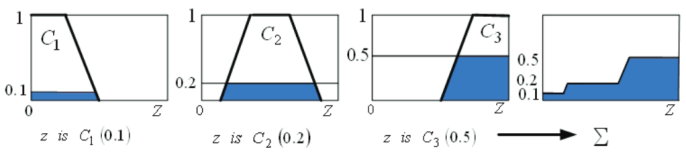
(a) *Fuzificación*



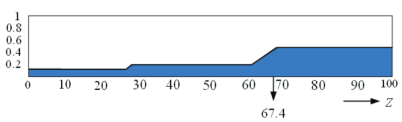
(b) Evaluación de Reglas



(c) Corte o Escalado



(d) Agregación de los valores de salida de las Reglas)



(e) *Defuzificación*

Figura 2.3: Método de Mamdani para inferencia difusa.

un valor más preciso, preservando la forma original del conjunto difuso. Este último se obtiene multiplicando todos los valores por el valor de verdad del antecedente.

3. **Agregación**: de las salidas y unificación de todas las reglas. Se combinan las funciones de pertenencia de todos los consecuentes para obtener un único conjunto difuso por cada variable de salida.

4. ***Defuzificación***: para expresar el resultado mediante un valor. Existen varios métodos para *defuzificar*, uno de los más usados es el centroide, que calcula el punto donde una línea vertical divide el conjunto en dos áreas con igual masa o área.

Ion Iancu [17] explica en su trabajo el funcionamiento de esta regla y muestra los gráficos de la Figura 2.3 que ejemplifican los pasos para la inferencia difusa.

2.2. Resúmenes Lingüísticos

Un resumen lingüístico es una manera intuitiva para los humanos de expresar información mediante texto generado en base a datos proporcionados [39]. Es decir, se crea una frase a partir de la información, la cual considera palabras del lenguaje natural tales como "mucho", "bajo" o "adecuado".

Las frases tienen la capacidad de proporcionar información de una forma natural para las personas, a diferencia de estadísticas o gráficos los cuales regularmente requieren de conocimiento experto para ser interpretados. Por esto es deseable que un sistema sea capaz de generar automáticamente los resultados en este lenguaje.

Para obtener dichos resúmenes se tiene como primer desafío el definir un conjunto de palabras que representen la información obtenida de ciertos atributos. Por ejemplo, "mucho" es un adjetivo que puede representar la característica de cantidad.

Primeramente, el trabajo de R.Yager [38] y a posterior extendido por J.Kacprzyk *et al* [20] definen la base de los resúmenes lingüísticos como lo siguiente:

- $Y = \{y_1, \ldots, y_i, \ldots, y_n\}$ es un conjunto de objetos (registros) o entidades en una base de datos. (Ej: Los registros de cada ciclo de riego)

- $A = \{A_1, \ldots, A_j, \ldots, A_m\}$ representa un conjunto de atributos que caracterizan los objetos de Y. (Ej: duración de riego o tiempo entre riegos son posibles atributos para los registros de riegos)

- $X_j = \{x_1, \ldots, x_k, \ldots, x_o\}$ son los posibles valores que puede tomar el atributo A_j (Ej: valores del universo de "duración de riegos"son los números reales tales que: $x_k \geq 0$)

Un resumen lingüístico del *set* de datos $D = \{A_1(y_1) \ldots A_j(y_i) \ldots A_m(y_n)\} = \{d_1, \ldots, d_{n*m}\}$ consiste en:

- Un descriptor S, una expresión lingüística o predicado que describe a las entidades, semánticamente representado como un conjunto difuso y definido en el dominio de A_j. (Ej: "bajo"para el atributo "salario").
- Un cuantificador de la cantidad Q, un cuantificador lingüístico que explica la cantidad considerada. (Ej: "la mayoría").
- Un calificador R, una expresión lingüística o predicado opcional que describe a las entidades, semánticamente representado como un conjunto difuso y definido en el dominio de A_j y que determina un subconjunto difuso de Y. (Ej: "joven"para el atributo "edad").
- Un valor de verdad T, una medida que representa la validez o veracidad de un resumen lingüístico. (Ej: 0.77)

De esta forma el núcleo de un resumen lingüístico, también llamado protoforma, es una proposición cuantificada como lo expresa Zadeh [40], y puede escribirse como:

- Q y_i son S
- Q R y_i son S

Un ejemplo de estas dos primeras protoformas se muestra en la Tabla 2.1.

Es importante obtener una medida que entregue algún valor sobre la veracidad o falsedad de la declaración, ya que sin un valor de verdad se podría asumir que todas las frases son siempre completamente verdaderas. Dado que los resúmenes están constituidos por cuantificadores, descriptores y calificadores en lenguaje natural, estos no cumplen con una lógica binaria (absoluta o matemática), donde las cosas son o no son. El lenguaje natural trabaja sobre la base de que las cosas pueden ser y no ser a la vez (lógica multivariada).

Tabla 2.1: Ejemplos de frases para protoformas Tipo 1 y Tipo 2 de resúmenes lingüísticos.

Protoforma	Estructura	Q	y_i	R	S	Resumen Lingüístico
Tipo 1	Qy_i son S	Muchos	Estudiantes	-	Ineficientes	Muchos Estudiantes son Ineficientes
Tipo 2	QRy_i son S	Muchos	Estudiantes	de Pregrado	Ineficientes	Muchos Estudiantes de Pregrado son Ineficientes

El valor de verdad T de un resumen lingüístico es una medida de este que entrega una cantidad numérica decimal entre 0 (completamente falso) y 1 (completamente verdadero), inclusives. Este valor permite saber el porcentaje de veracidad de los resúmenes con respecto a los datos y así ser capaz de comparar entre frases.

De las frases de la Tabla 2.1 se podrían tener como ejemplo los siguientes valores de verdad ficticios:

- $T(Muchos\ Estudiantes$ son $Ineficientes) = 0,8$
- $T(Muchos\ Estudiantes\ de\ Pregrado$ son $Ineficientes) = 0,65$

Al observar ambas frases se observa que, si se considera el mismo *set* de datos para realizar los cálculos, la primera frase es más verdadera que la segunda.

Dependiendo de la estructura del resumen (protoforma) existen diferentes formas de calcular el valor de verdad. Las protoformas Tipo 1 y Tipo 2 se calculan como muestran las ecuaciones 2.2 y 2.3, respectivamente.

$$T(Qy_i \text{ son } S) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_S(y_i)\right) \tag{2.2}$$

$$T(QRy_i \text{ son } S) = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_S(y_i)\wedge\mu_R(y_i)}{\sum_{i=1}^{n}\mu_R(y_i)}\right) \tag{2.3}$$

en donde $\wedge$ simboliza la operación “mínimo”, la que puede ser reemplazada por cualquier otra t-norma, y $\mu_Q(yi)$ es la función de pertenencia que representa el cuantificador lingüístico Q. Por ejemplo, para el cuantificador “muchos” la función de pertenencia podría estar

dada como:

$$\mu_Q(x) = \begin{cases} 1 & \text{para } x \geq 0,8 \\ 2x - 0,6 & \text{para } 0,3 < x < 0,8 \\ 0 & \text{para } x \leq 0,3 \end{cases} \tag{2.4}$$

Ya que las protoformas son solo una base para crear los resúmenes, se debe considerar qué tipo de frase se quiere entregar y cuál es necesaria para cada problema específico. Para ayudar a tomar la decisión es que existen diferentes tipos de métricas que permiten comprobar que los parámetros elegidos en la confección de las partes de los resúmenes son correctos.

Con el paso del tiempo se comenzó a utilizar una variación del método para generar resúmenes lingüísticos, esta variante consiste en generar un modelo en el cual cada una de sus partes abarcan diferentes aspectos del problema a diferentes generalidades. Este método se conoce como *Granular Linguistic Model of a Phenomenon* (GLMP) y permite abarcar todo el problema con mayor facilidad. Uno de los primeros en implementarlo es Triviño G. en su trabajo [33]

Granular Linguistic Model of a Phenomenon

El *GLMP* está basado en *Computational Theory of Perceptions (CTP)* que permiten desarrollar sistemas computacionales capaces de generar descripciones lingüísticas a partir de datos. El modelo es una red de *Perceptions Mapping (PMs)* como se observa en la Figura 2.4. Cada *PM* recibe un conjunto de *Computational Perceptions (CPs)* de entrada y/o datos, y devuelve un *CP* de salida. En esta red, cada *CP* cubre aspectos específicos del fenómeno con cierto grado de granularidad o generalidad. Utilizando diferentes funciones de agregación y expresiones lingüísticas, el paradigma *GLMP* permite modelar computacionalmente las percepciones.

Computational Perception (CP)

Un CP es el modelo computacional de una unidad de información adquirida por el diseñador sobre el fenómeno a modelar. Si fue generado a partir de solo datos es un CP de primer nivel (1-CP), en cambio si al menos tiene otro CP como entrada pasa a ser un CP de segundo nivel (2-CP). En general, los CP se corresponden con partes específicas del

fenómeno en ciertos grados de especificidad y están formados por una dupla (A, W) donde:

- $\mathbf{A} = (a_1, a_2, \ldots, a_n)$ es un vector de expresiones lingüísticas (palabras u oraciones en lenguaje natural) que representa todo el dominio lingüístico del *CP*. Cada a_i describe el valor del *CP* en cada situación. Estas oraciones pueden ser simples, por ejemplo a_i = "La velocidad del vehículo es alta", o compleja como a_i = "Durante la interacción, a veces la ejecución de la maniobra ha sido mala" .
- $\mathbf{W} = (w_1, w_2, \ldots, w_n)$ es un vector de grados de validez $w_i \in [0, 1]$ asignado a cada a_i en el contexto específico. w_i es el valor de verdad de a_i que sirve para validar la frase en una situación.

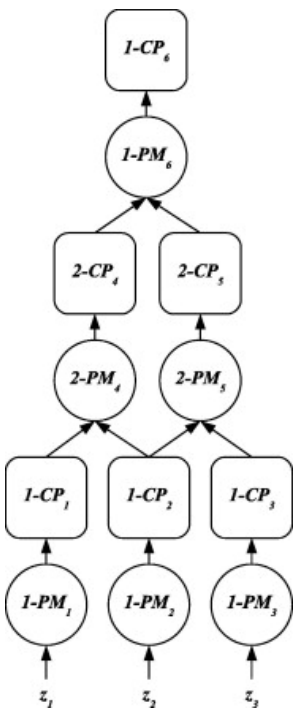


Figura 2.4: Ejemplo *GLMP*, donde los rectángulos/cuadrados son *CPs* y las círculos/óvalos *PMs*

Perception Mapping (PM)

Los *PM* se usan para crear y unir *CP*, están formados por la tupla (U, y, g, T) donde:

- **U** es un vector de *CP* de entrada, $U = (u_1, u_2, \ldots, u_n)$, donde $u_i = (A_{u_i}; W_{u_i})$. En el caso especial de un *PM* de primer orden (*1PM*), que son las entradas al *GLMP* y

representan valores $z \in R$ entregados por un sensor físico u obtenidos de una base de datos.

- **y** es el resultado o salida del *CP*, donde $y = (A_y, W_y) = (a_1, w_1), (a_2, w_2), \ldots, (a_{n_y}, w_{n_y})$.
- **g** es una función de agregación empleada para calcular el vector W asignados a cada elemento en y. Existen muchos tipos diferentes de funciones de agregación. Por ejemplo, g podría implementarse usando un conjunto de reglas difusas. En el caso de *1PM*, g se genera utilizando un conjunto de funciones de membresía.
- **T** es un algoritmo de generación de texto que permite producir las oraciones en A_y. En casos simples, T es solo una plantilla.

Métricas para Resúmenes Lingüísticos

Como primer acercamiento, están las cinco medidas de calidad para evaluar resúmenes lingüísticos para bases de datos descritas y resumidas por Kacprzyk, J. *et al* [20] las cuales son:

- **Valor de Verdad** (T): es el criterio de validación explicado anteriormente que fue introducido por Yager (ecuaciones 2.2 y 2.3), siendo el más conocido y empleado.
- **Valor de Imprecisión** (T_2): es un criterio de validez, que como su nombre menciona, busca resúmenes que entreguen información poco imprecisa (no evidente). Por ejemplo, el resumen "en casi todos los días de invierno, la temperatura es bastante fría" tiene un valor de verdad muy alto, pero no es útil o no genera información valiosa.

  Para cada uno de los descriptores S_j que componen la frase se tiene:

$$ins(s_j) = \frac{card\{x \in X_j : \mu_{s_j}(x) > 0\}}{card(X_j)} \tag{2.5}$$

  donde $card(.)$ denota la cardinalidad del conjunto (no difuso) correspondiente. Con lo que se obtiene el valor de imprecisión T_2 expresado en la ecuación 2.6 para el resumen.

$$T_2 = 1 - \sqrt[m]{\prod_{j=1,\ldots,m} ins(s_j)} \tag{2.6}$$

Se puede observar que el valor de imprecisión no depende de los datos, si no de la forma del resumen generado.

- **Valor de Cobertura (T_3)**: menciona cuántos objetos del *set* de datos "validan" lo expresado en el resumen que se forma con el descriptor S con respecto al total. Un ejemplo sería que si la cobertura tiene un valor igual a 0.15 significa que el 15 % de los objetos están consistentemente contenidos en el resumen en cuestión. El valor de esta medida depende del contenido de los datos y qué pertenencia tengan al compararlos con los *sets* difusos de los descriptores.

$$T_3 = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} h_i} \tag{2.7}$$

donde:

$$t_i = \begin{cases} 1 & \text{si } \mu_S(y_i) > 0 \text{ y } \mu_R(y_i) > 0 \\ 0 & \text{otro caso} \end{cases}$$

$$h_i = \begin{cases} 1 & \mu_R(y_i) > 0 \\ 0 & \text{otro caso} \end{cases}$$

- **Grado de Adecuación (T_4)**: describe qué tan característico es para el *set* de datos utilizado el resumen encontrado y se calcula como muestran las ecuaciones 2.8 y 2.9. Esto significa que si por ejemplo se tienen datos sobre trabajadores, si el 50 % de ellos son menores de 25 años y el 50 % son altamente calificados, entonces se puede esperar que los empleados que sean menores de 25 años y altamente calificados estén cerca del 25 %. Sin embargo, si el grado de adecuación es de, por ejemplo, 0,39 (es decir, el 39 % tiene menos de 25 años y está altamente calificado), entonces el resumen encontrado refleja una relación interesante, no esperada totalmente en nuestros datos. Un ejemplo más extremo y que demuestra su importancia es que un resumen trivial como "el 100 % de la venta es de cualquier artículo" tiene valor de verdad 1 pero su grado de adecuación es igual a 0. Para determinar su valor se tiene:

$$r_j = \frac{\sum_{i=1}^{n} h_i}{n} \qquad , j = 1, ..., n \tag{2.8}$$

donde:

$$h_i = \begin{cases} 1 & S_j(y_i) > 0 \\ 0 & \text{otro caso} \end{cases}$$

Obteniendo el valor del grado de adecuación de la ecuación 2.9:

$$T_4 = abs(\prod_{j=1,...,m} r_j - T_3) \tag{2.9}$$

- **Largo de un resumen (T_5)**: es relevante porque entre mayor cantidad de descriptores en un mismo resumen, es más difícil de comprender para una persona. Entre muchas formas de calcular esta métrica, los autores sugieren la forma que muestra la ecuación 2.10.

$$T_5 = 2(0,5^{card(S)}) \tag{2.10}$$

Otras medidas de interés que se pueden aplicar a los resúmenes, utilizadas por Kacprzyk [18] principalmente en resúmenes lingüísticos para series de tiempo, son las que se explican a continuación.

- **Grado de Especificidad**: la especificidad de un resumen (ecuación 2.11) es la medida en la que los conceptos del resumen definen e identifican de manera clara los datos involucrados. Es decir, entre mayor sea el valor de especificidad más claro se puede estar del subconjunto de datos al que se está apuntando. Entre mayor sea el valor de especificidad más específico es éste y más cercano al valor 1 estará, en caso contrario al ser menos específico se acercará más al valor 0.

$$Sp(A) = \int_0^{\alpha_{max}} \frac{1}{card(A_\alpha)} d\alpha \tag{2.11}$$

donde α_{max} es el mayor grado de pertenencia en A, A_α es el α-nivel de A (ej: $A_\alpha = \{x : A(x) \geq \alpha\}$) y card($A_\alpha$) es la cantidad de elementos en A_α.

En el caso particular de definiciones de forma trapezoidal como muestra la Figura 2.5, el grado de especificidad se puede calcular como la ecuación 2.12.

$$Sp(A) = 1 - \frac{(c+d) - (a+b)}{2} \tag{2.12}$$

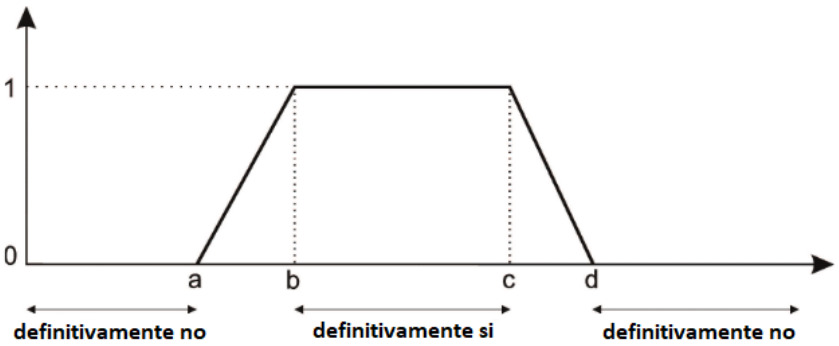


Figura 2.5: Ejemplo de función trapezoidal de pertenencia para un *set* difuso

- **Grado de Enfoque**: la forma extendida de resúmenes lingüísticos (con el calificador R) limita el espacio de búsqueda, ya que esta búsqueda se realiza en un subespacio limitado de todas las coincidencias (la mayoría) que cumplen una condición adicional dada por el calificador. El grado de enfoque mide cuántos registros cumplen con la propiedad R.

  Si el grado de enfoque es alto, entonces se tiene seguridad que el resumen se refiere a muchos objetos, por lo que es más general. En caso contrario, si el grado de enfoque es bajo, el resumen describe un patrón (local) que rara vez ocurre.

$$d_{foc}(QRy_i \text{ son } S) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_R(y_i)\right) \tag{2.13}$$

Ya que los resúmenes lingüísticos pueden desplegarse de diferentes formas, como por ejemplo en un reporte, es que es importante considerar métricas que también abarquen ese aspecto. De acuerdo con las máximas de Grice [16], un reporte de buena calidad debe contener cuatro máximos principales:

- **Máximo de Calidad**: tiene que ser verdad, no decir cosas con las que no se cuenta evidencia adecuada.
- **Máximo de la Manera**: tiene que ser claro, ordenado y evitar la ambigüedad.
- **Máximo de Relación**: tiene que ser relevante.
- **Máximo de Cantidad**: tiene que tener una extensión adecuada, evitar ser más informativo de lo que se requiere.

2.3. Trabajos relacionados

Los trabajos relacionados a resúmenes lingüísticos son cada vez más comunes y adquieren mayor relevancia, la mayoría aplicando la borrosidad para definirlos y generarlos.

En 1983, Zadeh L. [40] incorpora una de las bases para lo que se convertiría en los resúmenes lingüísticos al utilizar conjuntos difusos y definir conceptos y métricas claves para los cálculos relacionados a estos.

Somayajulu G. Sripada *et al* [32], entre los años 2001 y 2003, con el proyecto llamado SumTime[31], se buscó técnicas para mejorar la elaboración de resúmenes en lenguaje natural sobre datos meteorológicos basados en series de tiempo, deseando generar un método que produjera resúmenes efectivos sobre el clima.

En 2008, F. Pouzols, A. Barriga, D. Lopez y S. Sánchez [29] generan reportes en donde se incorporó mensajes para expresar el tráfico de red considerando parámetros como distribución para el rendimiento, duración, tamaño de transferencia, tamaño de paquete promedio, paquetes por flujo y para rendimiento calificado por tamaño de transferencia.

Van der Heide A., & Triviño G.[34], en 2009 intentan realizar resúmenes automáticos para la medición del consumo de energía, generando resúmenes que describen los datos de consumo eléctrico e intentando, además, aconsejar a usuarios sobre el comportamiento de consumo para reducir el gasto de dinero.

Kacprzyk J., Wilbik A. M., & Zadrożny S. [19] en 2008 , Kacprzyk J., & Wilbik A. M. [18] en 2010 y Kacprzyk J., & Zadrożny S [21] [22] [23] en 2009 , 2014 y 2016 realizan trabajos con resúmenes lingüísticos en series de tiempo, incorporando formas de medición y formalidades para su implementación, además de algunas pruebas en datos reales.

Una de las incorporaciones relevantes, es la realizada por J. Kacprzyk *et al* de una expresión temporal E_t, que para efectos de los cálculos se incorpora como un nuevo descriptor o calificador. El valor de verdad para un resumen Tipo 1 sería como el expresado en la ecuación 2.14 y un ejemplo podría ser “En los últimos semestres, la mayoría de los estudiantes

son ineficientes".

$$T(E_t, Qy_i \text{ son } S) = \mu_Q\left(\frac{\sum_{i=1}^{n} \mu_S(y_i) \wedge \mu_{E_t}(y_i)}{\sum_{i=1}^{n} \mu_{E_t}(y_i)}\right) \tag{2.14}$$

En 2010, Trivino G. *et al.* [33] intentan generar reportes de tráfico automovilístico para describir lo que sucede en una rotonda. Para esto modelaron el problema como un *Granular Linguistic Model of a Phenomenon (GLMP)* y utilizaron reglas *if-then* para inferir variables más complejas.

En 2010, Mendez Nunez S. *et al* [26] intentan generar reportes que contenga texto entendible por humanos para el análisis de riesgo y solvencia de finanzas. Para esto desarrollan un sistema que utiliza reglas *if-then*.

En 2011, Wilbik A. *et al* [36] trabajan generando resúmenes lingüísticos con datos temporales (series de tiempo) en el monitoreo de personas de la tercera edad que se encuentran en casas de reposo, permitiendo al personal a cargo tener una herramienta más rápida y natural de verificar el estado de los residentes.

En 2012, Alvarez Alvarez A. *et al* [7] generan reportes automáticos para la medición del tráfico vehicular en carreteras, en base a medidas básicas de parámetros de tráfico que se obtienen a partir de diferentes tipos de sensores, como cámaras de video, radar, mangueras presurizadas y bucles de entierro inductivo.

En 2012, Eciolaza L. *et al* [13] generan resúmenes para simulaciones de manejo mediante GLMP y reglas *if-then*. Para su trabajo utilizaron parámetros como velocidad del vehículo, posición lateral, ancho de vía, posición del volante, entre varios otros.

En 2012, Alvarez Alvarez A.. Trivino G., & Cordon O. [8] crean una máquina difusa de estados finitos basada en reglas *if-then* para modelar el caminar humano, en la que se utilizan valores de entrada comunes además del estado actual (o anterior) de la máquina como información. Además desarrollaron un procedimiento de aprendizaje genético para que la máquina difusa de estados finitos se ajuste a cada persona en particular.

En 2015, Novák V. [28] genera una caracterización de las series de tiempo en resúmenes lingüísticos considerando borrosidad, reglas *if-then* y un trabajo de identificación y segmentación de datos según patrones o tendencias.

En 2016, Sanchez Valdes D. *et al* [30] trabajan con series de tiempo modelando un *GLMP* para generar resúmenes lingüísticos de la actividad física de una persona considerando su tiempo de caminata. Esto lo consiguen, en mayor parte, gracias a datos de un sensor de movimiento (acelerómetros).

En 2016, Boran F. E., Akay D., & Yager R. R [10] generan una síntesis bastante completa del trabajo sobre resúmenes lingüísticos hasta la fecha, mencionando una recopilación de varios trabajos y métodos que se han utilizado en los últimos años en la creación e implementación de resúmenes lingüísticos, siendo algunos utilizados en casos reales.

Un resumen de algunos de los trabajos mencionados, en donde se diseñaron y aplicaron resúmenes lingüísticos para datos reales se puede apreciar en el Anexo A.

2.4. Metodología para minería de datos: CRISP-DM

Como menciona Moine J. *et al* [27] la minería de datos es una disciplina que ha crecido enormemente en los últimos años. Las organizaciones han comprendido que los grandes volúmenes de datos que residen en sus sistemas pueden ser analizados y explotados para obtener nuevo conocimiento a partir de los mismos.

La Minería de Datos es el proceso de extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos, siendo su principal objetivo encontrar información oculta o implícita, que no es posible obtener mediante métodos estadísticos convencionales. El proceso de minería se basa en el análisis de registros provenientes de bases de datos operacionales o bien *data warehouses*.

Son diversos los modelos de procesos que han sido propuestos para el desarrollo de proyectos de minería de datos tales como SEMMA (*Sample, Explore, Modify, Model, Assess*), KDD (*Knowledge Discovery Databases*), CRISP-DM (*Cross Industry Standard Process for Data Mining*), entre otros (una comparación entre sus etapas se puede ver en la Tabla 2.2). Uno de los modelos más utilizados es el modelo CRISP-DM [37], pues tiene en cuenta la aplicación al entorno de negocio de los resultados.

Tabla 2.2: Resumen de las correspondencias entre KDD, SEMMA y CRISP-DM

Modelos de Procesos	**KDD**	**CRISP-DM**	**SEMMA**
Número de pasos	9	6	5
Nombre de los pasos	Desarrollar y entender la aplicación	Comprensión del negocio	-
	Crear un conjunto de datos objetivo	Comprensión de los datos	Muestreo
	Limpiar y preprocesar los datos		Exploración
	Transformar los datos	Preparación de los datos	Modificación
	Elegir la tarea de minería de datos adecuada	Modelado	Modelación
	Elegir el algoritmo de minería de datos adecuado		
	Emplear el algoritmo de minería de datos		
	Interpretar patrones minados	Evaluación	Evaluación
	Usar el conocimiento descubierto	Despliegue	-

Como explica Gallardo J. en [9], CRISP-DM está conformado por una sucesión de fases no necesariamente rígida. Cada fase es estructurada en varias tareas generales que se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas. Las fases se observan en la Figura 2.6 y son las siguientes:

- **Comprensión del negocio** para entender objetivos y requerimientos del proyecto.
- **Comprensión de los datos**, recolectándolos e identificando posibles *subsets* de interés.
- **Preparación de los datos**, construyendo el *dataset* a utilizar.
- **Modelado**, en donde se seleccionan diferentes técnicas a utilizar y se aplican para luego optimizar los parámetros de los modelos.
- **Evaluación** de los modelos asegurándose que satisfagan los objetivos del negocio.
- **Despliegue** de lo obtenido, donde se define cómo se organiza y presenta el conocimiento de los modelos para su consumo.

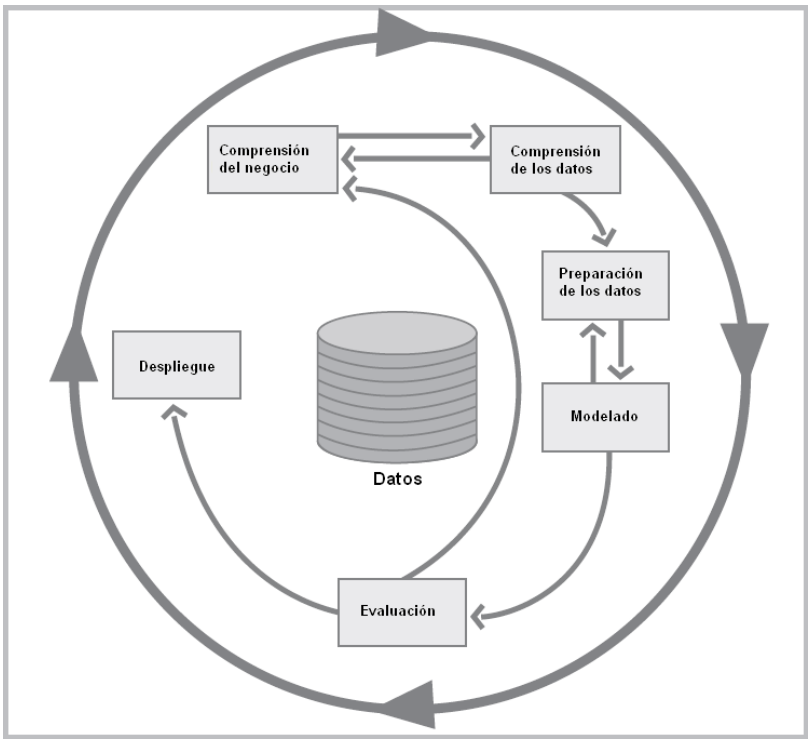


Figura 2.6: Modelo de proceso CRISP-DM.

Capítulo 3

Propuesta de Solución

En este capítulo se señalan y explican los modelos propuestos como solución a la problemática identificada respecto a riego de cultivos. Detallando las razones por elecciones de diseño y cálculos para la generación de resúmenes lingüístico. Además, para la realización del trabajo se sigue un esquema de trabajo tipo *CRISP-DM*.

3.1. Comprensión del negocio

Como se explicó en el capítulo 1, para la agronomía es muy importante la optimización del uso del agua para riego. Sabiendo esto es que se requiere satisfacer la necesidad de entregar información respecto al estado hídrico de las plantas y el cultivo en general, para así planificar la duración y frecuencia de los riegos.

Dado que ya se cuenta con otras formas de representar los datos, tales como gráficos, tablas con valores y estadísticas, se plantea como solución generar pequeños textos, frases o resúmenes para cada necesidad requerida por el usuario. Utilizando lo visto en el capítulo 2, se busca generar los resúmenes lingüísticos que entreguen mayor utilidad y veracidad, permitiendo ayudar a la toma de decisiones de una forma adicional a lo ya existente.

Tomando en cuenta las necesidades de los agrónomos se considera que las frases requeridas tienen que incorporar uno o más de los siguientes puntos:

- Duración del riego.
- Frecuencia de riego.
- Variaciones de humedad de los sensores.
- Percolación o drenaje de agua de riego.
- Comparación entre los valores de la humedad del campo y las líneas de gestión.

Las líneas de gestión (LG) utilizadas son principalmente nivel de riego y capacidad de campo, y éstas permiten un control agronómico al tener una noción de las cantidad de humedad (o agua) que debe contener la plantación. Ambas se pueden observar como rectas punteadas horizontales en el gráfico de la Figura 3.1 y se describen más a fondo a continuación:

- **Capacidad de Campo (CC)**: es la cantidad máxima de agua (o humedad) que puede retener la tierra en donde se encuentra el cultivo sin que en ésta ocurra una percolación profunda y el agua quede más abajo que las raíces de las plantas.
- **Nivel de Riego (NR)**: es la cantidad mínima deseada de humedad en el cultivo para obtener los resultados esperados por el agrónomo. Mantener la humedad por debajo de este margen puede provocar pérdida de productividad u otros efectos adversos.

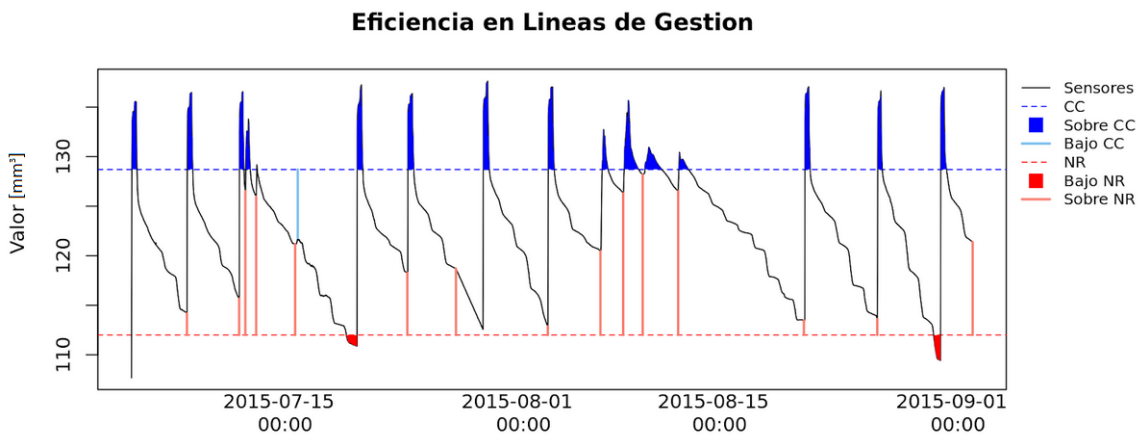


Figura 3.1: Gráfico de ejemplo de representación Líneas de Gestión en el Reporte

3.2. Compresión de los datos

Primeramente, los datos con los que se cuenta son los proporcionados por los sensores de humedad repartidos en los cultivos y, en algunos casos, datos de una estación meteorológica. Esto significa que en cada periodo de tiempo se guardan valores de humedad a diferentes profundidades como se aprecia en el Gráfico 3.2.

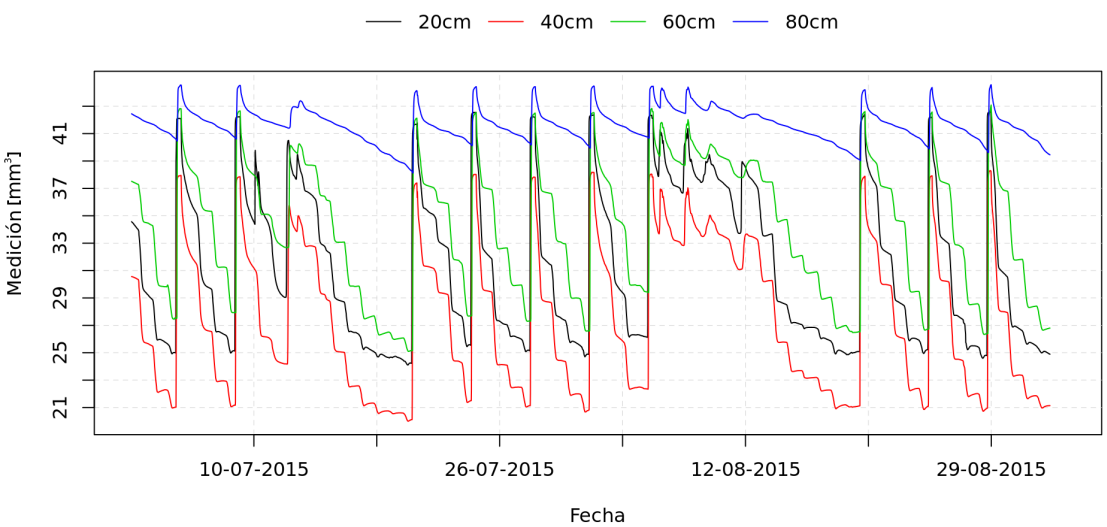


Figura 3.2: Gráfico de ejemplo de datos del sensor en un cultivo de paltos

Ya que dichos datos son muy simples y sin una interpretación adecuada no entregan mucho valor, se utilizan los resultados obtenidos luego del trabajo de Figueroa M. & Pope C. [14] en donde se identifican los riegos válidos además de limpiar, reducir, segmentar, modelar y clasificar los valores entregados por los sensores, lo que permite realizar la preparación de los datos a utilizar en los resúmenes lingüísticos. Los segmentos generados por su trabajo poseen atributos generales respecto a tiempos, tipos y variaciones de humedad. El resultado de este proceso puede apreciarse en los gráficos de la Figura 3.3.

Los datos obtenidos de la segmentación anterior son reducidos un poco más y reordenados para obtener valores más concretos que representen cada evento de riego, tales como duración de riego, cantidad de percolación profunda de agua, cantidad de agua absorbida, tiempo entre eventos de riego, cantidad de riegos posteriores al actual y tiempo transcurrido desde el inicio del riego. Además, conociendo los valores asignados a las líneas de

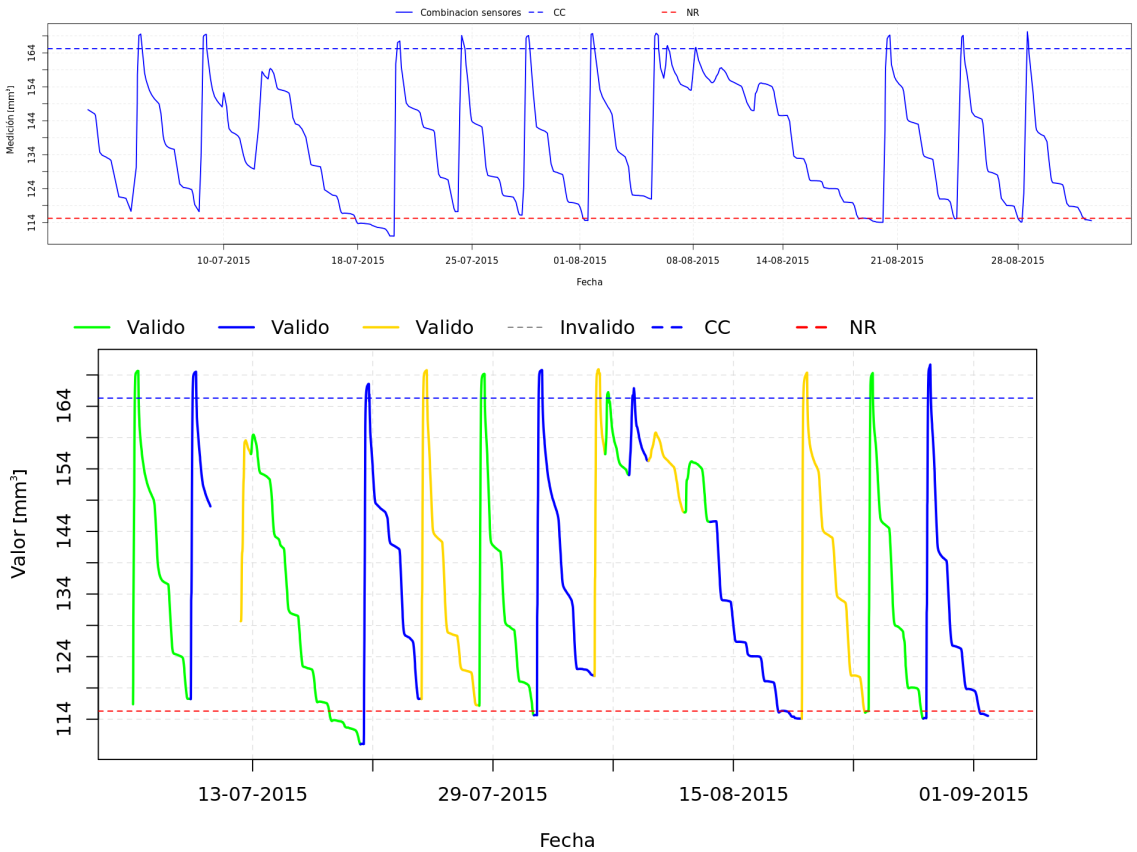


Figura 3.3: Gráficos de datos de sensores del cultivo de paltos agrupados e identificación de riegos

gestión, se calculan datos estadísticos de todos los eventos de riego como número de eventos de riego fuera de cada línea de gestión, la distancia absoluta promedio entre el máximo o mínimo valor de humedad y la línea de gestión (LG) superior o inferior, además de la pendiente de la recta o tendencia que se obtiene de esta distancia al considerar más de un riego. Estos datos son separados en 2 archivos.

Uno de los archivos se observa en la Figura 3.4, el cual contiene los datos que representan cada riego con sus respectivos campos de valores que se explican en la Tabla 3.1. Este archivo puede crecer con el tiempo y debiese tener un número de filas igual a los eventos de riegos que abarquen los datos iniciales.

riego_id	duracionriego_horas	tiempoentreriegos_horas	es_outlier	fechahora_inicio	riego_variacion_humedad	caida_variacion_humedad	consumo_variacion_humedad
1	6.25	NA	0	04-07-2015 7:00	19.45	-9.67	-6.71
2	6.25	88.75	0	08-07-2015 6:00	17.06	-10.1	-5.81
3	6	83.25	0	11-07-2015 23:30	15.92	NA	-9.3
4	4.25	23.5	0	13-07-2015 5:00	1.1	-1.05	-9.31
5	7.75	167.5	0	20-07-2015 8:45	19.32	-9.98	-5.85
6	7.75	79.25	0	23-07-2015 23:45	14.96	-9.33	-7.62
7	7.5	121	0	29-07-2015 8:30	17.5	-9.07	-7.88
8	8	103.75	0	02-08-2015 23:45	16.99	-9.6	-4.57
9	5.5	83.75	0	06-08-2015 19:30	9.21	-0.48	-4.01
10	8.25	32.5	0	08-08-2015 9:30	10.74	NA	-8.63
11	10.25	27	0	09-08-2015 20:45	2.06	NA	-3.37
12	11	48.25	0	12-08-2015 7:15	2.31	NA	-8.98
13	7.75	203.75	0	21-08-2015 6:00	15.99	-9.53	-7.04
14	5.75	118	0	26-08-2015 11:45	16.19	-8.29	-9.44
15	6.25	102	0	30-08-2015 23:30	17.89	-9.09	-3.73

Figura 3.4: Ejemplo del contenido del archivo de entrada con el detalle de los riegos

Tabla 3.1: Diccionario de datos del archivo de la Figura 3.4

Nombre del campo	**Tipo**	**Comentario**
riego_id	Int	Identificador del riego. Id más alto indica un riego más reciente
duracionriego_horas	Float	Tiempo que se estuvo regando, medido en horas
tiempoentreriegos_horas	Float	Tiempo que ha pasado desde que terminó el riego anterior e inició el actual, medido en horas
es_outlier	Int	Indica si se debe considerar el riego; puede tomar los valores 0 o 1
fechahora_inicio	Date	Fecha y hora en la que se inició el riego
riego_variacion_humedad	Float	Variación de humedad total que ocurrió en el riego
caida_variacion_humedad	Float	Variación de humedad percolada en un ciclo de riego
consumo_variacion_humedad	Float	Variación de humedad total que se consumió en un ciclo de riego (o entre dos riegos)

El segundo archivo almacena los valores estadísticos y de tendencia con respecto a los riegos y su comparación con las líneas de gestión (ver Figura 3.5). Cuenta con una única fila que tiene un valor para cada campo los cuales se describen en la Tabla 3.2.

pend_cc_reg	pend_nr_reg	prom_diff_abs_cc	prom_diff_abs_nr	cant_fuera_cc	cant_fuera_nr	cant_total_eve	...
-0.024500820	-0.11384231	6,73264E+14	1,13634E+14	50	2	56	

...	nr	cant_total_eventos_cc	cant_total_eventos_nr	CC	NR	pend_duracionriego_reg	pend_tiempoentreriegos_reg
...		56	56	130	108	-0.0864490772385509	0.378571428571428

Figura 3.5: Ejemplo del contenido del archivo de entrada con los valores estadísticos y de tendencia de todos los riegos considerados

Tabla 3.2: Elementos del archivo de entrada con los valores estadísticos y de tendencia de todos los riegos considerados que explica la imagen 3.5

Nombre del campo	**Tipo**	**Comentario**
pend_cc_reg	Float	Pendiente de la tendencia de la distancia de la capacidad de campo a los riegos.
pend_nr_reg	Float	Pendiente de la tendencia de la distancia del nivel de riego a los riegos.
prom_diff_abs_cc	Float	Promedio de la distancia entre la capacidad de campo y cada riego.
prom_diff_abs_nr	Float	Promedio de la distancia entre el nivel de riego y cada riego.
cant_fuera_cc	Int	Cantidad de riegos que sobrepasaron la capacidad de campo.
cant_fuera_nr	Int	Cantidad de riegos que estuvieron bajo el nivel de riego.
cant_total_eventos_cc	Int	Cantidad total de riegos que se comparan con la capacidad de campo.
cant_total_eventos_nr	Int	Cantidad total de riegos que se comparan con el nivel de riego.
CC	Int	Valor de humedad asignado a la línea de gestión Capacidad de Campo.
NR	Int	Valor de humedad asignado a la línea de gestión Nivel de Riego.
pend_duracionriego_reg	Float	Pendiente de la tendencia de la duración de riegos.
pend_tiempoentreriegos_reg	Float	Pendiente de la tendencia del tiempo entre riegos.

3.3. Preparación de los datos

Los datos recibidos son bastante más fáciles de utilizar que los entregados directamente por los sensores, por el trabajo de transformación y preparación para su posterior uso fue bajo. Estos cambios se resumen en el siguiente listado:

- Transformar los valores de los identificadores dejando un riego más reciente con un identificador de valor menor y un riego más antiguo con uno más alto. De esta forma el riego más reciente registrado queda con el identificador 1 y el más antiguo con el valor n-ésimo. Este cambio permite posteriormente acomodar los riegos en el *set* difuso de temporalidad "cantidad de riegos considerados".
- Convertir los valores de tiempo desde horas a días. De esta forma los valores se acomodan de mejor manera al orden de magnitud de los datos reales, y mejor a la forma de creación del *set* difuso.
- Calcular los días transcurridos desde cada riego hasta el riego más reciente almacenando este valor en una nueva columna. Esto permite acomodar los riegos en el *set* difuso de temporalidad "tiempo transcurrido".
- Convertir las distancias absolutas de riegos a las líneas de gestión a porcentuales. Para esto se considera la distancia entre ambas líneas de gestión como el 100 %, y se transforman los valores de forma proporcional.
- Convertir las veces que un riego sobrepasó alguna línea de gestión a porcentual. Esto se logra considerando la cantidad de riegos totales como el 100 %, y se transforman los valores de forma proporcional.
- Transformar los valores de las tendencias de pendiente de curva a grados. Para esto se aplica la arcotangente al valor de la pendiente de la curva.

3.4. Modelado

El modelado se separa en varias partes; primero, se debe ver como serán las protoformas de las frases deseadas y como se obtendrá una diferenciación entre las frases (métricas).

Después se debe pensar en cómo unirlas para generar párrafos entendibles, y también cómo será la estructura del programa que realizará las tareas.

Dadas las necesidades identificadas, y conociendo los datos iniciales y qué valores se podrían considerar, es que se opta por generar dos tipos de frases. Unas deben ser lo más simples posibles y permitir un rápido entendimiento e interpretación de su significado a la primera lectura. Las segundas deben agrupar características de interés de estas primeras frases simples y generar un valor adicional que no sea tan evidente en una simple vista de los datos.

Para esto se diseñó un *GLMP* como se observa en la Figura 3.6 en donde los requerimientos de información de los agrónomos forman parte de las frases sencillas que seleccionan solo datos como atributos de entrada. Todas estas frases simples permiten determinar un estado particular del cultivo y el riego, que al considerarlas en conjunto se entrega información suficiente para tener una idea del estado general del campo, pero aún así se requiere de interpretación humana para saber qué tan bien o mal se están realizando los procedimientos de hidratación.

Con la idea de obtener valor adicional es que las frases más complejas se enfocan en poder medir el nivel de desempeño que se ha tenido al regar, por lo que se generan PMs y CPs de segundo nivel que se forman solamente a partir de otros CPs. Estas frases hacen referencia a la eficiencia del riego comparándolas a las líneas de gestión y la "calidad de riego", que intentan ser una unificación de las demás frases y reflejar el estado general del riego.

Los resúmenes tipo 1-*CP* se generan y ajustan utilizando variaciones de las protoformas de resúmenes lingüísticos Tipo 1 y Tipo 2. Estas variaciones son pesadas para el caso particular que se trabajó y buscan cubrir diferentes cantidades de datos de acuerdo a la temporalidad que se le asigne. De esta forma se propuso los diseños de la fórmula 3.1 para frases que utilizan todos los datos sin discriminar y como la fórmula 3.2 para las frases que contemplen la temporalidad para describir los sucesos.

$$\underbrace{Atributo + Sujeto}_{A_i} + \underbrace{Descriptor}_{R} \tag{3.1}$$

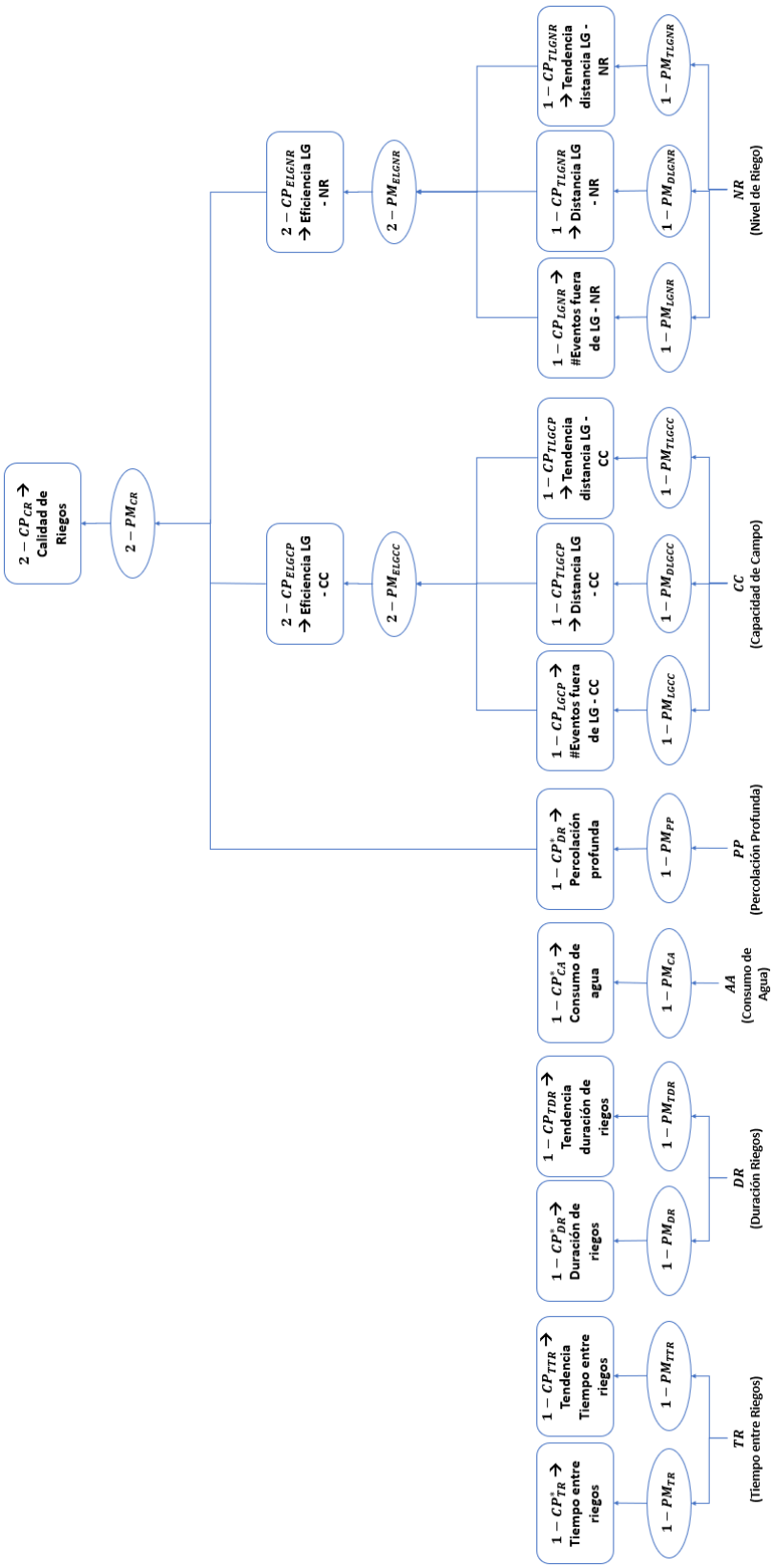


Figura 3.6: GLMP para el riego de cultivos. Los círculos representan PMs y los rectángulos CPs

$$\underbrace{Atributo + Sujeto}_{A_i} + \underbrace{Temporalizante}_{E_t} + \underbrace{Descriptor}_{R} \tag{3.2}$$

Estas variaciones de las protoformas convencionales no utilizan el cuantificador, por lo que las frases son creadas solo con descriptores y temporalizantes en vez de calificadores. El discriminante temporal (E_t) se refiere a frases que consideren la cantidad de riegos anteriores (ej: "los últimos 2 riegos") o el tiempo transcurrido (semanas, meses, etc.). La razón por la que no se utiliza el cuantificador es porque no se ajusta a la forma de las frases buscadas, incorporarlo significaría restringir datos que se intentan considerar.

Algunos ejemplos de frases de tipo 1-*CP* son:

- "Considerando todos los datos, la duración del riego ha sido **media**"
- "El tiempo entre riegos de **los últimos riegos** fue **muy bajo**"
- "La absorción de agua en **las últimas semanas** ha sido **baja**"
- "La percolación de agua de **los últimos 2 riegos** ha sido **alta** "
- "La distancia a la capacidad de campo fue **muy alta**"

Estas frases fueron creadas luego de iterar todas las combinaciones del *set* de palabras que se tiene para cada descriptor y temporalizante de cada prototipo de frase; se pueden apreciar todas las combinaciones atemporales posibles en el Anexo B. Cada palabra del *set* posee una función de pertenencia, también diseñada con ayuda de expertos en el tema, que verifica su validez en el contexto al aplicarles las fórmulas de verdad 3.3 y 3.4 basadas en la modificación y unión de las fórmulas de verdad de las protoformas Tipo 1 (ecuación 2.2), Tipo 2 (ecuación 2.3) y temporal (ecuación 2.14).

$$T(y_i \text{ fue } S) = \frac{1}{n}\sum_{i=1}^{n} \mu_S(y_i) \tag{3.3}$$

$$T(y_i \text{ de } E_t \text{ fue } S) = \frac{\sum_{i=1}^{n} \mu_S(y_i) \wedge \mu_{E_t}(y_i)}{\sum_{i=1}^{n} \mu_{E_t}(y_i)} \tag{3.4}$$

También se incorporó una versión modificada de la cobertura de cada frase, con lo que se puede distinguir cuántos datos abarca cada enunciado y qué tan representativo es en cantidad de datos. Las formas utilizadas son las de la ecuación 3.5 y ecuación 3.6 dependiendo si la frase contempla temporalidad.

$$C(y_i \text{ fue } S) = \frac{\sum_{i=1}^{n} h_i}{n} \tag{3.5}$$

$$C(y_i \text{ de } E_t \text{ fue } S) = \frac{\sum_{i=1}^{n} t_i}{n} \tag{3.6}$$

donde:

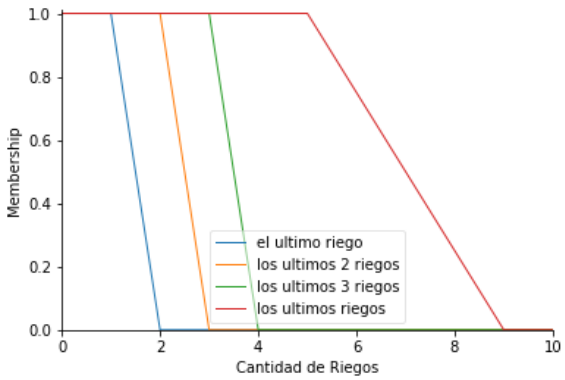
$$h_i = \begin{cases} 1 & \mu_S(y_i) > 0 \\ 0 & \text{otro caso} \end{cases}$$

$$t_i = \begin{cases} 1 & \text{si } \mu_S(y_i) > 0 \text{ y } \mu_{E_t}(y_i) > 0 \\ 0 & \text{otro caso} \end{cases}$$

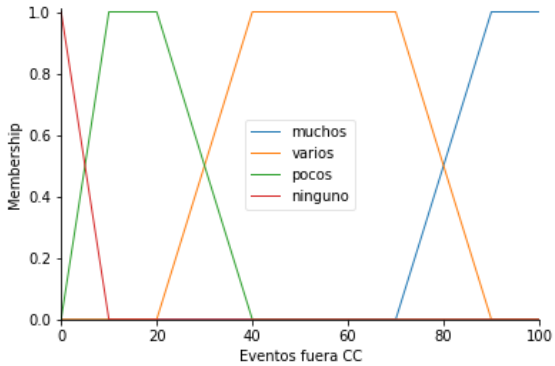
Las formas difusas que se utilizaron para definir todos los *sets* fueron trapezoides debido principalmente a tres factores: la facilidad con la que se pueden implementar, el hecho de que las palabras fueran simples y no se utilizan cuantificadores, y porque es más fácil explicárselo a los expertos y usuarios permitiéndoles hacer cambios posteriores si fuera necesario.

En algunos casos los extremos de los trapezoides terminan siendo "infinito", ya que se requiere que los valores máximos o mínimos puedan terminar en 1 en el límite del universo en vez de "bajar" a 0 en algún punto de este. Algunos ejemplos de *sets* difusos utilizados se pueden ver en la Figura 3.7, en donde se observa en la Figura 3.7a como se definió la cantidad de riegos como temporalidad, lo que se asemeja a lo realizado con el tiempo a considerar. Las Figuras 3.7b y 3.7c muestran las definiciones de 1-*CPs*, y de la misma forma, se debe definir también los valores para los 2-*CPs* como se muestra en el ejemplo de la calidad de riego de la figura 3.7d.

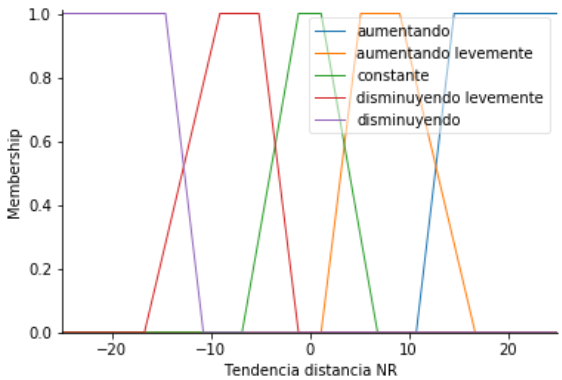
Estas definiciones se basan en conocimiento informado y se determinan para cierto tipo de tierra, tipo de cultivo y otros aspectos de planificación agronómica como los valores en



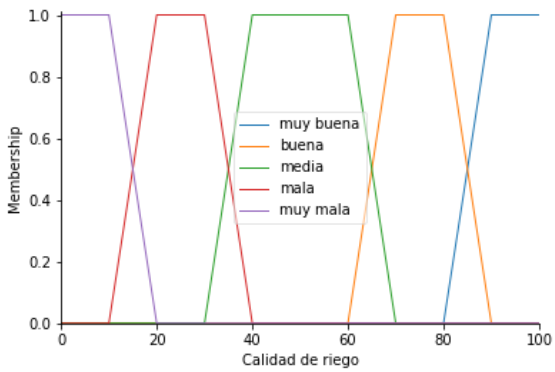
(a) Cantidad de riegos [unidades] (1-CP)



(b) Eventos fuera de la CC [%] (1-CP)



(c) Tendencia distancia al nivel de riego [°] (1-CP)



(d) Calidad de riego [%] (2-CP)

Figura 3.7: Definiciones de algunos *sets* difusos utilizados.

que se fijan las líneas de gestión, por lo que se deben tener varias consideraciones para definir de buena manera la forma difusa que representará cada palabra. Para este trabajo en particular, las que describen las líneas de gestión, la cantidad de agua absorbida, la duración de riegos, el tiempo entre riegos y la percolación profunda de agua son las que más ayuda y ajuste de expertos requieren.

Como se explicó en el capítulo 2, en el *GLMP* los *CPs* de segundo nivel pueden tener como entrada *PMs* de primer nivel y se asignan reglas difusas *if-then* para su generación. En este caso particular, las reglas utilizadas para las líneas de gestión se pueden observar en la Figura 3.8 y las reglas con las que se concluye la calidad de riego en la Figura 3.9.

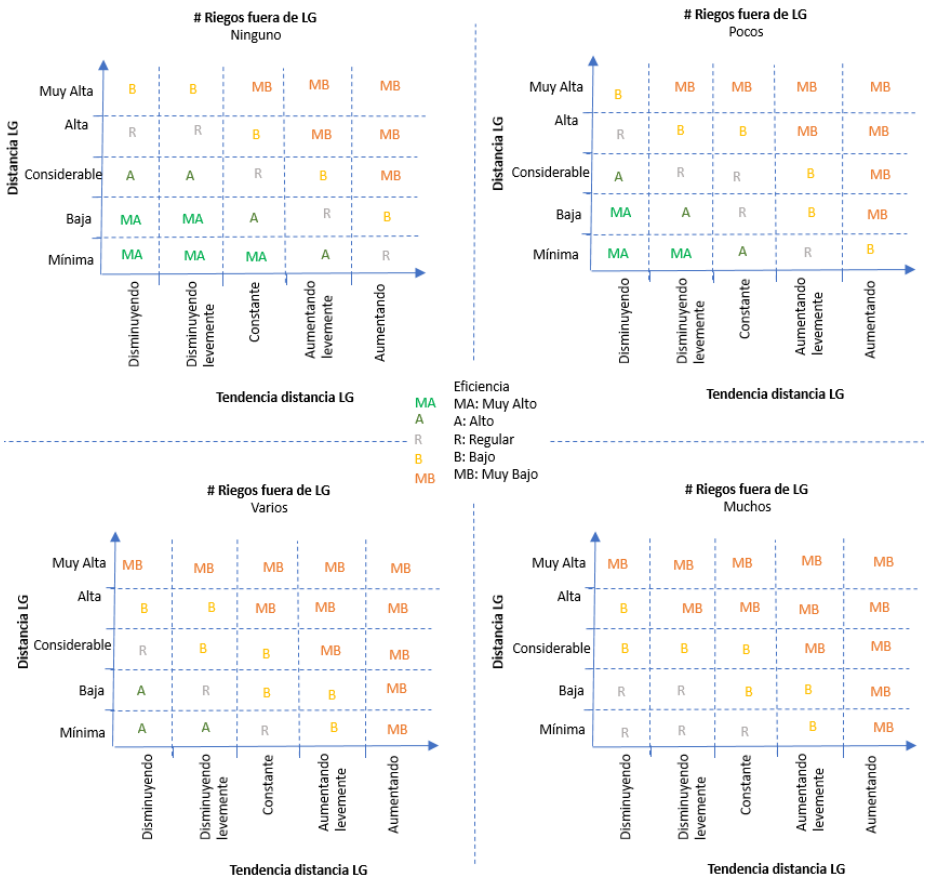


Figura 3.8: Matriz de reglas difusas para la deducción de los valores de las Líneas de Gestión

El diseño del software planteado para cumplir con estas funciones y requerimientos se aprecia en la Figura 3.10 y consta de las siguiente partes:

- **Generación de Mejores Resúmenes** es un *script* que genera el mejor resumen de cada aspecto del *GLMP*, además de considerar también las variaciones temporales posibles.
- **Sets Difusos** es el archivo en donde se guardan los valores a asignarle a los trapezoides, las palabras que representan cada una de las curvas y el tamaño del universo de entrada posible para dicho *set* difuso.
- **Reglas IF-THEN** es un archivo en donde quedan registradas todas los valores de las reglas para generar los de los 2-*CPs* y 2-*PMs*.

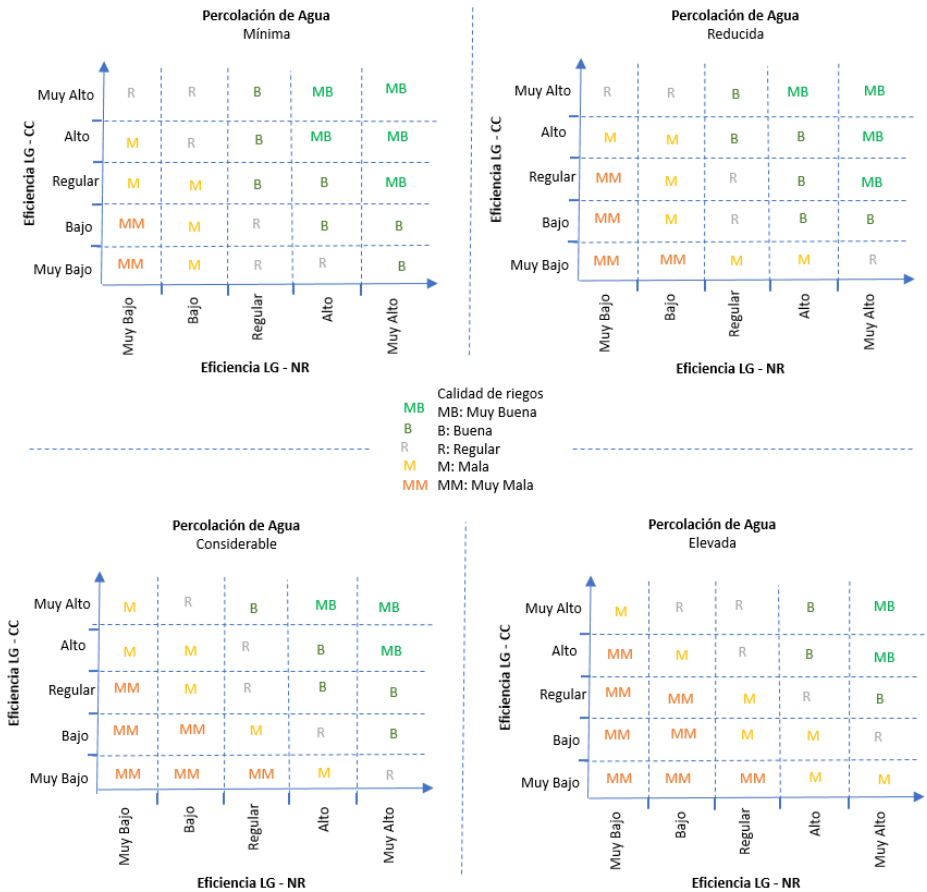


Figura 3.9: Matriz de reglas difusas para la deducción de los valores de la Calidad de Riego

- **Mejores Resúmenes** es el archivo de salida de este primer proceso en donde se tienen todas las mejores frases globales y temporales, además de los valores de verdad y cobertura de cada una de ellas.

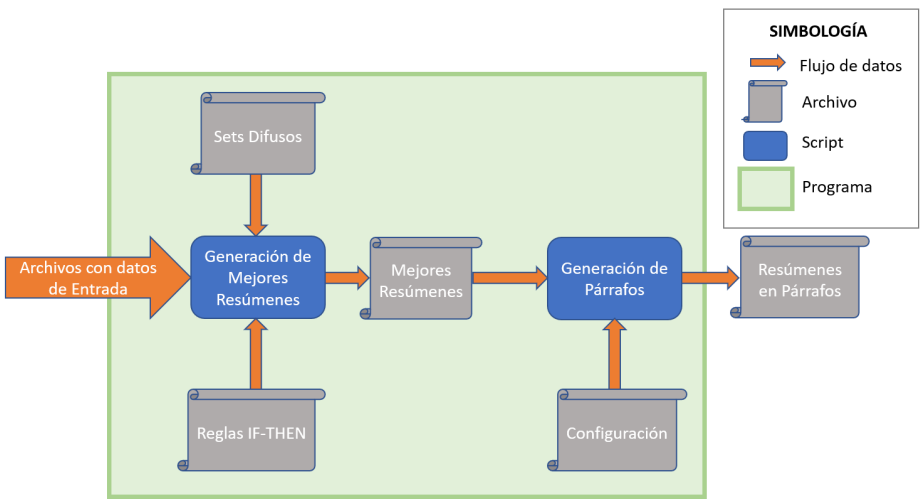


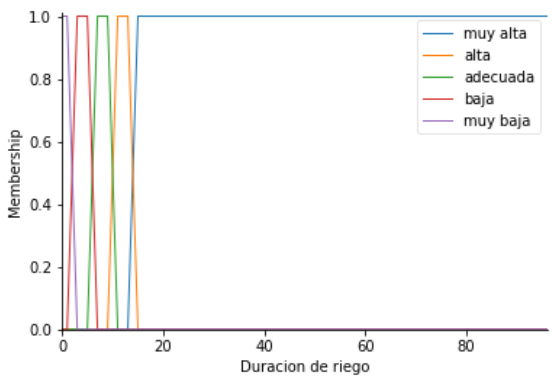
Figura 3.10: Estructura del programa encargado de la generación de resúmenes en párrafos.

- **Generación de Párrafos** es el *script* en donde se decide que resúmenes de cada tipo se desplegará (temporal o no), en que orden irán en cada párrafo y que palabras o frases conectoras las unirán.
- **Configuración** es el archivo en donde se define los órdenes y palabras para generar los párrafos.
- **Resúmenes en Párrafos** es el archivo resultado de todo el proceso y contiene los párrafos armados acompañados del valor de verdad del párrafo completo.

Los tres archivos de configuración se crearon con código JSON [2], ya que gracias a su compatibilidad facilita la implementación posterior, además de permitir una edición de los archivos más sencilla que otros tipos de formato como texto plano o archivo binario.


```
▼ array [17]
  ▼ 0  {6}
      etiqueta : Duracion de riego
      texto sujeto : de riego
      texto atributo : la duración
      Tipo CP : 1
    ▼ universo [3]
        0  : 0
        1  : 96.1
        2  : 0.1
    ▼ curvas [5]
      ▼ 0  {2}
          palabra : muy alta
        ▼ rango [4]
            0  : 13
            1  : 15
            2  : 24
            3  : inf
      ▶ 1  {2}
      ▶ 2  {2}
      ▶ 3  {2}
      ▶ 4  {2}
  ▶ 1  {6}
  ▶ 2  {6}
  ▶ 3  {6}
```


(a) Archivo de configuración de sets difusos.



(b) Representación gráfica del *set* difuso de duración de riego

Figura 3.11: Ejemplo de archivo de configuración de *sets* difusos y representación gráfica del *set* difuso de duración de riegos.

En el caso del archivo de ***Sets* Difusos**, cada *set* está representado en el primer nivel de la estructura del archivo JSON como se muestra en la Figura 3.11a, además pueden apreciarse todos sus componentes en la Tabla 3.3.

En este ejemplo se aprecia que los valores pueden ser “infinitos” como en el caso del límite superior de la curva “muy alta” (líneas de color azul) que se mantiene en valor 1 hasta el final del universo como se aprecia en la Figura 3.11b. Este “infinito” es más bien el máximo valor de un universo, por el bien del rendimiento computacional solo se consideran los valores que se definen dentro del universo.

Tabla 3.3: Elementos del archivo de configuración con los valores de los *sets* difusos que se muestran en la figura 3.11a

Nombre del campo	Tipo	Comentario
etiqueta	String	Nombre **clave** que describe el tipo de frase.
texto sujeto	String	Sujeto de la frase a generar.
texto atributo	String	Atributo del sujeto de la frase a generar.
Tipo CP	Int	1: si es un *set* difuso que tenga como input solo datos. 2: si es un *set* difuso que utiliza otro(s) *set(s)* difuso como entrada.
universo	Array de floats	Define el universo donde se mueven las curvas de cada palabra. el primer valor es el inicio del universo, el segundo valor es el fin del universo y el tercer valor es la cantidad de cifras significativas.
curvas	Array de objetos	Contiene las palabras en orden **descendente** y la definición de la curva de cada palabra. Cada curva (trapezoidal) está definida por cuatro valores numéricos.

En el caso del archivo con las **Reglas *IF-THEN***, este se organiza como un árbol como se observa en la Figura 3.12, en donde cada nivel es una dimensión que representa un *set* difuso de entrada. En el último nivel se tiene, además, el valor resultante de las combinaciones de cada rama.

Los strings con nombre “posible texto” que se encuentran en cada nivel son solo una forma de llevar un orden para la vista humana y configuración manual del archivo, ya que no son utilizados en el código.

Para dicho archivo es muy importante el orden de cada *array*, ya que representa el orden de las palabras de los *sets* difusos que representan, y siempre se suponen un ordenamiento descendente.

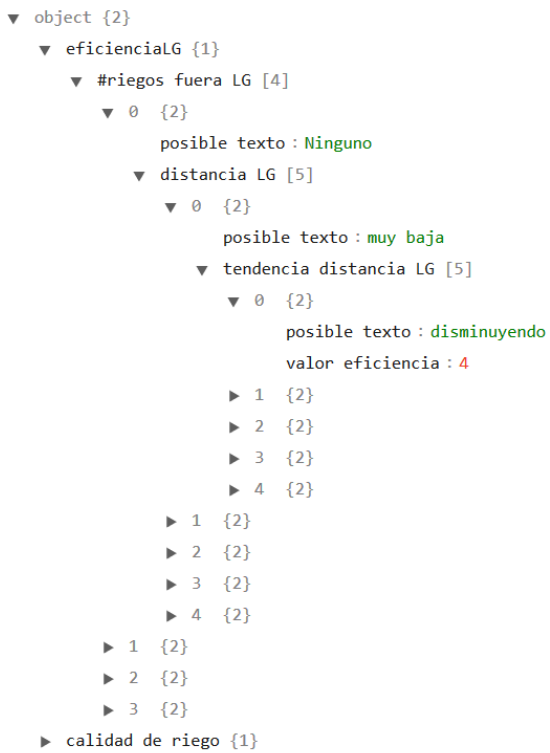


Figura 3.12: Archivo de configuración reglas IF-THEN.

Al igual que los dos anteriores, el archivo de **Configuración** de párrafos es un archivo JSON, cuyo aspecto se puede apreciar en la Figura 3.13. Se tienen tantos objetos en el primer nivel como párrafos se quieran crear. Cada objeto-párrafo cuenta con 3 subobjetos que se describen en la Tabla del Anexo C.

Para dicho archivo es importante el orden de los objetos “frases” y “conectores”, dado que es como serán escritos dentro del párrafo. El orden de los demás objetos es indiferente. Además se debe considerar que la forma de escoger que tipo de frases se mostrará es con el nombre de la etiqueta que la representa.

Para llevar un orden lógico dentro del código del programa desarrollado se hace uso de un sistema de identificadores numéricos (IDs) y etiquetas (strings) para cada tipo de frase y cada variación temporal, dependiendo de si es para el archivo de entrada (sets difusos) o el de salida (mejores resúmenes) que se presenta en el Anexo D.

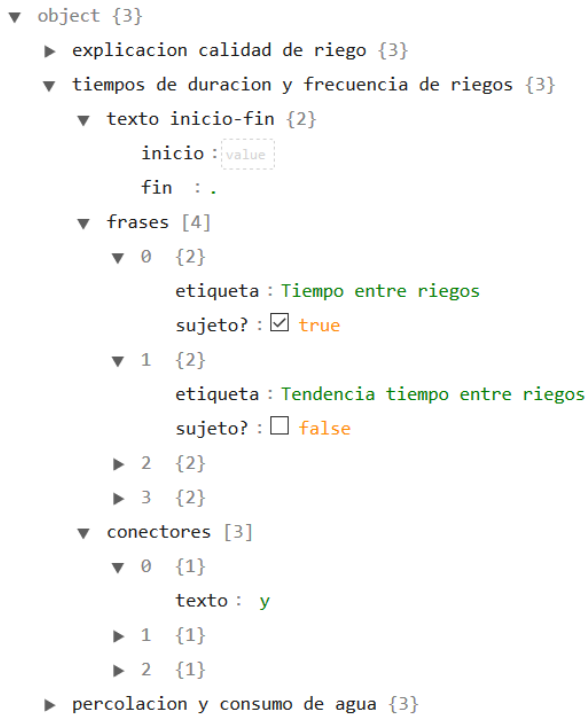


Figura 3.13: Archivo de configuración de párrafos.

Los IDs siguen las siguientes reglas:

- Cada tipo de frase comienza con un número entero.
- Si es que es una variación temporal de la frase se le asigna un decimal.
  - En el caso de temporalidad de fechas (semanas, meses, etc), se les asigna un 0,5.
  - En el caso de cantidad de últimos riegos considerados, se les asigna 0,6.
- Para el mismo tema agronómico de frases, el que tiene el menor valor decimal es más general.

Para saber qué resúmenes lingüísticos serán mostrados a los usuarios, primero se generan todas las posibles frases con sus combinaciones de palabras en el *script* de **Generaciones de Mejores Resúmenes**, calculando y guardando en el proceso su valor de verdad y cobertura. Todas aquellas que cumplan con los requisitos de veracidad mínimo (ajustado inicialmente en 0,001) son almacenadas y quedan como candidatas.

Por cada frase esperada del GLMP y sus variaciones temporales se escoge aquella con el mayor valor de verdad. En el caso de que ocurriese un empate, se intenta desempatar escogiendo la que tiene el mayor valor de cobertura. Si el empate persiste se elige la curva que esté graficada más a la derecha, considerando que sería un caso más pesimista o general.

Luego de completar este procedimiento y generar el archivo de mejores resúmenes, en el siguiente paso, realizado en el *script* **Generación de Párrafos**, se toma la decisión de cuál versión (temporal o global) de cada resumen se mostrará para los párrafos. Para esto se creó una nueva métrica llamada "Ponderador Vectorial de Métricas" (PVM), en donde se calcula el módulo de un vector cuyos componentes son el valor de verdad y la cobertura de cada frase como se muestra en la ecuación 3.7.

$$\text{PVM} = \sqrt{\text{Valor de Verdad}^2 + \text{Cobertura}^2} \tag{3.7}$$

Como última instancia, si es que al aplicar esta nueva métrica (PVM) se presenta algún empate, el *script* está diseñado para elegir la frase que sea más general, es decir, la que no presenta palabras con temporalidad.

Luego de ordenar los resúmenes dentro de un párrafo y añadirles los conectores entre ellos, se busca el valor de verdad mínimo entre todas las frases para cada párrafo y se almacena. Este número es el que representa el valor de verdad general del párrafo completo, ya que se considera el párrafo como un "resumen muy extenso" y se le aplica la *t-norm* del mínimo.

Finalmente, se espera que cada párrafo sea desplegado (o no) en un reporte como un resumen inicial de lo que contendrá posteriormente. Este reporte contiene aspectos relevantes del estado del campo y el riego como gráficos, tablas y valores estadísticos. La decisión de incorporar o no un párrafo dentro del reporte quedó en una fase dentro de la generación de dicho reporte que efectuará la empresa, ya que en algunos casos puede ser importante para sus clientes incorporar los párrafos a pesar del bajo valor de verdad que puedan poseer.

Capítulo 4

Implementación y Validación

En este capítulo se explica cómo se realizó la implementación de la propuesta, cómo se evaluó y qué resultados se obtuvieron.

Para el desarrollo de los *scripts* se utilizó Python 3.5.3 [6]. Las bibliotecas utilizadas en el código son varias de las estándares de Python como io, datetime y math; más otras conocidas como pandas (0.20.1) [5], numpy (1.12.1) [4] y jsonschema (2.6.0) [3]; y la más importante, que permitió el trabajo más rápido de la difusividad, fue scikit-fuzzy (0.3) [1].

Adicionalmente a los valores de verdad y cobertura, se implementaron en el código la mayoría de las métricas encontradas en trabajos anteriores mencionadas en el capítulo 2.

4.1. Evaluación

Como se mencionó en el capítulo 3, la forma principal de evaluar las frases es con el valor de verdad, y en menor medida la cobertura. Utilizando los modelos explicados y datos reales se generaron frases resúmenes y luego párrafos a partir de ellas. Estos datos fueron obtenidos gracias a la compañía, y corresponden a plantaciones de paltos y kiwis de campos ubicados en Chile.

Los resultados después de la generación de mejores resúmenes está en la Tabla 4.1 para las temáticas con variaciones temporales y la Tabla 4.2 para los demás. Estas frases son utilizadas en la generación de párrafos, obteniendo los mostrados en la Tabla 4.3.

Tabla 4.1: Ejemplos de resúmenes temporales obtenidos con su valor de verdad, cobertura y PVM agrupados por temática del resumen

Resumen	**Tipo *CP***	**Valor de verdad**	**Cobertura**	**PVM**
considerando todos los datos, la duración de riego ha sido media	1	0.62	0.87	1.07
la duración de riego de los últimos meses ha sido media	1	0.62	0.87	1.07
la duración de riego de los últimos 3 riegos ha sido media	1	0.67	0.20	0.70
considerando todos los datos, el tiempo entre riegos ha sido medio	1	0.60	0.60	0.85
el tiempo entre riegos de la ultima semana ha sido medio	1	0.84	0.13	0.85
el tiempo entre riegos de los últimos 2 riegos ha sido medio	1	1.00	0.13	1.01
considerando todos los datos, la percolación profunda de agua ha sido alta	1	0.58	0.60	0.83
la percolación profunda de agua de la ultima semana ha sido alta	1	0.85	0.20	0.87
la percolación profunda de agua de el ultimo riego ha sido alta	1	1.00	0.07	1
considerando todos los datos, el consumo de agua ha sido muy alto	1	0.31	0.33	0.45
el consumo de agua de la ultima semana ha sido muy alto	1	0.42	0.07	0.43
el consumo de agua de el ultimo riego ha sido bajo	1	1.00	0.07	1.00

Como se observa de los resúmenes de ejemplo de las Tablas 4.1 y 4.2, no todas las calificaciones de valor de verdad obtenidas superan los niveles deseables de 75 %. Esto debido a que en algunos casos la cantidad de los datos de entrada son muy bajos (solo 1 valor) teniendo como resultado solo el cálculo de la pertenencia del dato; o en otras en donde los datos de entrada son muchos y se busca agruparlos a todos en la misma frase, lo que produce un efecto adverso.

Si se observa el caso del 1-*CP* : duración de riego (Tabla 4.1), que considera las tres primeras frases coloreadas en gris, se nota que el rango de la frase temporal generada abarcó todos los datos (los últimos meses) resultado un valor de verdad y cobertura iguales a la frase atemporal que ya considera todos los datos. En cambio, para la frase temporal que toma en cuenta últimos riegos (últimos 3 riegos en este caso, siendo la tercera coloreada en gris), se logró una mejoría en el valor de verdad a costa de considerar una sección más

pequeña de los datos que se refleja en el menor valor de la cobertura.

Tabla 4.2: Ejemplos de resúmenes obtenidos con su valor de verdad. La cobertura es 1 para todas las frases

Resumen	**Tipo *CP***	**Valor de verdad**
los eventos sobre la capacidad de campo han sido muchos	1	1.00
la distancia a la capacidad de campo ha sido muy alta	1	0.84
la tendencia de la distancia a la capacidad de campo ha estado disminuyendo levemente	1	0.66
los eventos bajo el nivel de riego han sido varios	1	1.00
la distancia al nivel de riego ha sido media	1	1.00
la tendencia de la distancia al nivel de riego ha estado constante	1	1.00
la eficiencia de uso de la capacidad de campo ha sido baja	2	1.00
la eficiencia de uso del nivel de riego ha sido media	2	1.00
la calidad de riego ha sido mala	2	1.00
la tendencia de la duración de los riegos ha estado aumentando levemente	1	1.00
la tendencia del tiempo entre riegos ha estado aumentando	1	1.00

Tabla 4.3: Ejemplos de párrafos de resúmenes obtenidos, con sus correspondientes valores de verdad

Resumen	Valor de verdad
La percolación profunda de agua de el ultimo riego fue alta y el consumo de el último riego fue bajo.	1.00
El tiempo entre riegos de los últimos 2 riegos fue medio y la tendencia del tiempo ha ido aumentando. Además, considerando todos los datos, la duración de riego fue media y la tendencia de la duración ha ido aumentando levemente.	0.62
En general la calidad de riego fue mala, dado que la eficiencia de uso del nivel de riego fue media y la eficiencia de uso de la capacidad de campo fue baja. Los eventos sobre la capacidad de campo fueron muchos, la distancia fue muy alta y la tendencia de la distancia ha ido disminuyendo levemente. Los eventos bajo el nivel de riego fueron varios, la distancia fue media y la tendencia de la distancia ha ido constante.	0.66

Continuando con el mismo ejemplo, se aplica el algoritmo que toma la decisión de que frase es la mejor en base al valor del PVM. Si se observan los valores se tiene que existe un empate entre dos frases, por lo que se elige la más general (la primera), y es la que luego

aparece en los párrafos expuestos en la Tabla 4.3.

Otros casos relevantes son la obtención de los *2-CP*. Por ejemplo, para el cálculo de la Eficiencia de uso de la capacidad de campo, que se encuentra en la Tabla 4.2, se utilizan como entrada tres 1-*CPs*: los eventos sobre la capacidad de campo, la distancia a la capacidad de campo y la tendencia a la capacidad de campo. Aplicando el método de Mamdani [25] y defuzificando mediante la técnica del centroide se obtiene el *set* difuso que se aprecia en la Figura 4.1, en donde las partes coloreadas son el *set* difuso resultante luego de la agregación de todos los *sets* difusos obtenidos tras la evaluación de todas las reglas difusas, y la pertenencia de sus palabras con los valores de entrada. Para este caso "La eficiencia de uso de la capacidad de campo ha sido **baja**", ya que el centroide (recta vertical negra) coincidió con el máximo valor del trapezoide que representa la palabra "baja".

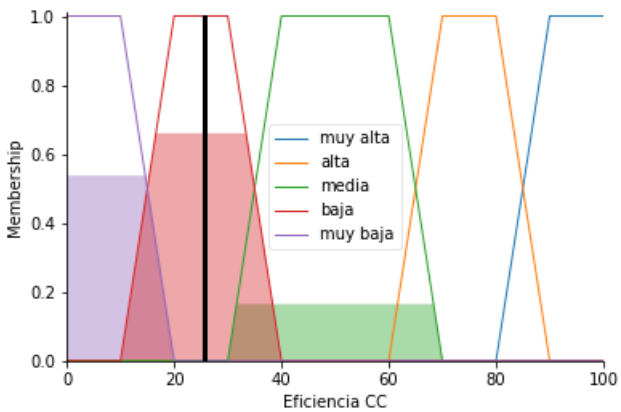


Figura 4.1: Ejemplo cálculo e interpretación de palabra con el método del centroide.

Encuestas para Validación

Además de evaluar cada frase por métricas que ya han sido ampliamente aceptadas y utilizadas (la pertenencia a un *set* difuso, el valor de verdad y la cobertura), se evaluaron las palabras utilizadas, las definiciones difusas de cada palabra y la aceptación de los párrafos mediante tres encuestas que fueron respondidas por expertos en el tema e involucrados en el proyecto.

La primera encuesta (Anexo E, Figuras E.1 y E.2) fue diseñada con el fin de ajustar y calibrar los *sets* difusos a la percepción de más personas. Por lo que se diseñó de forma que los consultados respondieran la palabra "correcta" para diferentes situaciones que les fueron mostradas con gráficos del mismo reporte. Las alternativas de etiqueta que podían escoger

eran las mismas descritas anteriormente, así se podría mantener una estandarización entre todas las respuestas. A pesar de limitar las palabras de respuestas, se adjuntó para todas las preguntas (y las tres encuestas) la posibilidad de entregar *feedback* adicional sobre agregar, eliminar o cambiar etiquetas de los *sets*. Un ejemplo de la encuesta puede observarse en la Figura 4.2.

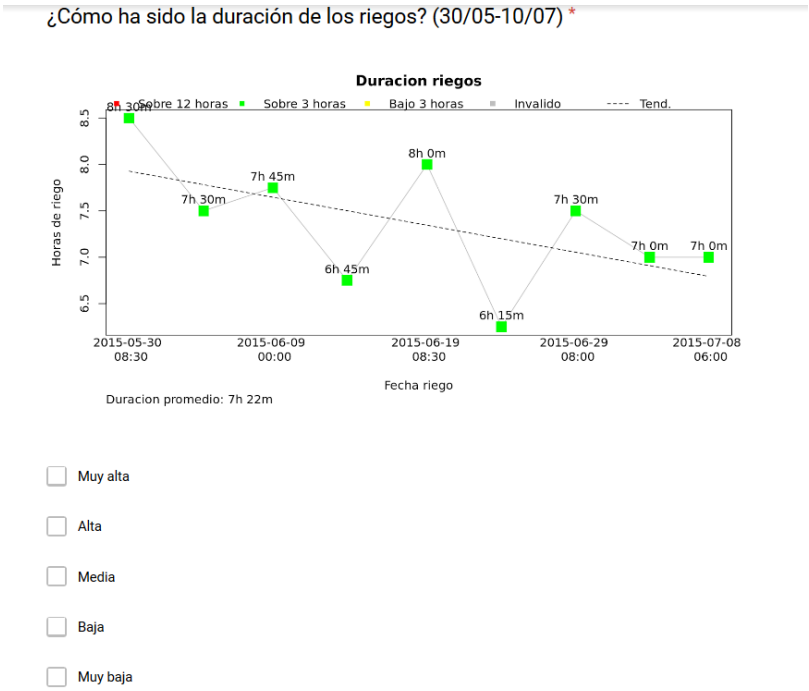


Figura 4.2: Ejemplo de pregunta sobre duración de riegos de la Encuesta 1.

La segunda encuesta (Anexo E, Figura E.3) es una modificación de la escala de Likert [24] y se pensó para evaluar las definiciones difusas ya existentes. Se les mostró a cada encuestado una frase generada sobre algún aspecto en específico de los riegos, además de adjuntar el/los gráficos y tablas necesarios para su comprobación. Luego de revisarlos, evaluaban en una escala de cuatro valores o etiquetas (mala, regular, bueno, excelente) que está diseñada con pocas opciones y en una cantidad par para evitar que las respuestas se concentren en el punto medio o más conocido como *"central tendency bias"*[12]. Un ejemplo de la encuesta puede observarse en la Figura 4.3.

Los promedios de las notas para cada tipo de frase se encuentran resumidos en la Tabla 4.4, donde se observa que las calificaciones obtenidas en general no son del todo buenas; si se consideran además los comentarios dejados por los encuestados en las dos primeras encuestas se puede determinar que:

- Algunas palabras (etiquetas) son difíciles de entender o interpretar porque en los gráficos no son fáciles de ver, tales como las tendencias, las duraciones de riego y los tiempos entre riegos.

- Algunas palabras (etiquetas) deben ser más descriptivas y tener una carga valórica, ya que se produjeron inquietudes sobre todo con las etiquetas “medio” y “media” debido a que fueron utilizadas indistintamente en diferentes tipos de frases. Pero en la realidad existen casos en donde “medio” es el resultado ideal, otros en donde es “menos malo” (como sería percolación profunda) y otros donde es una “calificación promedio” (como en los caso de eficiencia y calidad de riego).

- Considerar evitar la generación y utilización de las frases temporales, ya que el usuario tiene la capacidad de seleccionar de forma explícita la cantidad de datos/riegos con los que desea trabajar y que podrá visualizar al momento de generar el reporte.

- Muchas frases no se entienden bien por sí solas, por lo que se demuestra la importancia de juntar resúmenes para generar párrafos, que se expliquen entre sí y el evento general.

- Se demuestra la necesidad de configurar los *sets* difusos para cada cultivo y para los clientes específicos que utilizarán los reportes y que se adecuen a sus percepciones.

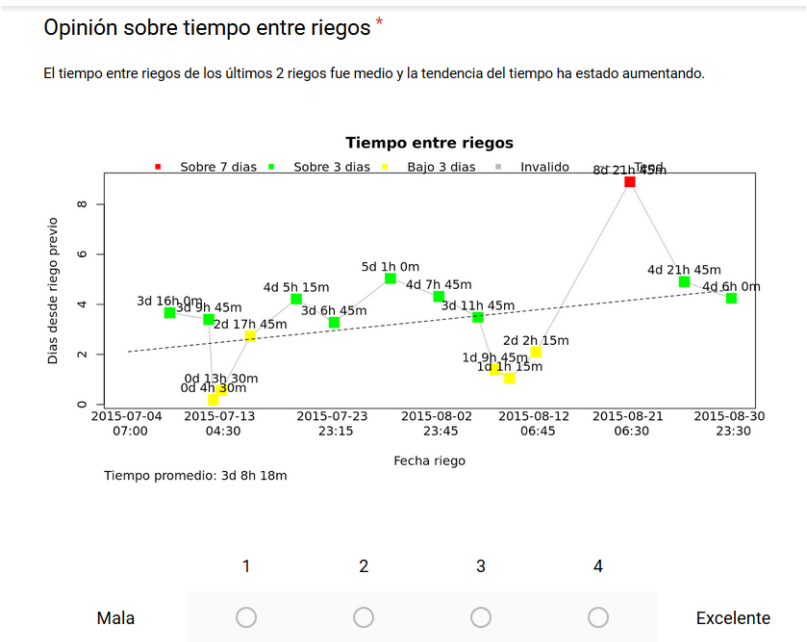


Figura 4.3: Ejemplo de pregunta sobre tiempo entre riegos de la Encuesta 2.

Tabla 4.4: Resultados de la evaluación de frases por cada tema

Tema de la frase	**Promedio de la evaluación [%]**
eficiencia capacidad de campo	68,9
consumo de agua	63,3
duración de riegos	60,0
eficiencia nivel de riego	59,0
tiempo entre riegos	58,3
Calidad de riego	57,1
tendencia distancia capacidad de campo	54,3
eventos bajo nivel de riego	54,3
eventos sobre capacidad de campo	53,3
tendencia distancia nivel de riego	53,3
percolación profunda de agua	51,7
distancia capacidad de campo	46,7
distancia nivel de riego	45,7

La tercera encuesta (Anexo E, Figura E.4) se creó con el fin de poder evaluar los párrafos dentro del reporte mediante las Máximas de Grice, por lo que se le mostraron diferentes reportes a los usuarios, para luego hacer las siguientes afirmaciones:

- Los párrafos reflejan lo que se muestra en los gráficos o tablas.
- La extensión de los párrafos es la adecuada.
- Los párrafos afirman eventos relevantes para el usuario.
- Los párrafos son precisos y evitan la ambigüedad.
- El formato y redacción de los párrafos son correctos y entendibles. (Agregado como adicional, no pertenece a las Máximas de Grice)

En cada una de estas se aplicó una escala de cinco opciones (muy en desacuerdo, en desacuerdo, indiferente, de acuerdo, muy de acuerdo) semejante a la escala de Likert [24].

Los resultados obtenidos, que pueden apreciarse en la Tabla 4.5 obtenida de la encuesta mostrado en la Figura 4.4, fueron bajos respecto a los objetivos. Esto debido a que al parecer no se cumplió la relevancia de lo expresado a los usuarios y no se evitó suficientemente

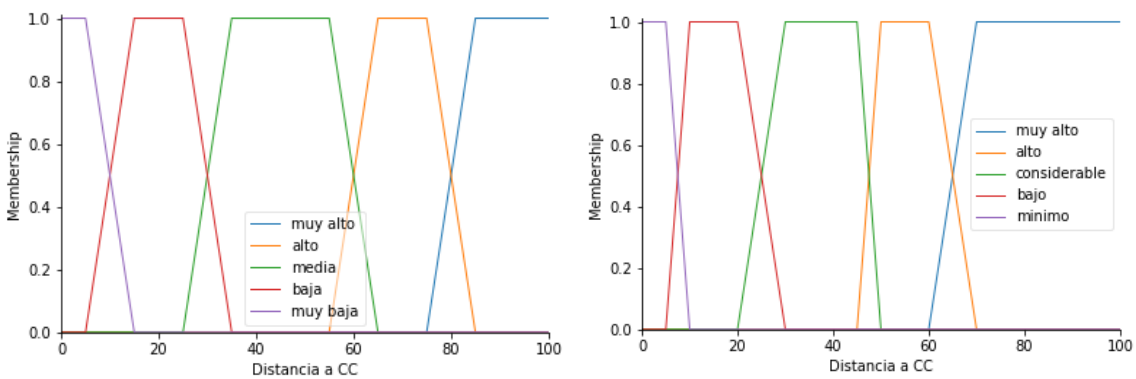
la ambigüedad de las palabras. Cabe destacar que, al igual que en la primera encuesta, los encuestados pudieron no entender del todo los gráficos, por lo que la forma en que perciben los resúmenes se ve influenciada de mala manera.

Tabla 4.5: Resultados de la evaluación de Máximas de Grice en el reporte

Máxima de Grice	Promedio de la evaluación [%]
Cantidad	75,0
Formato y redacción	56,3
Verdad	42,2
Relevancia	35,9
Manera (no ambigüedad)	35,9

Luego de revisar los resultados de las encuestas, sobre todo de las dos primeras, se realizaron los siguientes cambios de calibración a los *sets* difusos :

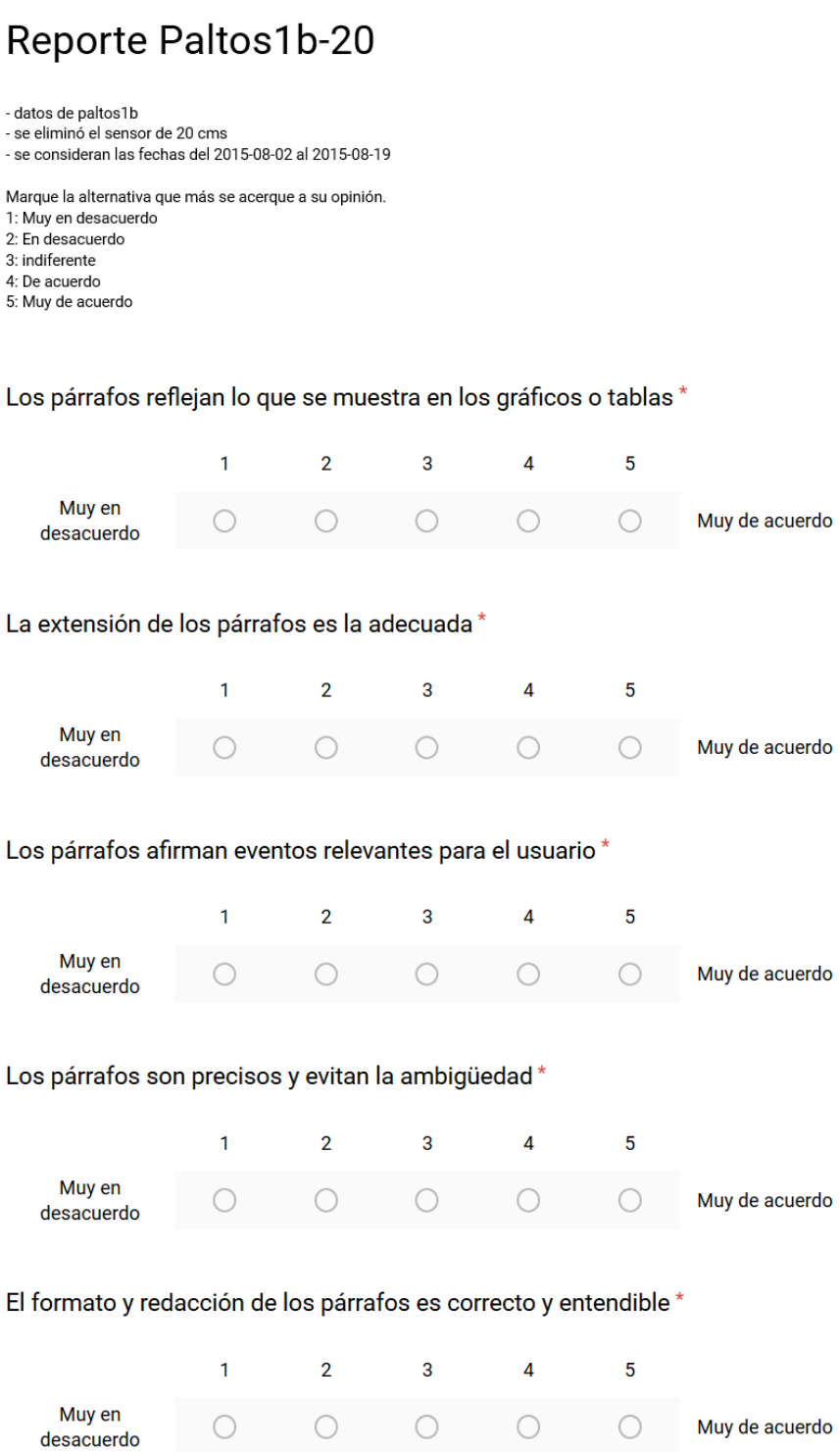
- *Set* difuso “distancia a una LG” (NR y CC):
  - Disminuir todos los valores de la distancia a LG (cerca del 80 % de lo inicial), es decir, se disminuyeron los tamaños que abarcan los *sets* menores y se le asignó mayor tamaño a la palabra “muy alta”.
  - Cambiar a “promedio de distancia a la LG” en vez de “distancia a la LG”.
  - Cambiar etiqueta “media” por “considerable”.
  - Cambiar etiqueta “muy baja” por “mínimo”.



(a) Set difuso antiguo.

(b) Set difuso nuevo.

Figura 4.5: Cambio de set difuso del promedio de la Distancia a la CC.



Reporte Paltos1b-20

- datos de paltos1b
- se eliminó el sensor de 20 cms
- se consideran las fechas del 2015-08-02 al 2015-08-19

Marque la alternativa que más se acerque a su opinión.
1: Muy en desacuerdo
2: En desacuerdo
3: indiferente
4: De acuerdo
5: Muy de acuerdo

Los párrafos reflejan lo que se muestra en los gráficos o tablas *

	1	2	3	4	5	
Muy en desacuerdo	○	○	○	○	○	Muy de acuerdo

La extensión de los párrafos es la adecuada *

	1	2	3	4	5	
Muy en desacuerdo	○	○	○	○	○	Muy de acuerdo

Los párrafos afirman eventos relevantes para el usuario *

	1	2	3	4	5	
Muy en desacuerdo	○	○	○	○	○	Muy de acuerdo

Los párrafos son precisos y evitan la ambigüedad *

	1	2	3	4	5	
Muy en desacuerdo	○	○	○	○	○	Muy de acuerdo

El formato y redacción de los párrafos es correcto y entendible *

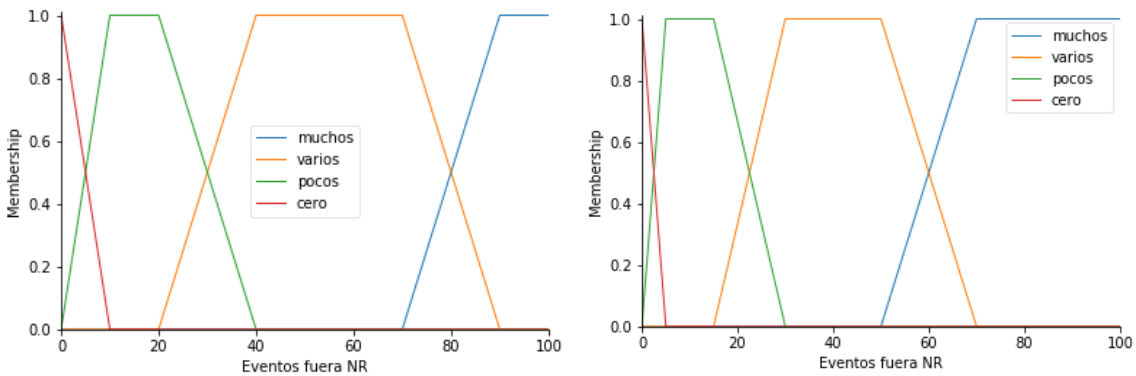
	1	2	3	4	5	
Muy en desacuerdo	○	○	○	○	○	Muy de acuerdo

Figura 4.4: Ejemplo de set de preguntas sobre los datos de Paltos1b de la Encuesta 3.

- *Set* difuso “eventos fuera de la LG” (NR y CC):
  - Cambiar definición de “ninguno” para que abarque un subconjunto menor del universo de entrada (pasar de valor de entrada máximo de la curva, o “d” en la

definición del trapezoide, de 10 % a 5 %).

- Cambiar definición de “pocos” para que abarque un subconjunto mayor del universo de entrada (pasar el valor de entrada mínimo de la curva, o “a” en la definición del trapezoide, de 10 % a 5 %).
- Cambiar etiqueta “ninguno”por “cero”.
- Cambiar valores de las curvas “muchos” y “varios”; los extremos de mayor valor de varios pasaron de 90 % a 70 % y de 70 % a 50 % , y lo mismo con los extremos inferiores de “muchos”.
- Cambiar valores de las curvas “varios” y “pocos”; los extremos de mayor valor de “pocos” pasaron de 40 % a 30 % y de 20 % a 15 %, y lo mismo con los extremos inferiores de “varios”.



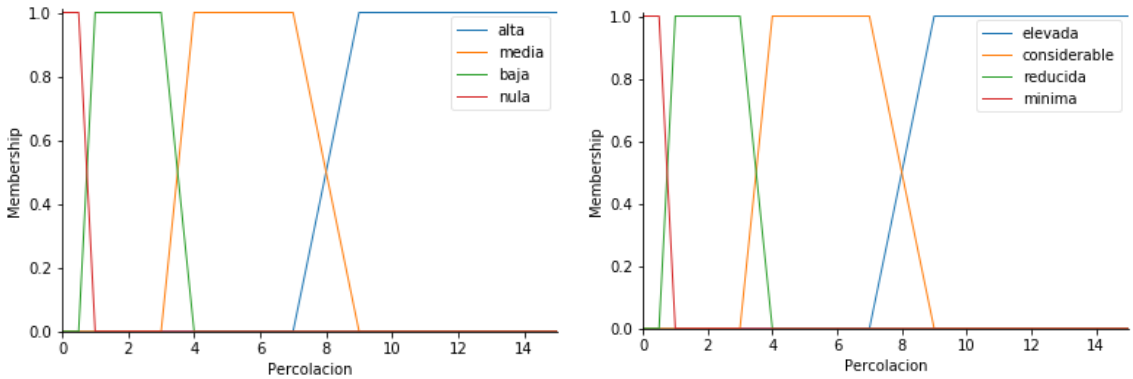
(a) Set difuso antiguo.

(b) Set difuso nuevo.

Figura 4.6: Cambio de set difuso de eventos bajo el NR.

- *Set* difuso “tiempo entre riegos”: cambiar etiqueta “medio” por “adecuado”.
- *Set* difuso “duración de riego”: cambiar etiqueta “media” por “adecuada”.
- *Set* difuso “percolación profunda de agua”:
  - Cambiar etiqueta “alta” por “elevada”.
  - Cambiar etiqueta “media” por “considerable”.
  - Cambiar etiqueta “baja” por “reducida”.
  - Cambiar etiqueta “nula” por “mínima”.

- *Set* difuso “consumo de agua”: cambiar etiqueta “medio” por “regular”.



(a) Set difuso antiguo.

(b) Set difuso nuevo.

Figura 4.7: Cambio de set difuso de percolación profunda.

- *Set* difuso “eficiencia de la LG (ambas)”: cambiar etiqueta “media” por “regular”.
- *Set* difuso “calidad de riego”: cambiar etiqueta “media” por “regular”.
- Cambiar los conectores “ha sido” por “fue” y “ha estado” por “ha ido”.

Las diferencias con los resultados iniciales debido al cambio de textos en las etiquetas se pueden observar en las Tablas 4.6 y 4.7 destacado con **negrita**, por otro lado las desigualdades respecto a la variación en la magnitud de las métricas por la reconfiguración de valores de los *sets* difusos se resalta en color rojo en la Tabla 4.7.

Debido al cambio en las frases consecuentemente los párrafos también sufrieron el cambio en las etiquetas a usar y sus valores de verdad como se aprecia en la Tabla 4.8 y se destacan de la misma manera que las tablas de frases. Se aprecian cambios menores en comparación a la Tabla 4.3 mostrada anteriormente.

Se observa que el ajuste que se aplicó al *set* difuso de “eventos sobre la capacidad de campo” produjo también un cambio en “eficiencia de la capacidad de campo”; el que se propagó hasta “calidad de riego”. A pesar de este cambio el valor de la eficiencia se mantuvo similar como se observa en la Figura 4.8. Donde sí cambió fue en el valor final del párrafo que habla sobre estos aspectos, ya que al disminuir su valor de verdad a 0.50 produjo que todo el párrafo de eficiencias y calidad de riego “disminuyera su credibilidad” a este mismo valor como se observa en la Tabla 4.8.

Tabla 4.6: Valores de la Tabla 4.1 luego de la modificación a partir de los resultados de las encuestas

Resumen	**Tipo *CP***	**Valor de verdad**	**Cobertura**	**PVM**
considerando todos los datos, la duración de riego fue **adecuada**	1	0.62	0.87	1.07
la duración de riego de los últimos meses fue **adecuada**	1	0.62	0.87	1.07
la duración de riego de los últimos 3 riegos fue **adecuada**	1	0.67	0.20	0.70
considerando todos los datos, el tiempo entre riegos fue **adecuado**	1	0.60	0.60	0.85
el tiempo entre riegos de la última semana fue **adecuado**	1	0.84	0.13	0.85
el tiempo entre riegos de los últimos 2 riegos fue **adecuado**	1	1.00	0.13	1.01
considerando todos los datos, la percolación profunda de agua fue **elevada**	1	0.58	0.60	0.83
la percolación profunda de agua de la última semana fue **elevada**	1	0.85	0.20	0.87
la percolación profunda de agua de el último riego fue **elevada**	1	1.00	0.07	1.00
considerando todos los datos, el consumo de agua fue muy alto	1	0.31	0.33	0.45
el consumo de agua de la última semana fue muy alto	1	0.42	0.07	0.43
el consumo de agua de el último riego fue bajo	1	1.00	0.07	1.00

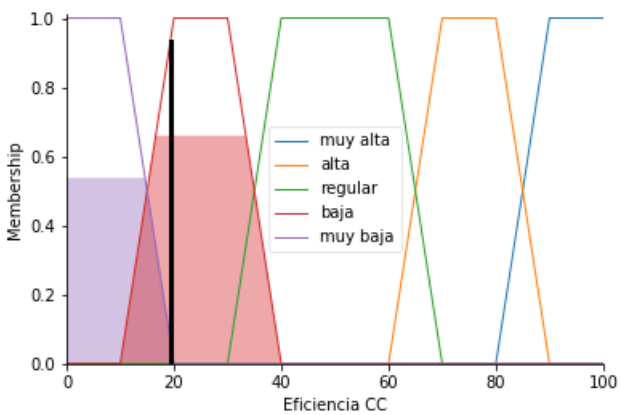


Figura 4.8: Ejemplo del cálculo e interpretación de etiqueta con el método del centroide luego del cambio en los *sets* difusos.

Tabla 4.7: Valores de la Tabla 4.2 luego de la modificación a partir de los resultados de las encuestas

Resumen	**Tipo *CP***	**Valor de verdad**
los eventos sobre la capacidad de campo fueron muchos	1	1.00
el promedio de la distancia a la capacidad de campo fue muy alto	1	1.00
la tendencia de la distancia a la capacidad de campo ha ido disminuyendo levemente	1	0.66
los eventos bajo el nivel de riego fueron muchos	1	0.50
el promedio de la distancia al nivel de riego fue **considerable**	1	1.00
la tendencia de la distancia al nivel de riego ha ido constante	1	1.00
la eficiencia de uso de la capacidad de campo fue baja	2	0.93
la eficiencia de uso del nivel de riego fue **regular**	2	1.00
la calidad de riego fue mala	2	1.00
la tendencia de la duración de los riegos ha ido aumentando levemente	1	1.00
la tendencia del tiempo entre riegos ha ido aumentando	1	1.00

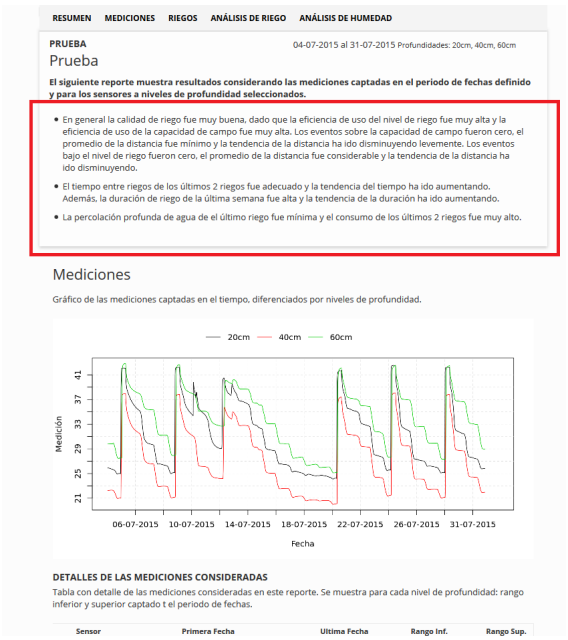
Tabla 4.8: Valores de la Tabla 4.3 después de las encuestas

Resumen	Valor de verdad
La percolación profunda de agua de el último riego fue **elevada** y el consumo de el último riego fue bajo.	1.00
El tiempo entre riegos de los últimos 2 riegos fue **adecuado** y la tendencia del tiempo ha ido aumentando. Además, considerando todos los datos, la duración de riego fue **adecuada** y la tendencia de la duración ha ido aumentando levemente.	0.62
En general la calidad de riego fue mala, dado que la eficiencia de uso del nivel de riego fue **regular** y la eficiencia de uso de la capacidad de campo fue baja. Los eventos sobre la capacidad de campo fueron muchos, el promedio de la distancia fue muy alto y la tendencia de la distancia ha ido disminuyendo levemente. Los eventos bajo el nivel de riego fueron muchos, el promedio de la distancia fue **considerable** y la tendencia de la distancia ha ido constante.	0.50

El nuevo modelo resultante debiese cumplir mejor con las expectativas de los usuarios que respondieron las encuestas, y a su vez un público más amplio que la primeras versiones, por lo que si se repitieran las encuestas también se debiesen tener mejores evaluaciones.

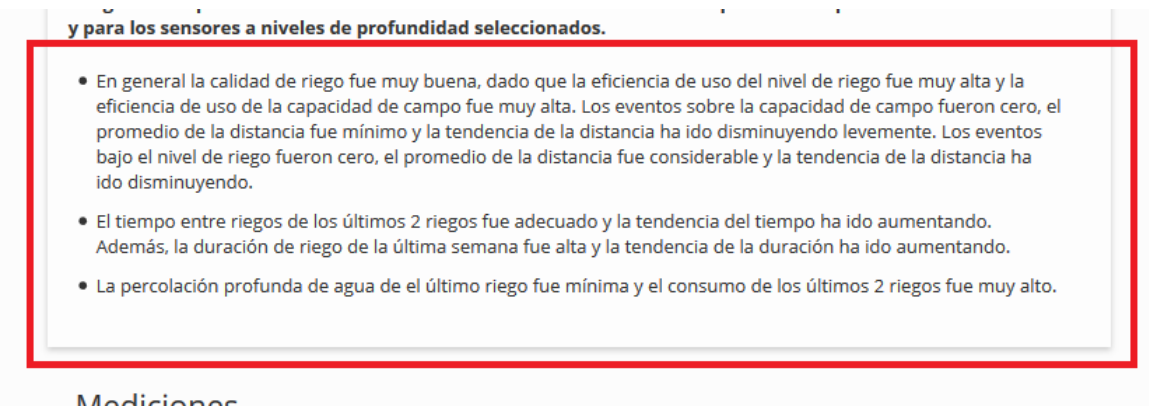
4.2. Despliegue

Como se explicó anteriormente, los párrafos fueron pensados como textos iniciales que resumieran aspectos relevantes del estado de los cultivos y los comportamientos de riego. Es así como siguiendo esta idea es que se añadieron a los reportes (que también seguían en construcción) resultando algo como se muestra en las Figuras 4.9a y 4.9b o el acercamiento a la sección de párrafos en la Figura 4.9c.



(a) Reporte (1/2)

(b) Reporte (2/2)



(c) *Zoom* párrafos

Figura 4.9: Ejemplo de párrafos en un reporte.

De esta forma, la persona encargada del campo podrá tener una descripción inicial de lo que muestran los gráficos, y sabrá en donde debe poner más atención para mejorar el

rendimiento del campo y el agua.

Otra forma de utilizar los resúmenes es mediante mensajes de alerta o avisos periódicos, los cuales pueden ser enviados a *smartphones* vía mensajes de texto (SMS) o desde la misma aplicación donde se generan los reportes. Esto permitiría a los usuarios tomar acción aún más rápidamente y no solo cuando se desee planificar un periodo largo de tiempo.

Conclusiones

Al realizar el trabajo, se obtuvo una visión más amplia y precisa de lo que significa codificar para generar lenguaje humano, no limitando las palabras a segmentos deterministas con valores rígidos. La lógica difusa es una herramienta poderosa para la creación y utilización de lenguaje humano (inexacto) en ambientes automatizados, entregando resultados favorables para el despliegue de información desde una perspectiva más cercana para el usuario que los gráficos, tablas y estadísticas.

Como se mencionó anteriormente, el objetivo de este trabajo es apoyar la toma de decisiones en un caso real de riego de cultivos y medir que la satisfacción de los usuarios sea mayor al 70 % considerando las Máximas de Grice, se puede concluir que éste se cumplió parcialmente, ya que no todas las máximas cumplieron con el porcentaje mínimo esperado. Esto debido principalmente a que las encuestas también tenían como objetivo aprender de los usuarios para configurar las frases según sus percepciones y entendimiento, por lo que se espera que existiesen diferencias con lo establecido inicialmente y así poder corregirlas.

A pesar que los resultados de valor de verdad obtenidos en algunos casos eran más bajos que lo esperado, existen ocasiones en donde es más importante para el negocio mostrar una frase que no es tan verdadera, en vez de no mostrar alguna información. Esto, por ejemplo, en el caso en donde se encontraban valores de 50 % de pertenencia entre dos palabras. Si ambas palabras en ese límite pueden ser medianamente válidas, es mejor mostrar alguna de las dos (la más pesimista para este trabajo), ya que es más importante que los usuarios aprecien un valor cercano a la realidad, en lugar de quedar con la duda por falta de información, lo que finalmente podría llevarlos a pensar en considerar como respuesta valores que ni siquiera estuvieron cerca de ser una afirmación válida (0 % de valor de verdad). Un ejemplo podría ser que se tuvieran duraciones de riego alta o muy altas (50 % de verdad cada una), lo que puede provocar que las plantas se ahoguen o se pierda mucha agua por

percolación profunda. Si no se muestra ninguna información o frase el usuario puede llegar a pensar que se está regando a duraciones correcta o incluso bajas, lo que lleva a tomar decisiones erróneas y llegar a producir un daño en los cultivos.

En el caso de los sensores, a pesar de los arduos trabajos anteriores de identificación de riegos y el hecho de intentar tener más de un sensor a diferentes profundidades, de todas formas existe ruido en los datos, lo que no ayuda mucho a generar los resúmenes correctos, perjudicando el reporte en general.

Las encuestas diseñadas fueron de gran ayuda para verificar y corregir las ideas iniciales de los resúmenes: estas fueron pensadas para abarcar lo más posible diferentes aspectos del problema. Los resultados demostraron satisfacción por lo que significa agregar los resúmenes dentro del reporte, ya que todos los encuestados concuerdan en que aumenta el valor del mismo, por lo que los bajos resultados conseguidos inicialmente se compensan de cierta forma. Entre los buenos comentarios generales, se encuentra el hecho de que se realizaron varios cambios según las sugerencias y respuestas obtenidas en las encuestas, por lo que una nueva evaluación de los párrafos debiese mejorar de buena manera sus evaluaciones.

El trabajo realizado, en general, impactó de buena forma en la empresa. Mientras se desarrollaban los resúmenes lingüísticos ya se les estaba ofreciendo a posibles clientes, los que se mostraron muy interesados en que los reportes contuvieran esta característica. Por otro lado, el hecho de encuestar a expertos en el tema permitió que algunos trabajadores recientes pudieran aprender más rápido el reporte que se está generando, y recibir críticas constructivas acerca de algunos aspectos del reporte que no tenían que ver con los resúmenes en sí.

Concretamente, el hecho de crear nuevos aspectos con los 2-*CP*, como las eficiencias a las líneas de gestión y la calidad de riego, aumentaron el valor general del reporte ya que no son temas que se vean a simple vista con solo un gráfico o tabla.

El haber creado el PVM permitió tomar decisiones sin estar basándonse únicamente en una métrica (valor de verdad) para definir la importancia de las frases. Esta propuesta puede aportar a futuros trabajos, pudiéndose incorporar más métricas y no solo las dos utilizadas en esta situación.

La metodología utilizada (CRISP-DM) fue útil, aunque hubo que flexibilizarla para este trabajo. Las etapas iniciales de comprensión del negocio y los datos son de mucha importancia, en lo posible deben realizarse durante todo el transcurso del trabajo para asegurar buenos resultados, y no solo en las primeras etapas como plantea la forma "convencional" o "estructurada" de la metodología CRISP-DM. No se tuvo reparos en volver a fases anteriores, incluso en etapas tardías del desarrollo, por ejemplo, estando en la etapa de modelado se pasó a verificar los objetivos del negocio, lo cual sirvió para evitar correcciones futuras o trabajo perdido.

También podría ser importante realizar encuestas en etapas tempranas del trabajo para verificar qué tipos de frases y palabras son las realmente necesarias y/o comprensibles para los usuarios del negocio. Además, que se podría acelerar la calibración de los *sets* difusos.

De acuerdo a los objetivos, el general consistía en apoyar la toma de decisiones en un caso real de riego de cultivos, mediante el diseño, implementación y evaluación de un algoritmo generador de resúmenes lingüísticos, logrando que la satisfacción de los usuarios finales y/o la empresa sea mayor al 70 % considerando las Máximas de Grice. Se cumplió parcialmente, ya que solo uno de estos cinco factores considerados tuvo una nota superior a lo deseado.

De los cuatro aspectos medidos que no lograron superar el 70 %, se encuentra la verdad de la frase. Este aspecto, aparte de ser evaluado por la encuesta, fue medido mediante la métrica de "valor de verdad", es por esto que se esperaba obtener a lo menos entre un 60 % y 70 % al igual que las peores mejores frases que se generaron.

Los otros tres factores medidos que no cumplieron con lo esperado de la evaluación ("relevancia", "manera" y "formato y redacción") debiesen tener un aumento significativo en la evaluación en una hipotética repetición de las encuestas, ya que se corrigieron percepciones de los encuestados como las palabras "medio" y "media" que no permitían una correcta comprensión del significado de las frases. Además de reajustar los valores de las curvas de las definiciones difusas, permitiendo más cercanía a sus propias percepciones.

Considerando lo aprendido de las encuestas, se puede destacar que lo mejor es ajustar los valores y palabras de las curvas de los *sets* difusos directamente con los usuarios finales que consumirán los párrafos del reporte.

En cuanto a los objetivos específicos de este trabajo, se definieron protoformas y fórmulas de valores de verdad para los resúmenes lingüísticos utilizados, y en varios de los casos estos superaron el valor de verdad mínimo de 75 %. Es un hecho que no en todos los casos se podía contar con este mínimo, ya que en algunas circunstancias se tenía un único valor de entrada que calzaba exactamente entre 2 curvas difusas, lo que producía que su pertenencia a ambas fuera cercana al 50 %. Aún así, en varias de las pruebas, al tener la posibilidad de utilizar variaciones temporales para un mismo tipo de resumen, se evita tener frases con valores de verdad muy bajos y no representativos.

Además, tal como se propuso se definieron los calificadores adecuados para la generación de las frases considerando como base la lógica difusa y las necesidades y entendimiento de los usuarios del negocio. Considerando también que las definiciones fueron mejoradas mediante las encuestas.

Por último, también se buscaba ayudar a planificar y optimizar el uso del agua y la calidad de las plantas, lo que se espera lograr mediante los datos obtenidos desde los sensores. Dado que los resúmenes lingüísticos generados entregan información concreta sobre la frecuencia y duración de riego, e información acerca de factores más complejos como la calidad de riego que se interpreta mediante aspectos definidos en la planificación del mismo agrónomo que revisará el reporte, se puede concluir que se aporta al menos con un valor agregado al entregar una nueva forma de ver la información, la que permite planificar y optimizar utilizando conocimiento experto.

Como trabajo futuro, se plantean los siguientes caminos:

- Generar un sistema que recomiende acciones a tomar utilizando reglas *if-then* y no solamente describa lo que sucedió. Para esto se requeriría mucha más ayuda de expertos y entrenamiento de usuarios para la interpretación.
- Incorporar un formato tipo encuesta u otro método para recibir datos de los propios usuario y entrenar las definiciones difusas, permitiendo que estas se configuren de forma automática de acuerdo a sus respuestas y percepciones.
- Ajustar los valores de los *sets* difusos y ver si es necesaria la incorporación de curvas más complejas que reemplacen los trapezoides actuales. Además de revisar si sería

conveniente utilizar otras métricas para ayudar a escoger los mejores resúmenes.

En cuanto al conocimiento adquirido gracias a la universidad antes de realizar esta memoria, las asignaturas que más me ayudaron fueron:

- **Bases de Datos Avanzadas**, ya que me permitieron familiarizarme con archivos con código JSON.
- **Diseño de Interfaces Usuarias**, me ayudó a tener una noción de la importancia del cómo se muestra el producto final, lo que influyó en el empeño para que los resúmenes fuesen lo más entendible posible.
- ***Machine Learning***, porque me ayudó a tener mejor conocimiento en Python y sobretodo en bibliotecas útiles como pandas.

Dentro de la universidad (Campus Valparaíso) se dicta un curso donde se enseñan aspectos de la lógica difusa, por lo que de haber podido tener acceso a él, el trabajo quizá se hubiese hecho más fácil en las etapas iniciales.

Bibliografía

[1] Fuzzy logic scikit (toolkit for scipy). https://github.com/scikit-fuzzy/scikit-fuzzy. Accessed: 2018-03-15.

[2] Introducing json. https://www.json.org/. Accessed: 2017-09-30.

[3] jsonschema 2.6.0. https://pypi.python.org/pypi/jsonschema. Accessed: 2018-03-15.

[4] numpy 1.12.1. https://pypi.python.org/pypi/numpy/1.12.1. Accessed: 2018-03-15.

[5] pandas 0.20.1. https://pypi.python.org/pypi/pandas/0.20.1/. Accessed: 2018-03-15.

[6] Python 3.5.3. https://www.python.org/downloads/release/python-353/. Accessed: 2018-03-15.

[7] Alberto Alvarez-Alvarez, Daniel Sanchez-Valdes, Gracian Trivino, Ángel Sánchez, and Pedro D Suárez. Automatic linguistic report of traffic evolution in roads. *Expert Systems with Applications*, 39(12):11293–11302, 2012.

[8] Alberto Alvarez-Alvarez, Gracian Trivino, and Oscar Cordon. Human gait modeling using a genetic fuzzy finite state machine. *IEEE Transactions on Fuzzy Systems*, 20(2):205–223, 2012.

[9] JAG Arancibia. Metodología para el desarrollo de proyectos en minería de datos crisp-dm. *Recuperado de http://oldemarrodriguez. com/-yahoo_site_admin/assets/docs/Documento_CRISP-DM*, 2385037, 2010.

[10] Fatih Emre Boran, Diyar Akay, and Ronald R Yager. An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications*, 61:356–377, 2016.

[11] Rodrigo Callejas Rodríguez, Juan Vera, and Cristián Kremer Fariña. *Riego de precisión en frutales con sensores de suelo*. Number 23 in Serie Ciencias Agronómicas. Universidad de Chile, Facultad de Ciencias Agronómicas, 2014.

[12] Igor Douven. A bayesian perspective on likert scales and central tendency. *Psychonomic bulletin & review*, pages 1–9, 2017.

[13] Luka Eciolaza, Martín Pereira-Fariña, and Gracian Trivino. Automatic linguistic reporting in driving simulation environments. *Applied Soft Computing*, 13(9):3956–3967, 2013.

[14] Manuel Figueroa and Christopher Pope. Root system water consumption pattern identification on time series data. *Sensors*, 17(6):1410, 2017.

[15] Morcillo Carlos González. Lógica difusa, una introducción práctica, 2011. `http://www.esi.uclm.es/www/cglez/downloads/docencia/2011_Softcomputing/LogicaDifusa.pdf`.

[16] H Paul Grice, Peter Cole, and Jerry L Morgan. Syntax and semantics. *Logic and conversation*, 3:41–58, 1975.

[17] Ion Iancu. A mamdani type fuzzy logic controller. In *Fuzzy Logic-Controls, Concepts, Theories and Applications*. InTech, 2012.

[18] Janusz Kacprzyk and ANNA Wilbik. Linguistic summaries of time series: on some extended aggregation techniques. *Studia i Materiały Polskiego Stowarzyszenia Zarzdzania Wiedza*.

[19] Janusz Kacprzyk, Anna Wilbik, and S Zadrożny. Linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Fuzzy Sets and Systems*, 159(12):1485–1499, 2008.

[20] Janusz Kacprzyk, Ronald R Yager, and Slawomir Zadrozny. Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support. *Knowledge discovery for business information systems*, pages 129–152, 2002.

[21] Janusz Kacprzyk and Sławomir Zadrożny. Linguistic data summarization: A high scalability through the. *Scalable Fuzzy Algorithms for Data Management and Analysis: Methods and Design: Methods and Design*, page 214, 2009.

[22] Janusz Kacprzyk and Sławomir Zadrożny. Linguistic summaries of time series: A powerful and prospective tool for discovering knowledge on time varying processes and systems. In *Towards the Future of Fuzzy Logic*, pages 65–77. Springer, 2015.

[23] Janusz Kacprzyk and Sławomir Zadrożny. Fuzzy logic-based linguistic summaries of time series: a powerful tool for discovering knowledge on time varying processes and systems under imprecision. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(1):37–46, 2016.

[24] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[25] Ebrahim H Mamdani and Sedrak Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, 7(1):1–13, 1975.

[26] Sheila Mendez-Nunez and Gracian Trivino. Combining semantic web technologies and computational theory of perceptions for text generation in financial analysis. In *Fuzzy systems (fuzz), 2010 IEEE international conference on*, pages 1–8. IEEE, 2010.

[27] Juan Miguel Moine, Silvia Ethel Gordillo, and Ana Silvia Haedo. Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. In *XVII Congreso Argentino de Ciencias de la Computación (CACIC 2011)*, 2011.

[28] Vilém Novák. Linguistic characterization of time series. *Fuzzy Sets and Systems*, 285:52–72, 2016.

[29] Federico Montesino Pouzols, Angel Barriga, Diego R Lopez, and Santiago Sánchez-Solano. Linguistic summarization of network traffic flows. In *Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference on*, pages 619–624. IEEE, 2008.

[30] Daniel Sanchez-Valdes, Alberto Alvarez-Alvarez, and Gracian Trivino. Dynamic linguistic descriptions of time series applied to self-track the physical activity. *Fuzzy Sets and Systems*, 285:162–181, 2016.

[31] Somayajulu G Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. Generating english summaries of time series data using the gricean maxims. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 187–196. ACM, 2003.

[32] Somayajulu G Sripada, Ehud Reiter, Jim Hunter, Jin Yu, and Ian P Davy. Modelling the task of summarising time series data using ka techniques. In *Applications and Innovations in Intelligent Systems IX*, pages 183–196. Springer, 2002.

[33] Gracian Trivino, Angel Sanchez, Antonio S Montemayor, Juan J Pantrigo, Raul Cabido, and Eduardo G Pardo. Linguistic description of traffic in a roundabout. In *Fuzzy systems (fuzz), 2010 IEEE international conference on*, pages 1–8. IEEE, 2010.

[34] Albert van der Heide and Gracián Triviño. Automatically generated linguistic summaries of energy consumption data. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, pages 553–559. IEEE, 2009.

[35] Marian Van Der Meulen, Robert H Logie, Yvonne Freer, Cindy Sykes, Neil McIntosh, and Jim Hunter. When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology*, 24(1):77–89, 2010.

[36] Anna Wilbik, James M Keller, and Gregory Lynn Alexander. Linguistic summarization of sensor data for eldercare. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 2595–2599. IEEE, 2011.

[37] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39, 2000.

[38] Ronald R Yager. A new approach to the summarization of data. *Information Sciences*, 28(1):69–86, 1982.

[39] Ronald R Yager. Fuzzy summaries in database mining. In *Artificial Intelligence for Applications, 1995. Proceedings., 11th Conference on*, pages 265–269. IEEE, 1995.

[40] Lotfi A Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with applications*, 9(1):149–184, 1983.

Anexos A

Tabla de trabajos recientes sobre resúmenes lingüísticos con aplicaciones a datos reales

Nombre trabajo	Área de aplicación	Medidas de calidad	¿Datos o frases temporales?	¿GLMP o reglas if-then?	¿Validado por usuarios o expertos?
Generating English Summaries of Time Series Data Using the Gricean Maxims [31]	Meteorología	Máximas de Grice	si	no	si
Linguistic summarization of time series using a fuzzy quantifier driven aggregation [19]	Finanzas	Valor de Verdad	si	no	si
Linguistic Summarization of Network Traffic Flows [29]	Redes de Computadores	Valor de Verdad	no	no	si
Automatically generated linguistic summaries of energy consumption data[34]	Consumo de Energía	Valor de Verdad, Heurística propia de Relevancia	si	no	si
Linguistic summaries of time series : on some extended aggregation techniques [18]	Finanzas	Valor de Verdad, Grado de Enfoque, Medida de la Informatividad	si	no	no
Linguistic description of traffic in a roundabout [33]	Tráfico Vehicular	Pertenencias a *sets* difusos	no	si	no
Combining Semantic Web technologies and Computational Theory of Perceptions for text generation in financial analysis [26]	Finanzas	Pertenencias a un *sets* difusos, Grado de Especificidad	si	si	no

Nombre trabajo	Área de aplicación	Medidas de calidad	¿Datos o frases temporales?	¿GLMP o reglas if-then?	¿Validado por usuarios o expertos?
Linguistic summarization of sensor data for eldercare [36]	Cuidado de Ancianos	Valor de Verdad, Grado de Enfoque	si	no	si
Automatic linguistic report of traffic evolution in roads [7]	Tráfico Vehícular	Grado de Validez Propios	si	si	no
Automatic Linguistic Reporting in Driving Simulation Environments [13]	Simulación de Manejo	Pertenencias a un *sets* difusos, Encuesta a expertos evaluando: relevancia, veracidad, vocabulario, cantidad adecuada de contenido.	no	si	si
Human gait modeling using a genetic fuzzy finite state machine [8]	Caminata de personas	-	no	si	si
Linguistic summaries of time series: a powerful tool for discovering knowledge on time varying processes and systems [22]	Finanzas y Registro Web de un servidor	Valor de Verdad (Yager)	si	no	no
Linguistic Characterization of Time Series	varios ejemplos explicativos [28]	Valor de Verdad Propio	si	no	no
Dynamic linguistic descriptions of time series applied to self-track the physical activity [30]	Actividad Física	- Máximas de Grice - Grado de Validez Propio (mediante media ponderada)	si	si	si
Fuzzy logic-based linguistic summaries of time series: A powerful tool for discovering knowledge on time varying processes and systems under imprecision [23]	Finanzas y Registro Web de un servidor	Valor de Verdad (Extensión con temporalidad)	si	no	no

Anexos B

Todas las posibles frases atemporales

1- PM_{TR}: Tiempo entre Riegos.

- $a_{TR_1} \longrightarrow$ El tiempo entre riegos fue **muy alto**
- $a_{TR_2} \longrightarrow$ El tiempo entre riegos fue **alto**
- $a_{TR_3} \longrightarrow$ El tiempo entre riegos fue **adecuado**
- $a_{TR_4} \longrightarrow$ El tiempo entre riegos fue **bajo**
- $a_{TR_5} \longrightarrow$ El tiempo entre riegos fue **muy bajo**

1- PM_{TTR}: Tendencia de Tiempo entre Riegos.

- $a_{TTR_1} \longrightarrow$ El tiempo entre riegos ha ido **aumentando**
- $a_{TTR_2} \longrightarrow$ El tiempo entre riegos ha ido **aumentando levemente**
- $a_{TTR_3} \longrightarrow$ El tiempo entre riegos ha ido **constante**
- $a_{TTR_4} \longrightarrow$ El tiempo entre riegos ha ido **disminuyendo levemente**
- $a_{TTR_5} \longrightarrow$ El tiempo entre riegos ha ido **disminuyendo**

1- PM_{CA}: Consumo de Agua.

- $a_{CA_1} \longrightarrow$ El consumo de agua fue **muy alta**
- $a_{CA_2} \longrightarrow$ El consumo de agua fue **alta**
- $a_{CA_3} \longrightarrow$ El consumo de agua fue **regular**
- $a_{CA_4} \longrightarrow$ El consumo de agua fue **baja**

- $a_{CA_5} \longrightarrow$ El consumo de agua fue **muy baja**

1- PM_{EFLGCC}: Eventos Fuera Capacidad de Campo.

- $a_{EFLGCC_1} \longrightarrow$ Los eventos bajo el nivel de riego fueron **muchos**
- $a_{EFLGCC_2} \longrightarrow$ Los eventos bajo el nivel de riego fueron **varios**
- $a_{EFLGCC_3} \longrightarrow$ Los eventos bajo el nivel de riego fueron **poco**
- $a_{EFLGCC_4} \longrightarrow$ Los eventos bajo el nivel de riego fueron **cero**

1- PM_{EFLGNR}: Eventos Fuera Nivel de Riego.

- $a_{EFLGNR_1} \longrightarrow$ Los eventos sobre la capacidad de campo fueron **muchos**
- $a_{EFLGNR_2} \longrightarrow$ Los eventos sobre la capacidad de campo fueron **varios**
- $a_{EFLGNR_3} \longrightarrow$ Los eventos sobre la capacidad de campo fueron **pocos**
- $a_{EFLGNR_4} \longrightarrow$ Los eventos sobre la capacidad de campo fueron **cero**

1- PM_{PP}: Percolación Profunda de agua.

- $a_{PP_1} \longrightarrow$ La percolación profunda de agua fue **elevada**
- $a_{PP_2} \longrightarrow$ La percolación profunda de agua fue **considerable**
- $a_{PP_3} \longrightarrow$ La percolación profunda de agua fue **reducida**
- $a_{PP_4} \longrightarrow$ La percolación profunda de agua fue **mínima**

1- PM_{DLGNR}: Distancia al Nivel de Riego.

- $a_{DLGNR_1} \longrightarrow$ El promedio de la distancia al nivel de riego fue **muy alto**
- $a_{DLGNR_2} \longrightarrow$ El promedio de la distancia al nivel de riego fue **alto**
- $a_{DLGNR_3} \longrightarrow$ El promedio de la distancia al nivel de riego fue **considerable**
- $a_{DLGNR_4} \longrightarrow$ El promedio de la distancia al nivel de riego fue **bajo**
- $a_{DLGNR_5} \longrightarrow$ El promedio de la distancia al nivel de riego fue **mínimo**

1- PM_{DLGCC}: Distancia a la Capacidad de Campo.

- $a_{DLGCC_1} \longrightarrow$ El promedio de la distancia a la capacidad de campo fue **muy alto**

- $a_{DLGCC_2} \longrightarrow$ El promedio de la distancia a la capacidad de campo fue **alto**
- $a_{DLGCC_3} \longrightarrow$ El promedio de la distancia a la capacidad de campo fue **considerable**
- $a_{DLGCC_4} \longrightarrow$ El promedio de la distancia a la capacidad de campo fue **bajo**
- $a_{DLGCC_5} \longrightarrow$ El promedio de la distancia a la capacidad de campo fue **mínimo**

1- PM_{DR}: Duración de Riego.

- $a_{DR_1} \longrightarrow$ La duración de los riegos fue **muy alta**
- $a_{DR_2} \longrightarrow$ La duración de los riegos fue **alta**
- $a_{DR_3} \longrightarrow$ La duración de los riegos fue **adecuada**
- $a_{DR_4} \longrightarrow$ La duración de los riegos fue **baja**
- $a_{DR_5} \longrightarrow$ La duración de los riegos fue **muy baja**

1- PM_{TDR}: Tendencia de la Duración de Riego.

- $a_{TDR_1} \longrightarrow$ La duración de los riegos ha ido **aumentando**
- $a_{TDR_2} \longrightarrow$ La duración de los riegos ha ido **aumentando levemente**
- $a_{TDR_3} \longrightarrow$ La duración de los riegos ha ido **constante**
- $a_{TDR_4} \longrightarrow$ La duración de los riegos ha ido **disminuyendo levemente**
- $a_{TDR_5} \longrightarrow$ La duración de los riegos ha ido **disminuyendo**

1- PM_{TLGNR}: Tendencia de la Distancia al Nivel de Riego.

- $a_{TLGNR_1} \longrightarrow$ La tendencia de la distancia al nivel de riego ha ido **aumentando**
- $a_{TLGNR_2} \longrightarrow$ La tendencia de la distancia al nivel de riego ha ido **aumentando levemente**
- $a_{TLGNR_3} \longrightarrow$ La tendencia de la distancia al nivel de riego ha ido **constante**
- $a_{TLGNR_4} \longrightarrow$ La tendencia de la distancia al nivel de riego ha ido **disminuyendo levemente**
- $a_{TLGNR_5} \longrightarrow$ La tendencia de la distancia al nivel de riego ha ido **disminuyendo**

1- PM_{TLGCC}: Tendencia de la Distancia a la Capacidad de Campo.

- $a_{TLGCC_1} \longrightarrow$ La tendencia de la distancia a la capacidad de campo ha ido **aumentando**
- $a_{TLGCC_2} \longrightarrow$ La tendencia de la distancia a la capacidad de campo ha ido **aumentando levemente**
- $a_{TLGCC_3} \longrightarrow$ La tendencia de la distancia a la capacidad de campo ha ido **constante**
- $a_{TLGCC_4} \longrightarrow$ La tendencia de la distancia a la capacidad de campo ha ido **disminuyendo levemente**
- $a_{TLGCC_5} \longrightarrow$ La tendencia de la distancia a la capacidad de campo ha ido **disminuyendo**

2- PM_{ELGCC}: Eficiencia de uso de la Capacidad de Campo.

- $a_{ELGCC_1} \longrightarrow$ La eficiencia de uso de la capacidad de campo fue **muy alta**
- $a_{ELGCC_2} \longrightarrow$ La eficiencia de uso de la capacidad de campo fue **alta**
- $a_{ELGCC_3} \longrightarrow$ La eficiencia de uso de la capacidad de campo fue **regular**
- $a_{ELGCC_4} \longrightarrow$ La eficiencia de uso de la capacidad de campo fue **baja**
- $a_{ELGCC_5} \longrightarrow$ La eficiencia de uso de la capacidad de campo fue **muy baja**

2- PM_{ELGNR}: Eficiencia de uso del Nivel de Riego.

- $a_{ELGNR_1} \longrightarrow$ La eficiencia de uso del nivel de riego fue **muy alta**
- $a_{ELGNR_2} \longrightarrow$ La eficiencia de uso del nivel de riego fue **alta**
- $a_{ELGNR_3} \longrightarrow$ La eficiencia de uso del nivel de riego fue **regular**
- $a_{ELGNR_4} \longrightarrow$ La eficiencia de uso del nivel de riego fue **baja**
- $a_{ELGNR_5} \longrightarrow$ La eficiencia de uso del nivel de riego fue **muy baja**

2- PM_{CR}: Calidad de Riego.

- $a_{ELGCR_1} \longrightarrow$ La calidad de los riegos fue **muy buena**
- $a_{ELGCR_2} \longrightarrow$ La calidad de los riegos fue **buena**
- $a_{ELGCR_3} \longrightarrow$ La calidad de los riegos fue **regular**
- $a_{ELGCR_4} \longrightarrow$ La calidad de los riegos fue **mala**
- $a_{ELGCR_5} \longrightarrow$ La calidad de los riegos fue **muy mala**

Anexos C

Elementos del archivo de configuración de párrafos

Nombre del Campo	Tipo	Comentario
texto inicio-fin	Objeto con Strings	Son las “mini frases” o conectores que aparecerán al principio y final del párrafo
frases	Objeto	Es donde se escoge cuáles de las mejores resúmenes aparecerán en el párrafo y en qué orden. Además, se elige si la frase aparecerá con el sujeto (sujeto+atributo) o solo el atributo.
conectores	Objeto con Strings	Son las palabras o símbolos con las que se unen las frases para que tengan sentido en el párrafo. Deben haber n-1 conectores, donde n es el número de frases del párrafo.

Anexos D

IDs y etiquetas utilizadas en el código de los tipos de frases generadas

ID Entrada	ID Salida	Etiqueta	Sets difusos utilizados
1	1.1	Duracion de riego	Duración de Riego
1 y 10	1.6	Duracion de riego-cr	Duración de Riego y Cantidad de Riegos
1 y 11	1.5	Duracion de riego-t	Duración de Riego y Tiempo considerado en días
2	2	Tiempo entre riegos	Tiempo entre riegos
2 y 10	2.6	Tiempo entre riegos-cr	Tiempo entre riegos y Cantidad de Riegos
2 y 11	2.5	Tiempo entre riegos-t	Tiempo entre riegos y Tiempo considerado en días
3	3	Eventos fuera CC	Porcentaje de eventos sobre la Capacidad de Campo
4	4	Distancia a CC	Promedio de la distancia a la Capacidad de Campo
5	5	Tendencia distancia CC	Pendiente Capacidad de Campo
6	6	Eventos fuera NR	Porcentaje de eventos bajo el Nivel de Riego
7	7	Distancia a NR	Promedio de la distancia al Nivel de Riego
8	8	Tendencia distancia NR	Pendiente Nivel de Riego
9	9	Percolacion	Percolación Profunda
9 y 10	9.6	Percolacion-cr	Percolación Profunda y Cantidad de Riegos
9 y 11	9.5	Percolacion-t	Percolación Profunda y Tiempo considerado en días
10	-	Cantidad de Riegos	Cantidad de Riegos
11	-	Tiempo considerado	Tiempo considerado[dias]
12	12	Consumo de agua	Consumo de Agua
12 y 10	12.6	Consumo de agua-cr	Consumo de Agua y Cantidad de Riegos
12 y 11	12.5	Consumo de agua-t	Consumo de Agua y Tiempo considerado en días
13	13	Eficiencia CC	Porcentaje de Eficiencia de la Capacidad de Campo
14	14	Eficiencia NR	Porcentaje de Eficiencia del Nivel de Riego
15	15	Calidad de riego	Calidad de Riego
16	16	Tendencia duracion riegos	Tendencia duracion de riegos
17	17	Tendencia tiempo entre riegos	Tendencia Tiempo entre riegos

Anexos E

Imágenes de resultados de encuestas

Comentarios duración de riegos

4 respuestas

"duración ha estado constante" > "ha sido constante". Con respecto a la duración del riego, no sé si poner "media" da alguna información realmente, capaz cuantificarla? (media para un agricultor puede ser larga para otro.... capaz ya lo están considerando..)

Entiendo que ha sido media porque se encuentra entre 3 y 12 con un promedio de unos 6, quizás también debería aclararse eso

1. La duración de riego ha sido media con respecto a que? riegos de paltos? riegos en general, comparados entre ellos?
2. No entiendo que significa que es tendencia constante... podría ser quizás bueno decir cuantos riegos se salieron de esa tendencia como "riegos malos o anómalos"

Se entiende la duración del riego, sin embargo no entiendo por que se diferencia los datos "Sobre 12 horas", "Sobre 3 horas", y "Bajo 3 horas", quizás se podría buscar otra forma de explicarlo y de visualizarlo. La leyenda "Invalido" no se alcanza a ver. En el resumen quizás se podría agregar que significa que la duración del riego ha sido media, y que significa que la tendencia ha sido constante, me explico quizás media significa que ha estado bien o regular o mal, y que la tendencia constante significa que he cumplido con un buen riego?

Comentarios tiempo entre riegos

4 respuestas

creo la segunda parte de la frase es la que tiene más valor: "y la tendencia del tiempo ha estado aumentando."... quizás sería más bien como: "el tiempo ha ido aumentando"

igual que en el anterior explicar "medio"

1. El tiempo entre riegos fue medio con respecto a que? no me dice mucho que sea medio...
2. No se si se entiende la tendencia quizás debería ser algo como "el tiempo entre riegos tiende a ir en aumento" o "los riegos se están distanciando entre sí"

Lo mismo que comente arriba que significa medio??, aquí se entiende que la tendencia de riego aumento, eso quiere decir que estoy regando más??

Comentarios consumo de agua

3 respuestas

comparado con X tiempo (ya que esta tomando desde julio) "el consumo de agua de la ultima semana ha sido muy alto"

Muy alto? se consumió toda el agua? necesita más info.

No entiendo en la gráfica como se representa la última semana.

Figura E.1: Encuesta 1: Comentarios.

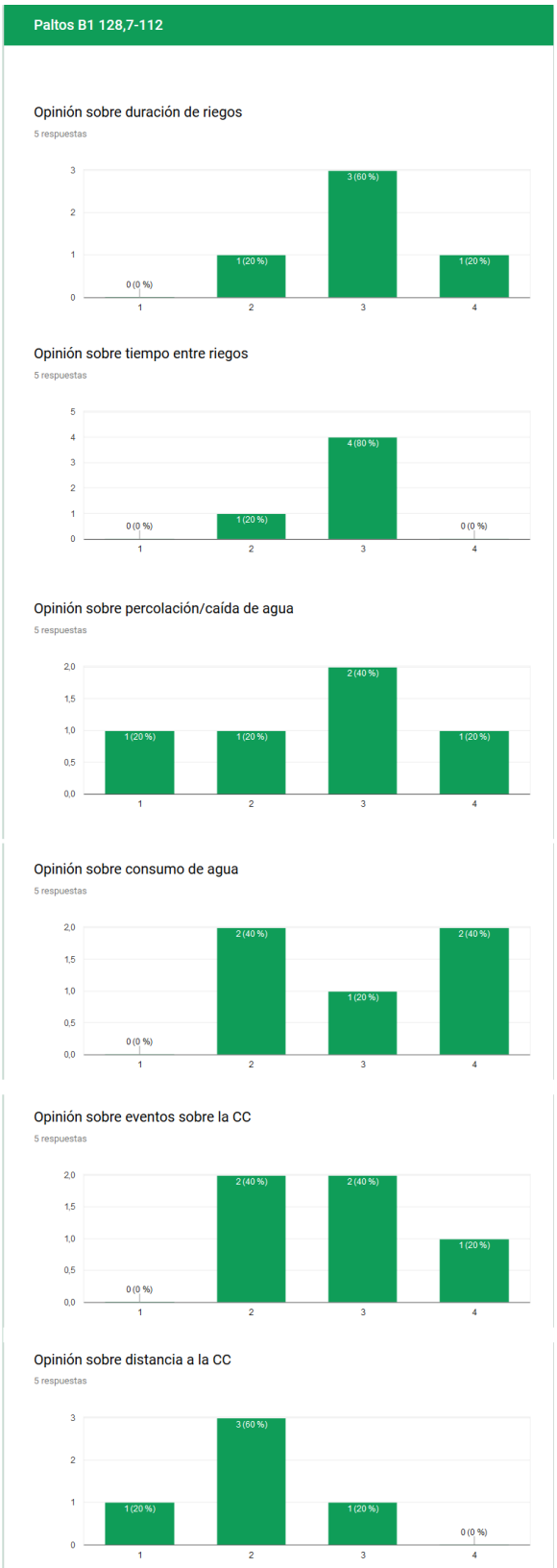


Figura E.2: Encuesta 1: Resultados encuesta de calibración definiciones difusas.

Figura E.3: Resultados encuesta para verificar *sets* difusos y palabras iniciales

Figura E.4: Resultados encuesta evaluar máximas de Grice